"Medidas de diferencia y clasificación automática no paramétrica de datos composicionales"

Josep A. Martín Fernández

TESIS DOCTORAL

codirigida por la Dra. Vera Pawlowsky Glahn y

por el Dr. Carles Barceló i Vidal

Programa de doctorado: Matemática Aplicada - U.P.C.

Girona, 2000

20 de diciembre de 2000

Prólogo

Es muy frecuente encontrar datos de tipo composicional en disciplinas tan dispares como son, entre otras, las ciencias de la tierra, la medicina, y la economía. También es frecuente en estos ámbitos el uso de técnicas de clasificación no paramétrica para la detección de agrupaciones naturales en los datos. Sin embargo, una búsqueda bibliográfica bastante exhaustiva y la presentación de resultados preliminares sobre el tema en congresos de ámbito internacional han permitido constatar la inexistencia de un cuerpo teórico y metodológico apropiado que permita desarrollar pautas y recomendaciones a seguir en el momento de realizar una clasificación no paramétrica de datos composicionales. Por estos motivos se ha elegido como tema de tesis la adaptación y desarrollo de métodos de agrupación adecuados a datos de naturaleza composicional, es decir, datos tales que el valor de cada una de sus componentes expresa una proporción respecto de un total. El título de la misma, "Medidas de diferencia y clasificación automática no paramétrica de datos composicionales", recoge no sólo este propósito, sino que añade la expresión "medidas de diferencia" con el propósito de reflejar el peso específico importante que tiene el estudio de este tipo de medida en el desarrollo del trabajo. La expresión "no paramétrica" se refiere a que en la misma no se considerarán técnicas de clasificación que presuponen la existencia de un modelo de distribución de probabilidad para las observaciones objeto de la agrupación.

En los inicios de este trabajo de investigación nos marcamos toda una serie de objetivos relacionados con los aspectos a tener en cuenta en un proceso de clasificación automática no paramétrica de datos composicionales. Los resultados del estudio de algunos de estos objetivos ya han sido publicados. Cuando esto suceda se indicará haciendo mención de la referencia bibliográfica correspondiente.

Antes de abordar el análisis de los métodos de clasificación, hemos profundizado en el estudio de las medidas de diferencia, de tendencia central y de dispersión. Estas medidas son elementos clave de los métodos de clasificación y están íntimamente relacionadas con la naturaleza de los datos. El estudio de estas medidas nos ha conducido, en primer lugar, al análisis de los inconvenientes que presentan las traslaciones como grupo de transformaciones de datos composicionales

(Martín-Fernández et al., 1998a). De este análisis ha surgido la necesidad de estudiar en profundidad las perturbaciones, definidas por primera vez en Aitchison (1986), como el grupo de transformaciones de datos composicionales (Martín-Fernández et al., 1998a). El hecho de tener definido un grupo de transformaciones diferente del usual nos ha conducido, de manera natural, a revisar los requisitos que debe cumplir cualquier medida de diferencia, de tendencia central y de dispersión entre datos composicionales (Martín-Fernández et al., 1998b). La introducción de estos requisitos ha centrado nuestro interés en el análisis de los inconvenientes que presentan las medidas de diferencia, de tendencia central y de dispersión habituales cuando son aplicadas sobre conjuntos de datos composicionales (Martín-Fernández et al., 1998b; Martín-Fernández et al., 2001). Siguiendo la línea iniciada por Aitchison (1992), en este trabajo de investigación proponemos nuevas medidas de diferencia entre datos composicionales (Martín-Fernández et al., 1998c; Martín-Fernández et al., 1999; Bren y Martín-Fernández, 1999) y exponemos medidas de tendencia central y de dispersión apropiadas para un conjunto de datos composicionales (Martín-Fernández et al., 1998b).

Después del estudio de las medidas de diferencia, de tendencia central y de dispersión, este trabajo de investigación ha proseguido con el análisis de los métodos de clasificación. En este análisis se ha puesto especial énfasis en la adaptación de los diferentes métodos de clasificación automática no paramétrica para su aplicación sobre datos composicionales (Martín-Fernández et al., 1998b). También se ha abordado el estudio de aspectos relacionados con la aplicación práctica de los métodos de clasificación. Entre estos aspectos, nuestro interés se ha centrado en la comparación de los resultados obtenidos al realizar clasificaciones automáticas cuando se usan las diferentes medidas de diferencia (Martín-Fernández et al., 1998a) y, de manera especial, en exponer los aspectos a tener en consideración cuando se aplican algunas de las técnicas de reducción de la dimensión de los datos (componentes principales, reescalado multidimensional, biplot,...) como herramientas de ayuda a la clasificación de datos composicionales.

El trabajo de investigación se había limitado, hasta el momento, a conjuntos de datos composicionales formados por observaciones que no contienen el valor cero. A continuación, se han planteado los aspectos a tener en cuenta cuando en el conjunto de datos a clasificar existen observaciones que contienen el valor cero. Una primera herramienta analizada ha sido la amalgamación de partes como etapa previa a la clasificación de conjuntos de datos con un elevado número de ceros absolutos (Martín-Fernández et al., 1997). Sin embargo, nuestro interés se ha centrado en el análisis de los inconvenientes que presentan los métodos usuales de reemplazamiento de ceros (Martín-Fernández et al., 2000). De los resultados de este análisis ha surgido

la propuesta de una nueva fórmula de reemplazamiento de los ceros por redondeo (Martín-Fernández et al., 2000). Para finalizar el estudio, se lleva a cabo el análisis de la sensibilidad del proceso de substitución de ceros en la clasificación resultante.

La estructura de la presente memoria responde a la persecución de los objetivos que acabamos de exponer. Ésta se inicia con un capítulo introductorio donde se presentan los elementos básicos de las técnicas de clasificación automática no paramétrica.

En el segundo capítulo se aborda el análisis de los conceptos más importantes en torno a los datos composicionales. Entre otros aspectos podemos destacar: las peculiaridades derivadas de su propia naturaleza, el espacio muestral, y las transformaciones y operaciones básicas. Hay que tener en cuenta que operaciones habituales en el espacio real multidimensional, como pueden ser la traslación y el producto por un escalar, asociadas a una concepción euclídea, no son admisibles para datos de tipo composicional. En este mismo capítulo, los esfuerzos se han concentrado principalmente en estudiar las medidas de diferencia entre datos composicionales compatibles con las características anteriores. Las medidas de diferencia no son sólo una herramienta esencial de las técnicas de clasificación, sino que constituyen, junto con las medidas de tendencia central y de dispersión, el elemento diferenciador clave entre las técnicas habituales. Una vez completado el estudio de las medidas de diferencia se prosigue con el análisis de las medidas de tendencia central y de dispersión para un conjunto de datos composicionales. Con ello se dispone de las herramientas necesarias para proceder al desarrollo de una metodología apropiada para la clasificación no paramétrica de datos composicionales, consistente en incorporar los elementos anteriores a las técnicas habituales y adaptarlas en la medida de lo necesario.

El tercer capítulo se dedica exclusivamente a proponer nuevas medidas de diferencia entre datos composicionales. Se exponen los elementos básicos de las medidas de divergencia y se analiza la aplicación de este tipo de medidas a datos de naturaleza composicional. El capítulo acaba con la presentación de una nueva medida de diferencia para datos composicionales basada en la medida de divergencia de Kullback-Leibler.

En el cuarto capítulo se acometen los objetivos relacionados con los métodos de clasificación. Es en este capítulo donde se incorporan las peculiaridades de los datos composicionales a las técnicas de clasificación y se exponen las pautas a seguir en el uso práctico de estas técnicas. El capítulo se completa con la aplicación de la metodología expuesta a un caso práctico.

En los capítulos anteriores se han desarrollado medidas que son únicamente aplicables en el supuesto que los datos composicionales motivo de la clasificación no contienen valores igual a cero. Sin embargo, en nuestro trabajo de investigación se ha podido constatar la existencia

de multitud de casos prácticos en los cuales existen observaciones con valores nulos. En el quinto capítulo de esta tesis se aborda el denominado *problema de los ceros*. Se analizan los inconvenientes de los métodos usuales de reemplazamiento y se propone una nueva fórmula de substitución de los ceros por redondeo. El capítulo finaliza con el estudio de un caso práctico.

En los apéndices finales de esta memoria se recogen los conjuntos de datos utilizados en los casos prácticos que se han desarrollado en la presente tesis.

Desde sus inicios, la elaboración de esta tesis ha significado un esfuerzo considerable en la selección de referencias apropiadas. Esta memoria se completa con la lista de las referencias bibliográficas más relevantes que se han consultado para llevar a cabo este trabajo de investigación.

En resumen, esta tesis aporta una revisión crítica de los métodos de clasificación automática no paramétrica más usuales desde la perspectiva de su aplicación a conjuntos de datos composicionales. Esta revisión nos ha conducido a proponer una metodología específica para la realización de clasificaciones automáticas de datos composicionales. Formando parte de este trabajo de investigación se han definido nuevas medidas de diferencia para datos composicionales. Entre estas medidas, basadas en las medidas de divergencia para distribuciones de probabilidad multinomiales, se destaca la disimilitud de Kullback-Leibler composicional. En esta memoria se recogen los principales aspectos a tener en cuenta en relación al *problema de los ceros* en el contexto de las clasificaciones de datos composicionales. Como contribución principal, destacamos la presentación de una nueva fórmula de reemplazamiento de los ceros por redondeo.

Índice General

Capítulo 1

Introducción a la clasificación automática no paramétrica

1.1 Introducción

En este capítulo se presentan los conceptos básicos que subyacen a la aplicación de una técnica de clasificación. Se resaltan dos elementos: el concepto de medida de diferencia entre dos observaciones y la estrecha relación de una medida de diferencia con las características matemáticas del soporte de los datos. Por otra parte, como es bien conocido, la forma y la disposición de los grupos existentes en un conjunto de datos afecta fuertemente al poder clasificador de los métodos de clasificación. Dedicamos las últimas secciones de este capítulo a exponer cómo afectan estas circunstancias a las diferentes técnicas de clasificación no paramétrica más utilizadas.

Finalmente, es necesario mencionar que con el fin de no hacer muy repetitivas las referencias a la bibliografía básica de esta memoria se ha optado por suprimirlas en su gran mayoría. Los aspectos relacionados con las técnicas de clasificación automática no paramétrica recogidos en este capítulo aparecen en la mayoría de textos de Análisis Multivariante. En nuestro trabajo se han consultado básicamente las obras de Everitt (1993), de Krzanowski (1988b), los dos volúmenes Krzanowski y Marriot (1994, 1995), y la más reciente de Gordon (1999).

1.2 Objetivo de una clasificación

En su acepción clásica, la *clasificación automática* (en inglés "cluster analysis") es una herramienta que pertenece a la familia de técnicas estadísticas denominadas *exploratorias* puesto que su ámbito de trabajo está centrado en el plano descriptivo de datos multivariantes. El objetivo

de esta técnica de análisis multivariante es realizar una clasificación. Es decir, a partir de una muestra representada por una matriz de datos (*individuos* $\times$ *variables*), asignar los individuos a grupos o *clusters*. Estos grupos, desconocidos a priori, serán sugeridos por los datos, y se entenderá que hemos obtenido una *buena* clasificación si los grupos creados son homogéneos en su interior y heterogéneos entre si. Es decir, una clasificación se considerará *razonable* si los individuos de un mismo grupo tienen valores parecidos en las variables observadas y, por el contrario, entre individuos pertenecientes a clases distintas pueden apreciarse características diferentes. El interés de una clasificación radica fundamentalmente en descubrir, analizar e interpretar la estructura de los datos. Aplicando esta técnica puede obtenerse una reducción del número de datos de la muestra asimilando cada individuo al representante de cada grupo, habitualmente el centroide y, además, la clasificación puede dar lugar a un análisis estadístico e interpretación de las características de cada grupo por separado. El proceso de la mayor parte de los diferentes tipos de clasificaciones puede plasmarse en un esquema como el siguiente:

$$\text{INDIVIDUOS} \Longrightarrow \text{ELECCIÓN de la MEDIDA DE DIFERENCIA} \Longrightarrow$$
$$\text{ELECCIÓN del MÉTODO DE CLASIFICACIÓN} \Longrightarrow \text{GRUPOS}$$

Este planteamiento comporta, como paso previo a la aplicación de las técnicas de clasificación automática no paramétrica, la necesidad de establecer una o varias de las siguientes medidas:

1. Una medida de diferencia entre dos datos.
2. Una medida de tendencia central de un conjunto de datos.
3. Una medida de dispersión de un conjunto de datos.

La medida de diferencia entre dos datos nos ha de permitir asignar individuos similares o cercanos a un mismo grupo, e individuos diferentes o alejados a grupos diferentes. Entre las técnicas de clasificación jerárquica ascendente se encuentran algunas técnicas que sólo requieren tener definida una medida de diferencia. Este es el caso de los métodos del *máximo*, del *mínimo* y de la *media*. Otras técnicas, como el método del *centroide*, requieren tener establecida además una medida de tendencia central. Finalmente, existen otras técnicas que requieren adicionalmente una medida de dispersión. Entre éstas se encuentra el método de Ward. En todo caso, un hecho relevante es que todas estas medidas deben ser establecidas teniendo en cuenta las características matemáticas del soporte de los datos a clasificar.

1.3 Notación

Como consideraciones generales es necesario detallar que a lo largo de todo este texto se entenderá que los términos *grupo* y *clase* se refieren al mismo concepto, así como también se consideran equivalentes las expresiones *individuo*, *elemento* y *observación*. Estas últimas expresiones, cuando se refieran a datos de tipo composicional, podrán reemplazarse por las expresiones *composición* y *dato composicional*. Por otra parte, también se utilizan indistintamente las palabras *característica* y *variable* que, en el caso de datos de tipo composicional, podrán ser substituidas por *parte* o *componente*.

La simbología específica, que se va introduciendo en el texto a medida que se exponen los diferentes conceptos, está basada fundamentalmente en la notación utilizada en la monografía de Aitchison (1986). Sin embargo, existe una serie de términos que se repetirán a lo largo de esta memoria y que es conveniente tener presente desde el inicio de su lectura.

En el proceso de una clasificación la información recogida de los elementos observados se organiza en una matriz de datos cuyos elementos se identificarán usando la notación siguiente:

- $\mathbf{X}$: vector aleatorio con soporte en $\mathbb{R}^D$ o matriz de datos cuyas filas corresponden a los elementos de la muestra o *individuos* y cuyas columnas corresponden a las características observadas o *variables*. A un elemento cualquiera de esta matriz lo indicaremos por $\mathbf{x}_{ik}$.
- n: número de individuos que forman la muestra o filas de la matriz de datos.
- D: número de variables contempladas en la muestra o columnas de la matriz de datos.
- G: número de grupos o clases de individuos a considerar en la clasificación.
- $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iD})$: i-ésima realización del vector aleatorio $\mathbf{X}$ o i-ésimo vector fila de la matriz $\mathbf{X}$, donde i toma los valores $i = 1, 2, ..., n$. Para simplificar la notación escribiremos simplemente $\mathbf{x}$ para indicar una observación cualquiera. Cuando pueda existir confusión se añadirá un '$\cdot$' al subíndice de $\mathbf{x}_i$. De esta manera, indicaremos por $\mathbf{x}_{i\cdot}$ la i-ésima observación, la cual se corresponde con la i-ésima fila de $\mathbf{X}$.
- $\mathbf{X}_k = (\mathbf{x}_{1k}, \mathbf{x}_{2k}, ..., \mathbf{x}_{nk})^t$: k-ésimo vector columna de la matriz $\mathbf{X}$ o valores observados de la k-ésima variable, donde k toma los valores $k = 1, 2, \ldots, D$. Cuando pueda existir confusión se añadirá un '$\cdot$' al subíndice. De esta manera, indicaremos por $\mathbf{x}_{\cdot k}$ la k-ésima componente la cual se corresponde con la k-ésima columna de $\mathbf{X}$.
- $\mathbf{C}_i$: i-ésimo grupo o clase que aparece en la clasificación. Cuando se haya finalizado la clasificación el índice i tomará los valores $i = 1, 2, \ldots, G$.

1.4 Tipos de datos y su tratamiento

1.4.1 Escalas de medidas y tipos de datos

Las técnicas de clasificación automática, como cualquier otra técnica de inferencia estadística, conllevan el manejo de datos reales y, por lo tanto, una serie de consideraciones a tener en cuenta según el tipo de datos que aparezcan en el estudio. Cuando se trata de agrupar individuos mediante una clasificación, surge el problema de establecer qué medida es la que nos proporciona el grado de similitud entre dos individuos cuyas características se han observado. En el momento de elegir la medida a utilizar hay que tener en cuenta la tipología de los datos que expresan las características estudiadas. Una presentación extensa relativa a escalas de medida se encuentra en la obra de Anderberg (1973).

En esta sección se presentan dos clasificaciones diferentes: una atendiendo al conjunto dominio de los datos; y la otra en referencia a la escala de medida:

- Tipos de datos según su conjunto dominio

  Se debe recordar que un conjunto de elementos se denomina *finito* si puede establecerse una correspondencia biunívoca, o uno-a-uno, entre el conjunto y un subconjunto de los números naturales del que puede identificarse el número mayor o máximo. Se dice que el conjunto es *infinito numerable* si la correspondencia puede establecerse con todo el conjunto de números naturales. En otro caso, al conjunto se le denomina *infinito no numerable*. Atendiendo a esta clasificación y en referencia al conjunto dominio de los datos del estudio, se dice que las variables observadas pueden ser de los tipos siguientes:

  1. Continuas. Son características cuyo dominio es infinito y no numerable. De manera intuitiva se dice que son características tales que entre los valores observados de dos individuos pueden existir infinitos valores intermedios. Son variables de este tipo las que miden aspectos como, entre otros: el *tiempo*; el *peso*; y la *temperatura*.
  2. Discretas. Su conjunto dominio es finito o, a lo sumo, infinito numerable. De este tipo son variables como, entre otras: el *número de días de baja laboral de un trabajador*; el *color de los ojos*; y la *calificación que merece una gestión política*.
  3. Binarios o dicotómicos. Son un caso particular de las anteriores. Las variables de este tipo miden características que se expresan mediante dos únicos valores: *presencia/ausencia*. Podemos considerar variables binarias, entre otras: el *fumar/no fumar*; el *sexo*; y la *calificación académica (apto/no apto)*.

- Tipos de datos según su escala de medida

  Este tipo de clasificación atiende, más que al tipo de observación en si misma, al tipo de información que sugiere un valor observado en referencia a otros valores de la muestra. Atendiendo a esta idea se establecen los siguientes tipos de datos según la escala de medida:

  1. Nominal. Se dice que la variable $\mathbf{X}_k$ es nominal si dadas las observaciones $\mathbf{x}_k$ y $\mathbf{x}_k^*$ sobre dos individuos, sólo podemos concluir que $\mathbf{x}_k = \mathbf{x}_k^*$, o bien que $\mathbf{x}_k \neq \mathbf{x}_k^*$. Son nominales características como, entre otras: el *estado civil*; el *voto emitido en una elección determinada*; y el *país de residencia.*
  2. Ordinal. En este caso puede afirmarse que si dos individuos son diferentes en la característica es porque, o bien $\mathbf{x}_k < \mathbf{x}_k^*$, o bien $\mathbf{x}_k > \mathbf{x}_k^*$. A este tipo pertenecen variables como, entre otras: la *gama de automóvil*; el *nivel de estudios*; y la *frecuencia de actividad física (nunca/esporádicamente/a menudo/diariamente).*
  3. Intervalo. Son variables en las que tiene sentido calcular la diferencia $\mathbf{x}_k - \mathbf{x}_k^*$ de los valores observados. Pertenecen a esta escala de medida variables como: la *temperatura media diurna en diferentes ciudades durante el mes de agosto*; y el *número de hijos por familia.*
  4. Razón. Se considera que una variable tiene esta escala de medida si es posible y tiene interés calcular la proporción $\mathbf{x}_k/\mathbf{x}_k^*$ entre los valores observados en los individuos. Como ejemplos que ilustren este tipo de escala de medida pueden citarse, entre otros: el *porcentaje del PIB que diferentes países dedican a educación*; el *precio en dólares del kilo de pan en los diferentes países*; y la *proporción de grasas en la composición de diferentes tipos de hamburguesas.*

  Con frecuencia a las variables nominales y a las ordinales se las denomina variables *cualitativas* o *categóricas* puesto que con ellas se expresan aspectos cualitativos de los individuos objeto del estudio que nos permiten separar los individuos en categorías diferentes. A una variable de los otros dos tipos de escala se la denomina variable *cuantitativa* debido a que, habitualmente, estas características tienen un conjunto dominio numérico.

Lógicamente, una misma variable se denominará de un tipo u otro atendiendo a la clasificación a que se hace referencia. Además, en la literatura se encuentran estrategias y técnicas que permiten transformar en determinados casos una variable de un tipo a otro. Este último aspecto toma relevancia en las técnicas de clasificación automática debido a la relación existente entre la medida de similitud entre individuos y los tipos de variables observadas en ellos.

Tal y como se justificará en la Sección 2.2, la naturaleza de los datos de tipo composicional objeto de este trabajo de investigación hace que podamos tratarlos como datos de tipo razón respecto a su escala de medida y, respecto a su espacio soporte, como datos de tipo continuo.

En las siguientes secciones se exponen los aspectos básicos de las medidas de similitud y cuáles son las más utilizadas con cada tipo de datos.

1.4.2 Similitud, medida de diferencia, disimilitud y distancia

En la introducción se ha expuesto la necesidad de establecer, en la mayoría de los métodos de clasificación, una medida de grado de similitud entre individuos. De hecho, siempre es posible fijar algunos de los siguientes tipos de medida, relacionadas o no entre si, para realizar la clasificación:

- Similitud
- Medida de diferencia
- Disimilitud
- Distancia

El contenido de esta sección se encuentra desarrollado con más profundidad en el libro de Everitt (1993) y en el de Kaufman y Rousseuw (1990). En el artículo de Gower (1983) se encuentra una lista completa de medidas. Cuadras (1989) presenta una exposición completa del *estado del arte* de las distancias en estadística.

Las medidas de diferencia, disimilitud y de distancia tienen, al contrario que la similitud, la misión de establecer el grado de separación o de diferencia entre dos individuos de la muestra. Una primera dificultad que aparece al realizar una clasificación es elegir la medida a utilizar. A continuación se exponen las propiedades de cada una de ellas y se describe de qué manera influye el tipo de datos, es decir, la tipología de las variables observadas en la muestra, en la elección de la similitud, la disimilitud o la distancia.

La medida de similitud entre dos observaciones puede establecerse de maneras muy distintas, desde la catalogación por un conjunto de expertos en una determinada materia, por ejemplo en inversiones económicas, hasta el uso de una función matemática de las D variables o características observadas. Se pretende que individuos con valores *parecidos* tengan una similitud alta y individuos con características *diferentes* tengan una similitud baja. En general se establece que toda similitud s entre dos individuos i, j de un conjunto de índices I, es función de los valores observados $\mathbf{x}_i$ y $\mathbf{x}_j$:

$$\begin{aligned} s : I \times I &\rightarrow \mathbb{R} \\ i\ ,\ j &\rightarrow s_{ij} = f(\mathbf{x}_i, \mathbf{x}_j). \end{aligned}$$

En la mayoría de los casos se procura que esta función f, y por ende la similitud s, cumpla algunas o todas de las tres propiedades siguientes:

1. Simetría: $s_{ij} = s_{ji}, \quad \forall\, i, j \in I.$
2. Máxima similitud: $s_{ij} \leq s_{ii}, \quad \forall\, i, j \in I.$
3. Interpretación: $0 \leq s_{ij} \leq 1, \quad \forall\, i, j \in I.$

La propiedad de la simetría puede parecer, en principio, de obligado cumplimiento. Sin embargo si se considera el caso del tiempo de desplazamiento en automóvil entre diversos puntos de una gran ciudad, se acepta que hay casos en los que puede establecerse una medida de similitud que no es necesariamente simétrica, puesto que por ejemplo el recorrido en uno y otro sentido no tiene porqué ser el mismo. La propiedad de máxima similitud recoge el concepto de igualdad entre individuos en el sentido que no puede existir otro individuo más similar que uno mismo, y que consideramos dos individuos totalmente similares si su similitud es igual a la de uno de ellos en si mismo. Con el objeto de mejorar la aplicación y la lectura de una similitud, se recomienda, si es posible, escalarla entre 0 y 1 mediante una transformación lineal del rango entre la similitud mínima y la máxima. De esta manera puede interpretarse la similitud como un tanto por ciento de semejanza entre individuos de la muestra. La similitud es una medida aplicada en clasificaciones de datos en las que, generalmente, las variables observadas son de tipo cualitativo.

Los conceptos de medida de diferencia, disimilitud y distancia son elementos clave en muchas técnicas estadísticas de análisis multivariante (Cuadras y Arenas, 1997). Entre estas técnicas, además de las técnicas de clasificación no paramétricas, se encuentran el análisis discriminante y las técnicas de reducción de la dimensionalidad como son, entre otras, las componentes principales, los diagramas biplot, y el reescalado multidimensional. Una lectura de Gower (1983) y Bren y Batagelj (1997) muestra los aspectos más básicos relacionados con el concepto de medida de diferencia. Estos aspectos se recogen en las definiciones siguientes:

Definición 1.1 Sea $\mathbf{E}$ el espacio muestral o adherencia del dominio de las observaciones. Decimos que una función $d : \mathbf{E} \times \mathbf{E} \longrightarrow \mathbb{R}$ es una *medida de diferencia* sobre el espacio $\mathbf{E}$ si verifica

las propiedades siguientes:

i. simetría: $d(\mathbf{x}, \mathbf{x}^*) = d(\mathbf{x}^*, \mathbf{x}), \quad \forall\, \mathbf{x}, \mathbf{x}^* \in \mathbf{E}$;

ii. mínima diferencia: $d(\mathbf{x}, \mathbf{x}) \leq d(\mathbf{x}, \mathbf{x}^*), \quad \forall\, \mathbf{x}, \mathbf{x}^* \in \mathbf{E}$. □

Definición 1.2 Una medida de diferencia se denomina *disimilitud* si cumple las propiedades siguientes:

i. $d(\mathbf{x}, \mathbf{x}^*) \geq 0, \quad \forall\, \mathbf{x}, \mathbf{x}^* \in \mathbf{E}$;

ii. $d(\mathbf{x}, \mathbf{x}) = 0, \quad \forall\, \mathbf{x} \in \mathbf{E}$.

Al par ordenado $(\mathbf{E}, d)$ se le denomina *espacio de disimilitud.* □

Las medidas de diferencia pueden verificar otras propiedades importantes recogidas en la siguiente definición:

Definición 1.3 Según las propiedades que verifique una medida de diferencia se la denomina de las siguientes formas:

- Una medida de diferencia d sobre $\mathbf{E}$ se dice que es:

  i. *definida*, si verifica que $\forall\, \mathbf{x}, \mathbf{x}^* \in \mathbf{E},\ d(\mathbf{x}, \mathbf{x}^*) = 0 \Longrightarrow \mathbf{x} = \mathbf{x}^*$;

  ii. *semi-definida*, si verifica que $\forall\ \mathbf{x}, \mathbf{x}^* \in \mathbf{E},\ d(\mathbf{x}, \mathbf{x}^*) = 0 \Longrightarrow \forall\ \mathbf{x}' \in \mathbf{E},\ d(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}^*, \mathbf{x}')$;

  iii. *métrica*, si verifica la *desigualdad triangular*: $\forall\ \mathbf{x}, \mathbf{x}^*, \mathbf{x}' \in \mathbf{E},\ d(\mathbf{x}, \mathbf{x}^*) \leq d(\mathbf{x}, \mathbf{x}') + d(\mathbf{x}^*, \mathbf{x}')$;

  iv. *ultramétrica*, si verifica que $\forall\, \mathbf{x}, \mathbf{x}^*, \mathbf{x}' \in \mathbf{E},\ d(\mathbf{x}, \mathbf{x}^*) \leq max\{d(\mathbf{x}, \mathbf{x}'), d(\mathbf{x}^*, \mathbf{x}')\}$.

- Decimos que una disimilitud d es una *semi-distancia* si verifica la desigualdad triangular. Al par ordenado $(\mathbf{E}, d)$ se le denomina *espacio semi-métrico.*

- Decimos que una disimilitud d es una *distancia*, o simplemente una *métrica*, si verifica la desigualdad triangular y es definida. Al par ordenado $(\mathbf{E}, d)$ se le denomina *espacio métrico.* □

La similitud y la disimilitud elegidas para analizar un conjunto de datos pueden aplicarse por separado y de manera independiente. Si se procede de este modo, muy probablemente se obtendrá una clasificación diferente usando una u otra. Considérese el ejemplo de los puntos de una gran ciudad y contémplese la similitud que se deriva del tiempo de recorrido en automóvil y la disimilitud que se obtiene con la distancia en línea recta entre los puntos. Es del dominio público que en las grandes ciudades existen puntos cercanos muy mal comunicados y viceversa. Sin embargo, en el caso de trabajar con una similitud (disimilitud) escalada entre 0 y 1, siempre es posible asociarle una disimilitud (similitud) a partir de la expresión:

$$d_{ij} = 1 - s_{ij}, \quad \forall\, i, j \in I. \tag{1.1}$$

En este caso las clasificaciones resultantes usando una u otra deberán coincidir. Otra transformación usual que nos relaciona similitudes y disimilitudes viene dada por la expresión:

$$d_{ij} = \sqrt{1 - s_{ij}}, \quad \forall\, i, j \in I. \tag{1.2}$$

En Cuadras (1989) se justifica que es preferible utilizar la transformación (1.2) a la (1.1) y que, en general, cuando la similitud no está escalada entre 0 y 1, la transformación más apropiada es

$$d_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}, \quad \forall\, i, j \in I. \tag{1.3}$$

En este caso, si la matriz de similitud entre individuos de un conjunto cualquiera $\mathbf{X}$ es una matriz semidefinida positiva, entonces la disimilitud d definida como (1.3) es una distancia y es *euclídea*. Decimos que es euclídea en el sentido que podemos representar los individuos del conjunto $\mathbf{X}$ por puntos de un espacio real $\mathbb{R}^D$ y que la matriz de distancia euclídea entre estos puntos del espacio real coincide con la matriz de disimilitudes (d_{ij}) entre las observaciones del conjunto $\mathbf{X}$.

Ejemplo 1.1 Si el conjunto de datos está formado por observaciones $\mathbf{x}_i$ cuyo dominio es el espacio real $\mathbb{R}^D$, puede considerarse como medida de disimilitud el coeficiente de correlación $r_{\mathbf{x}_i,\mathbf{x}_j}$ o r_{ij} entre los valores de dos individuos o filas de la matriz de datos. Su cálculo supone hacer medias y desviaciones en las que intervienen variables diferentes y por ello no es interpretable. Sin embargo, es útil como ejemplo de similitud puesto que a partir de él podemos construir varias disimilitudes y similitudes:

- $d_{ij} = \frac{1-r_{ij}}{2}$ y $s_{ij} = \frac{1+r_{ij}}{2}$;
- $d_{ij} = 1 - |r_{ij}|$ y $s_{ij} = |r_{ij}|$;

- $d_{ij} = 1 - r_{ij}^2$ y $s_{ij} = r_{ij}^2$.

Se puede observar como dos individuos con correlación altamente positiva tienen un grado de disimilitud cercano a 0 en todos los ejemplos y, por el contrario, dos individuos con alta correlación negativa tienen un grado de disimilitud cercano a 1 en la primera medida y cercana a cero en las otras dos. La clasificación resultante no sería pues equivalente.

1.4.3 Tipos de datos y medidas más usuales

- *Datos binarios*

  Si la muestra contiene n individuos de los que se ha observado la presencia, identificada con el número 1, o ausencia, número 0, de D características cualitativas, se considera que estamos trabajando con datos binarios. Con este tipo de datos lo más usual es trabajar con una medida de similitud. En el momento de establecer una medida de similitud entre dos individuos $\mathbf{x}_i$ y $\mathbf{x}_j$ la estrategia se basa en las frecuencias que se muestran en la tabla 1.1. Nótese que estas frecuencias representan los siguientes conceptos:

  Tabla 1.1: Frecuencias de presencias/ausencias comunes en las dos observaciones $\mathbf{x}_i$ y $\mathbf{x}_j$

		$\mathbf{x}_j$		Total
		1	0	
$\mathbf{x}_i$	1	a_{ij}	b_{ij}	presencias en i
	0	c_{ij}	d_{ij}	ausencias en i
Total		presencias en j	ausencias en j	D

  - a_{ij}: número de características presentes comunes a los dos individuos $\mathbf{x}_i$ y $\mathbf{x}_j$;
  - b_{ij}: número de características presentes en $\mathbf{x}_i$ y ausentes en $\mathbf{x}_j$;
  - c_{ij}: número de características ausentes en $\mathbf{x}_i$ y presentes en $\mathbf{x}_j$;
  - d_{ij}: número de características ausentes comunes a los dos individuos.

  Hay multitud de medidas de similitud propuestas en la bibliografía y no existe un criterio universal que permita elegir la medida más adecuada. Una de las más usuales es la basada en el *índice de coincidencias* o *matching coefficient*:

$$s_{ij} = \frac{a_{ij} + d_{ij}}{D}. \tag{1.4}$$

Esta similitud, que da el mismo peso a las coincidencias $1-1$ que a las ausencias $0-0$, tiene un rango de variación entre 0 y 1, alcanzándose la máxima similitud cuando $a_{ij}+d_{ij}=D$. Es decir, cuando i y j son dos filas idénticas de la matriz de datos. La mínima similitud se alcanzará cuando no haya ninguna coincidencia, es decir, si $a_{ij}+d_{ij}=0$ y por tanto, cuando $b_{ij}+c_{ij}=D$.

Si existe la preocupación que la posible información superflua observada desvirtúe la medida y, se pretende que las características que no están presentes en el conjunto de datos no afecten a la similitud, es deseable que la medida no sea función de las ausencias comunes d_{ij}. En este caso se puede construir una similitud en términos del *coeficiente de Jaccard* o simplemente *S-coefficient*:

$$s_{ij}=\frac{a_{ij}}{a_{ij}+b_{ij}+c_{ij}}.$$

Su rango de variación está entre 0 y 1, siendo 1 si $b_{ij}+c_{ij}=0$ y, por tanto, si $a_{ij}+d_{ij}=D$. Es decir, si los vectores fila de la matriz de datos son idénticos. El coeficiente tomará el valor 0 si no hay características presentes comunes en los dos individuos, sin tener en cuenta las ausencias comunes.

Desde otro punto de vista diferente, pero complementario, es posible establecer una medida de similitud entre dos individuos si se recurre al *coeficiente de correlación* entre individuos o de *Pearson* para medir el grado de relación entre los individuos $\mathbf{x}_i$ y $\mathbf{x}_j$:

$$s_{ij}=\frac{a_{ij}d_{ij}-b_{ij}c_{ij}}{\sqrt{(a_{ij}+c_{ij})(b_{ij}+d_{ij})(a_{ij}+b_{ij})(c_{ij}+d_{ij})}}.$$

En este caso estamos expresando que la mayor similitud corresponde al valor 1 y la máxima disimilitud se alcanza cuando $s_{ij}=-1$.

Como se ha mencionado, ninguna de las tres medidas de similitud presentadas es la más idónea en todos los casos; si se citan aquí es por ser las más utilizadas.

- *Datos nominales*

Se entiende que se trabaja con datos nominales cuando sobre los individuos de la muestra se han observado D variables cualitativas y cada una de ellas tiene más de dos niveles de valores. De entre las diferentes estrategias usadas para establecer una medida de similitud entre dos individuos la más utilizada consiste en una generalización del *índice de coincidencias* definido en (1.4). Se considera s_{ijk} un coeficiente binario que toma el valor 1 si los individuos i y j coinciden en el valor de la característica k, y que toma el valor 0 en

caso contrario. Se construye la medida de similitud realizando la operación siguiente:

$$s_{ij} = \frac{\sum_{k=1}^{D} s_{ijk}}{D},$$

donde el numerador representa el número total de coincidencias entre los dos individuos. Este coeficiente tiene un rango de variación entre 0 y 1. El valor 0 se da cuando los dos individuos no tienen ninguna coincidencia, y el valor 1 se da para individuos que toman los mismos valores para las D variables consideradas.

- *Datos ordinales*

  En este caso los niveles de variación de D variables nominales siguen un orden lógico o graduación. En este supuesto, puede optarse por establecer una medida de similitud análoga al caso de datos nominales. Sin embargo, este procedimiento tiene el defecto de estar considerando igualmente disimilares dos individuos que en una característica determinada toman valores en niveles contiguos y otros dos individuos que toman valores en niveles muy separados.

  Otra estrategia diferente pasa por aplicar directamente a los valores codificados numéricamente una medida de diferencia de las que se utilizan cuando los datos son de tipo cuantitativo. Con el objetivo de facilitar su manejo se realiza una recodificación numérica de los niveles de las variables. Se considera que cada característica k de las D observadas puede tomar el valor desde el nivel 1 a su nivel máximo N_k. Entonces, si se desea aplicar esta estrategia, es útil realizar un escalado previo de los rangos de variación para pasarlos todos al intervalo cero-uno. Considérese v_{ik} el valor del individuo i en la variable k y calcúlese la siguiente expresión para todo individuo y toda variable:

$$z_{ik} = \frac{v_{ik} - 1}{N_k - 1}.$$

  Nótese que con este procedimiento se ha obtenido una matriz de datos cuantitativos de valores entre 0 y 1. A partir de esta matriz de datos transformados se establece la disimilitud entre individuos calculando la medida de diferencia entre los vectores fila de la matriz.

- *Datos multinomiales*

  Este tipo de datos es muy diferente de los anteriores debido a que, habitualmente, más que recoger información sobre el valor de D características objeto del estudio son datos donde se contemplan las frecuencias en que aparecen los D diferentes valores de una característica

determinada. De este modo, se considera que el conjunto de datos está formado por datos multinomiales si cada observación $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iD})$ se compone de las frecuencias o número de presencias de los D estados o valores de la variable objeto del estudio. Además, la suma $\sum_{k=1}^{D} x_{ik} = t_i$, puede ser igual o diferente para todos los individuos de la muestra.

Con este tipo de datos es habitual definir una disimilitud basada en la distribución de probabilidad χ^2. Esta disimilitud, que es la más usual, viene definida por la expresión siguiente:

$$d_{\chi^2}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{f=i,j} \sum_{k=1}^{D} \frac{\left[x_{fk} - t_f \frac{x_{ik}+x_{jk}}{t_i+t_j}\right]^2}{t_f \frac{x_{ik}+x_{jk}}{t_i+t_j}}, \tag{1.5}$$

donde f representa la fila de la matriz de datos o el individuo de la muestra. Se observa que la expresión $t_f \frac{\mathbf{x}_{ik}+\mathbf{x}_{jk}}{t_i+t_j}$ realiza el papel de frecuencia teórica o esperada.

En el Capítulo 3 se introducirán otras disimilitudes adecuadas para los datos multinomiales y se tratará la relación entre los datos multinomiales y los datos composicionales.

- *Datos cuantitativos*

Los datos se denominan cuantitativos cuando las D características observadas son cuantitativas. Para establecer una medida de diferencia entre los individuos la estrategia más utilizada es recurrir a considerar una disimilitud d_{ij}. De hecho, si se denomina $M = \max\{d_{ij}; 1 \leq j, i \leq n\}$, siendo n el tamaño de la muestra, siempre puede definirse una medida de similitud mediante la expresión:

$$s_{ij} = 1 - \frac{d_{ij}}{M}.$$

Análogamente al caso de datos binarios, en el caso de datos cuantitativos existen multitud de medidas de diferencia distintas, no existiendo un criterio absoluto que permita decidir la disimilitud más adecuada. Por este motivo a continuación se exponen las medidas más utilizadas y sus características principales:

  - Distancia euclídea:

$$d_{\text{Euc}}(\mathbf{x}, \mathbf{x}^*) = \left[\sum_{k=1}^{D} (x_k - x_k^*)^2\right]^{1/2}. \tag{1.6}$$

    De todas las distancias es la más usada. Sin embargo, tal y como justificaremos en las siguientes secciones, esta distancia no es una medida de diferencia adecuada entre dos datos de tipo composicional.

    La distancia euclídea tiene la propiedad de ser invariante por traslaciones, pero tiene el defecto de ser muy dependiente de los cambios de escala de las variables. Si cada

una de las variables se cambia de escala con un factor diferente, como, por ejemplo, al proceder a su estandarización, la distancia no se conserva.

Por estandarizar los datos $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ entendemos realizar la siguiente transformación:

$$\tilde{\mathbf{x}}_{ik} = \frac{\mathbf{x}_{ik} - \overline{\mathbf{x}_{\cdot k}}}{s_{\cdot k}}, \qquad 1 \leq i \leq n \quad 1 \leq k \leq D,$$

donde $\overline{\mathbf{x}_{\cdot k}}$ y $s_{\cdot k}$ representan, respectivamente, la media aritmética y la desviación típica de la k-ésima variable.

De hecho, si las características observadas son cuantitativas, pero tienen unos rangos de variación y unidades diferentes, es recomendable estandarizar las variables antes de calcular las distancias. Esto sucede por ejemplo al estudiar las variables edad (años) y altura (cm) en un conjunto de personas. Sin embargo, si las D variables observadas tienen rangos de variación similares el hecho de estandarizar las variables antes de calcular la distancia euclídea puede distorsionar la clasificación resultante – véase el Ejemplo 1.2.

– Distancias L_q (q-métricas de Minkowski):

$$d_{\mathrm{Min}}(\mathbf{x}, \mathbf{x}^*) = \left[\sum_{k=1}^{D} |\mathbf{x}_k - \mathbf{x}_k^*|^q \right]^{1/q}, \; q > 1 \; .$$

Esta familia de distancias generalizan la distancia euclídea (caso $q = 2$). Asimismo, son invariantes por traslación pero no por cambios de escala.

Se destaca la distancia del caso $q = 1$, llamada distancia *City block* o *Manhattan*. Recibe este nombre debido a que representa la distancia entre dos puntos de una ciudad en la que hay que desplazarse por calles perpendiculares entre si y paralelas a los ejes coordenados. Esta distancia es aplicable en los casos que interesa tener sólo en cuenta la variación en términos absolutos. Por ejemplo, cuando interese que el punto $\mathbf{x}_1 = (2, 2)$ tenga el mismo grado de disimilitud con el punto $\mathbf{x}_2 = (3, 4)$ que con el punto $\mathbf{x}_3 = (2, 5)$.

– Distancia de Mahalanobis d_{Mah}:

Consideramos $\boldsymbol{\Omega}$ una población caracterizada por el vector aleatorio $\mathbf{X}$ con vector de esperanzas $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$. Se define la distancia de Mahalanobis entre dos individuos $\mathbf{x}$ y $\mathbf{x}^*$ como

$$d_{\mathrm{Mah}}(\mathbf{x}, \mathbf{x}^*) = \sqrt{(\mathbf{x} - \mathbf{x}^*)\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^*)^t}. \tag{1.7}$$

Esta distancia puede extenderse al caso en que $\mathbf{X}$ represente un conjunto de observaciones $\mathbf{x}_1$,$\mathbf{x}_2$,...,$\mathbf{x}_n$. Basta reemplazar, respectivamente, $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ por el vector de medias $\mathbf{m}$ y la matriz de covarianzas $\mathbf{S}$ siguientes:

$$\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad \text{y} \quad \mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})^t(\mathbf{x}_i - \mathbf{m}). \tag{1.8}$$

Entonces se define la distancia de Mahalanobis como

$$d_{\text{Mah}}(\mathbf{x}, \mathbf{x}^*) = \sqrt{(\mathbf{x} - \mathbf{x}^*)\mathbf{S}^{-1}(\mathbf{x} - \mathbf{x}^*)^t}. \tag{1.9}$$

En la práctica esta distancia siempre viene asociada a un conjunto de datos. Por este motivo la trataremos más extensamente en la Sección 1.4.4 como una de las distancias más habituales para medir la disimilitud entre dos conjuntos de observaciones de tipo cuantitativo. En la expresión (1.9) se observa que para el caso $\mathbf{S} = \mathbf{Id}_{D\times D}$, es decir, cuando las D variables observadas sean incorrelacionadas dos a dos y con varianza igual a uno, la distancia de Mahalanobis entre dos individuos no es más que su distancia euclídea. Este hecho permite considerar la distancia euclídea como un caso particular de la d_{Mah} e indica que para muestras de variables correlacionadas o con distinta variabilidad se eligirá la distancia de Mahalanobis.

Esta distancia d_{Mah} posee la buena propiedad de ser invariante por transformaciones lineales no singulares de los datos. Por tanto, en particular, es invariante por cambios de escala de las variables –véase el Ejemplo 1.2.

En la definición de la distancia de Mahalanobis se observa que esta medida requiere el cálculo de una matriz inversa y, en consecuencia, aparece el problema de las matrices de covarianzas singulares. En este caso puede recurrirse a la matriz *inversa generalizada* o *pseudo-inversa* de la matriz de covarianzas. A pesar de que la matriz inversa generalizada de una matriz singular no es única, se cumple (Cuadras, 1991) que la distancia de Mahalanobis entre dos individuos no depende de la pseudo-inversa utilizada.

Por otra parte, el hecho que esta distancia contenga matrices de covarianzas en su definición se convierte en un problema cuando se está realizando una clasificación puesto que no se conocen los grupos existentes. Pueden adoptarse entonces diferentes estrategias que van desde calcular las distancias usando la matriz de covarianzas global –ignorando la estructura en grupos– hasta, en el caso de datos que pueda suponerse que proceden de una distribución normal multivariante, usar el algoritmo

de estimación de la matriz de covarianzas ponderada intragrupos introducido por Art, Gnanadesikan, y Kettenring (1982).

– Distancia general:

$$d(\mathbf{x}, \mathbf{x}^*) = \sqrt{(\mathbf{x} - \mathbf{x}^*)\mathbf{M}(\mathbf{x} - \mathbf{x}^*)^t}.$$

Esta distancia generaliza la distancia de Mahalanobis para una matriz $\mathbf{M}$ definida positiva. La matriz $\mathbf{M}$ se denomina la matriz de la métrica utilizada para establecer la distancia. De hecho si $\mathbf{M} = \mathbf{Id}_{D\times D}$ se obtiene la distancia euclídea. Este caso general de la distancia tiene la dificultad de la interpretabilidad asociada a la matriz $\mathbf{M}$ escogida.

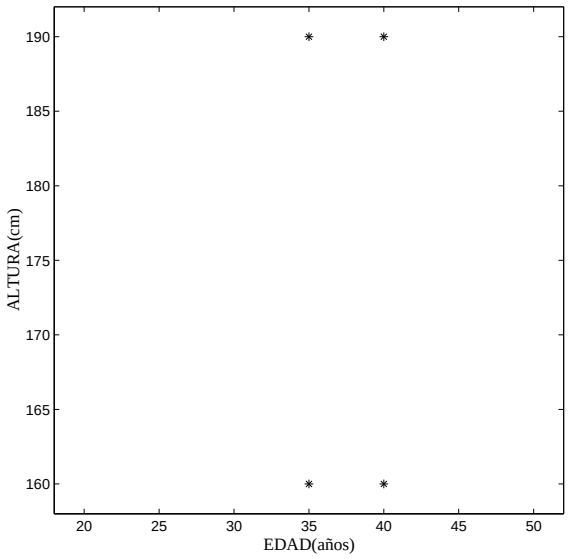
Ejemplo 1.2 Efecto del cambio de escala y la estandarización:

Los datos de este ejemplo han sido extraídos del libro de Kaufman y Rousseuw (1990). En la tabla 1.2 se muestra la edad y la altura de cuatro personas simbolizadas por $\mathbf{x}_1$,$\mathbf{x}_2$,$\mathbf{x}_3$, y $\mathbf{x}_4$.

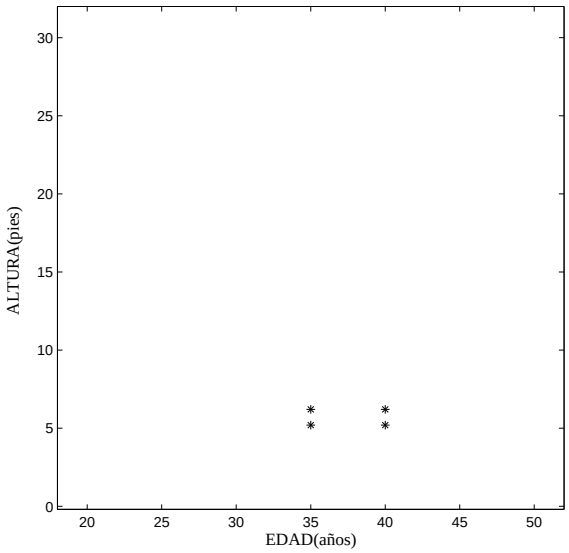
Tabla 1.2: Edad y Altura de cuatro personas en valores originales y sus transformados.

	Edad(años)	Altura(cm)	Altura(pies)	Edad(estand.)	Altura(estand.)
$\mathbf{x}_1$	35	190	6.2	-1	1
$\mathbf{x}_2$	40	190	6.2	1	1
$\mathbf{x}_3$	35	160	5.2	-1	-1
$\mathbf{x}_4$	40	160	5.2	1	-1

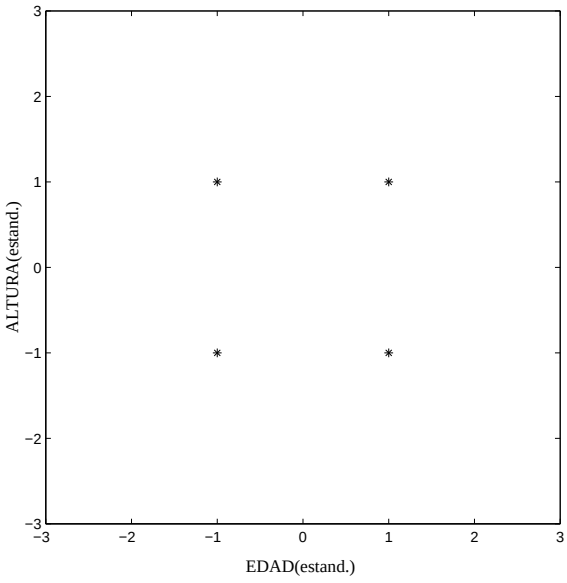
En las figuras 1.1(a) y 1.1(b) se observa como el diferente cambio de escala en las variables afecta a la distancia euclídea entre los individuos. Compárese la distancia d_{23} de las matrices de la tabla 1.3 antes y después del cambio de escala con la variación sufrida por la distancia d_{12}. En la figura 1.1(c) se observa que el efecto de la estandarización ha sido fulminante. Ahora dos individuos de la misma edad y que difieren en 30 cm de altura, $\mathbf{x}_1$ y $\mathbf{x}_3$, están a la misma distancia euclídea –véase la tabla 1.3– que dos individuos igual de altos que difieren en 5 años de edad, $\mathbf{x}_1$ y $\mathbf{x}_2$. Al estandarizar se pasa a trabajar con variables sin unidades de medida, con media cero y desviación uno.



(a)



(b)



(c)

Figura 1.1: Representación de los datos de los individuos del Ejemplo 1.2: (a) *Edad (años)* × *Altura (cm);* (b) *Edad (años)* × *Altura (pies);* (c) *Edad* × *Altura, ambas estandarizadas.*

Tabla 1.3: Matrices de distancias euclídea entre los individuos del Ejemplo 1.2: (a) edad en años y altura en cm.; (b) edad en años y altura en pies ; (c) edad y altura estandarizadas.

$$\begin{pmatrix} 0 & 5 & 30 & 30.41 \\ 5 & 0 & 30.41 & 30 \\ 30 & 30.41 & 0 & 5 \\ 30.41 & 30 & 5 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 5 & 1 & 5.1 \\ 5 & 0 & 5.1 & 1 \\ 1 & 5.1 & 0 & 5 \\ 5.1 & 1 & 5 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 2 & 2 & 2\sqrt{2} \\ 2 & 0 & 2\sqrt{2} & 2 \\ 2 & 2\sqrt{2} & 0 & 2 \\ 2\sqrt{2} & 2 & 2 & 0 \end{pmatrix}$$

(a) (b) (c)

Por el contrario, si consideramos la distancia de Mahalanobis d_{Mah} podemos calcular la matriz de covarianzas en cada caso y comprobar que nos produce la misma matriz de distancias –véase la tabla 1.4.

Tabla 1.4: Cálculos en base a los datos de los individuos del Ejemplo 1.2. Matriz de: (a) covarianzas con edad en años y altura en cm; (b) covarianzas con edad en años y altura en pies; (c) covarianzas con edad y altura estandarizadas; (d) distancias de Mahalanobis.

$$\begin{pmatrix} 6.25 & 0 \\ 0 & 225 \end{pmatrix} \quad \begin{pmatrix} 6.25 & 0 \\ 0 & 0.25 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 2 & 2 & 2\sqrt{2} \\ 2 & 0 & 2\sqrt{2} & 2 \\ 2 & 2\sqrt{2} & 0 & 2 \\ 2\sqrt{2} & 2 & 2 & 0 \end{pmatrix}$$

(a) (b) (c) (d)

- *Datos mixtos*

  En este capítulo se ha expuesto qué medidas de diferencia son las más utilizadas según el tipo de datos que contiene la muestra, siempre pensando que la muestra contiene datos de un solo tipo. Pero, ¿qué medida se elegirá si en una muestra se observan variables de diferente tipología (*datos mixtos*)?

  La similitud más utilizada (Everitt, 1993; Cuadras y Arenas, 1997) es el coeficiente de similitud de Gower entre dos observaciones $\mathbf{x}_i$ y $\mathbf{x}_j$ definido por

$$s_{ij} = \frac{\sum_{k=1}^{D} \delta_{ijk} s_{ijk}}{\sum_{k=1}^{D} \delta_{ijk}}.$$

El factor δ_{ijk} toma el valor cero cuando un dato de alguno de los dos individuos es declarado *missing* o *ausente*; en otro caso toma el valor 1. El factor s_{ijk} representa la contribución de la k-ésima variable a la similitud entre los individuos. Si la k-ésima variable es binaria o nominal se define como

$$s_{ijk} = \begin{cases} 0, & \text{si } \mathbf{x}_{ik} \neq \mathbf{x}_{jk}; \\ 1 & \text{si } \mathbf{x}_{ik} = \mathbf{x}_{jk}. \end{cases}$$

Si la k-ésima variable es cuantitativa se define como

$$s_{ijk} = 1 - \frac{|\mathbf{x}_{ik} - \mathbf{x}_{jk}|}{R_k},$$

donde el factor R_k representa el rango de la variable k.

Se observa que la similitud que se ha definido para datos mixtos siempre está comprendida entre 0 y 1, por lo que, se le puede asociar directamente una disimilitud usando la transformación propuesta en la expresión (1.2).

1.4.4 Medidas de diferencia entre dos conjuntos de datos

Bien como ayuda a la interpretación de la clasificación obtenida, bien como paso obligado de algunos métodos de clasificación, se necesita definir el grado de disimilitud de un elemento a un grupo de individuos y, en general, el nivel de disimilitud entre dos grupos de individuos de la muestra. Es importante remarcar que, en general, una vez se ha definido una medida de diferencia entre dos conjuntos, de ella se deduce directamente la medida de diferencia entre una observación y un conjunto de datos. Basta con considerar que uno de los dos conjuntos está formado por un solo individuo.

A continuación introducimos las disimilitudes más usuales teniendo en cuenta que en la Sección 2.5 se expondrán las peculiaridades que pueden presentar estas disimilitudes cuando los datos son de tipo composicional. Ciertamente, no hay que perder de vista que algunas de las disimilitudes que presentamos son la base de algunas de las técnicas de clasificación automática no paramétrica que se presentarán en la Sección 1.5.

Anteriormente hemos expuesto que, para establecer la disimilitud entre dos elementos, debe tenerse en cuenta la tipología de las variables observadas. Para establecer la disimilitud entre grupos hay que proceder de la misma manera. De manera general, podemos resumir la definición de una medida de diferencia entre dos conjuntos en el siguiente esquema:

- Para datos cualitativos:

  De acuerdo con Everitt (1993), si la muestra está formada por individuos cuyas D características observadas son todas cualitativas, puede definirse una disimilitud llamada *distancia genética* del siguiente modo:

  Definición 1.4 Si se tienen dos grupos de individuos, $\mathbf{A}$ y $\mathbf{B}$, se define la distancia genética entre ellos como

  $$d_{\mathbf{AB}} = \sum_{k=1}^{D} \left(1 - \sum_{l=1}^{N_k} \left(p_{\mathbf{A}kl}\, p_{\mathbf{B}kl} \right)^{1/2} \right)^{1/2} ,$$

  donde D es el número de variables categóricas que caracterizan a los individuos, N_k es el número de niveles o estados de la k-ésima variable categórica y, $p_{\mathbf{A}kl}$ y $p_{\mathbf{B}kl}$ son, respectivamente, las proporciones de individuos de los grupos $\mathbf{A}$ y $\mathbf{B}$ que en la k-ésima variable tienen el nivel l. □

- Para datos cuantitativos:

  Se sigue una de las dos estrategias siguientes:

  1. Se elige un representante de cada uno de los dos grupos, habitualmente el *centroide* o punto medio, y se define la diferencia entre los dos conjuntos como la disimilitud entre los representantes. Por ejemplo, si se escogen los centroides $\overline{\mathbf{x}}_{\mathbf{A}}$ y $\overline{\mathbf{x}}_{\mathbf{B}}$ de dos grupos $\mathbf{A}$ y $\mathbf{B}$ –véase la figura 1.2– y la distancia euclídea como medida de disimilitud, se obtiene:

     $$d_{\mathrm{Euc}}(\mathbf{A},\mathbf{B}) = \sqrt{\sum_{k=1}^{D} \left(\overline{\mathbf{x}}_{\mathbf{A}_k} - \overline{\mathbf{x}}_{\mathbf{B}_k} \right)^2}.$$

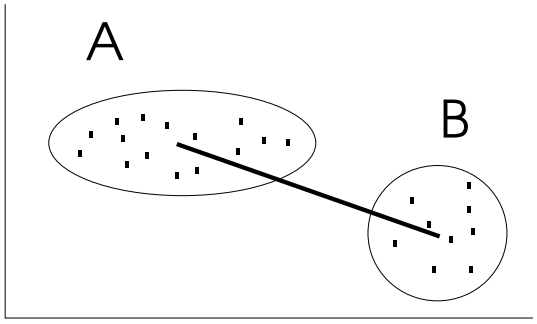
     

     Figura 1.2: Disimilitud entre centroides

     Si se eligiera la distancia de Mahalanobis d_{Mah} como medida de disimilitud, entonces se obtendría:

     $$d_{\mathrm{Mah}}(\mathbf{A},\mathbf{B}) = \sqrt{(\overline{\mathbf{x}}_{\mathbf{A}} - \overline{\mathbf{x}}_{\mathbf{B}})\mathbf{W}^{-1}(\bar{\mathbf{x}}_{\mathbf{A}} - \bar{\mathbf{x}}_{\mathbf{B}})^t} \quad ,$$

donde $\mathbf{W}$ es la matriz ponderada de covarianzas intragrupos de $\mathbf{A}$ i $\mathbf{B}$. Recordemos que si las variables están incorrelacionadas dos a dos, la distancia de Mahalanobis coincide con la euclídea. En todo caso, es necesario suponer a priori que los dos grupos son homogéneos respecto la varianza. Cuando esto no ocurre, Everitt (1993) sugiere el uso de la disimilitud llamada *information radius* de Jardine i Sibson definida por la expresión:

$$R_{\mathbf{AB}} = \log_2 \left(\frac{\det\left(\frac{1}{2}\mathbf{S_B}\right)}{\sqrt{\det(\mathbf{S_A})\det(\mathbf{S_B})}} \right) + \frac{1}{2}\log_2\left(1 + \frac{1}{4}D^2_{\mathbf{AB}}\right),$$

donde

$$D^2_{\mathbf{AB}} = (\overline{\mathbf{x}}_{\mathbf{A}} - \overline{\mathbf{x}}_{\mathbf{B}}) \left(\frac{1}{2}(\mathbf{S_A} + \mathbf{S_B})^{-1}\right) (\overline{\mathbf{x}}_{\mathbf{A}} - \overline{\mathbf{x}}_{\mathbf{B}})^t,$$

y $\mathbf{S_A}$, $\mathbf{S_B}$ son, respectivamente, las matrices de covarianzas dentro de cada grupo.

2. La segunda estrategia para definir una medida entre dos conjuntos de datos se basa en definir la diferencia entre los dos grupos a partir de las disimilitudes interindividuales: *mínima disimilitud, máxima disimilitud, o disimilitud media.* Las tres se basan en la idea que una vez elegida la disimilitud d a utilizar para medir la diferencia entre dos individuos, se define entonces un criterio para calcular la diferencia entre los dos grupos.

   Presentamos a continuación estas tres posibilidades, que son las más usuales, y que a su vez constituyen el núcleo básico de tres de las técnicas de clasificación automática no paramétrica más utilizadas: el método del vecino más próximo, el método del vecino más alejado y el método de la media:

   – Una primera medida de diferencia entre dos grupos $\mathbf{A}$ y $\mathbf{B}$ puede definirse como la disimilitud entre los *vecinos más cercanos* o disimilitud del *mínimo* –véase la figura 1.3: $d_{\min}(\mathbf{A},\mathbf{B}) = \min\{d(\mathbf{x},\mathbf{x}^*) |\ \mathbf{x} \in \mathbf{A}\ ,\ \mathbf{x}^* \in \mathbf{B}\}$.

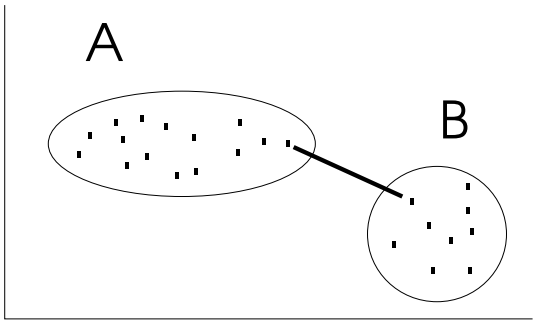


Figura 1.3: Disimilitud entre los vecinos más cercanos.

– Otra posibilidad consiste en definir la disimilitud entre dos grupos como la disimilitud existente entre los *vecinos más alejados* o disimilitud del *máximo* –véase la figura 1.4: $d_{\max}(\mathbf{A},\mathbf{B}) = \max\{d(\mathbf{x},\mathbf{x}^*)|\ \mathbf{x}\in\mathbf{A}\ ,\ \mathbf{x}^*\in\mathbf{B}\}$.

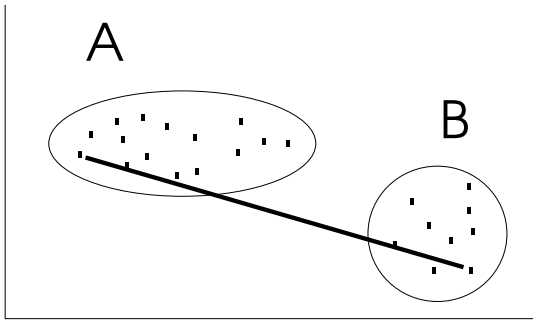


Figura 1.4: Distancia entre los vecinos más alejados.

– Finalmente, la tercera posibilidad consiste en definir la disimilitud entre dos grupos **A** y **B** como la media de las disimilitudes interindividuales o disimilitud *media* –véase la figura 1.5: $d_{\text{media}}(\mathbf{A},\mathbf{B}) = \frac{1}{n_{\mathbf{A}}n_{\mathbf{B}}}\sum_{\mathbf{x}\in\mathbf{A}}\sum_{\mathbf{x}^*\in\mathbf{B}} d(\mathbf{x},\mathbf{x}^*)$, donde $n_{\mathbf{A}}$ y $n_{\mathbf{B}}$ son, respectivamente, el número de individuos de cada grupo.

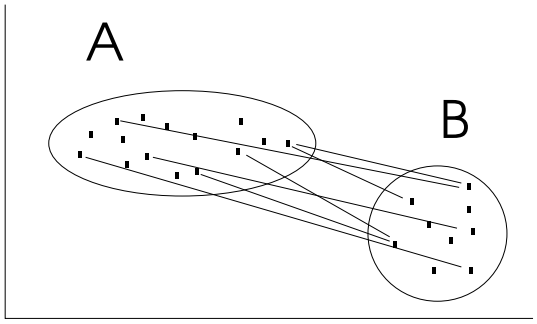


Figura 1.5: Distancias entre individuos.

De hecho, análogamente al caso de las disimilitudes entre individuos, tampoco existe en todos los casos un criterio global que nos permita escoger la disimilitud entre grupos más idónea. Como veremos en la Sección 1.5, diferentes disimilitudes entre grupos producen clasificaciones diferentes, resultando que unas son más adecuadas que otras según la estructura interna existente en los datos.

1.5 Métodos de clasificación no paramétrica

Una vez elegida una medida de diferencia o de similitud con el fin de establecer el grado de disimilitud entre los individuos del conjunto, debe iniciarse la clasificación propiamente dicha.

En esta sección realizamos una exposición somera de los diferentes métodos de clasificación. No consiste en una presentación exhaustiva de todos los métodos existentes ni en una relación de las técnicas más adecuadas, sino en una exposición de las técnicas más utilizadas junto con algunos comentarios sobre su aplicabilidad, sus virtudes y defectos principales. Estos comentarios se acompañan de indicaciones sobre los aspectos a tener en cuenta cuando se desea aplicar uno de estos métodos sobre conjuntos de datos composicionales.

El contenido de esta sección ha sido desarrollado de manera extensa en las monografías de Everitt (1993) y de Kaufman y Rousseuw (1990). También se encuentra un buen desarrollo de las técnicas de clasificación no paramétricas en los libros de Krzanowski (1988b), de Krzanowski y Marriot (1995) y de Jobson (1992).

1.5.1 Técnicas de clasificación: su diferenciación

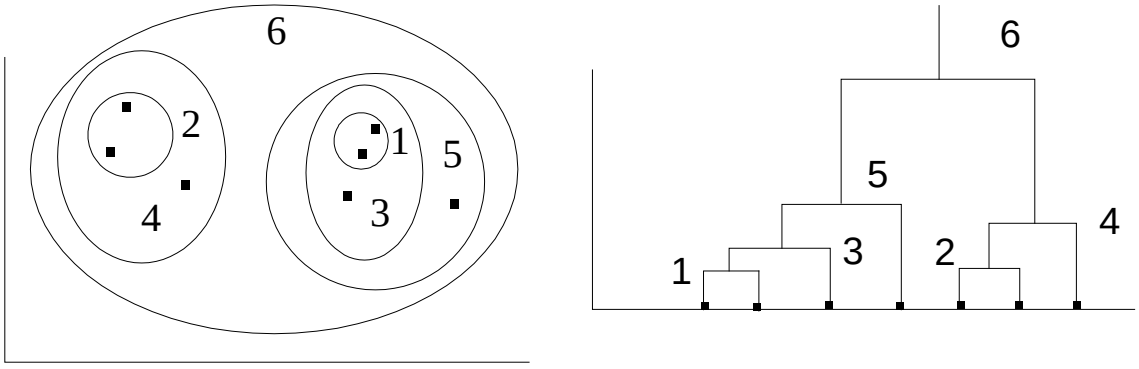
El orden en el que se presentan las diferentes técnicas de clasificación en esta sección está íntimamente ligado a la naturaleza de cada método de clasificación. Por otro lado, somos conscientes que existe la posibilidad de trabajar con técnicas que permiten el solapamiento de grupos, es decir, métodos en los que puede limitarse, o no, el número de grupos diferentes a los que puede pertenecer un mismo elemento. En esta tesis centramos nuestro trabajo en el estudio de métodos de clasificación sin solapamiento, dejando para futuros trabajos de investigación el análisis de las otras técnicas.

Dentro de las técnicas de clasificación sin solapamiento se consideran dos grandes familias: *técnicas jerárquicas* y *técnicas no jerárquicas*. A continuación, y a modo de introducción, se exponen las características más relevantes y los métodos más usuales de cada una de las dos familias anteriores.

- *Técnicas jerárquicas*

  A partir de los n individuos del conjunto $\mathbf{X}$ que forman las observaciones se construye una estructura jerárquica dentro del conjunto $\mathcal{P}(\mathbf{X})$ de partes de $\mathbf{X}$ –véase la figura 1.6(a). Es el usuario de este tipo de técnica el responsable de decidir a qué nivel de la jerarquía construye la partición del conjunto $\mathbf{X}$ de individuos para obtener la clasificación final. Esta estructura de conjuntos de la muestra se suele visualizar mediante un gráfico llamado *dendrograma*, del término griego *dendros* cuyo significado es *árbol* –véase la figura 1.6(b). Se observa, a través de la estructura jerárquica, que cada grupo está, o no, totalmente incluido en otro grupo superior.

Dentro de esta familia de métodos se distingue entre los *jerárquicos aglomerativos o ascendentes* y los *jerárquicos divisivos o descendentes* según si comienzan, respectivamente, a construir la jerarquía desde el nivel en que cada individuo es un grupo o desde el nivel en que sólo existe el grupo conjunto total. Puede observarse que la numeración que se ha utilizado en las figuras 1.6(a) y 1.6(b) representa una estructura obtenida con un método ascendente.



(a) (b)

Figura 1.6: (a) *Estructura jerárquica de subconjuntos;* (b) *Dendrograma correspondiente a la estructura.*

- *Técnicas no jerárquicas*
  - *Grupos disjuntos*

    Cuando se utiliza una técnica de este tipo, en todo el proceso de clasificación, y a pesar de producirse multitud de intercambios, cada individuo pertenece sólo a un grupo. Este tipo de métodos tienen el defecto de necesitar conocer el número de grupos a formar antes de iniciar la técnica. Suelen combinarse con los jerárquicos para realizar las clasificaciones llamadas "*mixtas*".
  - *Particiones estocásticas (mixture models)*

    Esta es la única familia de métodos de clasificación que trabaja los datos mediante el estudio del modelo de distribución de probabilidad que los sustenta. De hecho la estrategia más usada es la que consiste en utilizar modelos de mixtura de distribuciones. En esta tesis no se trata este tipo de técnicas por ser de las denominadas "*paramétricas*".
  - *Métodos de optimización*

El carácter diferenciador de este tipo de técnica consiste en que estos métodos están basados en algoritmos de optimización numérica. Para poder aplicar una técnica de este tipo se necesita conocer a priori el número de grupos a formar. Pueden establecerse equivalencias entre alguno de estos métodos y métodos jerárquicos, y también pueden establecerse conexiones con los métodos de particiones estocásticas.

– *Clasificación borrosa (fuzzy cluster)*

En este tipo de métodos los grupos vienen definidos por un índice de pertenencia de cada elemento a cada grupo. Una vez finalizada la aplicación de una técnica de clasificación borrosa, cada individuo tendrá un índice de pertenencia a cada uno de los grupos. Será el usuario del método el responsable de elegir el umbral de pertenencia o el criterio de asignación de un individuo a un grupo. Los grupos resultantes pueden ser disjuntos, jerárquicos o con solapamiento. El estudio de este tipo de métodos de clasificación supera los objetivos inicialmente marcados para esta tesis. En consecuencia, el análisis de este tipo de métodos será abordado en futuros trabajos de investigación.

1.6 Técnicas de clasificación jerárquicas

Antes de exponer los diferentes métodos de clasificación jerárquicos debe especificarse con detalle el significado de la palabra jerarquía. Para ello es importante recordar que trabajamos con datos de una muestra, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, de n individuos de una población $\mathbf{\Omega}$, sobre los que se han observado D variables y, además, se ha elegido una medida de diferencia d para medir el grado de disimilitud entre los individuos con el objetivo de obtener una clasificación en grupos o subconjuntos pertenecientes a $\mathcal{P}(\mathbf{\Omega})$.

En estas condiciones puede establecerse la siguiente definición de jerarquía:

Definición 1.5 Decimos que $\mathcal{H} \subset \mathcal{P}(\mathbf{\Omega})$ es una jerarquía si se verifica que $\forall\ \mathbf{A},\ \mathbf{B} \in \mathcal{H}$ se cumple una de las tres relaciones siguientes:

$$\mathbf{A} \cap \mathbf{B} = \emptyset, \quad \mathbf{A} \subset \mathbf{B}, \quad \mathbf{B} \subset \mathbf{A}.$$

□

Según la definición anterior, se tiene construida una jerarquía en un conjunto de individuos cuando dados dos grupos cualesquiera $\mathbf{A}$, $\mathbf{B}$ de individuos, o bien un grupo forma una estructura *superior* al otro o bien los dos grupos no tienen ningún individuo en común.

Para clasificar la muestra mediante un método jerárquico se necesita que tanto el conjunto total $\mathbf{X}$ como los individuos $\mathbf{x}_i$ por separado, estén integrados en la jerarquía. Con este objetivo se define el concepto de *jerarquía total*:

Definición 1.6 Decimos que $\mathcal{H}$ es una jerarquía total sobre la muestra $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, si $\mathbf{X} \in \mathcal{H}$ y $\{\mathbf{x}_i\} \in \mathcal{H}$, $\forall\, i = 1, 2, \ldots, n$. □

En la práctica, la jerarquía sobre los datos de la muestra se construye mediante dos familias de procedimientos iterativos diferentes:

- *Algoritmos divisivos o descendentes*
- *Algoritmos aglomerativos o ascendentes*

1.6.1 Algoritmos jerárquicos divisivos

Los algoritmos divisivos, como su nombre indica, consisten en sucesivas particiones de conjuntos. En el primer estado se considera el conjunto total $\mathbf{X}$ (se tiene un solo grupo). Si se ha fijado algún criterio de parada del algoritmo y no se satisface, se procede a dividir el conjunto $\mathbf{X}$, en dos subconjuntos de individuos sin intersección, de acuerdo con un criterio específico. En este nivel de jerarquía tenemos la muestra dividida en dos clases. Si se considera conveniente, se dividen a su vez cada una, o una sola, de las clases. Se repite el proceso hasta que, como máximo, cada individuo de la muestra forme un grupo (se tienen n clases). Los criterios de parada, si existen, y los criterios de partición, generalmente basados en la disimilitud elegida, son los que distinguen un método divisivo de otro. Visualmente, los algoritmos divisivos construyen la estructura jerárquica de arriba a abajo (empiezan por la parte alta del dendrograma y van descendiendo).

Entre los algoritmos divisivos se considera que hay de dos tipos: los *monotéticos* y los *politéticos*. Los algoritmos monotéticos, usados con conjuntos de datos binarios, basan las particiones que realizan en la presencia/ausencia de una característica en los individuos. Los algoritmos politéticos, más usados con conjuntos de datos cuantitativos, tienen en cuenta todas las variables observadas.

- Datos binarios

  Para cada par de variables $\mathbf{X}_k$ y $\mathbf{X}_m$, de las D características observadas, se considera la tabla de contingencia que aparece en la tabla 1.5.

Tabla 1.5: Tabla de contingencia entre las variables $\mathbf{X}_k$ y $\mathbf{X}_m$

		$\mathbf{X}_m$		Total
		1	0	
$\mathbf{X}_k$	1	a_{km}	b_{km}	indiv. con carac. $\mathbf{X}_k$
	0	c_{km}	d_{km}	indiv. sin carac. $\mathbf{X}_k$
Total		indiv. con carac. $\mathbf{X}_m$	indiv. sin carac. $\mathbf{X}_m$	n

A partir de las frecuencias que aparecen en la tabla 1.5 se calcula el estadístico de independencia

$$\chi^2(\mathbf{X}_k, \mathbf{X}_m) = \frac{(a_{km}d_{km} - b_{km}c_{km})^2 n}{(a_{km} + b_{km})(a_{km} + c_{km})(b_{km} + d_{km})(c_{km} + d_{km})}.$$

Para cada variable $\mathbf{X}_k$ se calcula

$$\sum_{m \neq k} \chi^2(\mathbf{X}_k, \mathbf{X}_m).$$

El criterio de partición consiste en dividir los individuos por presencia/ausencia en la variable $\mathbf{X}_k$ que posea mayor valor en la suma antes calculada, es decir, con más *dependencia estocástica* de las otras, puesto que se entiende que ésta es la variable que más diferencia entre si a los individuos de la muestra. Se repite sucesivamente el proceso, en cada conjunto por separado, excluyendo la variable utilizada en la partición anterior.

La propia naturaleza de los datos nos hace concluir que este método de clasificación no es una herramienta adecuada para llevar a cabo una agrupación de datos composicionales.

- Datos cuantitativos

  En este caso, a pesar que usualmente suele escogerse la distancia euclídea, puede utilizarse cualquier disimilitud, d, entre individuos.

  En el primer paso del algoritmo, se calcula la media de las disimilitudes de cada individuo a los otros, $\bar{d}_i = \frac{1}{n}\sum_{j \neq i} d_{ij}$. Se escoge el individuo, supongamos $i = 1$, que tiene la mayor de las medias y se separa para formar un grupo él solo. A continuación, dentro del grupo que tiene $n-1$ individuos, se calcula la media de las disimilitudes de cada individuo a los otros del grupo, $\bar{d}_i = \frac{1}{n-1}\sum_{j \neq i,1} d_{ij}$, y se resta de la disimilitud al elemento separado: $\bar{d}_i - d_{i1}$. Si todas estas diferencias son negativas el elemento $\mathbf{x}_1$ forma un grupo él solo y se procede a partir la otra clase comenzando el algoritmo otra vez. Si alguna diferencia es positiva se escoge el individuo que proporciona la mayor diferencia y se asigna al grupo del individuo

$\mathbf{x}_1$. De este modo, se obtienen dos grupos uno con dos individuos y otro con el resto. Para cada individuo del grupo con $n-2$ individuos se calculan las medias de las disimilitudes entre individuos y la media de las disimilitudes a los dos individuos del otro grupo. Se restan las dos medias y, si todas las diferencias son negativas se considera que tenemos una clase de dos individuos y se procede a dividir el grupo de $n-2$ individuos iniciando el algoritmo sobre él. Si existen diferencias positivas se asigna al grupo de los 2 individuos la observación que proporciona diferencia positiva máxima. Se sigue el procedimiento hasta obtener dos grupos para los que todas las diferencias de medias respectivas son negativas. A partir de aquí se procede a dividir cada uno de los dos grupos por separado iniciando el algoritmo sobre cada uno de ellos.

Recordemos que los datos composicionales pueden tratarse, respecto a su espacio soporte, como datos de tipo continuo. En consecuencia, con este tipo de datos podemos utilizar este algoritmo. En este caso, obsérvese que el elemento clave es la elección de una disimilitud d que sea adecuada para los datos composicionales. La cuestión de cuáles son las características que debe poseer una disimilitud d para poder ser considerada adecuada será analizada en profundidad en el Capítulo 2 de esta tesis.

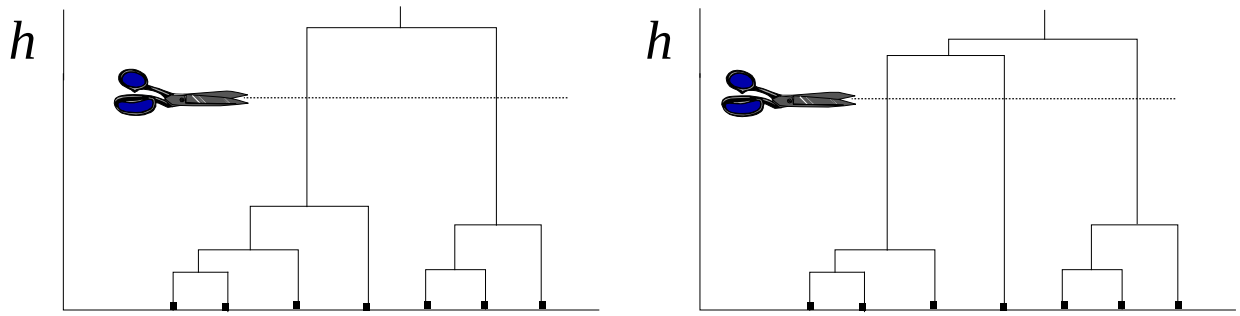
1.6.2 Algoritmos jerárquicos aglomerativos

Los algoritmos aglomerativos son en su mayoría de tipo *politético*. Al contrario de lo que sucede en los algoritmos divisivos, las técnicas jerárquicas aglomerativas consisten en sucesivas uniones de conjuntos siguiendo un criterio específico. En el primer estado se consideran n conjuntos o clases cada uno de ellos compuesto por un solo individuo. Si no existe limitación o si se ha especificado un criterio de parada y éste no se cumple, dos de estos grupos individuales se unen, siguiendo un criterio específico, para formar un solo grupo que a partir de ahora se considerará como un solo *individuo*. Si existe, se vuelve a evaluar el criterio de parada y, se unen los dos individuos, de los $n-1$ que quedan, que indique el criterio de clasificación. Se itera el algoritmo hasta que se verifique el criterio de parada, en caso de que exista o, como máximo, hasta que quede un solo individuo: el conjunto total $\mathbf{X}$. Gráficamente puede observarse que un algoritmo aglomerativo construye el dendrograma de abajo a arriba, siendo recomendable su uso cuando se prevea un número grande de clases. Un criterio de parada habitualmente usado es el de imponer un mínimo de grupos en la clasificación. De esta manera sólo se construye el dendrograma hasta el nivel de jerarquía que los contemple y se evita que el coste computacional sea grande.

El funcionamiento de estos algoritmos permite establecer una medida, h, de la jerarquía

establecida en el conjunto de nuestros datos. Esta medida h, denominada índice de jerarquía, representa el nivel de similitud o de disimilitud al que se realiza la unión de dos clases. El estudio de este tipo de índices es muy útil en el sentido que permiten detectar en que momento la jerarquía que se está estableciendo en los datos empieza a tener una componente de artificialidad elevada. Es por ello que se utiliza para determinar de una manera totalmente exploratoria el número de grupos que forman la clasificación. Este índice de jerarquía, h, se sitúa en el eje vertical del dendrograma.

Se puede observar como, según a que nivel se produzca el salto en el índice de jerarquía, se decide la existencia de un número u otro de clases (*corte del árbol*). Por ejemplo, en el primer caso del dendrograma de la figura 1.7(a), se supondrá que existen dos grupos. Por el contrario, en el dendrograma de la figura 1.7(b) se considerará que existen tres.



(a) (b)

Figura 1.7: (a) *Dendrograma indicando dos grupos;* (b) *Dendrograma indicando tres grupos.*

Esta herramienta tan útil se define por la siguiente expresión:

Definición 1.7 Decimos que h es un índice de jerarquía sobre una jerarquía $\mathcal{H}$ si es una función real no negativa tal que $h(\mathbf{A}) \leq h(\mathbf{B})$ siempre que $\mathbf{A} \subset \mathbf{B}$. Además, se tiene que $h(\{\mathbf{x}_i\}) = 0$ para todo individuo $\mathbf{x}_i$ de la muestra. □

Se puede establecer una relación directa entre el índice de jerarquía y la disimilitud elegida como medida de diferencia entre los individuos de la muestra. De hecho, la relación que liga estos dos elementos de la clasificación es la que permite distinguir entre un método u otro, entre los algoritmos jerárquicos ascendentes. El proceso de clasificación mediante un algoritmo jerárquico ascendente puede resumirse esquemáticamente en los siguientes pasos:

1. Elección de la disimilitud a aplicar entre individuos y entre grupos. Cálculo de la matriz $n \times n$ de disimilitudes entre individuos. En este estado se tienen n grupos (los individuos).

2. Unión en un solo grupo de los dos grupos cuya disimilitud sea mínima. Esta disimilitud es el nivel de jerarquía del grupo unión. A este grupo se le considera como uno solo. El número de grupos o clases disminuye en uno. Si el número de grupos es igual a 1 o al número mínimo deseado, se para el algoritmo.

3. Cálculo de la nueva matriz de disimilitudes siguiendo el criterio elegido en la disimilitud entre grupos.

4. Se repiten los dos pasos anteriores hasta finalizar.

En el segundo paso, donde se realiza la elección de los grupos cuya disimilitud es mínima, pueden presentarse empates. Este hecho, que se da con mayor frecuencia cuando se trabaja con variables cualitativas que con variables cuantitativas, supone en la práctica que la clasificación obtenida no puede considerarse única. Una estrategia para salvar este escollo consiste en reordenar los datos y volverlos a clasificar. Entonces, se comparan los resultados obtenidos y, si es factible, se decide la estructura definitiva de la clasificación.

A continuación se presenta una breve descripción de los métodos aglomerativos más utilizados.

- Método del mínimo

  Este método también es conocido con los nombres, en inglés, *single linkage, minimum distance, nearest neighbour*, y, *connectedness method*. Estas diferentes maneras de denominar al método no son más que intentos de recoger en el nombre su particularidad. En este método se establece la disimilitud entre dos clases $\mathbf{C}_k$ y $\mathbf{C}_l$, como la disimilitud entre los vecinos más próximos o disimilitud mínima entre los grupos:

  $$d_m(\mathbf{C}_k, \mathbf{C}_l) = \min_{\mathbf{x}_i \in \mathbf{C}_k; \mathbf{x}_j \in \mathbf{C}_l} \{d(\mathbf{x}_i, \mathbf{x}_j)\}.$$

  El nivel de jerarquía del grupo unión o nivel de fusión es:

  $$h_m(\mathbf{C}_v \cup \mathbf{C}_w) = \min_{k \neq l} \{d_m(\mathbf{C}_k, \mathbf{C}_l)\}.$$

  El método del mínimo tiene tendencia a construir grupos con forma alargada. Este hecho, que puede ser una virtud en determinados casos, produce el efecto de encadenamiento artificial o clases unidas por puentes. Por lo tanto, no se recomienda la aplicación de este método a conjuntos de datos en los que se adivinen clases muy cercanas.

- Método del máximo

  Análogamente al método anterior, el método del máximo puede encontrarse en la bibliografía con diferentes nombres: *complete linkeage, maximum distance, farthest neighbor*, y, *diameter method*. La disimilitud elegida para medir la disimilitud entre grupos es la de los vecinos más alejados o disimilitud máxima:

  $$d_M(\mathbf{C}_k, \mathbf{C}_l) = \max_{\mathbf{x}_i \in \mathbf{C}_k; \mathbf{x}_j \in \mathbf{C}_l} \{d(\mathbf{x}_i, \mathbf{x}_j)\}.$$

  El nivel de fusión queda fijado uniendo las clases más cercanas:

  $$h_M(\mathbf{C}_v \cup \mathbf{C}_w) = \min_{k \neq l} \{d_M(\mathbf{C}_k, \mathbf{C}_l)\}.$$

  Este método tiende a formar clases muy compactas, de reducido diámetro. Debido a la disimilitud que se utiliza, sólo se asignarán al mismo grupo en las primeras iteraciones elementos muy cercanos. Se recomienda su uso cuando en la muestra se encuentran clases muy poco separadas. Si entre los datos existen individuos alejados del resto o atípicos (en inglés, *outliers*), puede desvirtuarse la clasificación.

- Método de la media o *average linkeage method*

  Si se desea una opción intermedia entre los dos métodos anteriores puede escogerse este método. En este caso la disimilitud entre dos clases es la media aritmética de las disimilitudes entre sus individuos. Para su cálculo usaremos la fórmula siguiente:

  $$d_a(\mathbf{C}_k, \mathbf{C}_l) = \frac{1}{n_k n_l} \sum_{\mathbf{x}_i \in \mathbf{C}_k} \sum_{\mathbf{x}_j \in \mathbf{C}_l} d(\mathbf{x}_i, \mathbf{x}_j),$$

  donde n_k y n_l son, respectivamente, el número de individuos de cada grupo.

  El nivel de jerarquía de la unión viene dado por la expresión siguiente:

  $$h_a(\mathbf{C}_v \cup \mathbf{C}_w) = \min_{k \neq l} \{d_a(\mathbf{C}_k, \mathbf{C}_l)\}.$$

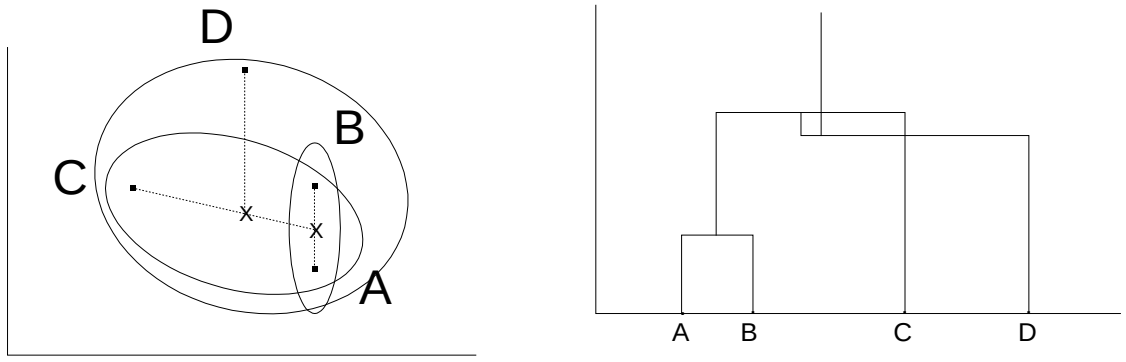
  Este método, como el del máximo, no es recomendable frente a conjuntos de datos que contengan datos claramente alejados del resto.

Los tres métodos anteriores tienen en común que operan directamente sobre la matriz $n \times n$ de disimilitudes entre individuos de la muestra, (d_{ij}), sin precisar de la matriz de datos $\mathbf{X}$. También poseen la propiedad de la *invarianza monótona*. Esta propiedad significa que si se aplica uno de estos métodos sobre la matriz (d_{ij}) o sobre la matriz $(\hat{d}_{ij})$ resultante de realizar una transformación monótona sobre (d_{ij}), se obtiene la misma clasificación. Ello hace que,

por ejemplo, se obtengan las mismas clases usando la distancia euclídea o usando la distancia euclídea al cuadrado.

Nótese que es factible escoger cualquiera de estos tres métodos para realizar una clasificación de datos composicionales. Análogamente al caso del método jerárquico divisivo, en la aplicación de estos tres métodos únicamente debe tenerse en cuenta que la disimilitud d entre individuos sea una medida adecuada para los datos composicionales.

Por otro lado, al realizar una clasificación es posible que el nivel de fusión en una determinada iteración sea inferior al nivel de fusión de reuniones anteriores. Si esto sucede, se dice que se ha producido una *inversión* (figuras 1.8(a) y 1.8(b)).



(a) (b)

Figura 1.8: (a) *Clasificación con una inversión;* (b) *Dendrograma con una inversión.*

Este efecto es un serio inconveniente en el momento de la interpretación de la estructura jerárquica obtenida al clasificar. Ninguno de los tres métodos anteriores produce inversiones, pero sí pueden producirse inversiones en una clasificación realizada con el método siguiente:

- Método del centroide

  Este método aplica un tipo de estrategia diferente a la que aplican los métodos anteriores en relación a la medida de disimilitud entre clases que se utiliza. La idea consiste en representar cada clase, $\mathbf{C}_k$, por su centro de gravedad, punto medio o *centroide*, que se representa por el vector de D componentes, $\bar{\mathbf{x}}_k$. Entonces, la disimilitud entre grupos se mide por la disimilitud entre centroides, es decir:

$$d_c(\mathbf{C}_k, \mathbf{C}_l) = d(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_l).$$

El nivel de fusión entre dos clases queda fijado en la expresión siguiente:

$$h_c(\mathbf{C}_v \cup \mathbf{C}_w) = \min_{k \neq l}\{d_c(\mathbf{C}_k, \mathbf{C}_l)\}.$$

La propia naturaleza del método hace su uso aconsejable en el caso de datos cuantitativos. A pesar de ello, también se utiliza con los otros tipos de datos. El hecho de trabajar con los centroides de las clases de individuos implica la necesidad de actuar sobre la matriz de datos $\mathbf{X}$ y el cálculo, en cada iteración, de una fila o columna de la matriz de disimilitudes. Este método no posee la propiedad de la invarianza monótona y tiene el grave defecto de poder producir inversiones. Debido a ello, el nivel de fusión de la unión de dos clases no es un índice de jerarquía tal y como se ha definido. Su comportamiento no queda afectado al trabajar con conjuntos de datos con datos anormalmente alejados del resto.

Análogamente a los métodos descritos anteriormente, si se desea utilizar este método para realizar una clasificación de datos composicionales debe elegirse una disimilitud adecuada. Sin embargo, a diferencia de los métodos analizados hasta ahora, la aplicación del método del centroide requiere, además, la elección de una medida de tendencia central que sea coherente con la naturaleza de los datos composicionales. Esta cuestión se analiza en profundidad en el Capítulo 2 de esta tesis.

- Método de Ward

  A pesar de que este método se asemeja a los anteriores en el hecho de construir una estructura jerárquica, difiere de todos ellos en su planteamiento. Se basa en la idea, presente en los métodos de optimización, que una buena clasificación significa establecer clases *heterogéneas* entre si, y, que cada clase esté compuesta por un conjunto *homogéneo* de individuos, es decir, que la varianza dentro del grupo sea mínima. El algoritmo parte de un estado inicial donde se tienen n grupos, cada uno formado por un solo individuo y, por tanto, homogéneos totalmente. A partir de aquí, se realizan las fusiones con la idea de perder la mínima homogeneidad. Este planteamiento comporta que el uso del método se restrinja a datos cuantitativos. Es más, exige la existencia de una escala métrica de los datos.

  Se define la variabilidad dentro de la clase $\mathbf{C}_k$ como:

  $$V_k = \sum_{i \in \mathbf{C}_k} ||\mathbf{x}_i - \bar{\mathbf{x}}_k||^2,$$

  donde $\bar{\mathbf{x}}_k$ representa el centroide de la clase y $||\cdot||$ la norma definida sobre los datos.

La homogeneidad de una partición en G clases la calculamos mediante la siguiente expresión:

$$H_G = \sum_{k=1}^{G} V_k.$$

Si en una iteración del método pretendemos la fusión de dos clases v y w, la homogeneidad de las $G-1$ clases resultantes es:

$$H_{G-1} = \sum_{k \neq v,w} V_k \quad + \quad V_{\{\mathbf{C}_v \cup \mathbf{C}_w\}}.$$

Se puede demostrar que la pérdida de homogeneidad que ello conlleva viene dada por la expresión siguiente:

$$H_{G-1} - H_G = \frac{n_v n_w}{n_v + n_w} ||\bar{\mathbf{x}}_v - \bar{\mathbf{x}}_w||^2.$$

Es precisamente esta pérdida de homogeneidad el valor que el método Ward utiliza como medida de disimilitud entre clases. Es decir, se considera que

$$d_W(\mathbf{C}_v, \mathbf{C}_w) = \frac{n_v n_w}{n_v + n_w} ||\bar{\mathbf{x}}_v - \bar{\mathbf{x}}_w||^2,$$

es la disimilitud entre las dos clases $\mathbf{C}_v$ y $\mathbf{C}_w$. Puede observarse que ésta es función de los centroides de los dos grupos. Se fija el nivel de fusión mediante la expresión:

$$h_W(\mathbf{C}_v \cup \mathbf{C}_w) = \min_{k \neq l} \{d_W(\mathbf{C}_k, \mathbf{C}_l)\}.$$

Por lo tanto, se realiza la unión de grupos que minimiza la pérdida de homogeneidad.

Al igual que la mayoría de métodos jerárquicos, el de Ward tiende a crear clases con forma *esférica*, lo que es un inconveniente delante de grupos con otras formas como, por ejemplo, las elípticas de las poblaciones normales multivariantes. Sin la intención de despreciar los otros métodos, puede afirmarse que el método de Ward, al igual que el de la media, es recomendable para un primer intento en el análisis exploratorio de la estructura de la muestra. Sin embargo, hay que tener presente que en presencia de datos atípicos su comportamiento es peor que el método del mínimo y el del centroide.

Si se desea utilizar el método de Ward para clasificar conjuntos de datos composicionales deberá tenerse en cuenta que el concepto de variabilidad de un conjunto de datos es el elemento clave de este método de clasificación. Esta cuestión, que se analiza en detalle en el Capítulo 2 de esta tesis, ha sido motivo de estudio en Martín-Fernández et al. (1998b).

- Método flexible

  En realidad el método flexible no es un método como tal si no más bien la generalización de todos los métodos jerárquicos aglomerativos anteriores. Se basa en la fórmula recurrente introducida por Lance y Williams (1967) que define la disimilitud entre una clase $\mathbf{C}_k$ y la clase resultante de fusionar otras dos clases $\mathbf{C}_v \cup \mathbf{C}_w$. La fórmula propuesta es la siguiente:

$$d(\mathbf{C}_k, \mathbf{C}_v \cup \mathbf{C}_w) = \alpha_v d(\mathbf{C}_k, \mathbf{C}_v) + \alpha_w d(\mathbf{C}_k, \mathbf{C}_w) + \beta d(\mathbf{C}_v, \mathbf{C}_w) + \gamma |d(\mathbf{C}_k, \mathbf{C}_v) - d(\mathbf{C}_k, \mathbf{C}_w)|, \tag{1.10}$$

  donde d representa una disimilitud definida entre grupos.

  Una vez calculadas las disimilitudes entre todos los grupos, se fusionan las clases v, w que presentan mínima disimilitud. El nivel de jerarquía se hace igual a esta disimilitud mínima. Entonces se usa la fórmula (1.10) para calcular de manera eficiente la nueva matriz de disimilitudes entre los grupos existentes después de la fusión.

  Lance y Williams (1967) sugieren que el método flexible de clasificación se formule con valores de los parámetros que cumplan las restricciones siguientes:

$$\alpha_v + \alpha_w + \beta = 1 \; ; \alpha_v = \alpha_w \; ; \beta < 1 \; ; \gamma = 0.$$

  Por ejemplo, los autores proponen elegir valores negativos pequeños para el parámetro β y señalan que $\beta = -0.25$ es un valor útil en las clasificaciones.

  Utilizando la fórmula de recurrencia (1.10) pueden generarse los métodos jerárquicos expuestos préviamente en esta sección. En la tabla 1.6 se muestran los valores que nos proporcionan cada uno de los diferentes métodos. En general, en la tabla 1.6 se considera

Tabla 1.6: Valores de los coeficientes α_v, α_w, β, y γ en la fórmula (1.10).

método	α_v	α_w	β	γ
mínimo	$1/2$	$1/2$	0	$-1/2$
máximo	$1/2$	$1/2$	0	$1/2$
media	$\frac{n_v}{n_v+n_w}$	α_v	0	0
centroide	$\frac{n_v}{n_v+n_w}$	α_v	$-\alpha_v\alpha_w$	0
ward	$\frac{n_k+n_v}{n_k+n_v+n_w}$	$\frac{n_k+n_w}{n_k+n_v+n_w}$	$\frac{-n_k}{n_k+n_v+n_w}$	0

  que para los métodos del mínimo, del máximo y de la media, la expresión $d(\mathbf{C}_v, \mathbf{C}_w)$ representa tanto una disimilitud como una similitud. Sin embargo, cuando en la tabla 1.6 se hace referencia al método del centroide, en la expresión $d(\mathbf{C}_v, \mathbf{C}_w)$ se considera la distancia

euclídea, y para el método de Ward se considera la distancia euclídea al cuadrado. Esta limitación para el método del centroide y para el método de Ward se deriva del supuesto que el método flexible se utiliza para clasificar datos cuyo espacio soporte es $\mathbb{R}^D$ y que, en este espacio, se tiene definida una estructura de espacio métrico inducida por la distancia euclídea. Si se desea utilizar el método flexible para conjuntos de datos composicionales deberá tenerse en cuenta que el concepto de espacio métrico es un elemento fundamental. La estructura de espacio métrico sobre el espacio soporte de los datos composicionales, que se analiza en el Capítulo 2 de esta tesis, ha sido estudiada de manera extensa en Barceló-Vidal (2000).

Por otro lado, es importante resaltar que, exceptuando el método del centroide, el resto de métodos poseen la propiedad de la invarianza monótona (Cuadras, 1991). Esta propiedad se utiliza en el Capítulo 3 de esta tesis cuando se analiza la relación entre las dos medidas de diferencia para datos composicionales que se presentan.

1.7 Técnicas de clasificación no jerárquicas

1.7.1 Grupos disjuntos

Los métodos que se exponen en esta sección consisten en agrupar los individuos de la muestra en una clasificación simple formada por G grupos sin ningún elemento en común, donde el número de grupos a obtener, G, es conocido a priori. El proceso común de estos métodos se inicia en la elección de una primera partición de los individuos y continúa con toda una serie de intercambios de sus miembros con el objetivo de encontrar una partición mejor. Los aspectos que distinguen entre un método u otro son el modo cómo eligen la partición inicial y qué se entiende por *partición mejor*. El hecho de no necesitar calcular y guardar la matriz de similitudes o disimilitudes hace más recomendables estos métodos que los jerárquicos cuando se trabaja con una muestra con un número elevado de individuos.

Los métodos que siguen este planteamiento se encuentran en la bibliografía bajo los nombres, entre otros, de *centros móviles*, k-medias (en inglés *k-means*), de Forgy, y clasificaciones rápidas (en inglés *quickcluster* o *fastcluster*). En general, en todos ellos, la manera de construir la partición inicial pasa por la elección de G individuos de la muestra o por generar G datos que sean los representantes de los G grupos iniciales. A partir de aquí se asigna cada individuo de la muestra al grupo cuyo representante le es más próximo. Pueden elegirse multitud de maneras diferentes de establecer los G representantes iniciales. A continuación exponemos las

más utilizadas:

- Escoger los G primeros individuos de la muestra.
- Escoger los individuos situados en las posiciones $\frac{n}{G}$, $\frac{2n}{G}$, ... , $\frac{(G-1)n}{G}$, y, n.
- Escoger G individuos de la muestra mediante el uso de números aleatorios.
- Generar aleatoriamente G observaciones. El valor aleatorio de cada componente de una observación se obtiene mediante un número aleatorio perteneciente al rango de la variable correspondiente.
- Con el objetivo de encontrar G representantes iniciales que estén separados entre si pero que tengan bastantes individuos cerca de ellos, se intentan elecciones basadas en la idea de *densidad*. Una de las más simples consiste en elegir como primer representante el centroide $\bar{\mathbf{x}}$ de toda la muestra y, mediante un recorrido por los datos, elegir los otros representantes con la condición que estén como, mínimo, a una disimilitud δ de los representantes ya elegidos. Este método es lo suficientemente sencillo como para permitir intentar dos o tres valores de δ hasta conseguir una primera partición adecuada.

Una vez se ha establecido de algún modo la partición inicial, el proceso que se sigue consiste en calcular el centroide de cada grupo y reasignar cada individuo al grupo cuyo centroide le queda más próximo. Se itera este proceso hasta que no deba realizarse ningún cambio, es decir, hasta que todas las clases sean estables.

Este proceso se conoce como *método de Forgy*. Se observa que, en cada iteración y recolocación de los individuos los centroides permanecen fijos. Existe una variación introducida en MacQueens's (1967) en la que se siguen los siguientes pasos:

1. Escoger los G primeros individuos de la muestra como representantes de los G grupos iniciales.
2. Asignar el resto de individuos al grupo cuyo representante le sea más cercano. Después de cada asignación recalcular el centroide del grupo.
3. Cuando se han asignado todos los individuos de la muestra se considera a los centroides de los grupos como puntos fijos y se reasigna cada individuo al grupo cuyo centro le es más cercano. Este paso es similar al de una iteración del método de Forgy.

Se observa que la gran diferencia consiste en que después de cada asignación de un individuo a un grupo se recalcula su centroide. Se recomienda usar este método pero con la siguiente mejora:

recalcular el centroide del grupo donde ha sido asignado el individuo y, también, el centroide del grupo del cual se ha cogido el individuo.

La virtud de estos métodos es tener un coste computacional inferior a los jerárquicos, aspecto a tener en cuenta cuando la muestra es de gran tamaño. Un defecto que poseen es que la clasificación obtenida depende de la partición inicial elegida. Pero el gran handicap radica en la necesidad de conocer a priori el número de grupos que se quieren formar. Se recomienda, por tanto, un estudio previo del número de grupos de la muestra mediante las técnicas que se describen en la sección siguiente. Existen otras alternativas para salvar la dificultad de desconocer a priori el número de clases. Una de ellas consiste en realizar una clasificación con un valor de G muy elevado y a partir de las G clases obtenidas aplicar un algoritmo de clasificación jerárquico que permiten decidir el número de grupos de la muestra. Otra estrategia es la denominada *clasificación mixta*. Esta técnica consiste en realizar unas pocas (3 o 4) clasificaciones iniciando el algoritmo con particiones iniciales diferentes y un número de grupos G pequeño. Las clases obtenidas en las diferentes clasificaciones se intersecan para obtener así un número elevado de clases (como mucho G^3 o G^4, respectivamente). Estas clases serán el punto de partida de un método jerárquico que permitirá, mediante el nivel de fusión, decidir el número de clases de la muestra.

Análogamente al método del centroide, si se desea utilizar el método de grupos disjuntos para realizar una clasificación de datos composicionales debe elegirse una medida de tendencia central y una disimilitud que sean coherentes. En el Capítulo 2 de esta tesis se exponen los requisitos para que una medida de diferencia sea adecuada y se presenta una medida de tendencia central coherente con el carácter composicional de los datos.

El siguiente ejemplo ilustra el funcionamiento de estas técnicas de clasificación no jerárquicas:

Ejemplo 1.3 La muestra contiene 8 elementos de los que se han observado 2 características obteniéndose los datos siguientes:

$$\mathbf{x}_1 = (1, 1.5),\ \mathbf{x}_2 = (5, 5),\ \mathbf{x}_3 = (2, 3),\ \mathbf{x}_4 = (6, 4),$$

$$\mathbf{x}_5 = (3, 2),\ \mathbf{x}_6 = (2, 0.75),\ \mathbf{x}_7 = (7, 4),\ \mathbf{x}_8 = (6, 3).$$

Se desea obtener una clasificación de los elementos en $G = 2$ grupos. Inicialmente, escogemos al azar los dos primeros centros de cada grupo:

$$\bar{\mathbf{x}}_1^1 = \mathbf{x}_6 = (2, 0.75) \quad \text{y} \quad \bar{\mathbf{x}}_2^1 = \mathbf{x}_5 = (3, 2),$$

donde el superíndice indica que son los centroides de la primera iteración del algoritmo.

Se asigna, sin recalcular el centro, cada individuo al grupo cuyo centro le es más cercano y se obtienen los grupos siguientes: $\mathbf{C}_1^1 = \{\mathbf{x}_1, \mathbf{x}_6\}$, $\mathbf{C}_2^1 = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_7, \mathbf{x}_8\}$.

Una vez calculados los centroides de estos grupos –puntos simbolizados por círculos '$\circ$' en la figura 1.9(a)– , se reasignan, sin recalcular los centros, los individuos al grupo cuyo centroide le es más cercano –véase la figura 1.9(b). Geométricamente, la reasignación pasa por observar a qué semiespacio pertenece cada punto, de los dos semiespacios creados por el hiperplano mediatriz de los dos centros. A continuación se iteran estos pasos hasta no obtener ninguna reasignación –véase la figura 1.9(c). La clasificación acaba en 3 iteraciones y se obtienen los dos grupos

$$\mathbf{C}_1^3 = \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_6\}, \quad \mathbf{C}_2^3 = \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_7, \mathbf{x}_8\}.$$

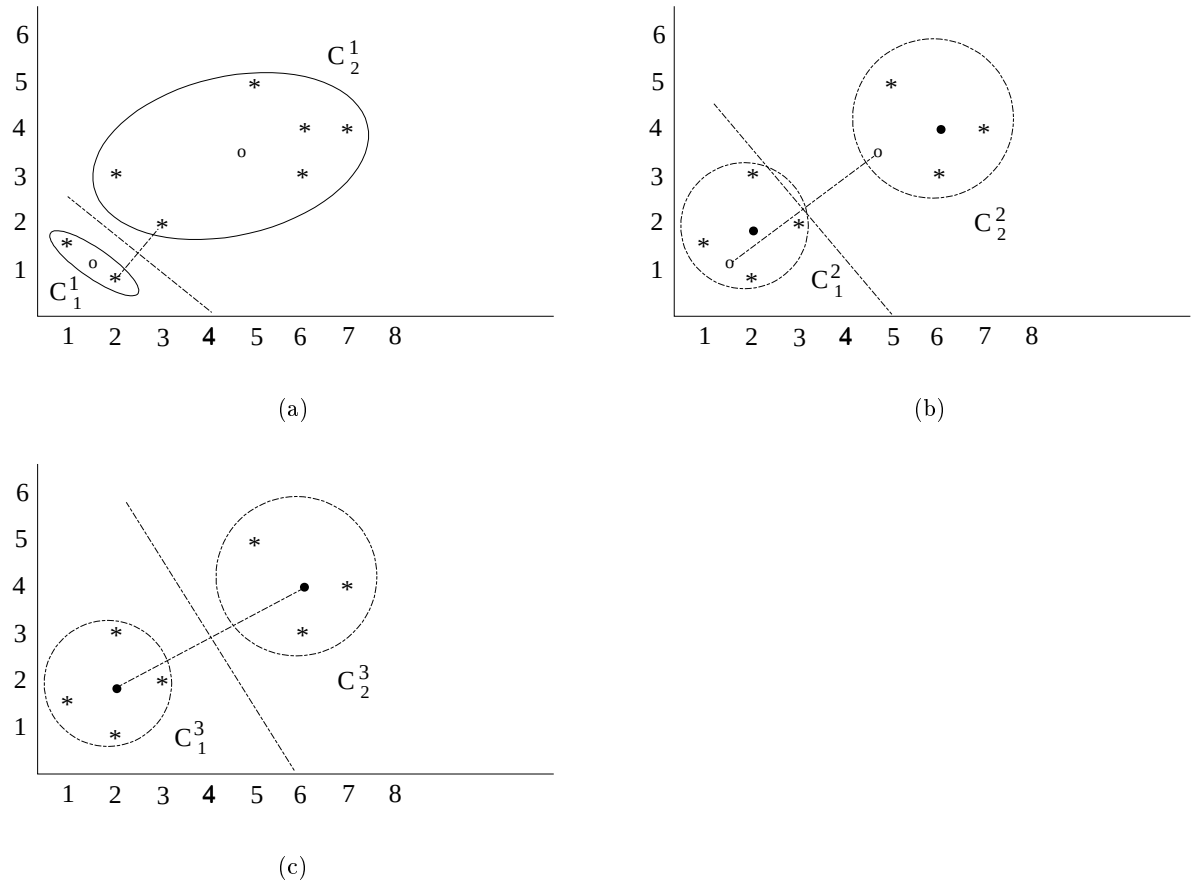


Figura 1.9: Ejemplo de clasificación automática mediante una técnica de grupos disjuntos: (a) *grupos iniciales y sus centros '$\circ$';* (b) *grupos en la segunda iteración y sus centros '$\bullet$';* (c) *grupos en la última iteración y sus centros '$\bullet$'.*

□

1.7.2 Métodos de optimización

Los métodos de clasificación que se consideran en esta sección tienen en común el hecho de producir una partición de los individuos mediante la maximización o la minimización de algún criterio numérico. La estructura resultante de aplicar estos métodos no se asemeja a la obtenida mediante los algoritmos jerárquicos sino más bien a la estructura formada por los métodos de grupos disjuntos y los métodos de particiones estocásticas. Es decir, los grupos resultantes de la clasificación no contienen subgrupos estructurados jerárquicamente. De hecho, los procedimientos numéricos propuestos en estos métodos se relacionan con los criterios obtenidos para el caso de la distribución normal multivariante en los métodos de particiones estocásticas. El número de grupos G en que se divide la muestra se debe conocer a priori, por lo que se sugiere la realización de un estudio previo del número de clases a obtener.

En el espíritu de todos estos métodos se encuentra la idea de establecer grupos heterogéneos entre si y homogéneos en su interior. Precisamente, es según el criterio de medida de la homogeneidad utilizado cómo se distinguen los métodos entre si. De la multitud de criterios diferentes que existen para el caso de variables cuantitativas, los más utilizados se basan en el estudio de la homogeneidad de los grupos mediante las tres matrices siguientes:

- Matriz de dispersión total:

$$\mathbf{T} = \frac{1}{n}\sum_{k=1}^{G}\sum_{i=1}^{n_k}(\mathbf{x}_{ik} - \bar{\mathbf{x}})^t(\mathbf{x}_{ik} - \bar{\mathbf{x}});$$

- Matriz de dispersión *intra-grupos* (en inglés **W***ithin-group*):

$$\mathbf{W} = \frac{1}{n-G}\sum_{k=1}^{G}\sum_{i=1}^{n_k}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)^t(\mathbf{x}_{ik} - \bar{\mathbf{x}}_k);$$

- Matriz de dispersión *entre-grupos* (en inglés **B***etween-group*):

$$\mathbf{B} = \sum_{k=1}^{G} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^t(\mathbf{x}_k - \bar{\mathbf{x}}),$$

donde G es el número de grupos a formar, n el número de individuos de la muestra, n_k el número de individuos del grupo k-ésimo, $\bar{\mathbf{x}}$ el centroide de la muestra, $\bar{\mathbf{x}}_k$ el centroide del k-ésimo grupo, y, $\mathbf{x}_{ik}$ es el vector fila que recoge el valor de las D variables para el i-ésimo individuo de la k-ésima clase.

Estas tres matrices, de dimensión $D \times D$, satisfacen la ecuación siguiente:

$$\mathbf{T} = \mathbf{W} + \mathbf{B}.$$

De esta expresión se observa que los criterios numéricos que se elijan han de ir dirigidos a que el papel que juega la matriz **W** en la ecuación sea mínimo y que, por lo tanto, el papel de **B** sea el mayor posible. De este planteamiento se derivan, entre otros, los dos criterios siguientes:

- Minimización de la traza de la matriz **W**:

  Los D elementos de la diagonal de la matriz **W** son *promedios* de las varianzas por grupos de las D variables observadas en la muestra. Se puede demostrar que minimizar la suma de estos elementos es equivalente a minimizar la suma de la distancia euclídea al cuadrado de cada individuo al centroide del grupo al que ha sido asignado. Los efectos de esta característica son una tendencia a formar grupos de forma esférica. Además, el método no es invariante por cambios de escala. Esta idea está presente de una manera más o menos implícita en el método jerárquico de Ward, en el método k-means, y en el criterio de la mínima varianza de los métodos de particiones estocásticas. Minimizar la traza de la matriz **W** es el método más utilizado.

- Minimización del determinante de **W**:

  Se puede demostrar que este criterio equivale a maximizar el cociente de los determinantes de las matrices **T** y **W**. En análisis de la varianza multivariante uno de los tests para contrastar las diferencias entre los vectores de medias de los diferentes grupos se basa en la ratio de estos dos determinantes. Valores grandes de este cociente indican que los grupos son diferentes. Se busca entonces formar grupos de individuos que maximicen esta ratio o, lo que es equivalente, que minimicen el valor del determinante de **W**, puesto que el determinante de **T** es constante.

  Este método, a diferencia del criterio de minimizar la traza, no tiende a formar grupos con forma esférica y, tiene la propiedad de ser invariante por cambios de escala. Sin embargo, presenta el inconveniente de no poderse aplicar cuando **W** sea una matriz singular ($|\mathbf{W}|=0$), y tiene el defecto de tender a crear grupos con la misma forma.

Por lo que se refiere al número de individuos de los grupos resultantes de la clasificación, los dos métodos anteriores tienen la particularidad de formar clases equilibradas. Con el objetivo de salvar este defecto, y los expuestos anteriormente, diversos autores han propuesto diferentes alternativas. En el libro de Everitt (1993) se encuentra una exposición detallada de las alternativas más usuales.

En la práctica, una clasificación mediante un método de optimización se traduce en la búsqueda de un máximo o un mínimo de una determinada función. Es por este motivo que

el núcleo del algoritmo que se aplique en una clasificación de este tipo debe constar de los pasos siguientes:

1. Elección del punto de partida: debe elegirse una partición inicial de los individuos de la muestra en G grupos. Por ejemplo, una estrategia puede consistir en aplicar préviamente un método de clasificación jerárquico.

2. Elección del siguiente punto: consiste en calcular, según el criterio de optimización elegido, el cambio que supone mover un individuo de un grupo a otro y elegir el cambio que sea óptimo para el criterio.

3. Criterio de interrupción o parada: se itera el segundo paso del algoritmo hasta que no exista ningún cambio de grupo que mejore el criterio elegido.

Obviamente, el algoritmo debe adaptarse y desarrollarse según el criterio numérico con el que se trabaje. Cuando se desee utilizar un método de optimización para realizar clasificaciones de datos composicionales deberá tenerse en cuenta que la medida de homogeneidad escogida sea adecuada. En el caso de los métodos de optimización basados en las tres matrices de dispersión, $\mathbf{T}, \mathbf{W}$, y $\mathbf{B}$, será necesario disponer de una medida de variabilidad coherente con las características matemáticas del soporte de los datos composicionales. Esta cuestión se analiza en el Capítulo 2 de esta tesis.

El ejemplo siguiente ilustra el funcionamiento de este tipo de técnicas.

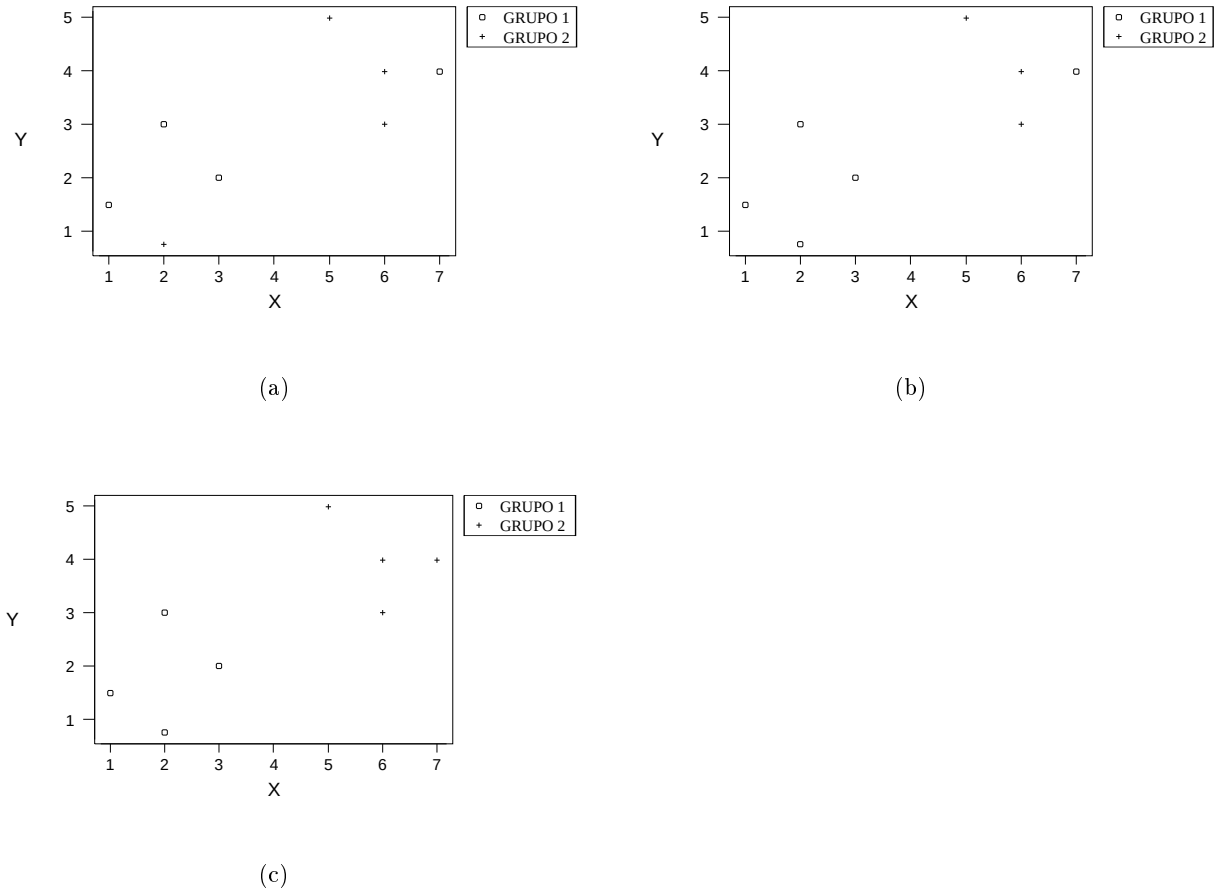
Ejemplo 1.4 Consideramos nuevamente los datos del Ejemplo 1.3 donde se ha ilustrado una técnica de clasificación de grupos disjuntos. La muestra consta de 8 elementos de los que se han observado 2 características obteniéndose los datos siguientes:

$$\mathbf{x}_1 = (1, 1.5),\ \mathbf{x}_2 = (5, 5),\ \mathbf{x}_3 = (2, 3),\ \mathbf{x}_4 = (6, 4),$$

$$\mathbf{x}_5 = (3, 2),\ \mathbf{x}_6 = (2, 0.75),\ \mathbf{x}_7 = (7, 4),\ \mathbf{x}_8 = (6, 3).$$

Se pretende realizar una clasificación de los elementos en $G = 2$ grupos utilizando el método de minimización de la traza de la martriz $\mathbf{W}$. Se consideran los grupos iniciales siguientes (figura 1.10(a)): $\mathbf{C}_1^1 = \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_7\}$, $\mathbf{C}_2^1 = \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_6, \mathbf{x}_8\}$. Unos cálculos elementales proporcionan que Traza$(\mathbf{W}^{(1)}) = 7.52$, donde el superíndice indica el número de iteración a la que corresponde la matriz. Se asigna el elemento $\mathbf{x}_6$ al primer grupo (figura 1.10(b)) y se obtiene que en esta segunda iteración Traza$(\mathbf{W}^{(2)}) = 5.19$. Si en la tercera iteración se asigna el elemento $\mathbf{x}_7$ al segundo grupo (figura 1.10(c)) y se recalcula la matriz de covarianza intragrupos se obtiene que

Traza($\mathbf{W}^{(3)}$) = 1.44. Con el objetivo de no complicar y alargar el ejemplo se ha omitido el paso de calcular cual es la asignación, dentro de cada iteración, que maximiza el decremento de la traza y el paso final de comprobar si la tercera iteración es la última supuesto que se cumpliría si ninguna reasignación disminuyese la traza. □



(a) (b)

(c)

Figura 1.10: Ejemplo de clasificación automática mediante una técnica de optimización: (a) *grupos en la iteración inicial;* (b) *grupos en la segunda iteración;* (c) *grupos en la última iteración.*

1.8 Ayudas a la clasificación

De la exposición realizada en los capítulos anteriores debe extraerse la conclusión que el uso de las técnicas de clasificación automática no se limita a una mera aplicación de una herramienta concreta a los datos objeto del estudio, sino más bien consiste en una serie de fases encadenadas donde las decisiones a tomar en cada fase dependen de los resultados de las fases anteriores.

Cuando un investigador se plantea realizar una clasificación no puede decidir, *a priori* y sin más, si deberá considerar todas o sólo algunas de las variables observadas, si escogerá una similitud u otra, o qué técnica concreta de clasificación usará para agrupar los datos. Es necesario que el investigador haga pruebas: eliminando variables, cambiando la medida de disimilitud, usando diferentes algoritmos de clasificación, escogiendo un subconjunto de individuos, etc. Si se sigue este modo de proceder se llega a una fase en la que el investigador debe plantearse toda una serie de cuestiones fundamentales como son, entre otras: ¿Hasta qué punto las agrupaciones obtenidas son naturales o son un producto artificial de la misma técnica de clasificación? ¿Los grupos obtenidos reflejan alguna interpretación convincente? De las diferentes clasificaciones obtenidas, ¿existe alguna que podamos catalogar como la que *mejor* recoge la estructura de los datos?

En esta sección no se pretende dar una respuesta teórica a todas estas preguntas, simplemente se exponen una serie de herramientas y resultados que pueden utilizarse como guía tanto en las fases intermedias de la clasificación, como en la fase final de interpretación de resultados.

En el libro de Everitt (1993) y en el libro, más reciente, de Gordon (1999) se describen en detalle las ayudas a la clasificación más usuales. Los libros de Everitt y Dunn (1991) y de Mardia et al. (1992), dedicados al Análisis Multivariante, analizan en profundidad las principales características de estas ayudas. Sin embargo, es en Gordon (1998) donde se encuentra un verdadero estado del arte en el tema de validación de clasificaciones automáticas.

1.8.1 Coeficiente de correlación cofenética

Este coeficiente se utiliza como medida de calidad de la clasificación jerárquica realizada. Debido a que una técnica de clasificación jerárquica impone una estructura en los datos, se requiere una medida que analice si esta estructura se encuentra de manera natural en el conjunto de datos o si ha sido construida artificialmente por la clasificación.

En la Sección 1.6.2, referente a *Algoritmos jerárquicos aglomerativos*, se ha explicado que este tipo de técnicas de clasificación permiten definir una medida o jerarquía, h, que representa el nivel de similitud o disimilitud al que se efectúa la unión de dos grupos. Sin pérdida de generalidad, supóngase que en la clasificación se ha elegido una disimilitud. Se denomina matriz *cofenética*, $C = (c_{ij})$, a la matriz $n \times n$ donde c_{ij} representa el índice de jerarquía de la primera clase que contiene los individuos $\mathbf{x}_i$ y $\mathbf{x}_j$. Por lo tanto, c_{ij} viene a representar el nivel del dendrograma en el que los dos individuos pasan a formar parte de la misma clase.

El coeficiente de *correlación cofenética* mide el grado de relación entre los valores c_{ij}, fruto

de la clasificación realizada, y los valores d_{ij} o disimilitudes originales entre los individuos. La medida del grado de relación entre estos dos conjuntos de valores se basa en el cálculo del coeficiente de correlación lineal. De hecho, no es necesario analizar todos los valores y se trabaja sólo con los $\frac{n(n-1)}{2}$ valores inferiores o superiores a la diagonal de cada matriz. Valores cercanos a cero en el coeficiente de correlación cofenética son indicadores de una artificialidad notable en la estructura de la clasificación resultante. Una clasificación totalmente coincidente con una estructura jerárquica natural en los datos originales daría como resultado el valor uno en el coeficiente de correlación cofenética.

Con el propósito de ilustrar la utilización del coeficiente de correlación cofenética como medida de calidad de la clasificación obtenida presentamos el siguiente ejemplo.

Ejemplo 1.5 Recuérdese el ejemplo ilustrativo de la clasificación mediante el método de grupos disjuntos donde la muestra consta de 8 elementos de los que se han observado 2 características obteniéndose los datos siguientes: $\mathbf{x}_1 = (1, 1.5)$, $\mathbf{x}_2 = (5, 5)$, $\mathbf{x}_3 = (2, 3)$, $\mathbf{x}_4 = (6, 4)$, $\mathbf{x}_5 = (3, 2)$, $\mathbf{x}_6 = (2, 0.75)$, $\mathbf{x}_7 = (7, 4)$, $\mathbf{x}_8 = (6, 3)$.

Si se emplea el método jerárquico de Ward para clasificar las 8 observaciones anteriores se obtiene el dendrograma que se muestra en la figura 1.11.

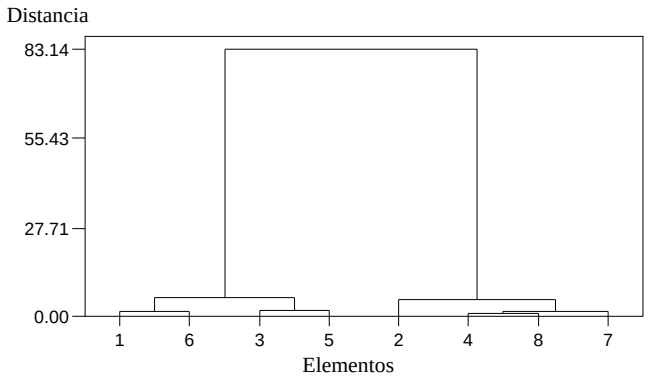


Figura 1.11: Dendrograma con el método Ward y la distancia euclídea al cuadrado.

Nótese que en la aplicación del método de Ward se ha utilizado su formulación equivalente mediante el método flexible y la distancia euclídea al cuadrado.

Si se desea calcular el coeficiente de correlación cofenética debe considerarse la matriz co-

fenética C siguiente:

$$C = \begin{pmatrix} 0 & & & & & & & \\ 83.410 & 0 & & & & & & \\ 5.781 & 83.141 & 0 & & & & & \\ 83.141 & 5.333 & 83.141 & 0 & & & & \\ 5.781 & 83.141 & 2 & 83.141 & 0 & & & \\ 1.562 & 83.141 & 5.781 & 83.141 & 5.781 & 0 & & \\ 83.141 & 5.333 & 83.141 & 1.667 & 83.141 & 83.141 & 0 & \\ 83.141 & 5.333 & 83.141 & 1 & 83.141 & 83.141 & 1.667 & 0 \end{pmatrix},$$

y la matriz D de distancias euclídeas al cuadrado entre observaciones:

$$D = \begin{pmatrix} 0 & & & & & & & \\ 28.250 & 0 & & & & & & \\ 3.250 & 13 & 0 & & & & & \\ 31.250 & 2 & 17 & 0 & & & & \\ 4.250 & 13 & 2 & 13 & 0 & & & \\ 1.563 & 27.063 & 5.063 & 26.563 & 2.563 & 0 & & \\ 42.250 & 5 & 26 & 1 & 20 & 35.563 & 0 & \\ 27.250 & 5 & 16 & 1 & 10 & 21.063 & 2 & 0 \end{pmatrix}.$$

Al calcular el coeficiente de correlación entre las dos matrices se obtiene un valor de 0.829 que puede considerarse indicador de que la estructura jerárquica recoge satisfactoriamente las disimilitudes originales entre las observaciones.

De la observación del dendrograma se propone considerar los dos grupos siguientes:

$$C_1 = \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_6\}, \quad C_2 = \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_7, \mathbf{x}_8\}.$$

1.8.2 Reducción de la dimensión de datos multivariantes

En la actualidad, la práctica totalidad de análisis estadísticos de datos se realiza con la ayuda de los paquetes estadísticos informáticos. Este hecho, que posibilita la realización de análisis estadísticos de complejidad elevada con relativa facilidad, comporta una tendencia a olvidar la exploración de los datos con técnicas más simples. Estas técnicas elementales deberían utilizarse en los estadios iniciales del proceso de clasificación. Todas las particularidades de la distribución de los datos que se aprecien con su aplicación deberán ser corroboradas por las herramientas más complejas que se usarán en las siguientes fases del estudio.

Como ejemplo de exploraciones iniciales a realizar en un conjunto de datos multivariantes se puede considerar la construcción y estudio de las distribuciones marginales de las D variables

incluidas en la clasificación. La observación de estas distribuciones unida al cálculo de algunos estadísticos marginales básicos pueden sugerir, entre otros aspectos, la presencia de datos *erróneos*, de datos excesivamente alejados del resto y la existencia de asimetrías que pueden indicar la conveniencia de una transformación de los datos.

También puede ser de utilidad la representación de los diagramas bivariantes de las D variables tomadas dos a dos, puesto que en ellos es posible detectar variables altamente correlacionadas y vislumbrar estructuras internas en los datos que serán de gran ayuda en la fase de interpretación de los grupos obtenidos.

Lo deseable sería poder disponer de una técnica gráfica que nos permitiera visualizar las D variables a la vez de todos los individuos. Desgraciadamente este deseo no es factible para $D \geq 4$. A pesar de esta dificultad, existen técnicas gráficas que intentan plasmar en un gráfico bidimensional la información recogida en las D variables. Un ejemplo de estos gráficos lo encontramos en los diagramas de *Andrews* que se basan en representar cada individuo $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iD})$ de la muestra $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ como la función suma de los D términos siguientes:

$$\mathbf{x}_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(t) + x_{i3}\cos(t) + x_{i4}\sin(2t) + x_{i5}\cos(2t) + ... \quad ,$$

donde $t \in [-\pi, \pi]$. Este tipo de función tiene la buena propiedad de conservar la distancia euclídea, en el sentido que individuos cercanos de la muestra producen funciones cuyos gráficos resultan cercanos. Este hecho permite detectar posibles grupos de individuos si se reconocen en el diagrama de Andrews agrupaciones de curvas. El gran inconveniente de esta técnica es el grado notable de confusión que aparece en el diagrama si el número de individuos o curvas a representar es elevado.

Todas las técnicas gráficas de este tipo, como son los diagramas de Andrews, que intentan reflejar los datos multivariantes en un diagrama bidimensional tienen aspectos que hacen recomendable su uso y tienen un inconveniente u otro que hacen necesario confrontar la estructura de los datos que sugieren con otras técnicas de clasificación. Son, por lo tanto, herramientas de ayuda a la clasificación.

Existe otro tipo de estrategia, mucho más utilizada y compleja, que es útil como ayuda a la resolución del problema de representar los datos multivariantes: la *reducción de la dimensión de los datos multivariantes*. De una manera u otra, se intenta transformar o resumir los datos multivariantes de dimensión D de la muestra $\mathbf{X}$ en datos pertenecientes a espacios de dimensiones menores. Como el objetivo es facilitar su representación gráfica y, por ende, su interpretabilidad, suele reducirse la dimensión hasta, como mínimo, conseguir trabajar en un espacio tridimensional. A pesar de estar todas ellas íntimamente relacionadas podemos diferenciar las técnicas más

utilizadas que permiten este tipo de estrategia en tres clases:

- *las componentes principales, gráficos biplot y la búsqueda proyectiva (del inglés "Projection Pursuit");*
- *el análisis factorial y la clasificación por variables;*
- *las técnicas de reescalado multidimensional (del inglés "Multidimensional Scaling").*

La técnica de *componentes principales* intenta simplificar el análisis de datos multivariantes mediante la consideración de un número pequeño de combinaciones lineales de las D variables originales. Estas combinaciones lineales, que constituyen a su vez nuevas variables llamadas las componentes principales, se construyen ortogonales entre si y con el objetivo de recoger las direcciones o ejes de máxima varianza de los datos. De esta manera, se habla de la primera componente principal como el eje de máxima varianza, la segunda componente principal expresa la segunda dirección en cuanto a varianza, y así sucesivamente, hasta la D-ésima componente principal. Siguiendo el criterio de retener sólo aquellas componentes principales que en conjunto recojan un porcentaje elevado de la varianza total de los datos se reduce la dimensión para pasar habitualmente a trabajar, como máximo, con las tres primeras componentes principales. Una vez obtenidas las coordenadas de los datos en la nueva base de variables se realizan los gráficos bivariantes por pares de componentes principales. Si esta técnica se utiliza en los estadios preliminares de una clasificación, permite aplicar los algoritmos de clasificación a las coordenadas de los datos en la base de componentes principales, habiendo o no reducido el número de variables a tener en cuenta, y posibilita contrastar las agrupaciones obtenidas con las clases resultantes de trabajar con datos originales. Esta confirmación se hace muy interesante si se recuerda el trabajo de Chang (1983) donde se muestra que no necesariamente son las primeras componentes principales las que mejor clasifican entre los individuos, presentando un ejemplo en el que, gráficamente, es en el diagrama bivariante entre la primera y la quinceava componente principal donde se distinguen perfectamente los grupos que forman el conjunto de datos considerados.

El método de componentes principales también puede usarse en las fases finales del estudio, simplemente como una ayuda gráfica que permite obtener una representación de los grupos fruto de aplicar una técnica de clasificación a los datos originales.

A diferencia de la técnica de componentes principales, los gráficos biplot se basan en la descomposición en valores singulares de la matriz de datos $\mathbf{X}$. Sin embargo, las dos técnicas tienen en común el hecho de realizar un cambio de base y de representar los datos de la matriz $\mathbf{X}$

en función de las coordenadas en la nueva base. En el método de los gráficos biplot la elección de la nueva base no puede realizarse únicamente de un solo modo. Es el usuario el que decide como construye la nueva base, existiendo la posibilidad de obtener la misma base que se obtendría aplicando el método de las componentes principales. Por este motivo, puede considerarse que los gráficos biplot generalizan la técnica de componentes principales. Una particularidad muy útil que poseen los gráficos biplot es que en el mismo gráfico se representan simultáneamente los n individuos, usualmente como puntos, y las D variables originales, usualmente como vectores. La obra de Gower y Hand (1996) recoge un análisis en profundidad de esta técnica de representación en baja dimensión.

La técnica de *búsqueda proyectiva* puede considerarse como una generalización de las componentes principales puesto que se basa en la búsqueda de un espacio de dimensión menor referido a una base ortogonal de manera que la proyección de los datos sobre él optimice un determinado valor numérico que, en el caso de las componentes principales, corresponde a la varianza. Atendiendo a esta idea, las consideraciones expuestas para las componentes principales pueden extenderse a la búsqueda proyectiva. En los trabajos de Friedman (1987), de Jones y Sibson (1987), y de Nason (1995) se trata en profundidad las principales características de esta técnica.

El *análisis factorial*, via componentes principales o via estimación máximo verosímil, y la *clasificación por variables* son técnicas que inciden en la búsqueda de relaciones entre las variables que forman el estudio. Estas relaciones han de permitir reducir la dimensión mediante la transformación, eliminación o la agrupación de variables. Por lo tanto, al igual que las técnicas anteriores, se pueden utilizar o bien como ayuda previa a la formación de los grupos o bien como ayuda a la interpretación de las clases obtenidas. Respecto a la clasificación por variables, técnica no desarrollada en este texto, hay que tener presente que requiere, a su vez, la elección de una medida de similitud entre variables y de un algoritmo de clasificación para obtener grupos de variables.

El *reescalado multidimensional* difiere notablemente de los métodos anteriores puesto que no trabaja directamente con los datos observados. Esta técnica se basa en los valores de las similitudes, disimilitudes o distancias entre los n individuos objeto de la clasificación. Según el tipo de medida de diferencia escogida se califica a los métodos de reescalado multidimensional como *no métricos* o como *métricos*. En ambos casos, la técnica de reescalado multidimensional construye un conjunto de valores *coordenados*, a partir de los $\frac{n(n-1)}{2}$ valores, inferiores o superiores a la diagonal, de la matriz de disimilitudes. Habitualmente se usan dos coordenadas de cada individuo para reducir la dimensionalidad y poder realizar un gráfico donde representar

individuos y sus disimilitudes. Estas características del reescalado lo hacen especialmente indicado para todas las fases de un estudio de clasificación y, especialmente indicado si se prueban diferentes medidas de disimilitud en un mismo conjunto de datos multivariantes.

1.8.3 El problema del número de grupos

Estudiar préviamente el número de grupos que forman el conjunto de datos es condición necesaria para muchos métodos de clasificación. Este problema, al igual que toda la clasificación en su conjunto, no tiene solución única ni tampoco existe un criterio que permita decidir en cada caso cual es la manera óptima de decidir el número de grupos a construir. Existe una serie de estrategias y herramientas que pueden ser útiles para indicar las clases que pueden contener los datos.

Si se utiliza un método de clasificación jerárquico puede detectarse el número de grupos si se observa en el dendrograma un salto grande, en relación a los otros saltos, en los niveles de jerarquías en los que se unen las clases. Este salto significa un incremento en la disimilitud entre grupos y por lo tanto una fusión a un nivel superior puede atribuirse a una estructura artificial aportada por la técnica de clasificación jerárquica. Mientras más bajo sea el nivel de disimilitud donde se inicia el salto observado y más grande sea éste, más homogéneas serán las clases en su interior y más heterogéneos serán los grupos entre si. Se procede, entonces, a *cortar* el dendrograma por el salto de nivel detectado y se obtiene la propuesta del número clases que forman el conjunto. Este número de grupos debe compararse con el obtenido si a los datos se les aplica otra técnica de clasificación jerárquica y otra medida de disimilitud.

En Everitt (1993) se recoge el método de Mojena para decidir de manera automática el nivel de corte del dendrograma. Este método consiste en considerar los niveles de fusión $\alpha_0,\ \alpha_1,\ ,\ldots,\ \alpha_{n-1}$ correspondientes a las etapas del algoritmo jerárquico en las que se tienen clasificadas las n observaciones del conjunto en $n,\ n-1,\ \ldots,\ 1$ grupos, respectivamente. Se calcula la media $\overline{\alpha}$ y la desviación típica s_α del conjunto de niveles de fusión y se establece como nivel de corte del dendrograma el primer nivel de fusión α_j que satisface

$$\alpha_j > \overline{\alpha} + ks_\alpha, \tag{1.11}$$

donde k es una constante cuyo valor más adecuado es 1.25 (Everitt, 1993). Este método, al ser automático, tiene la ventaja de ser *objetivo*. Sin embargo, Cooper y Milligan (1988) argumentan que no es uno de los mejores métodos para decidir de manera automática el número de grupos en una clasificación. En el mismo trabajo estos autores exponen que el índice C, sugerido por Calinski y Harabasz, tiene un comportamiento más adecuado.

El índice C se obtiene mediante la siguiente expresión:

$$C(g) = \frac{\frac{\text{Traza}(\mathbf{B})}{g-1}}{\frac{\text{Traza}(\mathbf{W})}{n-g}}, \tag{1.12}$$

donde g es el número de grupos, $\mathbf{B}$ es la matriz de varianzas entre-grupos y $\mathbf{W}$ es la matriz de varianzas intra-grupos. Para la clasificación obtenida con un método de clasificación determinado, este índice C debe calcularse considerando que en el conjunto de datos deben construirse g grupos, para $g = 1, 2, \ldots, n$. Entonces, se escogerá como número de grupos óptimo G el que maximice el valor del índice C:

$$G = \max\{C(g)/g = 1, 2, ..., n\}.$$

Se observa que el índice está persiguiendo encontrar una clasificación donde los grupos sean heterogéneos entre si, o un *numerador grande*, y cada grupo sea homogéneo en su interior, o un *denominador pequeño*.

1.8.4 Validación de la clasificación

Una vez realizada una clasificación es interesante disponer de una serie de pautas para valorar los grupos obtenidos. Por lo que respecta a la valoración de una clasificación en si misma, una herramienta evaluadora es la *tasa de clasificación errónea*, del inglés *misclassification rate*, propia del *análisis discriminante*. Esta técnica de análisis de datos multivariantes no se desarrolla en este trabajo de investigación por alejarse de los objetivos marcados. A parte de esta posibilidad, existen otras estrategias alternativas y complementarias que pueden considerarse como herramientas para evaluar la estabilidad de la clasificación. Una de ellas consiste en dividir aleatoriamente el conjunto de los datos en dos subgrupos, repetir a continuación la agrupación en cada subgrupo por separado y, comprobar que se obtiene en cada subgrupo la misma estructura que en el global. También hay que recordar que si, préviamente a la aplicación del algoritmo de clasificación, se han eliminado algunas variables con el objeto de reducir la dimensión de los datos, se hace aconsejable rehacer la clasificación considerando todas las variables y comprobar que se obtiene la misma estructura.

Debido a considerar diferentes disimilitudes y métodos de clasificación, puede plantearse la conveniencia de comparar las dos clasificaciones diferentes obtenidas con, eso sí, el mismo número G de grupos. En el caso de que el conjunto de datos esté formado por una cantidad no muy elevada de individuos puede realizarse una comparación por inspección de los grupos, a partir del dendrograma, y detectar las diferencias entre ambas. En otro caso puede utilizarse el

índice R_G citado en el libro de Everitt (1993). Este coeficiente representa un intento de medir las *coincidencias* entre dos clasificaciones sobre los mismos datos mediante la siguiente expresión:

$$R_G = \frac{T_G - \frac{1}{2}P_G - \frac{1}{2}Q_G + \binom{n}{2}}{\binom{n}{2}},$$

donde

$$T_G = \sum_{i=1}^{G}\sum_{j=1}^{G} m_{ij}^2 \; - \; n; \quad P_G = \sum_{j=1}^{G} m_{i\cdot}^2 \; - \; n; \text{y} \;\; Q_G = \sum_{i=1}^{G} m_{\cdot j}^2 \; - \; n.$$

El término m_{ij} representa el número de individuos coincidentes entre el i-ésimo grupo de la primera clasificación y la j-ésima clase de la segunda clasificación. Los términos $m_{i\cdot}$ y $m_{\cdot j}$ son las acumulaciones marginales de cada una de las clasificaciones. Puede demostrarse que este índice toma valores entre 0 y 1, cogiendo el valor máximo $R_G = 1$ cuando existe coincidencia total entre dos clasificaciones.

1.8.5 Comentarios finales

Las medidas de diferencia, de tendencia central, y de variabilidad forman parte, de una manera implícita o explícita, de las ayudas a la clasificación expuestas en esta sección. Por este motivo, cuando se realice una clasificación de datos composicionales estas ayudas a la clasificación han de ir acompañadas de medidas adecuadas. Las características que deben poseer estas medidas para ser adecuadas sobre datos composicionales son tratadas en el Capítulo 2 de esta tesis.

Por otro lado, es necesario resaltar que, en general, es imposible anticipar qué elección de variables principales, qué similitud o distancia y qué técnica de agrupación es la que nos dará la clasificación más informativa e interesante. A pesar de ello, sí suelen recomendarse una serie de pasos a seguir a la hora de realizar una clasificación automática de datos. Estos pasos son los siguientes:

- Representar gráficamente los datos originales, utilizando si se cree conveniente técnicas como, entre otras, *componentes principales*, *búsqueda proyectiva* y *reescalado multidimensional*.

- Decidir si se realiza algún tipo de transformación de los datos originales. Si se considera oportuno se estandarizan los datos, se reduce el número de variables, se reescala alguna variable, etc.

- Escoger la técnica a aplicar. Si se elige un método jerárquico hay que decidir la disimilitud y el algoritmo a aplicar. A pesar de que ningún método es mejor que los otros, entre los jerárquicos se recomienda el de Ward y el de la media; entre los optimizadores, el de minimizar el determinante de la matriz ponderada de varianzas intragrupo W; y, en caso de tener datos que sigan un modelo de distribución de probabilidad, utilizar los algoritmos de particiones estocásticas.

El seguimiento de estos pasos no debe ser puramente secuencial sino que se procederá a combinarlos y repetirlos de manera que se trabaje iterativamente entre ellos.

1.8.6 Aplicaciones informáticas (*Software*)

La gran mayoría de técnicas recogidas en este texto son inabordables si no se dispone de la ayuda de un ordenador. Afortunadamente, en la actualidad, la totalidad de paquetes estadísticos más utilizados incorporan procedimientos para la clasificación de datos multivariantes. A continuación se describen las posibilidades que ofrecen algunas de las principales aplicaciones informáticas existentes en el mercado comercial:

- *MINITAB*

  El MINITAB versión 12 (2000) para Windows es un paquete estadístico totalmente integrado en el entorno de trabajo Windows con todo lo que ello comporta: aplicaciones gráficas, compatibilidad con otros productos, facilidad de introducción en el ambiente MINITAB, sistema interactivo de ayuda al usuario, etc.

  Incluye las técnicas principales de análisis de datos multivariantes: análisis discriminante, componentes principales, análisis factorial, clasificación por variables, clasificación individuos, etc. Sólo permite incorporar las distancias y similitudes más habituales e incluye los algoritmos de clasificación más utilizados. Sin embargo, el hecho de no ser un paquete especializado en datos multivariantes le supone algunas limitaciones, como puede ser que no permite realizar gráficos biplot y gráficos de reescalado muldidimensional.

- *SAS*

  En nuestra experiencia con su versión 6.12 para Windows hemos observado que el producto del instituto SAS constituye una herramienta completísima y de primer orden en todos los aspectos. Entre los muchos módulos que integran este paquete se destaca el módulo *STAT* por las muchísimas herramientas y técnicas estadísticas que contiene. Además de encontrar una serie de procedimientos que permiten el estudio de los datos multivariantes,

como son las componentes principales, el análisis factorial y el análisis discriminante, pueden aplicarse una larga lista de programas para realizar una clasificación automática. Encontramos la mayoría de métodos jerárquicos, con las distancias más habituales, y el método de grupos disjuntos *Fastclus*.

Este paquete estadístico contiene un procedimiento interesante si se desea utilizar la distancia de Mahalanobis para la clasificación. El procedimiento, llamado *ACECLUS* realiza una estimación de la matriz ponderada de covarianzas intra-grupos sin necesidad de conocer, a priori, el número de grupos a formar.

El paquete trabaja mediante la ejecución de programas incorporados por el instituto SAS y también permite la elaboración de procedimientos personales en un lenguaje de programación propio, sin que este aspecto suponga una dificultad añadida.

La documentación que acompaña al soporte lógico es realmente extensa. Cada procedimiento contiene multitud de ejemplos y referencias bibliográficas, tanto en lo que se refiere a libros de texto como a artículos de investigación publicados en revistas especializadas.

Como consecuencia de trabajar en entorno Windows sus facilidades gráficas son muy manejables además de compatibles con otros productos.

- *S-PLUS*

  Algunos de los ejemplos que ilustran esta tesis han estado elaborados con el paquete estadístico MINITAB y con el paquete S-PLUS (2000) tanto en su versión para Windows como en su versión para Unix.

  La mayoría de los adjetivos calificativos del paquete SAS son útiles para describir este producto. Cabe destacar el hecho de estar más integrado en el entorno de Windows con lo que gana en cuanto a manejabilidad. También se diferencia del anterior producto en ser más *amigable*. Es decir, bastan unas pocas sesiones prácticas para introducirse en su uso.

  Por lo que se refiere a algoritmos de clasificación debe mencionarse que permite la inclusión de múltiples disimilitudes y distancias. Estos algoritmos están basados en procedimientos desarrollados en el libro de Kaufman y Rousseuw (1990) bajo los nombre de *PAM, AGNES, CLARA*, etc.

  Del S-PLUS destaca que, además de un completo tratamiento de la clasificación no paramétrica –es decir, jerárquica y de grupos disjuntos– se incluyen procedimientos para realizar clasificaciones mediante las particiones estocásticas (en inglés, *model-based cluster*). Contiene incluso un procedimiento para el estudio previo del número de grupos a

formar. También, en su versión UNIX, permite realizar *clasificación borrosa* (en inglés, *fuzzy cluster*).

Otras características que se detectan en el uso de este paquete estadístico son la posibilidad de una exploración gráfica completísima, su facilidad de programación mediante un lenguaje propio estructurado similar al C++ y su conexión con el lenguaje FORTRAN, lo que permite aprovechar multitud de procedimientos de libre acceso.

- *SPSS*

  La descripción del SPSS versión 7.5 (1999) no puede diferir mucho de la realizada para el producto MINITAB. Simplemente merece mención que en el SPSS el tratamiento de los datos de tipo cualitativo es mucho más completo y que permite incorporar muchas más distancias y disimilitudes en los algoritmos de clasificación.

- *MATLAB*

  El paquete MATLAB en su versión 5.3.1 (1999) es un paquete que incluye multitud de facilidades para el cálculo matemático. Por lo tanto, no es un paquete exclusivamente diseñado como herramienta para realizar análisis estadísticos. Sin embargo, la mayor parte del tratamiento de datos composicionales desarrollado en esta tesis se ha realizado con funciones programadas en MATLAB. En Aitchison (1986) se presenta el paquete CODA como una herramienta específica para el análisis de datos composicionales. El mismo autor ha desarrollado un trabajo de conversión del paquete CODA, originalmente programado en lenguaje BASIC, a funciones que se ejecutan en MATLAB. A partir de estas funciones hemos implementado en MATLAB muchas otras nuevas facilidades que permiten un tratamiento más completo de conjuntos de datos composicionales. Al ser un producto totalmente diseñado para ser compatible con todas las aplicaciones que trabajen en el entorno Windows, resulta sumamente sencillo conectar las posibilidades del MATLAB con las del paquete S-PLUS y las del programa MINITAB.

Capítulo 2

Datos composicionales

2.1 Introducción

En el capítulo anterior se ha insistido en la importancia del papel que desempeña el concepto de medida de diferencia entre dos observaciones a la hora de aplicar las técnicas de clasificación no paramétricas. También se ha recordado que la elección de una medida de diferencia adecuada depende fuertemente del tipo de datos que intervienen en la clasificación. Por ello, en este capítulo nos centramos en presentar de manera detallada los conceptos básicos referentes a los datos composicionales: contexto de utilización de este tipo de datos, el espacio muestral $\mathcal{S}^D$, las subcomposiciones, las amalgamas, las perturbaciones y los requerimientos de una medida de diferencia. La mayoría de estos conceptos se encuentran desarrollados en Aitchison (1986) y en Barceló (1996). Ponemos un énfasis especial en exponer los inconvenientes de las medidas habitualmente utilizadas con datos cuantitativos cuando se aplican sobre datos composicionales y presentamos la definición de una medida de diferencia entre dos datos composicionales basada en el concepto introducido recientemente por Aitchison (1992, 1997, 2000, 2001) y que ha sido motivo de estudio en otros trabajos de Martín-Fernández et al. (1998b) y de Aitchison (2000, 2001). Finalmente, en este capítulo ilustramos el comportamiento de esta medida y de las medidas de diferencia más habituales, analizando, para dimensión $D = 3$, la forma que adquieren los entornos de equidistancia según la posición del centro de los mismos.

El concepto de medida de tendencia central de un conjunto de observaciones y el de medida de dispersión juegan un papel clave en algunas técnicas de clasificación no paramétrica. Al igual que las medidas de diferencia, las medidas de tendencia central y de dispersión de un conjunto de datos deben tener en cuenta la naturaleza de los mismos. En este capítulo mostramos los inconvenientes de las medidas habituales de tendencia central y de dispersión cuando se aplican

sobre datos composicionales y, presentamos unas medidas alternativas basadas en las propuestas de Aitchison (1997).

Como referencias introductorias a la temática de los datos composicionales se ha utilizado la monografía de Aitchison (1986). El concepto de medida de diferencia entre dos datos composicionales y las medidas de tendencia central y de dispersión de un conjunto de datos composicionales están recogidas en Aitchison (1992, 1997) y Martín-Fernández et al. (1998b, 1998a).

2.2 Definiciones y propiedades básicas

2.2.1 Contexto

El estudio de datos composicionales aparece en multitud de situaciones diferentes. Se contemplan datos composicionales en disciplinas muy diversas: en petrología; en sedimentología; en bioquímica; en el estudio de las propiedades de muchos objetos o substancias –tales como carburantes, aleaciones metálicas, etc.– que dependen fundamentalmente de la proporción relativa en que están mezclados sus ingredientes; en sociología, cuando se aborda el estudio del tiempo dedicado por las personas a las diferentes actividades diarias; o en economía, al analizar la distribución del presupuesto de las familias entre las diferentes partidas del gasto familiar.

Los dos últimos ejemplos ilustran dos tipos de situaciones diferentes que aparecen en el estudio de problemas donde se consideran datos composicionales. El estudio de las actividades diarias corresponde a problemas en los cuales los datos observados ya tienen suma constante. En cambio, el estudio de los presupuestos familiares corresponde a problemas en los que el investigador transforma a posteriori los datos iniciales para que la suma de sus componentes tenga en todos ellos el mismo valor. Es importante destacar este último caso puesto que en él es el investigador quién decide trabajar con datos de tipo composicional. Estos y otros ejemplos pueden encontrarse, entre otros muchos, en Aitchison (1986), Bohling et al. (1996), Davis et al. (1995), Martín-Fernández et al. (1997), Vives y Villarroya (1996) y Zhou, Chen, y Lou (1991).

Cuando los datos composicionales puedan entenderse como realizaciones de un fenómeno aleatorio, resulta factible la aplicación del análisis estadístico para llevar a cabo investigaciones e interpretaciones adecuadas de estos conjuntos de datos. Sin embargo, la restricción de suma constante de las componentes es a menudo ignorada o interpretada incorrectamente a la hora de proponer modelos estadísticos, de realizar inferencias o de aplicar determinadas técnicas de análisis multivariante de datos.

2.2.2 Definiciones básicas

La definiciones siguientes presentan los elementos más básicos en referencia a los datos de tipo composicional: los datos *composicionales*, las *composiciones*, y el *espacio soporte* o *espacio muestral* que se conoce como *símplex*.

Definición 2.1 Un *dato composicional* o una D-*parte* $\mathbf{x}$ es un $D \times 1$ vector cuyas *componentes* $x_1, x_2, \ldots, x_D$ son todas estrictamente positivas

$$x_1 > 0, x_2 > 0, \ldots, x_D > 0, \tag{2.1}$$

y tal que la suma de todas ellas es igual a 1:

$$x_1 + x_2 + \ldots + x_D = 1. \tag{2.2}$$

□

Definición 2.2 El símplex $\mathcal{S}^D$ es el subconjunto del espacio real $\mathbb{R}^D$ definido por

$$\mathcal{S}^D = \{(x_1, x_2, \ldots, x_D) : x_1 > 0, x_2 > 0, \ldots, x_D > 0; x_1 + x_2 + \ldots + x_D = 1\}.$$

□

Definición 2.3 Una *composición* de D-partes es cualquier vector aleatorio $\mathbf{X}$ cuyo recorrido esté contenido en el símplex $\mathcal{S}^D$. □

Nótese que, en un principio, sólo se consideran datos cuyas componentes sean todas diferentes de cero. Cuando queramos distinguir un vector aleatorio de una de sus realizaciones, simbolizaremos aquél mediante letras mayúsculas –$\mathbf{X}, \mathbf{X}^*, \ldots$–. Sin embargo, cuando el contexto no se preste a confusión, utilizaremos también letras minúsculas –$\mathbf{x}, \mathbf{x}^*, \ldots$– para su simbolización. En lo que resta de esta sección, no distinguiremos *composición* de *dato composicional* ya que todas las definiciones que siguen pueden aplicarse indistintamente a ambos conceptos.

2.2.3 El operador clausura

Observemos que a partir de un vector cualquiera $\mathbf{w} \in \mathbb{R}^D_+$, es posible obtener una composición $\mathbf{x}$ de $\mathcal{S}^D$ sin más que escalar convenientemente las componentes de $\mathbf{w}$ de modo que su suma sea igual a la unidad. Ello nos conduce de forma natural a la definición del "*operador clausura*":

Definición 2.4 Cada observación $\mathbf{w} \in \mathbb{R}_+^D$ tiene asociada de forma unívoca un tamaño $t = w_1 + w_2 + \ldots + w_D$, que es una variable aleatoria definida en $\mathbb{R}^+$, y una composición $\mathbf{x} = \mathbf{w}/t$ de $\mathcal{S}^D$.

Llamaremos *operador clausura* $\mathcal{C}$ a la aplicación de $\mathbb{R}_+^D$ en $\mathcal{S}^D$ que hace corresponder a cada observación $\mathbf{w}$ la composición $\mathbf{w}/t$ asociada:

$$\mathcal{C}(\mathbf{w}) = \mathbf{w}/(w_1 + w_2 + \ldots + w_D) \ , \quad \mathbf{w} \in \mathbb{R}_+^D.$$

□

Nótese que, si bien una observación $\mathbf{w}$ tiene asociada una única composición $\mathcal{C}(\mathbf{w})$, una composición $\mathbf{x} \in \mathcal{S}^D$ tiene asociadas infinitas observaciones –véase la figura 2.1.

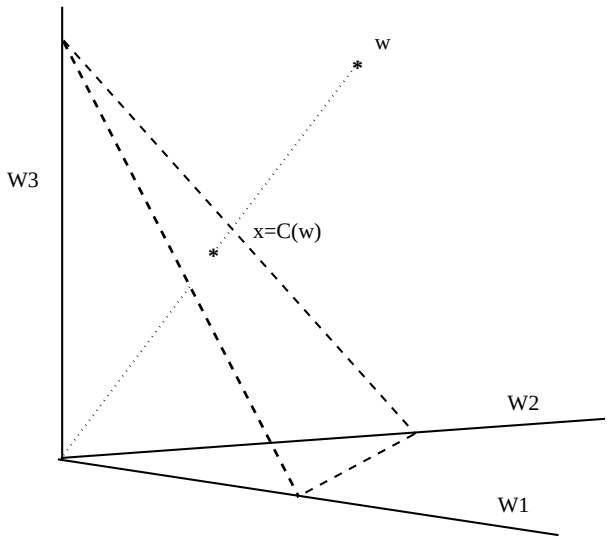


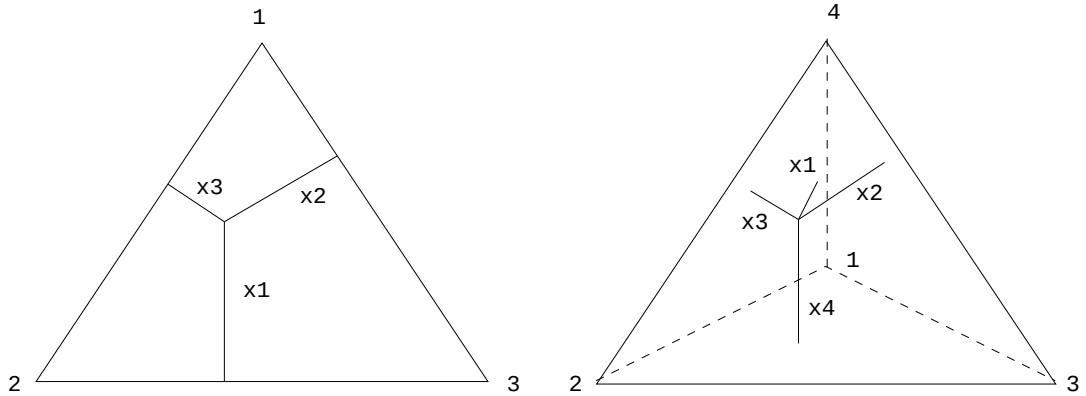
Figura 2.1: *La composición* $\mathbf{x}$ *y una de sus observaciones* $\mathbf{w}$ *asociadas.*

Este operador se aplicará en las situaciones en las que el investigador ha decidido tratar sus datos como composicionales y los datos recogidos son vectores de $\mathbb{R}_+^D$ cuya suma de componentes no es constante. Aparecerá en situaciones en las que se considera que el valor absoluto de las componentes no aporta información por sí mismo y, en cambio, la información radica en los cocientes o *ratios* entre componentes, dejando sin importancia el valor de la suma de las componentes del vector de observaciones.

La restricción de suma unitaria (2.2) hace que una D-parte sea, en esencia, un vector de dimensión $D-1$. Así, un dato composicional queda completamente especificado si se conocen $D-1$ cualesquiera de sus componentes. En el fondo, la restricción de suma unitaria podría substituirse por cualquier otra restricción de suma constante $x_1 + x_2 + \ldots + x_D = k$, dependiendo de la escala utilizada para expresar los datos composicionales. Por tanto, un hecho muy

importante a tener en cuenta es que un dato composicional tan sólo aporta información sobre las magnitudes relativas x_i/x_j $(i, j = 1, 2, \ldots, D; i \neq j)$ de las componentes que lo integran. Por ello, la atención del investigador que aplica cualquier técnica estadística a este tipo de datos deberá dirigirse hacia estos cocientes o *ratios* entre las componentes de un dato composicional más que hacia las magnitudes absolutas de aquéllas. Con este planteamiento queda plenamente justificada la catalogación –véase la Sección 1.4.1– de los datos composicionales como datos de tipo continuo y de tipo razón. Por otra parte, fácilmente puede deducirse que un dato composicional también queda plenamente determinado si se conocen $D-1$ cocientes del tipo $x_i/x_k \quad (i = 1, 2, \ldots, D;\ i \neq k)$. Observemos que en este hecho radica la necesidad fundamental de que cualquier técnica y/o medida que se aplique a datos composicionales sea invariante por cambios de escala de los datos.

En el caso $D = 3$, el símplex $\mathcal{S}^3$ puede representarse por el triángulo equilátero de altura unidad de la figura 2.2(a): cada dato composicional $\mathbf{x} = (x_1, x_2, x_3) \in \mathcal{S}^3$ se corresponde con el punto interior al triángulo que dista x_1, x_2 y x_3, respectivamente, de los lados l_{23}, l_{13} y l_{12} del triángulo. A esta representación se la conoce como *diagrama ternario*. En el caso $D = 4$, el símplex $\mathcal{S}^4$ se representa por el interior de un tetraedro de altura unidad –véase la figura 2.2(b). En este caso a la representación la denominamos *diagrama cuaternario*.



(a) (b)

Figura 2.2: (a) *Representación de un dato composicional* (x_1, x_2, x_3) *del símplex* $\mathcal{S}^3$ *en el diagrama ternario;* (b) *Representación de un dato composicional* (x_1, x_2, x_3, x_4) *del símplex* $\mathcal{S}^4$ *en el diagrama cuaternario.*

2.2.4 Subcomposiciones y amalgamas

Definición 2.5 Si s es un subconjunto cualquiera de las partes $x_1, x_2, \ldots, x_D$ de una composición $\mathbf{x}$ de $\mathcal{S}^D$, y $\mathbf{x}_s$ simboliza el subvector formado por las componentes correspondientes de $\mathbf{x}$, entonces $\mathcal{C}(\mathbf{x}_s)$ recibe el nombre de *subcomposición* de las s partes de $\mathbf{x}$. □

De este modo, la formación de una subcomposición puede considerarse como una transformación de $\mathcal{S}^D$ en $\mathcal{S}^s$ que a cada composición $\mathbf{x} \in \mathcal{S}^D$ le hace corresponder la composición $\mathcal{C}(\mathbf{x}_s) \in \mathcal{S}^s$. La transformación *subcomposición* aplicada a datos composicionales juega un papel análogo a la operación *proyección* de datos multivariantes de $\mathbb{R}^D$ sobre espacios de dimensión inferior. Un hecho remarcable es que la transformación subcomposición conserva las magnitudes relativas o cocientes de las partes que integran la subcomposición. Es por esta razón que, por comodidad y cuando no exista posibilidad de confusión, denominaremos a una subcomposición simplemente por $\mathbf{x}_s$.

Así, por ejemplo, las subcomposiciones de $\mathcal{S}^2$ que se obtienen al prescindir de la 3ª componente y tener tan sólo en cuenta las componentes 1ª y 2ª de la composición, se obtienen geométricamente proyectando los puntos de $\mathcal{S}^3$ sobre el lado l_{12}, desde el vértice 3 –véase la figura 2.3.

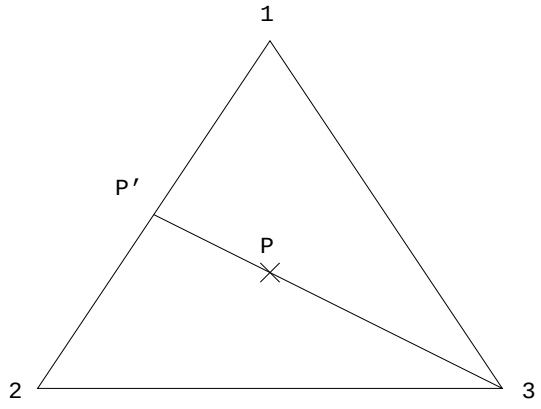


Figura 2.3: *Subcomposición $P' \in \mathcal{S}^2$ representada como una proyección lineal de $P \in \mathcal{S}^3$.*

En las situaciones en las que se detecta una sobre-fragmentación o sobre-dimensionalidad de los datos puede ser conveniente acumular o *amalgamar* algunas partes entre si. También puede interesar aglutinar las componentes de una composición para formar una nueva composición amalgamada cuando estas componentes contienen una gran cantidad de ceros (Martín-Fernández et al., 1997). Así, por ejemplo, a partir de la composición $\mathbf{x} = (x_1, x_2, \ldots, x_9) \in \mathcal{S}^9$, nuestro interés puede centrarse en la composición amalgama $\mathbf{x}_4 = (x_2 + x_3, x_1 + x_4, x_5 + x_6 + x_7, x_8 + x_9)$ de $\mathcal{S}^4$.

Definición 2.6 Si el conjunto de las D partes de una composición de $\mathcal{S}^D$ está partido en C $(C < D)$ subconjuntos mutuamente disjuntos y sumamos las componentes pertenecientes a un mismo subconjunto, la composición de las C partes que resulta recibe el nombre de *amalgama*.

□

En el Capítulo 5 de esta tesis se presentará una definición de la operación amalgama en términos matriciales y se analizará la utilidad de esta operación en el contexto del *problema de los ceros* en los datos composicionales.

Para completar este apartado de definiciones restan por introducir dos operaciones que dotan al símplex de estructura de espacio vectorial. Estas dos operaciones son la *perturbación* y la *transformación potencia* (del inglés, *power transformation*).

2.2.5 El espacio vectorial $(\mathcal{S}^D, \circ, \cdot)$

Definición 2.7 Sea $\mathbf{x}$ una composición de D partes y $\mathbf{p}$ un vector perteneciente a $\mathbb{R}_+^D$. Entonces, la operación

$$\mathbf{p} \circ \mathbf{x} = \mathcal{C}\left(p_1 x_1, p_2 x_2, \ldots, p_D x_D\right) = \left(\frac{p_1 x_1}{\sum p_i x_i}, \frac{p_2 x_2}{\sum p_i x_i}, \ldots, \frac{p_D x_D}{\sum p_i x_i}\right)$$

se denomina *perturbación.*

Diremos que la composición original $\mathbf{x}$ se ve alterada por el *vector de perturbaciones* $\mathbf{p}$ para formar una *composición perturbada* $\mathbf{p} \circ \mathbf{x}$. □

Como se verá más adelante, las perturbaciones sobre las composiciones de $\mathcal{S}^D$ vienen a jugar un papel equivalente al que desempeña el grupo de las traslaciones en $\mathbb{R}^D$ y, dentro de un contexto paramétrico, similar al papel del *ruido blanco* aditivo de las distribuciones normales multivariantes.

Propiedad 2.1 Sea $\mathbf{x}$ una composición de D partes y $\mathbf{p}$ un vector perteneciente a $\mathbb{R}_+^D$. Entonces se cumple que $\mathbf{p} \circ \mathbf{x} = \mathcal{C}(\mathbf{p}) \circ \mathbf{x}$. □

La propiedad anterior, cuya demostración omitimos por su simplicidad, nos indica que, sin pérdida de generalidad, podemos restringir la transformación perturbación a elementos $\mathbf{p}$ pertenecientes al símplex. Por lo tanto, puede considerarse la perturbación como una operación interna del espacio $\mathcal{S}^D$.

La propiedad siguiente nos indica que la operación perturbación es totalmente compatible con la transformación subcomposición. Esta propiedad nos será de utilidad para establecer en las secciones siguientes una medida de tendencia central adecuada al carácter composicional.

Propiedad 2.2 Sean $\mathbf{x}$, $\mathbf{x}^* \in \mathcal{S}^D$ y $\mathbf{x}_s$, $\mathbf{x}_s^*$ sus dos subcomposiciones resultantes de contemplar un mismo subconjunto de s partes. Entonces se cumple que $\mathbf{x}_s \circ \mathbf{x}_s^* = (\mathbf{x} \circ \mathbf{x}^*)_s$. □

Queremos resaltar el hecho de que considerar las traslaciones como transformación en el símplex carece de sentido. Si sumamos dos composiciones como si fueran vectores de $\mathbb{R}^D$, el resultado $\mathbf{x} + \mathbf{x}^*$ nunca pertenecerá a $\mathcal{S}^D$. Para solventar este problema se podría pensar en aplicar el operador clausura al resultado, $\mathcal{C}(\mathbf{x}+\mathbf{x}^*)$. Esta operación tiene, entre otros, el inconveniente de no ser compatible con la transformación subcomposición, es decir, $\mathcal{C}(\mathbf{x}_s+\mathbf{x}_s^*) \neq (\mathcal{C}(\mathbf{x}+\mathbf{x}^*))_s$. Sin embargo, esta operación traslación tiene su utilidad como expresión matemática de la operación natural *mezcla* de dos composiciones. Es decir, dadas $\mathbf{w}$, $\mathbf{w}^*$ dos observaciones de $\mathbb{R}_+^D$, su *mezcla* $\mathbf{w} + \mathbf{w}^*$ satisface que $\mathcal{C}(\mathbf{w} + \mathbf{w}^*) = \mathcal{C}(\mathcal{C}(\mathbf{w}) + \mathcal{C}(\mathbf{w}^*))$. En la Sección 2.6, donde se tratarán las medidas de tendencia central, volveremos a referirnos a la transformación traslación.

A continuación presentamos una serie de propiedades sencillas que nos dan idea del importante papel que juega la operación perturbación entre datos composicionales.

Propiedad 2.3 Sean $\mathbf{x}$, $\mathbf{x}^*$, $\mathbf{x}'$ tres composiciones de D partes y sea $\mathbf{e} = (\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D})$ el *baricentro* del símplex $\mathcal{S}^D$. Entonces se cumple que

i. $\mathbf{x} \circ \mathbf{x}^* = \mathbf{x}^* \circ \mathbf{x}$;

ii. $(\mathbf{x} \circ \mathbf{x}^*) \circ \mathbf{x}' = \mathbf{x} \circ (\mathbf{x}^* \circ \mathbf{x}')$;

iii. $\mathbf{e} \circ \mathbf{x} = \mathbf{x}$;

iv. $\mathbf{x} \circ \mathbf{x}^{-1} = \mathbf{e} \iff \mathbf{x}^{-1} = \mathcal{C}(\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_D})$. □

Omitimos la demostración de las propiedades anteriores debido a su extrema simplicidad. Nótese que el símplex $\mathcal{S}^D$ con la operación interna perturbación tiene estructura de grupo conmutativo.

La definición y propiedad siguientes presentan la operación *transformación potencia*. Consiste en una operación *externa* de un elemento del símplex por un escalar de $\mathbb{R}$. Nuevamente se omiten las demostraciones por su simplicidad.

Definición 2.8 Sea $\mathbf{x} \in \mathcal{S}^D$ y α un escalar perteneciente a $\mathbb{R}$. Definimos en $\mathcal{S}^D$ una operación externa del producto de un escalar por una composición a partir de la expresión

$$\alpha \cdot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha).$$

□

Propiedad 2.4 Sean $\mathbf{x}$, $\mathbf{x}^* \in \mathcal{S}^D$ y sean α, β dos escalares. Entonces se cumple que

i. $\alpha \cdot (\mathbf{x} \circ \mathbf{x}^*) = (\alpha \cdot \mathbf{x}) \circ (\alpha \cdot \mathbf{x}^*)$;

ii. $(\alpha + \beta) \cdot \mathbf{x} = (\alpha \cdot \mathbf{x}) \circ (\beta \cdot \mathbf{x})$;

iii. $\alpha \cdot (\beta \cdot \mathbf{x}) = (\alpha\beta) \cdot \mathbf{x}$, donde $\alpha\beta$ representa el producto habitual en $\mathbb{R}$;

iv. $1 \cdot \mathbf{x} = \mathbf{x}$. □

Queremos poner énfasis en el hecho de que con las Propiedades 2.3 y 2.4 se ha dotado al símplex $\mathcal{S}^D$ de una estructura de espacio vectorial sobre el cuerpo de los números reales $\mathbb{R}$. En la figura 2.4 queda plasmada esta *visión vectorial* de los datos composicionales. Se representa un mismo vector perturbación con origen en $\mathbf{x}$ y extremo $\mathbf{p} \circ \mathbf{x}$ y, con origen en $\mathbf{e}$ y extremo en $\mathbf{p}$.

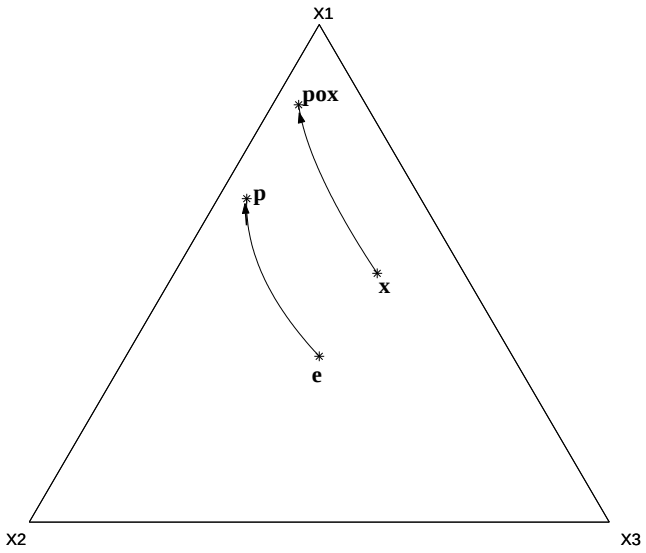


Figura 2.4: *Representación "vectorial" de una perturbación en* $\mathcal{S}^3$ *para* $\mathbf{x} = (0.5, 0.15, 0.35)$, $\mathbf{p} = (0.65, 0.3, 0.05)$, *y* $\mathbf{e} = (1/3, 1/3, 1/3)$.

En la figura 2.5 mostramos un ejemplo de la operación producto por escalar. En la figura 2.6 se observa la representación de la perturbación que nos transforma la composición $\mathbf{x}$ en la composición $\mathbf{x}^*$ cuya expresión matemática es

$$\mathbf{x}^* \circ \mathbf{x}^{-1} = \mathcal{C}\left(\frac{x_1^*}{x_1}, \frac{x_2^*}{x_2}, \ldots, \frac{x_D^*}{x_D}\right). \tag{2.3}$$

2.2.6 Matrices elementales

Recogemos a continuación una serie de *matrices elementales* (Aitchison, 1986) que utilizaremos posteriormente en las demostraciones de algunas de las propiedades de ésta y de las siguientes

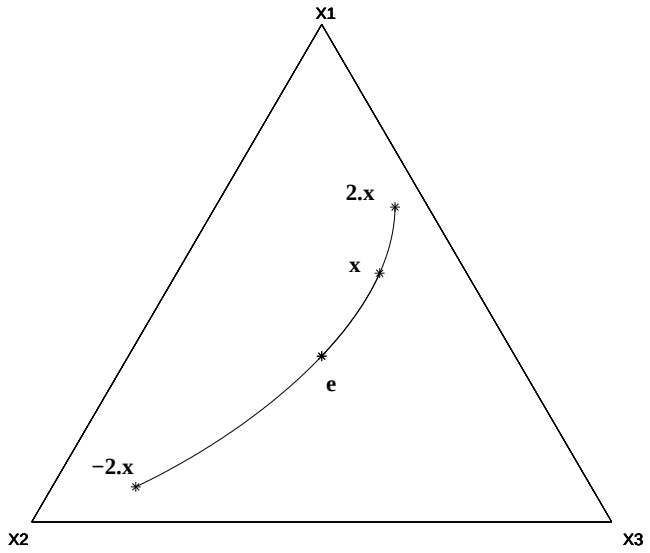


Figura 2.5: *Representación del producto* $\alpha \cdot \mathbf{x}$, *para* $\mathbf{x} = (0.5, 0.15, 0.35)$, $\alpha = 2$ *y* $\alpha = -2$.

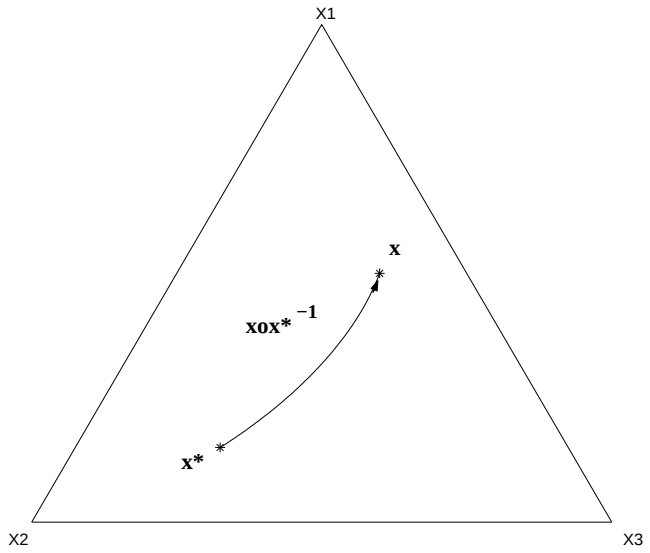


Figura 2.6: *Representación vectorial de la perturbación* $\mathbf{x} \circ \mathbf{x}^{*-1}$ *para* $\mathbf{x} = (0.5, 0.15, 0.35)$ *y* $\mathbf{x}^* = (0.15, 0.6, 0.25)$.

secciones.

Definición 2.9 Definimos las *matrices elementales* siguientes:

a. $\mathbf{I}_{D-1}$: matriz identidad de orden $D-1$.

b. $\mathbf{J}_{D-1}$: matriz cuadrada de orden $D-1$ con todos sus elementos iguales a 1.

c. $\mathbf{j}_{D-1}$: vector columna $\{D-1\} \times 1$, cuyas $D-1$ componentes son todas iguales a 1.

d. $\mathbf{F}_{D-1,D}$: matriz $\{D-1\} \times D$ que se obtiene al añadir a la matriz identidad $\mathbf{I}_{D-1}$ una última columna igual a $-\mathbf{j}_{D-1}$: $\mathbf{F}_{D-1,D} = [\mathbf{I}_{D-1} : -\mathbf{j}_{D-1}]$.

e. $\mathbf{H}_{D-1}$: matriz cuadrada de orden $D-1$ igual a $\mathbf{I}_{D-1} + \mathbf{J}_{D-1}$.

f. $\mathbf{G}_D$: matriz cuadrada de orden D cuyos elementos son todos igual a $-\frac{1}{D}$ excepto los de la diagonal que toman el valor $1-\frac{1}{D}$. Esta matriz puede expresarse como $\mathbf{G}_D = \mathbf{I}_D - \frac{1}{D}\mathbf{J}_D$.

□

2.3 Las transformaciones logcociente aditiva y logratio centrada

La forma tradicional y extensamente aplicada de abordar el estudio de la interdependencia entre las partes de una composición $\mathbf{x}$, ha sido a través de las covarianzas y las correlaciones de las componentes $x_1, x_2, \ldots, x_D$, tal cual aparecen en $\mathbf{x}$. Desde Pearson (1897) hasta Chayes (1983), se ha puesto repetidamente de relieve la dificultad de interpretar correctamente las covarianzas y los coeficientes de correlación entre las componentes directas de una composición. Algunas de estas dificultades de interpretación se deben al sesgo negativo y la incoherencia con las subcomposiciones (Aitchison, 1986; Barceló, 1996). Todas las dificultades de interpretación vienen motivadas por el hecho ya comentado de centrar equivocadamente la atención en las magnitudes absolutas de las componentes de una composición $\mathbf{x} = (x_1, x_2, \ldots, x_D)$. Nuestra atención debe centrarse, en cambio, por los motivos mencionados anteriormente en la magnitud relativa de las componentes, es decir, en los cocientes x_k/x_c $(k, c = 1, 2, \ldots, D; k \neq c)$. Las definiciones que presentamos a continuación se basan en la concepción dada por Aitchison (1986) en torno al concepto de variabilidad de un conjunto de datos composicionales. La idea consiste en transformar los datos composicionales de manera que el espacio soporte de los datos transformados sea el espacio real multidimensional. Presentamos dos transformaciones: la logcociente aditiva y la logratio centrada. Vamos a presentar sus diferencias, sus ventajas y sus inconvenientes, sin olvidar que son dos transformaciones íntimamente relacionadas.

2.3.1 La transformación logcociente aditiva

Definición 2.10 Dado un dato composicional $\mathbf{x}$ de D-partes, la *transformación logcociente aditiva*, alr, de $\mathbf{x} \in \mathcal{S}^D$, en $\mathbf{y} \in \mathbb{R}^{D-1}$ se define por

$$\mathbf{y} = \text{alr}(\mathbf{x}) = (\log \frac{x_1}{x_D}, \log \frac{x_2}{x_D}, \ldots, \log \frac{x_{D-1}}{x_D})$$

que abreviaremos por

$$\mathbf{y} = \text{alr}(\mathbf{x}) = \log \frac{\mathbf{x}_{-D}}{x_D}. \tag{2.4}$$

□

Aunque en la definición podría usarse cualquier base para la función logaritmo, la función log más utilizada es el logaritmo neperiano.

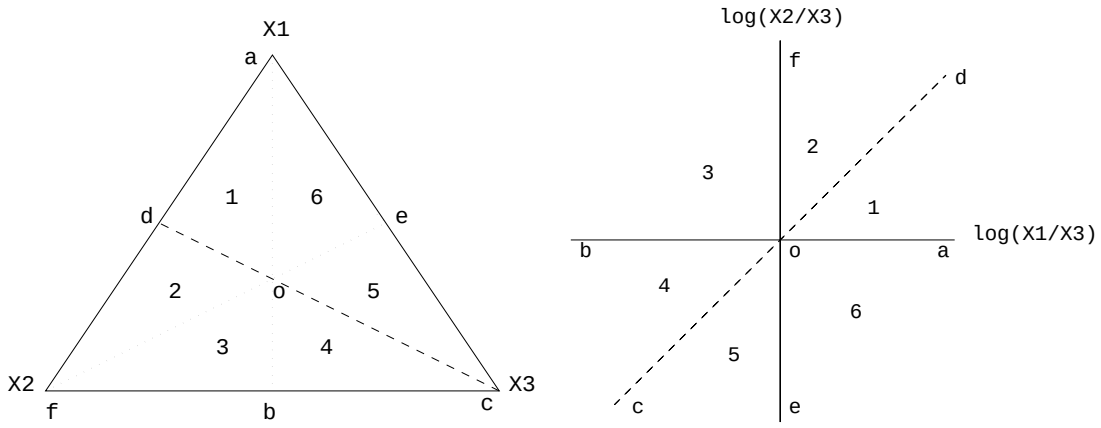
Propiedad 2.5 La transformación alr es una transformación biyectiva, cuya transformación inversa no es otra que la *transformación logística aditiva generalizada* agl que transforma el vector $\mathbf{y} \in \mathbb{R}^{D-1}$ en el vector $\mathbf{x} \in \mathcal{S}^D$ definido por

$$\begin{aligned} x_k &= \frac{\exp(y_k)}{\sum_{c=1}^{D-1} \exp(y_c) + 1} \qquad (k = 1, \ldots, D-1), \\ x_D &= 1 - x_1 - x_2 \ldots - x_{D-1} = \frac{1}{\sum_{c=1}^{D-1} \exp(y_c) + 1}. \end{aligned}$$

□

En el caso de dimensión $D = 3$, la transformación alr aplica las 6 regiones en que las tres alturas dividen al triángulo de la figura 2.7(a), en las regiones del plano $\mathbb{R}^2$ que se muestran en la figura 2.7(b).

Un inconveniente de la transformación alr es su falta de simetría, dado que una de las partes de la composición –la parte x_D que figura en el denominador– adquiere un protagonismo especial frente al resto de componentes. Este hecho implica, por ejemplo, que no podemos definir una medida de diferencia entre dos observaciones composicionales $\mathbf{x}$ y $\mathbf{x}^*$ basándonos en la distancia euclídea entre las respectivas observaciones transformadas $\mathbf{y}$ e $\mathbf{y}^*$. Sin embargo, este inconveniente no es tan grave como pudiera parecer en un principio puesto que las medidas de diferencia, de tendencia central y de dispersión se definirán de manera que sean compatibles con las permutaciones de las partes de los datos.



(a) (b)

Figura 2.7: (a) *Regiones en que queda dividido el símplex* $\mathcal{S}^3$ *por las bisectrices del triángulo;* (b) *Regiones de* $I\!R^2$ *en correspondencia con las regiones de* $\mathcal{S}^3$ *por la transformación* alr.

A la transformación alr se le asocia la denominada *matriz de covarianzas de logcocientes* de una composición $\mathbf{x} \in \mathcal{S}^D$ definida por

$$\mathbf{\Sigma} = [\sigma_{kc}] = [\mathrm{cov}\{\log(\frac{x_k}{x_D}), \log(\frac{x_c}{x_D})\} : k, c = 1, 2, \ldots, D-1].$$

Observamos que la matriz de covarianzas de logcocientes $\mathbf{\Sigma}$ no es más que la matriz de covarianzas del vector aleatorio $\mathbf{y}$ de $\mathbb{R}^{D-1}$ que se obtiene al aplicar la transformación alr a la composición $\mathbf{x}$. La matriz de covarianzas $\mathbf{\Sigma}$ no posee ninguna restricción más que la de ser simétrica y definida no negativa, por tratarse de una matriz de covarianzas.

2.3.2 La transformación logratio centrada

Definición 2.11 Dado un dato composicional $\mathbf{x}$ de D-partes, la *transformación logcociente centrada* clr de $\mathbf{x} \in \mathcal{S}^D$ en $\mathbf{z} \in \mathbb{R}^D$ se define por

$$\mathbf{z} = \mathrm{clr}(\mathbf{x}) = \log \frac{\mathbf{x}}{g(\mathbf{x})}, \tag{2.5}$$

donde $g(\mathbf{x})$ es la media geométrica $(x_1 x_2 \ldots x_D)^{1/D}$ de las D componentes de $\mathbf{x}$. □

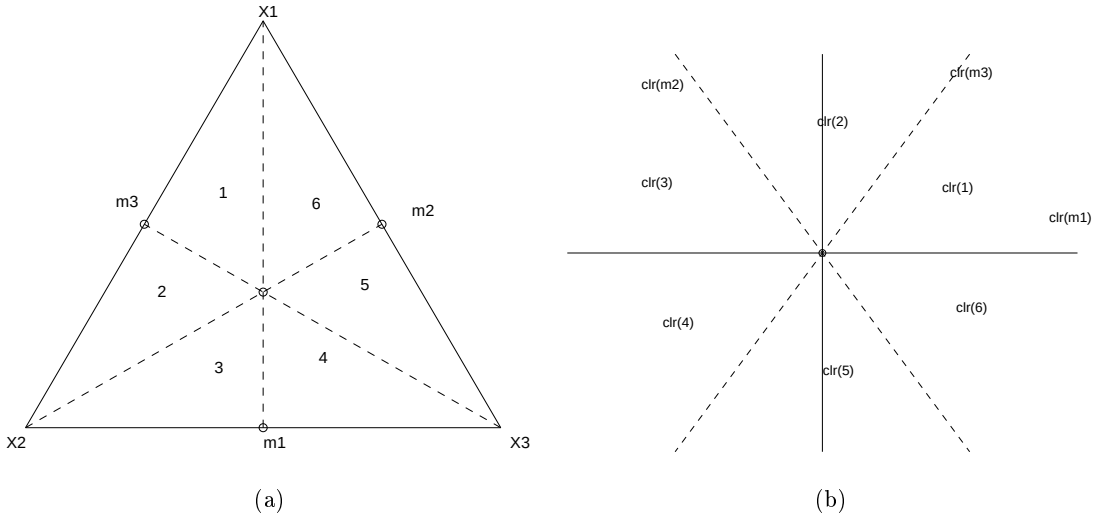
Propiedad 2.6 La transformación clr es biyectiva y su transformación inversa, que denominamos ilc, viene definida por

$$x_k = \frac{\exp(z_k)}{\sum_{c=1}^{D} \exp(z_c)} \qquad (k = 1, 2, \ldots, D). \tag{2.6}$$

□

Las transformaciones clr e ilc establecen una relación biunívoca entre el símplex y el hiperplano $\mathrm{clr}(\mathcal{S}^D) = \{ \mathbf{z} \in \mathbb{R}^D \ / \ \sum z_k = 0\}$. Este hiperplano de $\mathbb{R}^D$, de dimensión $D-1$, puede representarse, para el caso $D=3$ en un gráfico bidimensional mediante la realización previa de un cambio de base.

Consideremos las 6 regiones en que las tres alturas o mediatrices $-m1, m2, m3-$ dividen al triángulo $\mathcal{S}^3$ de la figura 2.8(a). La aplicación clr transforma esas regiones en las regiones del plano $\mathbb{R}^2$ que se muestran en la figura 2.8(b).



(a) (b)

Figura 2.8: (a) *Regiones en que queda dividido el símplex $\mathcal{S}^3$ por las mediatrices del triángulo;* (b) *Regiones de $\mathbb{R}^2$ en correspondencia con las regiones de $\mathcal{S}^3$ por la transformación* clr.

El cambio de base realizado en el hiperplano $\mathrm{clr}(\mathcal{S}^3)$ viene dado por la matriz:

$$\begin{pmatrix} \frac{\sqrt{6}}{3} & -\frac{\sqrt{6}}{6} & -\frac{\sqrt{6}}{6} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \end{pmatrix}. \tag{2.7}$$

Obsérvese que la condición $\sum z_k = 0$ que satisfacen los puntos del hiperplano $\mathrm{clr}(\mathcal{S}^3)$ es equivalente a exigir que $< \mathbf{z}, \mathbf{j}_3 >= 0$, donde $\mathbf{j}_3 = (1,1,1) \in \mathbb{R}^3$. Por lo tanto, el vector $\mathbf{j}_3$ es ortogonal al hiperplano $\mathrm{clr}(\mathcal{S}^3)$. Entonces, para obtener la matriz del cambio de base, partimos del vector $\frac{1}{\sqrt{3}}\mathbf{j}_3$ y aplicamos el método de ortogonalización de Gram-Schmidt.

Mediante la transformación clr puede establecerse una relación entre las estructuras de espacio vectorial de los conjuntos $\mathcal{S}^D$ y $\mathbb{R}^D$. Esta relación se muestra en la propiedad siguiente cuya demostración omitimos dada su simplicidad.

Propiedad 2.7 Sean $\mathbf{x}, \mathbf{x}^*$ dos composiciones de D partes y sean α, β dos escalares. Entonces

se cumple que

$$\mathrm{clr}((\alpha \cdot \mathbf{x}) \circ (\beta \cdot \mathbf{x}^*)) = \alpha \mathrm{clr}(\mathbf{x}) + \beta \mathrm{clr}(\mathbf{x}^*). \tag{2.8}$$

□

Es importante poner énfasis en el hecho de que no sólo se ha dotado al símplex $\mathcal{S}^D$ de una estructura de espacio vectorial –veánse las Propiedades 2.3 y 2.4– sino que adicionalmente podemos transportar nuestra intuición euclídea al campo de los datos composicionales mediante la relación (2.8). Para ilustrar esta idea podemos observar en la figura 2.9 de qué manera rectas en el espacio real –figura 2.9(a)– se transforman en *rectas composicionales* en el símplex –figura 2.9(b)– a partir de la transformación ilc. Se han numerado puntos en las dos figuras con el propósito de resaltar la correspondencia biunívoca entre los dos espacios.

El hecho de que las rectas sean geodésicas en un espacio euclídeo nos proporciona la posibilidad de definir *geodésicas composicionales*. Esta definición se basará en una definición de *distancia composicional* que esté relacionada con la distancia euclídea mediante la transformación clr. En la sección siguiente se tratarán en profundidad los conceptos relacionados con las medidas de diferencia para datos composicionales.

Por otra parte, si $\mathbf{x}$ es una composición, denominamos *matriz de covarianzas de logcocientes centrados* de $\mathbf{x}$ a la matriz $D \times D$

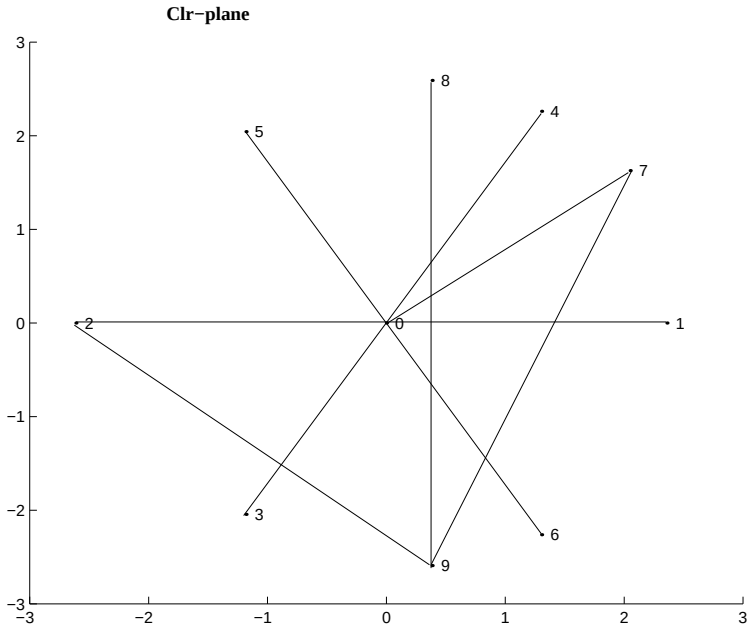
$$\mathbf{\Gamma} = [\gamma_{ij}] = \Big[\mathrm{cov}[\log \frac{x_i}{g(\mathbf{x})}, \log \frac{x_j}{g(\mathbf{x})}] : i, j = 1, 2, \ldots, D\Big].$$

Obsérvese que la matriz $\mathbf{\Gamma}$ no es más que la matriz de covarianzas del vector aleatorio $\mathbf{z}$ de $\mathbb{R}^D$ que se obtiene al aplicar la transformación clr a la composición $\mathbf{x}$. Se trata de una matriz singular dado que, al igual que la matriz de covarianzas directas, la suma de los elementos de una fila cualquiera de $\mathbf{\Gamma}$ es siempre igual a cero. El inconveniente de la falta de simetría que se nos presenta en la definición de la transfomación alr no aparece en la definición de la transformación clr. Sin embargo, la estructura de covarianza asociada a esta transformación y recogida en la matriz $\mathbf{\Gamma}$ no queda liberada del inconveniente de la singularidad de la matriz de covarianzas.

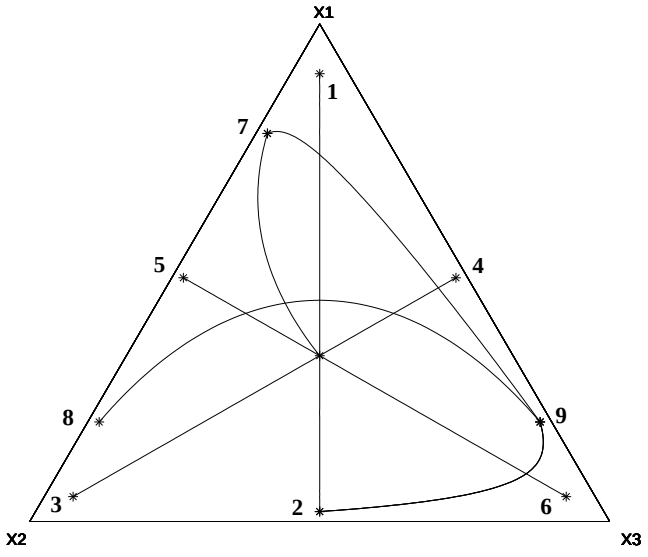
Si se conoce una cualquiera de dos matrices –$\mathbf{\Sigma}$ o $\mathbf{\Gamma}$– es posible calcular la otra (Aitchison, 1986), puesto que si consideramos que $\mathbf{z} = \mathrm{clr}(\mathbf{x})$, $\mathbf{y} = \mathrm{alr}(\mathbf{x})$ y que las matrices $\mathbf{F}$ y $\mathbf{H}$ son las dadas en la Definición 2.9 pueden establecerse las relaciones:

$$\mathbf{y}^t = \mathbf{F}\mathbf{z}^t; \quad \mathbf{\Gamma} = \mathbf{F}^t\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}\mathbf{F}; \text{ y } \quad \mathbf{\Sigma} = \mathbf{F}\mathbf{\Gamma}\mathbf{F}^t. \tag{2.9}$$

Por tanto, para conocer la estructura de covarianza de una composición bastará disponer de una cualquiera de las matrices de covarianza $\mathbf{\Sigma}$ o $\mathbf{\Gamma}$.



(a)



(b)

Figura 2.9: (a) *Rectas en el plano* clr-*transformado.* (b) *Rectas composicionales correspondientes por la transformación* ilc *en* $\mathcal{S}^3$.

2.4 Medida de diferencia entre dos datos composicionales

2.4.1 Requerimientos para una medida de diferencia

Reconocemos de antemano que el hecho de utilizar una medida de diferencia no adecuada a la tipología de los datos no es impedimento para que en determinadas ocasiones puedan obtenerse clasificaciones con resultados razonables. También somos conscientes que el uso de una disimilitud adecuada a la naturaleza de los datos no garantiza, por si misma, una agrupación libre de errores.

Sin embargo, tal y como se ilustrará en esta sección, entendemos que el hecho de utilizar una medida de diferencia compatible con el carácter composicional de los datos supone dar coherencia al análisis de los datos. Es por este motivo que, adicionalmente a los aspectos establecidos en las Definiciones 1.1, 1.2, y 1.3, en Aitchison (1992) se definen otras tres propiedades que debe cumplir cualquier medida de diferencia entre dos D-partes de $\mathcal{S}^D$ como consecuencia de la propia naturaleza de los datos composicionales. Adicionalmente, a lo largo de nuestra investigación, ha surgido la necesidad de reflejar la coherencia entre las medidas de diferencia entre datos composicionales y la estructura de espacio vectorial del símplex $\mathcal{S}^D$. Es por este motivo que hemos añadido un cuarto requisito referente a la operación del producto por escalar. Presentamos a continuación estos cuatro requisitos.

Propiedad 2.8 Si $\mathbf{x}, \mathbf{x}^*$ simbolizan dos D-partes cualesquiera de $\mathcal{S}^D$, cualquier medida de diferencia d debe cumplir las siguientes propiedades:

R1. Invariación respecto de las permutaciones de sus partes:

$$d(\mathbf{P}\mathbf{x}^t, \mathbf{P}\mathbf{x}^{*t}) = d(\mathbf{x}, \mathbf{x}^*), \quad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D,$$

sea cual sea la permutación $\mathbf{P}$ que se aplique a las partes de una composición.

R2. Dominación respecto de las subcomposiciones:

$$d_D(\mathbf{x}, \mathbf{x}^*) \geq d_s(\mathcal{C}(\mathbf{x}_s), \mathcal{C}(\mathbf{x}^*{}_s)), \quad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D,$$

sea cual sea la subcomposición escogida (d_D y d_s simbolizan, respectivamente, la medida de diferencia d aplicada a los datos composicionales de $\mathcal{S}^D$ y $\mathcal{S}^s$).

R3. Invariación respecto de las perturbaciones:

$$d(\mathbf{p} \circ \mathbf{x}, \mathbf{p} \circ \mathbf{x}^*) = d(\mathbf{x}, \mathbf{x}^*), \quad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D, \ \forall \mathbf{p} \in \mathbb{R}_+^D.$$

R4. Coherencia respecto al producto por un escalar:

$$d(\alpha \cdot \mathbf{x}, \alpha \cdot \mathbf{x}^*) = |\alpha| \, d(\mathbf{x}, \mathbf{x}^*), \quad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D, \ \forall \alpha \in \mathbb{R}.$$

□

Observamos que el requisito R1 exige que la medida de diferencia entre dos D-partes no dependa del orden en que las componentes estén dispuestas.

El requerimiento R2 exige que la medida de diferencia entre dos D-partes $\mathbf{x}$ y $\mathbf{x}^*$ sea mayor o igual que la medida de diferencia entre cualquiera de sus subcomposiciones respectivas $\mathbf{x}_s$ y ${\mathbf{x}^*}_s$. Por lo tanto, se exige que la diferencia entre dos composiciones no debe aumentar por el hecho de comparar menos partes de los datos. Recordemos que, por su definición la transformación subcomposición juega un papel análogo a la operación proyección de datos multivariantes de $\mathbb{R}^D$ sobre espacios de dimensión inferior. La propiedad R2 pone de manifiesto que la medida diferencia entre dos composiciones debe ser mayor o igual a la medida de diferencia entre sus *proyecciones* respectivas –véase la figura 2.10.

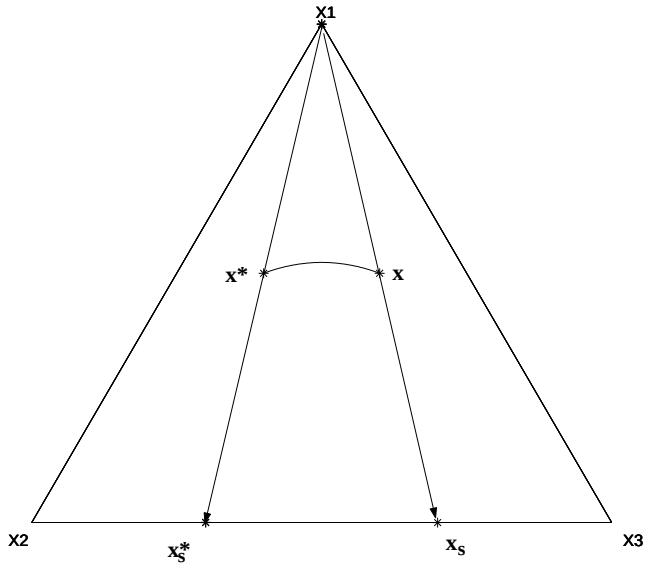


Figura 2.10: *Las composiciones* $\mathbf{x} = (0.5, 0.15, 0.35)$, $\mathbf{x}^* = (0.5, 0.35, 0.15)$ *y sus subcomposiciones* $\mathbf{x}_s = (0.3, 0.7)$, $\mathbf{x}_s^* = (0.7, 0.3)$.

Los requisitos R3 y R4 expresan que una medida de diferencia entre datos composicionales debe ser compatible con la estructura de espacio vectorial que posee el símplex $\mathcal{S}^D$ con las operaciones perturbación y transformación potencia. El requisito R3 nos exige que si aplicamos

una misma perturbación $\mathbf{p}$ a dos datos composicionales $\mathbf{x}$ y $\mathbf{x}^*$ cualesquiera, entonces las D-partes transformadas $\mathbf{p} \circ \mathbf{x}$ y $\mathbf{p} \circ \mathbf{x}^*$ deben diferir en la misma cantidad que las D-partes iniciales $\mathbf{x}$ y $\mathbf{x}^*$ –véase la figura 2.11.

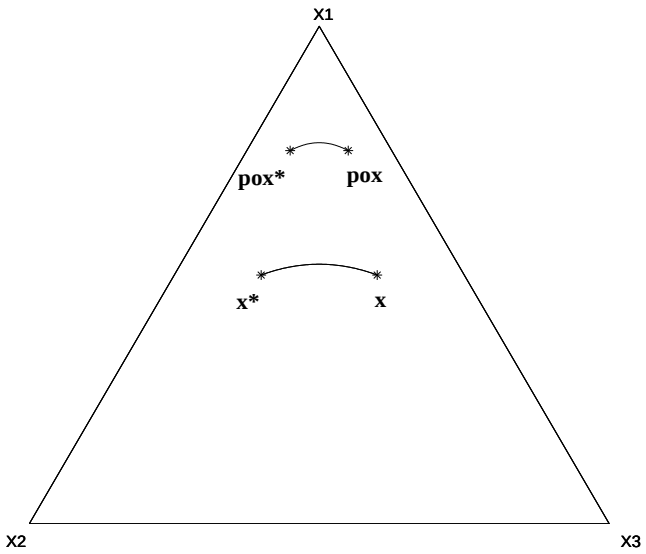


Figura 2.11: *Representación de los elementos* $\mathbf{x} = (0.5, 0.15, 0.35)$, $\mathbf{x}^* = (0.5, 0.35, 0.15)$ *y de sus perturbaciones* $\mathbf{p} \circ \mathbf{x}$ *y* $\mathbf{p} \circ \mathbf{x}^*$, *donde* $\mathbf{p} = (0.6, 0.2, 0.2)$.

Si consideramos las perturbaciones como el grupo de transformaciones en el símplex $\mathcal{S}^D$, parece lógico exigir que una medida de diferencia definida en $\mathcal{S}^D$ sea invariante por perturbaciones, al igual que se exige que una medida de diferencia definida en un espacio real sea invariante por traslaciones. En la práctica la necesidad de realizar una perturbación puede aparecer en aquellos problemas en los que el investigador observa que las D partes de una observación de $\mathbb{R}_+^D$ están expresadas en unidades de medida diferentes. Si se desea expresar todas las partes en la misma unidad de medida será necesario aplicar una perturbación cuyas partes serán los factores de cambio de unidad. Si el investigador utiliza una medida de diferencia que no sea invariante por perturbaciones, las diferencias entre los datos antes o después de perturbar pueden ser totalmente dispares y, por lo tanto, el análisis estadístico de los datos puede ser totalmente erróneo.

Finalmente, el requisito R4 exige que la medida de diferencia sea coherente con la transformación potencia. Obsérvese que este requisito no es más que el conocido "*Teorema de Thales*". La figura 2.12 nos ilustra este requisito para el caso del símplex $\mathcal{S}^3$.

En las figuras 2.10, 2.11, y 2.12 se observa cómo estos requisitos pueden llegar a parecer poco intuitivos. Nótese que en las tres figuras los puntos escogidos para ilustrar las propiedades

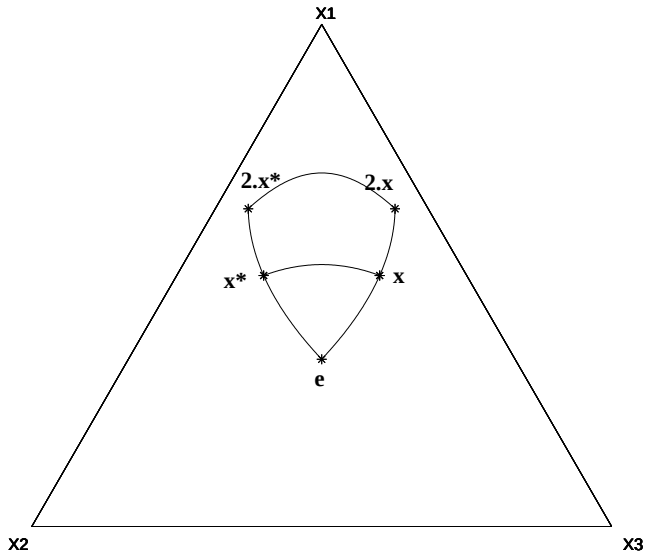


Figura 2.12: *Las composiciones* $\mathbf{x} = (0.5, 0.15, 0.35)$, $\mathbf{x}^* = (0.5, 0.35, 0.15)$ *y sus correspondientes* $2 \cdot \mathbf{x}$, $2 \cdot \mathbf{x}^*$.

aparecen unidos mediante *geodésicas composicionales*. De esta manera, desde un punto de vista euclídeo, puede parecer que la diferencia entre las subcomposiciones $\mathbf{x}_s$ y $\mathbf{x}_s^*$, y que la diferencia entre las composiciones perturbadas $\mathbf{p} \circ \mathbf{x}$ y $\mathbf{p} \circ \mathbf{x}^*$, no verifican, respectivamente, los requisitos R2 y R3. En realidad este hecho es una indicación de que la distancia euclídea no es una distancia adecuada para medir las diferencias entre datos composicionales. Todas estas propiedades han sido objeto de estudio en varios trabajos recientes: Aitchison (1997), Martín-Fernández et al. (1998a, 1998c), y Aitchison (2000, 2001).

2.4.2 Medidas de diferencia más usuales entre dos observaciones

El hecho de considerar las perturbaciones como el grupo de transformaciones sobre $\mathcal{S}^D$ implica que cualquier medida de diferencia entre $\mathbf{x}$ y $\mathbf{x}^*$ que sea compatible con el carácter composicional debe basarse en la perturbación que nos pasa de una composición a la otra –véase la expresión (2.3). En Aitchison (1992) se demuestra que cualquier medida de diferencia entre dos datos composicionales compatible con la naturaleza composicional de estos, debe poder expresarse en función de los ratios entre las componentes de las observaciones, es decir,

$$d(\mathbf{x}, \mathbf{x}^*) = f(\frac{x_1}{x_1^*}, \frac{x_2}{x_2^*}, ..., \frac{x_D}{x_D^*}). \tag{2.10}$$

Este resultado ratifica que lo realmente importante en el momento de establecer diferencias entre observaciones son las proporciones entre sus componentes.

En la bibliografía científica se encuentra un considerable número de medidas de diferencia susceptibles de ser aplicadas a los datos composicionales. En Martín (1996) se analiza el comportamiento de algunas de las disimilitudes más habituales en estudios de clasificación automática no paramétrica de datos composicionales. La tabla 2.1 contiene una relación de algunas de las disimilitudes más usuales que pueden definirse sobre el espacio $\mathcal{S}^D$. Es importante destacar que, a diferencia del resto de disimilitudes de la tabla 2.1, las distancias de Mahalanobis(raw) y de Mahalanobis(clr) en la práctica vienen siempre asociadas a un conjunto de datos –véanse las Definiciones 1.7 y 1.9 del Capítulo 1. Por este motivo, estas dos medidas se analizan en la subsección siguiente.

A continuación exponemos los aspectos más relevantes de cada una de las disimilitudes que aparecen en la tabla 2.1 poniendo énfasis en su compatibilidad composicional. Al finalizar esta exposición se presenta un ejemplo para mostrar a modo de contraejemplo qué disimilitudes no verifican algunos o todos estos requisitos.

- *La distancia de Aitchison*

  Resulta sencillo demostrar que la medida de diferencia

  $$\Lambda(\mathbf{x},\mathbf{x}^*) = \Big[\sum_{\substack{k,c=1 \\ k<c}}^{D} \Big[\log\Big(\frac{x_k/x_c}{x_k^*/x_c^*}\Big)\Big]^2 \Big]^{1/2},$$

  definida en Aitchison (1992) es una distancia que cumple las propiedades R1-R4 enunciadas en la Propiedad 2.8.

  Es también simple demostrar que esta distancia $\Lambda(\mathbf{x},\mathbf{x}^*)$ puede expresarse equivalentemente de forma simétrica respecto de las componentes que integran las D-partes $\mathbf{x}$ y $\mathbf{x}^*$. Más concretamente:

  $$\Lambda(\mathbf{x},\mathbf{x}^*) = D^{1/2}\Big[\sum_{k=1}^{D}\Big(\log(\frac{x_k}{g(\mathbf{x})}) - \log(\frac{x_k^*}{g(\mathbf{x}^*)})\Big)^2\Big]^{1/2},$$

  donde $g(\mathbf{x})$, $g(\mathbf{x}^*)$ representan, respectivamente, la media geométrica de las partes de las observaciones $\mathbf{x}$ y $\mathbf{x}^*$. Observemos que, salvo la constante multiplicativa $D^{1/2}$, $\Lambda(\mathbf{x},\mathbf{x}^*)$ no es más que la distancia euclídea en $\mathbb{R}^D$ entre los datos clr-transformados:

  $$\Lambda(\mathbf{x},\mathbf{x}^*) = D^{1/2} d_{\mathrm{Euc}}(\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{x}^*)) = D^{1/2} d_{\mathrm{Euc}}\left(\log(\frac{\mathbf{x}}{g(\mathbf{x})}),\log(\frac{\mathbf{x}^*}{g(\mathbf{x}^*)})\right).$$

  Siguiendo lo expuesto en Martín-Fernández et al. (1998b), en este trabajo de investigación proponemos eliminar la constante multiplicativa $D^{1/2}$ de la definición. Así, tal y como aparece en la tabla 2.1, definimos la *distancia de Aitchison* como

  $$d_{\mathrm{Ait}}(\mathbf{x},\mathbf{x}^*) = d_{\mathrm{Euc}}(\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{x}^*)) = \Big[\sum_{k=1}^{D}\Big(\log(\frac{x_k}{g(\mathbf{x})}) - \log(\frac{x_k^*}{g(\mathbf{x}^*)})\Big)^2\Big]^{\frac{1}{2}}. \qquad (2.11)$$

Tabla 2.1: Algunas disimilitudes entre datos composicionales.

Disimilitud	$d(\mathbf{x},\mathbf{x}^*)$	
Aitchison	$d_{\text{Ait}}(\mathbf{x},\mathbf{x}^*) =$	$\left[\sum_{k=1}^{D}\left(\log(\frac{x_k}{g(\mathbf{x})})-\log(\frac{x_k^*}{g(\mathbf{x}^*)})\right)^2\right]^{\frac{1}{2}}$
Angular	$d_{\text{Ang}}(\mathbf{x},\mathbf{x}^*) =$	$\arccos\left(\sum_{k=1}^{D}\sqrt{\frac{x_k^2}{\sum x_c^2}}\sqrt{\frac{x_k^{*2}}{\sum x_c^{*2}}}\right)$
Bhattacharyya (arccos)	$d_{\text{B-a}}(\mathbf{x},\mathbf{x}^*) =$	$\arccos\left(\sum_{k=1}^{D}\sqrt{x_k}\sqrt{x_k^*}\right)$
Bhattacharyya(log)	$d_{\text{B-l}}(\mathbf{x},\mathbf{x}^*) =$	$-\log\left(\sum_{k=1}^{D}\sqrt{x_k}\sqrt{x_k^*}\right)$
City Block	$d_{\text{CB}}(\mathbf{x},\mathbf{x}^*) =$	$\sum_{k=1}^{D}\lvert x_k - x_k^*\rvert$
Euclídea	$d_{\text{Euc}}(\mathbf{x},\mathbf{x}^*) =$	$\left[\sum_{k=1}^{D}(x_k - x_k^*)^2\right]^{\frac{1}{2}}$
J-Divergencia	$d_{\text{J-D}}(\mathbf{x},\mathbf{x}^*) =$	$\left[\sum_{k=1}^{D}(\log(x_k)-\log(x_k^*))(x_k - x_k^*)\right]^{\frac{1}{2}}$
Logarítmica	$d_{\text{Log}}(\mathbf{x},\mathbf{x}^*) =$	$\left[\sum_{k=1}^{D}(\log(x_k)-\log(x_k^*))^2\right]^{\frac{1}{2}}$
Mahalanobis (raw)	$d_{\text{M-r}}(\mathbf{x},\mathbf{x}^*) =$	$\left[(\mathbf{x}-\mathbf{x}^*)^t\mathbf{K}^-(\mathbf{x}-\mathbf{x}^*)\right]^{\frac{1}{2}}$
Mahalanobis (clr)	$d_{\text{M-c}}(\mathbf{x},\mathbf{x}^*) =$	$\left[(\text{clr}(\mathbf{x})-\text{clr}(\mathbf{x}^*))^t\mathbf{\Gamma}^-(\text{clr}(\mathbf{x})-\text{clr}(\mathbf{x}^*))\right]^{\frac{1}{2}}$
Matusita	$d_{\text{Mat}}(\mathbf{x},\mathbf{x}^*) =$	$\left[\sum_{k=1}^{D}(\sqrt{x_k}-\sqrt{x_k^*})^2\right]^{\frac{1}{2}}$
Minkowski	$d_{\text{Min}}(\mathbf{x},\mathbf{x}^*) =$	$\left[\sum_{k=1}^{D}\lvert x_k - x_k^*\rvert^q\right]^{\frac{1}{q}}$

Es evidente que el cumplimiento de los requisitos R1, R3 y R4 –véase la Propiedad 2.8– no queda afectado por el hecho de haber eliminado el factor $D^{1/2}$. La propiedad de la dominación por subcomposiciones queda demostrada en la siguiente propiedad.

Propiedad 2.9 Si $\mathbf{x}, \mathbf{x}^*$ simbolizan dos D-partes cualesquiera de $\mathcal{S}^D$, la distancia d_{Ait} cumple que:

$$d_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) \geq d_{\text{Ait}}\left(\mathcal{C}(\mathbf{x}_s), \mathcal{C}(\mathbf{x}^*{}_s)\right), \quad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D,$$

sea cual sea la subcomposición de s partes escogida.

Demostración

Dado que la distancia d_{Ait} es invariante por permutaciones la demostración de esta propiedad se reduce sin pérdida de generalidad al caso $\mathbf{x}_s = \mathcal{C}(x_1, x_2, \ldots, x_{D-1})$. Podemos simplificar aún más la demostración si consideramos que la distancia d_{Ait} es invariante por perturbaciones y que las perturbaciones son compatibles con la transformación subcomposición –véase la Propiedad 2.2. De esta manera, aplicando la perturbación $\mathbf{p} = \mathbf{x}^{*-1}$ a la desigualdad

$$d_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) \geq d_{\text{Ait}}\left(\mathcal{C}(\mathbf{x}_{D-1}), \mathcal{C}(\mathbf{x}^*{}_{D-1})\right), \quad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D,$$

se obtiene que es totalmente equivalente a la desigualdad

$$d_{\text{Ait}}(\mathbf{x}, \mathbf{e}_D) \geq d_{\text{Ait}}\left(\mathcal{C}(\mathbf{x}_{D-1}), \mathbf{e}_{D-1}\right), \quad \forall \mathbf{x} \in \mathcal{S}^D,$$

donde $\mathbf{e}_D$ y $\mathbf{e}_{D-1}$ representan respectivamente el baricentro de los espacios $\mathcal{S}^D$ y $\mathcal{S}^{D-1}$. A partir de la definición dada en (2.11), la expresión anterior elevada al cuadrado puede expresarse como

$$\sum_{k=1}^{D}\left(\log(x_k) - \frac{1}{D}\sum_{c=1}^{D}\log(x_c)\right)^2 \geq \sum_{k=1}^{D-1}\left(\log(x_k) - \frac{1}{D-1}\sum_{c=1}^{D-1}\log(x_c)\right)^2. \tag{2.12}$$

Fijémonos en el primer término de la desigualdad. Podemos interpretarlo cómo la *variabilidad total* T del conjunto de valores $\{\log(x_1), \log(x_2), \ldots, \log(x_D)\}$. De acuerdo con Krzanowski (1988b), en general, si un conjunto de valores se parte en G grupos, la variabilidad total T del conjunto se descompone en la suma de dos términos no negativos $T = B + W$, donde B es la variabilidad entre los G grupos y W es suma de las variabilidades interiores de cada grupo. En nuestro caso podemos considerar que el conjunto $\{\log(x_1), \log(x_2), \ldots, \log(x_D)\}$ lo partimos en dos grupos: uno con $D-1$ elementos,

$\{\log(x_1), \log(x_2), \ldots, \log(x_{D-1})\}$ y el otro con un único elemento, $\{\log(x_D)\}$. Es evidente que este último subconjunto, con un único elemento, tiene variabilidad nula. Entonces podemos interpretar el segundo término de la desigualdad (2.12) como la suma W de la *variabilidad* dentro del subconjunto de valores $\{\log(x_1), \log(x_2), \ldots, \log(x_{D-1})\}$ y de la variabilidad nula del subconjunto $\{\log(x_D)\}$. Teniendo en cuenta que el término B de la descomposición de la variabilidad tiene siempre un valor no negativo queda demostrada la propiedad. □

Es inmediato comprobar como la distancia $d_{\mathrm{Ait}}(\mathbf{x}, \mathbf{x}^*)$ puede también expresarse en función de las componentes de los datos alr-transformados $\mathrm{alr}(\mathbf{x}) = \log(\mathbf{x}/x_D)$ y $\mathrm{alr}(\mathbf{x}^*) = \log(\mathbf{x}^*/x_D^*)$ de $\mathbb{R}^{D-1}$, resultado de aplicar a las D-partes $\mathbf{x}$ y $\mathbf{x}^*$ la transformación logcociente aditiva alr de $\mathcal{S}^D$ en $\mathbb{R}^{D-1}$. Así, se cumple que

$$\begin{aligned} d^2_{\mathrm{Ait}}(\mathbf{x}, \mathbf{x}^*) &= D^{-1/2} \{\mathrm{alr}(\mathbf{x}) - \mathrm{alr}(\mathbf{x}^*)\} \begin{bmatrix} d & -1 & \ldots & -1 \\ -1 & d & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ldots & d \end{bmatrix} \{\mathrm{alr}(\mathbf{x}) - \mathrm{alr}(\mathbf{x}^*)\}^t = \\ &= D^{-1/2} \{\mathrm{alr}(\mathbf{x}) - \mathrm{alr}(\mathbf{x}^*)\} \mathbf{H}^{-1} \{\mathrm{alr}(\mathbf{x}) - \mathrm{alr}(\mathbf{x}^*)\}^t, \end{aligned}$$

donde $\mathbf{H}$ es la matriz elemental presentada en la Definición 2.9.

Para el caso D=3 la figura 2.13 muestra el comportamiento de la distancia d_{Ait}. En la figura se observa que la forma de los entornos en el símplex que se obtienen con la distancia de Aitchison es fuertemente dependiente de la posición de su centro (Martín-Fernández et al., 1998b). Los entornos de centro cercano al baricentro del símplex –punto $\mathbf{o} = (0.3, 0.4, 0.3)$– tienen una cierta forma esférica similar a los que se obtendrían con la distancia euclídea. Sin embargo, si el centro de los entornos está cerca de un vértice –punto $\mathbf{o} = (0.8, 0.1, 0.1)$– o de una cara del símplex –punto $\mathbf{o} = (0.45, 0.1, 0.45)$– los entornos aparecen más apretados y distorsionados con respecto a una circunferencia. Ello es debido a que cerca de los vértices o de las caras del símplex algunas componentes son aproximadamente nulas y, un pequeño cambio en el valor de una componente aproximadamente nula se traduce en una gran diferencia desde el punto de vista composicional. Esta gran diferencia se produce debido a que la distancia $d_{\mathrm{Ait}}(\mathbf{x}, \mathbf{x}^*)$ puede expresarse en función de los ratios entre componentes –véase, por ejemplo, la expresión (2.11)– y, en consecuencia, es compatible con la naturaleza composicional de los datos. Por lo tanto, se explica que un pequeño cambio en una componente aproximadamente nula, que figura en el denominador de una ratio, provoque

una gran diferencia en el valor de la distancia. Este comportamiento de la distancia d_{Ait} es, a nuestro entender, un comportamiento deseable para cualquier medida de diferencia entre dos datos composicionales. De esta manera, los entornos de centro cercano a un vértice o a una cara del símplex tendrán una forma coherente con la naturaleza composicional.

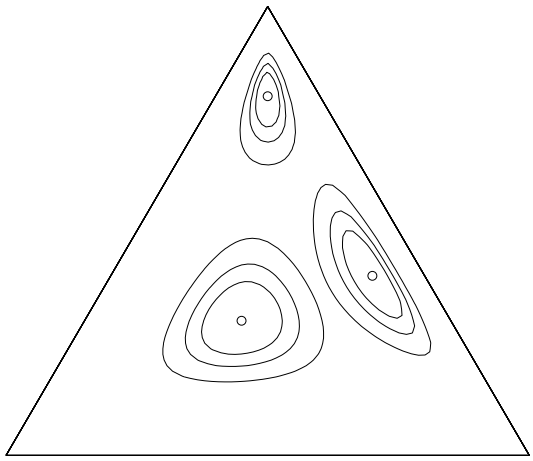
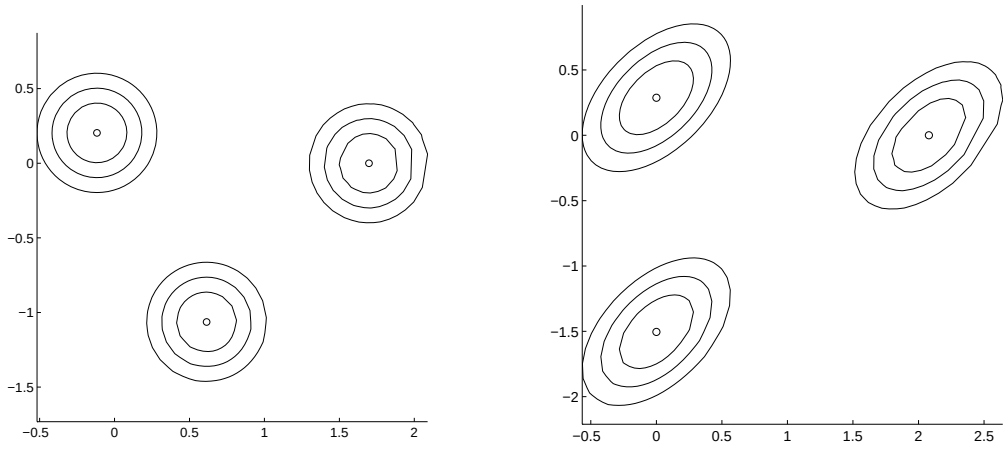


Figura 2.13: *Entornos en $\mathcal{S}^3$ con la distancia de Aitchison.*

Por otra parte, de la expresión (2.11) se desprende claramente que estos mismos entornos representados en el plano clr-transformado serán todos ellos circunferencias. El efecto de cambiar de centro se traduce en una simple traslación de las circunferencias –véase la figura 2.14(a). Por lo que se refiere a la transformación alr en Barceló (1996) se muestra con detalle que los entornos de la figura 2.13 representados en el plano alr-transformado se convierten en elipses. Cuando se cambia el centro de los entornos en el plano alr se observa una simple traslación de las elipses –véase la figura 2.14(b).



(a) (b)

Figura 2.14: Entornos con la distancia de Aitchison: (a) *en el plano* clr*;* (b) *en el plano* alr.

Todas las propiedades expuestas nos conducen a afirmar que, la distancia de Aitchison d_{Ait} es una medida de diferencia adecuada para los datos composicionales. Barceló-Vidal (2000) establece que la expresión de la distancia d_{Ait} en la fórmula (2.11) refleja que, la transformación clr es una isometría entre el símplex $\mathcal{S}^D$ y el hiperplano $\mathrm{clr}(\mathcal{S}^D) \in \mathbb{R}^D$. Por lo tanto, el espacio $(\mathcal{S}^D, \circ, \cdot, d_{\mathrm{Ait}})$ tiene estructura de espacio métrico. En consecuencia, la distancia de Aitchison es una medida de diferencia adecuada para la realización de una clasificación automática no paramétrica de datos composicionales. La compatibilidad de esta distancia con las medidas de tendencia central y de dispersión de un conjunto de datos composicionales, que ha sido objeto de estudio en otros trabajos (Martín-Fernández et al., 1998b; Martín-Fernández et al., 1998a), se analiza en las secciones siguientes.

Con el propósito de establecer comparaciones entre el comportamiento de la d_{Ait} y el de las otras medidas de diferencia que se expondrán en esta misma sección, los entornos que se obtendrán con las otras medidas tendrán los mismos centros que los de la figura 2.13.

- *La distancia angular*

  La distancia angular

$$d_{\mathrm{Ang}}(\mathbf{x}, \mathbf{x}^*) = \arccos\left(\sum_{k=1}^{D} \sqrt{\frac{x_k^2}{\sum x_c^2}} \sqrt{\frac{x_k^{*2}}{\sum x_c^{*2}}}\right), \tag{2.13}$$

  definida en el símplex nos proporciona el ángulo –entre 0 y $\frac{\pi}{2}$– que forman los vectores de $\mathbb{R}^D$ cuyos afijos son las observaciones $\mathbf{x}$ y $\mathbf{x}^*$ proyectadas en la esfera D-dimensional de radio unidad y centrada en el origen de $\mathbb{R}^D$. De esta observación se deduce fácilmente que la distancia d_{Ang} es una medida de diferencia invariante por el grupo de rotaciones en la esfera. Esta distancia es defendida por Watson y Philip (1989) y provocó una larga e interesante polémica entre Watson-Philip y Aitchison –véanse Watson y Philip (1989, 1990, 1991) y Aitchison (1990, 1991). Como se ilustrará en el Ejemplo 2.1 esta distancia no es invariante por perturbaciones y no respeta la dominación por subcomposiciones.

  La figura 2.15 nos muestra los entornos que se obtienen con la distancia d_{Ang} en el símplex $\mathcal{S}^3$. Puede observarse que el comportamiento de la distancia angular no es "coherente" con el carácter composicional puesto que al acercarnos a los lados y los vértices del símplex los entornos no aparecen más apretados. Estos entornos no son más que la intersección con el símplex de conos con vértice en el origen de coordenadas. La superficie lateral de estos conos está formada por todas las semirectas de $\mathbb{R}_+^D$ que forman un mismo ángulo con la semirecta que pasa por el centro de los entornos.

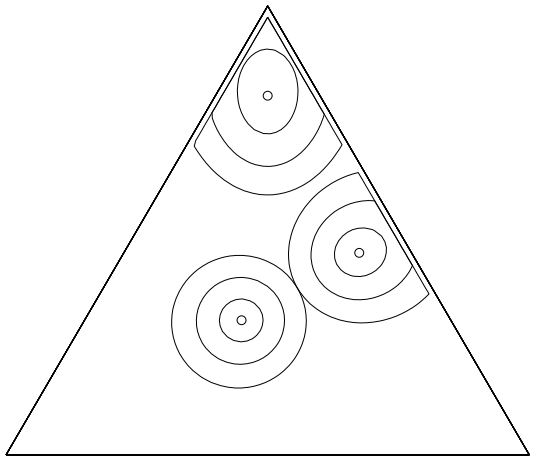


Figura 2.15: *Entornos en $\mathcal{S}^3$ con la distancia Angular.*

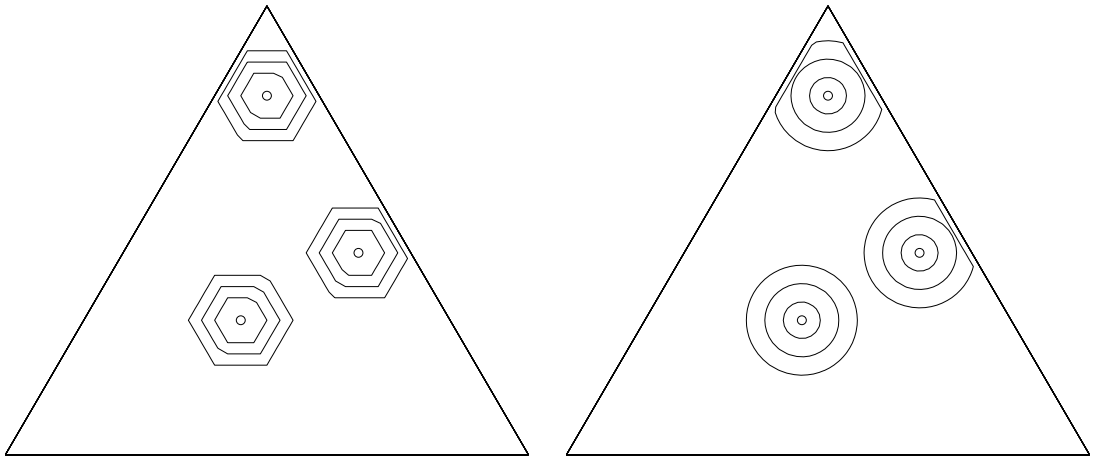
- *La distancia Minkowski*

  En la tabla 2.1 se observa que las distancias City Block y euclídea son casos particulares de la distancia de Minkowski para los valores $q = 1$ y $q = 2$, respectivamente. La distancia euclídea es la distancia más utilizada de entre todas las disimilitudes aplicadas a observaciones del espacio $\mathbb{R}^D$. Un hecho remarcable es que todas las distancias de la familia Minkowski tienen la propiedad de ser invariantes por traslaciones. Este hecho, que puede ser una virtud en multitud de casos, resulta ser un inconveniente cuando estas medidas se aplican a datos composicionales. Los requisitos de invariación por perturbaciones y de dominación por subcomposiciones no son satisfechos por ninguna de las distancias de Minkowski. Todos estos aspectos serán tratados en el Ejemplo 2.1.

  Las figuras 2.16(a) y 2.16(b) muestran, respectivamente, el comportamiento de las distancias City Block y Euclídea por lo que se refiere a entornos en el símplex $\mathcal{S}^3$. Podemos observar que el hecho de ser invariantes por traslaciones provoca que los entornos no se vean afectados por la posición del centro. Estas formas hexagonales y circulares de los entornos se obtienen como resultado de intersecar el símplex $\mathcal{S}^3$ con los entornos de $\mathbb{R}^3$ definidos por las distancias City Block y Euclídea. Estos entornos de $\mathbb{R}^3$ son, respectivamente, octoedros regulares y esferas.

- *Las disimilitudes entre distribuciones multinomiales*

  Las disimilitudes Bhattacharyya, J-Divergencia, Logarítmica y Matusita son útiles como medidas de diferencia entre dos distribuciones de probabilidad multinomial. Entre los estudios teóricos relacionados con estos conceptos destacamos los de Burbea y Rao (1982,



(a) (b)

Figura 2.16: Entornos en el símplex $\mathcal{S}^3$: (a) *con la distancia City Block;* (b) *con la distancia euclídea.*

1983) y los de Rao (1982, 1995). Todas estas medidas están relacionadas con el concepto de entropía y han sido utilizadas en disciplinas tan diversas como la teoría de la información, la economía, la genética, la antropología, y la biología, entre otras. En el Capítulo 3 se tratarán en profundidad todos los aspectos relacionados con este tipo de medidas de diferencia.

La observación detallada de las fórmulas de la tabla 2.1 nos conduce a subrayar las características siguientes:

- La disimilitud J-Divergencia puede entenderse como una distancia a medio camino entre la distancia euclídea y la distancia logarítmica si se tiene en cuenta que, para valores reales cercanos a 1, la función $\log x$ puede aproximarse por la función $x - 1$. En Martín (1996) se explora en profundidad la aplicación práctica de esta medida de diferencia en problemas de clasificación automática no paramétrica de datos de tipo composicional. El autor realiza esta exploración mediante un estudio comparativo de los resultados obtenidos al utilizar diferentes medidas de disimilitud en la clasificación. Entre estas medidas se encuentran todas las medidas entre distribuciones multinomiales que hemos incluido en la tabla 2.1. En sus conclusiones, el autor destaca que la disimilitud J-Divergencia es la medida que produce los mejores resultados en relación al reconocimiento de la estructura de grupos existente en el conjunto de datos considerado.
- Recuérdese que la distancia de Aitchison puede entenderse como la distancia euclídea aplicada a los datos clr-transformados. De manera análoga, las distancias logarítmica

y de Matusita pueden considerarse como el resultado de calcular la distancia euclídea entre los datos transformados, respectivamente, por la función $\log(x)$ y $\sqrt{x}$. Estas dos distancias aparecen tratadas en profundidad en Burbea y Rao (1982).

– Las disimilitudes de Bhattacharyya, $d_{\text{B-a}}$ con la función $\arccos(x)$ y $d_{\text{B-l}}$ con la función $-\log(x)$, presuponen en su definición que se aplican a un par de observaciones composicionales normalizadas. De esta manera, la expresión $\sum_{k=1}^{D}\sqrt{x_k}\sqrt{x_k^*}$ toma valores entre cero y uno. Si observamos la fórmula de la disimilitud $d_{\text{B-a}}$ vemos que la podemos interpretar como el ángulo que forman los vectores unitarios $\sqrt{\mathbf{x}} = (\sqrt{x_1}, \sqrt{x_2}, \ldots, \sqrt{x_D})$ y $\sqrt{\mathbf{x}^*} = (\sqrt{x_1^*}, \sqrt{x_2^*}, \ldots, \sqrt{x_D^*})$, cuyos afijos están sobre la esfera D-dimensional de radio igual unidad y centrada en el origen de $\mathbb{R}^D$. Por lo tanto, esta medida no es más que la distancia angular entre los vectores $\sqrt{\mathbf{x}}$ y $\sqrt{\mathbf{x}^*}$.
La disimilitud $d_{\text{B-a}}$ se ha utilizado como medida de diferencia en el estudio de Vives y Villarroya (1996) donde se realiza una clasificación automática no paramétrica de un conjunto de datos composicionales obteniendo, según los propios autores, una agrupación razonable y coherente. Esta distancia también aparece en la literatura como la distancia de Rao entre dos distribuciones de probabilidad multinomiales (Rao, 1995). Tiene la propiedad de ser la distancia geodésica en el espacio de parámetros de las distribuciones de probabilidad multinomiales.
Por otra parte, en Burbea (1983) se expone la relación que existe entre las disimilitudes de Bhattacharyya y la distancia de Matusita. Esta relación se desarrolla en las igualdades siguientes:

$$\begin{aligned} d^2_{\text{Mat}}(\mathbf{x}, \mathbf{x}^*) &= \sum_{k=1}^{D} (\sqrt{x_k} - \sqrt{x_k^*})^2 \\ &= \sum_{k=1}^{D} x_k + x_k^* - 2\sqrt{x_k}\sqrt{x_k^*} \\ &= 2 - 2\sum_{k=1}^{D} \sqrt{x_k}\sqrt{x_k^*} \\ &= 2 - 2\cos\left(d_{\text{B-a}}(\mathbf{x}, \mathbf{x}^*)\right) \\ &= 2 - 2\exp\left(-d_{\text{B-l}}(\mathbf{x}, \mathbf{x}^*)\right). \end{aligned} \tag{2.14}$$

En el fondo, las tres medidas tienen un origen común en la similitud entre dos observaciones composicionales dada por la expresión:

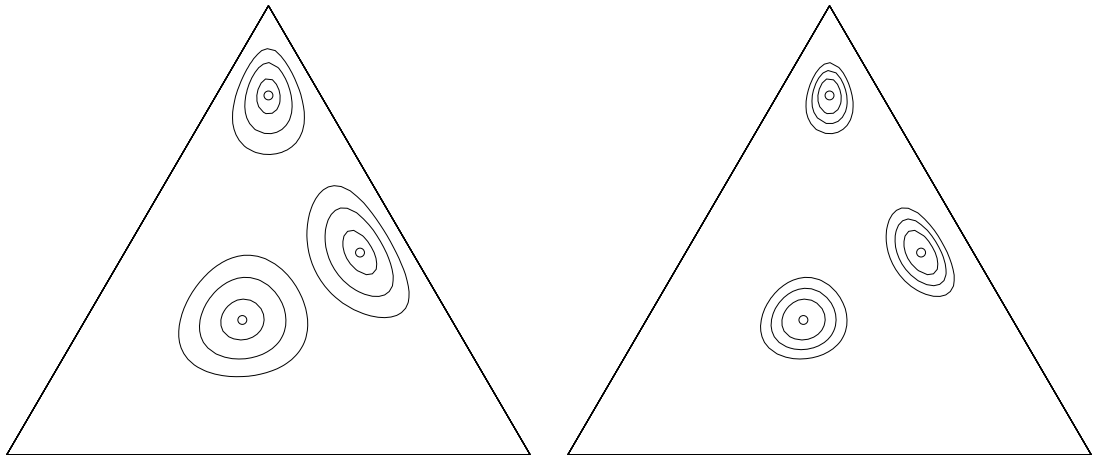
$$s(\mathbf{x}, \mathbf{x}^*) = \sum_{k=1}^{D} \sqrt{x_k}\,\sqrt{x_k^*},$$

que representa el *coseno* del ángulo que forman los vectores unitarios $\sqrt{\mathbf{x}}$ y $\sqrt{\mathbf{x}^*}$. Las tres medidas no son más que alternativas de convertir esta similitud en una disimilitud. Merece mención especial la medida de Matusita puesto que está basada en la transformación de similitud a disimilitud

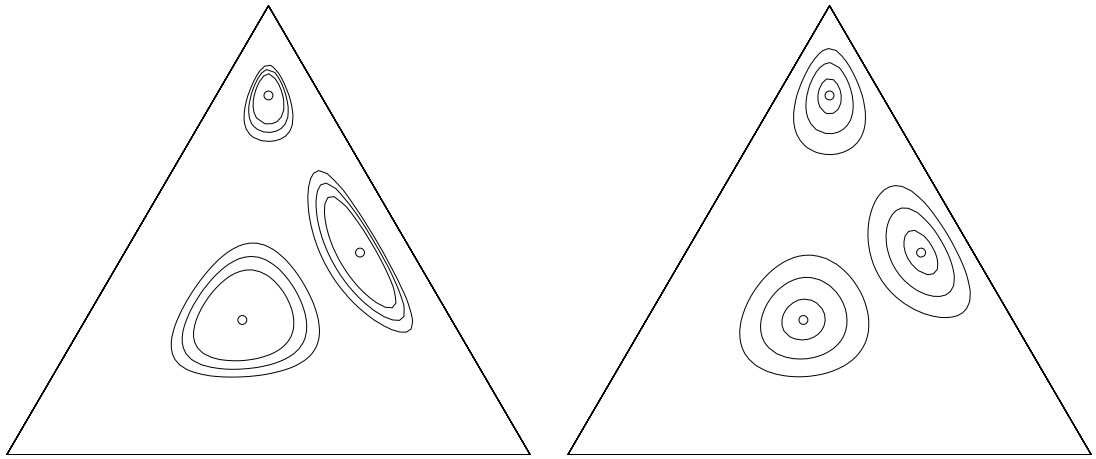
$$d_{\mathrm{Mat}}(\mathbf{x}, \mathbf{x}^*) = \sqrt{2 - 2s(\mathbf{x}, \mathbf{x}^*)}.$$

Esta transformación, propuesta por Gower y citada en Everitt (1993), ya ha sido comentada en la Sección 1.4.2. Por otra parte, observemos que las relaciones (2.14) se basan en transformaciones monótonas. Ello hace que estas tres medidas den lugar a la misma agrupación de observaciones cuando se apliquen técnicas de clasificación automática no paramétricas que tengan la propiedad de ser invariantes por transformaciones monótonas de la matriz de distancias entre individuos. Como consecuencia de esta relación, los entornos obtenidos con estas tres medidas serán equivalentes –véanse las figuras 2.17. Esta equivalencia se refiere a que los puntos de un entorno definido por una de las disimilitudes también constituyen un entorno con cualquier de las otras disimilitudes. Únicamente estos entornos difieren en la constante definida como radio del entorno.

Ninguna de estas medidas basadas en disimilitudes entre distribuciones de probabilidad multinomiales verifica los requisitos de invariación por perturbaciones y de dominación por subcomposiciones –véase el Ejemplo 2.1. Sin embargo, en la figura 2.17 puede observarse que todas estas medidas tienen un buen comportamiento por lo que se refiere a la dependencia de la forma de los entornos respecto la posición del centro. Así, si el centro se sitúa próximo a las caras o los vértices del símplex, los entornos aparecen más apretados, tal y como sucede con la distancia de Aitchison.



(a) (b)



(c) (d)

Figura 2.17: Entornos en el símplex $\mathcal{S}^3$: (a) *con la distancia Bhattacharyya (arccos);* (b) *con la distancia J-Divergencia;* (c) *con la distancia logarítmica;* (d) *con la distancia Matusita.*

2.4.3 Medidas de diferencia entre dos observaciones en relación a un conjunto de datos

En la práctica, el cálculo de las distancias de Mahalanobis entre dos observaciones viene asociado a un conjunto de datos. La distancia de Mahalanobis(raw) $d_{\text{M-r}}$, expuesta en la tabla 2.1, presupone el conocimiento de la matriz de covarianza $\mathbf{K}$ de un conjunto de datos composicionales $\mathbf{X}$. La distancia de Mahalanobis(clr) $d_{\text{M-c}}$ presupone el conocimiento de la matriz de covarianza $\mathbf{\Gamma}$ del conjunto de datos clr-transformado $\text{clr}(\mathbf{X})$. Por lo tanto, la distancia $d_{\text{M-r}}$ es una distancia de Mahalanobis definida directamente en el símplex y, la distancia $d_{\text{M-c}}$ es una distancia de Mahalanobis definida en el espacio clr-transformado. Debido a que estas distancias tienen su origen en la definición de una distancia de una observación a un conjunto de datos, analizaremos sus aspectos teóricos más relevantes en la Sección 2.5 de distancias entre un dato composicional y una composición.

En Barceló (1996) se expone en detalle que la distancia de Mahalanobis $d_{\text{M-r}}$ calculada directamente sobre datos composicionales no es una distancia adecuada. Ciertamente, recordemos que, análogamente a lo que sucede con las distancias de Minkowski, la $d_{\text{M-r}}$ es también una medida invariante por traslaciones. En el Ejemplo 2.1 mostramos que esta distancia no verifica los requisitos de invariación por perturbaciones y de dominación por subcomposiciones.

En Barceló-Vidal et al. (1999) se demuestra que el cálculo de la distancia $d_{\text{M-c}}$ puede realizarse de dos maneras diferentes totalmente equivalentes: bien a partir del cálculo de la matriz pseudo-inversa $\mathbf{\Gamma}^-$, tal y como aparece en la tabla 2.1, bien prescindiendo de una componente cualquiera del conjunto de datos clr$-$transformados. En este segundo modo de proceder hay que ser consciente que se ha reducido la dimensión al valor $D-1$ y la matriz de covarianzas será no singular. Por lo tanto, en la fórmula se usará la matriz inversa en lugar de una pseudo-inversa. En la siguiente propiedad analizamos en profundidad esta característica de la distancia $d_{\text{M-c}}$.

Propiedad 2.10 Sea $\mathbf{X}$ un conjunto de observaciones composicionales pertenecientes a $\mathcal{S}^D$. Simbolizamos por $\mathbf{Z}_D = \text{clr}(\mathbf{X})$ el conjunto de $\mathbb{R}^D$ formado por las observaciones clr-transformadas y, simbolizamos por $\mathbf{z}_D = \text{clr}(\mathbf{x})$ y $\mathbf{z}_D^* = \text{clr}(\mathbf{x}^*)$ dos observaciones clr-transformadas cualesquiera. La propiedad de invariación por permutaciones nos permite suponer, sin pérdida de generalidad, que la parte de la cual se prescinde es la última. Entonces, llamamos $\mathbf{Z}_{D-1}$ al conjunto de datos perteneciente a $\mathbb{R}^{D-1}$ resultante de prescindir de la última componente de todas las observaciones del conjunto $\mathbf{Z}_D$ y, en correspondencia, usamos los símbolos $\mathbf{z}_{D-1}$ y $\mathbf{z}_{D-1}^*$. Sea $\mathbf{\Gamma}_D$ la matriz de covarianza del conjunto de datos $\mathbf{Z}_D$ y $\mathbf{\Gamma}_{D-1}$ la matriz de covarianza

del conjunto de datos $\mathbf{Z}_{D-1}$. Entonces se cumple que

$$(\mathbf{z}_D - \mathbf{z}_D^*)\mathbf{\Gamma}_D^-(\mathbf{z}_D - \mathbf{z}_D^*)^t = (\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)\mathbf{\Gamma}_{D-1}^-(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)^t.$$

Demostración

Consideramos $\mathbf{F}$ la matriz elemental $\mathbf{F}_{D-1,D} = [\mathbf{I}_{D-1} : -\mathbf{j}_{D-1}]$ presentada en la Definición 2.9. Sea $\mathbf{E}_{D-1,D} = [\mathbf{I}_{D-1} : \mathbf{0}_{D-1}]$ la matriz donde las $D-1$ primeras columnas forman la matriz identidad y la última columna contiene el vector nulo. Unas simples manipulaciones algebraicas permiten demostrar las igualdades siguientes:

$$\mathbf{I}_{D-1} = \mathbf{E}\mathbf{F}^t = \mathbf{F}\mathbf{E}^t, \tag{2.15}$$

$$\mathbf{z}_D - \mathbf{z}_D^* = (\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)\mathbf{F}, \tag{2.16}$$

$$\mathbf{\Gamma}_D = \mathbf{F}^t\mathbf{\Gamma}_{D-1}\mathbf{F}, \tag{2.17}$$

$$\mathbf{\Gamma}_{D-1} = \mathbf{E}^t\mathbf{\Gamma}_D\mathbf{E}. \tag{2.18}$$

A partir de la igualdad (2.16) puede deducirse la siguiente secuencia de igualdades:

$$\begin{aligned}(\mathbf{z}_D - \mathbf{z}_D^*)\mathbf{\Gamma}_D^-(\mathbf{z}_D - \mathbf{z}_D^*)^t &= [(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)\mathbf{F}]\mathbf{\Gamma}_D^-[(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)\mathbf{F}]^t \\ &= (\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)(\mathbf{F}\mathbf{\Gamma}_D^-\mathbf{F}^t)(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*).\end{aligned}$$

Si comparamos el último término de esta cadena de igualdades y el segundo término de la propiedad que deseamos demostrar entonces, resta por probar que

$$(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)(\mathbf{F}\mathbf{\Gamma}_D^-\mathbf{F}^t)(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*) = (\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)\mathbf{\Gamma}_{D-1}^-(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)^t. \tag{2.19}$$

El valor de la forma cuadrática

$$(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)\mathbf{\Gamma}_{D-1}^-(\mathbf{z}_{D-1} - \mathbf{z}_{D-1}^*)^t,$$

no depende (Mardia et al., 1992) de la matriz pseudo-inversa que se utilice. En consecuencia, la demostración de la igualdad (2.19) se puede obtener si probamos que la matriz resultante del producto de matrices $\mathbf{F}\mathbf{\Gamma}_D^-\mathbf{F}^t$ es una pseudo-inversa de la matriz $\mathbf{\Gamma}_{D-1}$. Como es bien conocido (Mardia et al., 1992), si una matriz $\mathbf{M}$ de dimensiones $D \times D$ es singular entonces, una matriz $\mathbf{M}^-$ es una pseudo-inversa de $\mathbf{M}$ si satisface que

$$\mathbf{M}\mathbf{M}^-\mathbf{M} = \mathbf{M}.$$

Para el caso de matrices $\mathbf{M}$ no singulares la matriz pseudo-inversa $\mathbf{M}^-$ es única y coincide con la matriz inversa $\mathbf{M}^{-1}$. La siguiente secuencia de igualdades nos demuestra que $\mathbf{F}\boldsymbol{\Gamma}_D^-\mathbf{F}^t$ es una pseudo-inversa de la matriz $\boldsymbol{\Gamma}_{D-1}$:

$$\begin{aligned}
\boldsymbol{\Gamma}_{D-1}(\mathbf{F}\boldsymbol{\Gamma}_D^-\mathbf{F}^t)\boldsymbol{\Gamma}_{D-1} &= \mathbf{I}_{D-1}\boldsymbol{\Gamma}_{D-1}\mathbf{F}\boldsymbol{\Gamma}_D^-\mathbf{F}^t\boldsymbol{\Gamma}_{D-1}\mathbf{I}_{D-1} \\
&= \mathbf{E}\mathbf{F}^t\boldsymbol{\Gamma}_{D-1}\mathbf{F}\boldsymbol{\Gamma}_D^-\mathbf{F}^t\boldsymbol{\Gamma}_{D-1}\mathbf{F}\mathbf{E}^t \\
&= \mathbf{E}(\mathbf{F}^t\boldsymbol{\Gamma}_{D-1}\mathbf{F})\boldsymbol{\Gamma}_D^-(\mathbf{F}^t\boldsymbol{\Gamma}_{D-1}\mathbf{F})\mathbf{E}^t \\
&= \mathbf{E}\boldsymbol{\Gamma}_D\boldsymbol{\Gamma}_D^-\boldsymbol{\Gamma}_D\mathbf{E}^t \\
&= \mathbf{E}\boldsymbol{\Gamma}_D\mathbf{E}^t \\
&= \boldsymbol{\Gamma}_{D-1}.
\end{aligned}$$

En la práctica, la matriz $\boldsymbol{\Gamma}_{D-1}$ será probablemente una matriz no singular y, por lo tanto, la distancia de Mahalanobis podrá ser calculada mediante la expresión

$$d_{\text{M-c}}(\mathbf{x},\mathbf{x}^*) = [(\mathbf{z}_{D-1}-\mathbf{z}^*_{D-1})\boldsymbol{\Gamma}_{D-1}^{-1}(\mathbf{z}_{D-1}-\mathbf{z}^*_{D-1})^t]^{1/2}.$$

□

Procediendo de la misma manera que en el caso de la transformación clr, existe la posibilidad de definir una distancia de Mahalanobis en el espacio alr-transformado a partir de la expresión siguiente:

$$d_{\text{M-a}}(\mathbf{x},\mathbf{x}^*) = \left[(alr(\mathbf{x})-alr(\mathbf{x}^*))\boldsymbol{\Sigma}^{-1}(alr(\mathbf{x})-alr(\mathbf{x}^*))^t\right]^{\frac{1}{2}}, \tag{2.20}$$

donde la matriz $\boldsymbol{\Sigma}$ representa la matriz de covarianzas de los datos alr-transformados –véase la Definición 2.10. Si no se ha incluido esta distancia en la tabla 2.1 ha sido porque las distancias $d_{\text{M-c}}$ y $d_{\text{M-a}}$ son (Barceló-Vidal et al., 1999) exactamente la misma.

Propiedad 2.11 Sea $\mathbf{X}$ un conjunto de observaciones composicionales pertenecientes a $\mathcal{S}^D$ y sean $\mathbf{x}$, $\mathbf{x}^*$ dos composiciones de $\mathcal{S}^D$. Entonces se cumple que

$$d_{\text{M-a}}(\mathbf{x},\mathbf{x}^*) = d_{\text{M-c}}(\mathbf{x},\mathbf{x}^*)$$

Demostración

Consideramos las observaciones transformadas $\mathbf{y} = \text{alr}(\mathbf{x})$, $\mathbf{y}^* = \text{alr}(\mathbf{x}^*)$, $\mathbf{z} = \text{clr}(\mathbf{x})$, y $\mathbf{z}^* = \text{clr}(\mathbf{x})$. Sean $\boldsymbol{\Gamma}$ y $\boldsymbol{\Sigma}$ las matrices de covarianzas asociadas a los conjuntos de datos transformados

$\mathbf{Z} = \text{clr}(\mathbf{X})$) y $\mathbf{Y} = \text{alr}(\mathbf{X})$, respectivamente. Recordemos que, según se ha expuesto en la expresión (2.9), se satisfacen las relaciones

$$\mathbf{y}^t = \mathbf{F}\mathbf{z}^t; \quad \mathbf{\Gamma} = \mathbf{F}^t\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}\mathbf{F}; \text{ y } \quad \mathbf{\Sigma} = \mathbf{F}\mathbf{\Gamma}\mathbf{F}^t. \tag{2.21}$$

Usando la primera de estas relaciones se obtiene la siguiente igualdad

$$(\mathbf{y} - \mathbf{y}^*)\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{y}^*)^t = (\mathbf{z} - \mathbf{z}^*)\mathbf{F}^t\mathbf{\Sigma}^{-1}\mathbf{F}(\mathbf{z} - \mathbf{z}^*)^t. \tag{2.22}$$

En consecuencia, la propiedad quedará demostrada si probamos que el producto de matrices $\mathbf{F}^t\mathbf{\Sigma}^{-1}\mathbf{F}$ es una matriz pseudo-inversa de la matriz $\mathbf{\Gamma}$. Es decir, si se cumple que

$$\mathbf{\Gamma}\mathbf{F}^t\mathbf{\Sigma}^{-1}\mathbf{F}\mathbf{\Gamma} = \mathbf{\Gamma}. \tag{2.23}$$

Utilizando la relación $\mathbf{\Gamma} = \mathbf{F}^t\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}\mathbf{F}$ y la propiedad $\mathbf{H} = \mathbf{F}\mathbf{F}^t$ (Aitchison, 1986) se obtiene la siguiente secuencia de igualdades que nos demuestra identidad (2.23)

$$\begin{aligned}
\mathbf{\Gamma}\mathbf{F}^t\mathbf{\Sigma}^{-1}\mathbf{F}\mathbf{\Gamma} &= \mathbf{F}^t\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}\mathbf{F}\mathbf{F}^t\mathbf{\Sigma}^{-1}\mathbf{F}\mathbf{F}^t\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}\mathbf{F} \\
&= \mathbf{F}^t\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}\mathbf{F} \\
&= \mathbf{\Gamma}.
\end{aligned}$$

En consecuencia, la distancia de Mahalanobis(clr) d_{Mc} presentada en la tabla 2.1 puede ser calculada como la distancia de Mahalanobis en el espacio alr-transformado. □

2.4.4 Ejemplo

Con el propósito de ilustrar con contraejemplos los casos en los que las disimilitudes anteriores no cumplen los requerimientos expuestos en la Propiedad 2.8, dedicamos esta sección a presentar y comentar extensamente un ejemplo muy ilustrativo que aparece por primera vez en Martín-Fernández et al. (1998a).

Ejemplo 2.1 Consideremos el conjunto $\mathbf{X}$ de datos composicionales formado por las siguientes cuatro observaciones de $\mathcal{S}^3$:

$$\mathbf{x}_1 = (0.1, 0.2, 0.7), \quad \mathbf{x}_2 = (0.2, 0.1, 0.7), \quad \mathbf{x}_3 = (0.3, 0.4, 0.3) \quad \text{y} \quad \mathbf{x}_4 = (0.4, 0.3, 0.3).$$

Para cada $i = 1, 2, \ldots, 4$, usamos el símbolo $\mathbf{x}_i^*$ para indicar la composición perturbada $\mathbf{p} \circ \mathbf{x}_i$, donde $\mathbf{p} = (0.8, 0.1, 0.1)$. De manera análoga, $\mathbf{s}_i$ representa la subcomposición de la observación $\mathbf{x}_i$ formada por las dos primeras componentes. La figura 2.18 muestra la posición de estos elementos en el diagrama ternario, y la tabla 2.2 resume los valores de las disimilitudes de la tabla 2.1 entre algunas de las composiciones de $\mathbf{X}$.

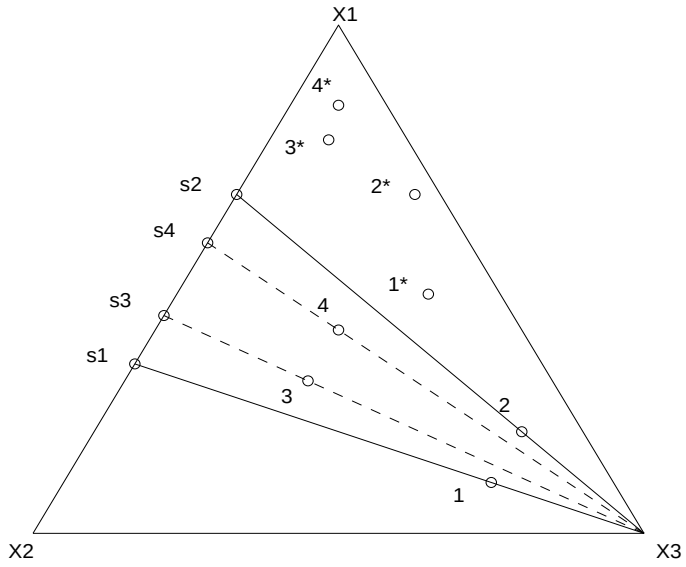


Figura 2.18: Las observaciones 1, 2, 3, y 4, sus subcomposiciones, $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$, y $\mathbf{s}_4$, y sus perturbadas 1^*, 2^*, 3^*, y 4^*.

Tabla 2.2: Disimilitudes de la tabla 2.1 entre algunas composiciones de la figura 2.18.

Disimilitud	$d(\mathbf{x}_1, \mathbf{x}_2)$	$d(\mathbf{x}_1^*, \mathbf{x}_2^*)$	$d(\mathbf{s}_1, \mathbf{s}_2)$	$d(\mathbf{x}_3, \mathbf{x}_4)$	$d(\mathbf{x}_3^*, \mathbf{x}_4^*)$	$d(\mathbf{s}_3, \mathbf{s}_4)$
Aitchison	0.98	0.98	0.98	0.41	0.41	0.41
Angular	0.19	0.33	0.64	0.24	0.08	0.28
Bhattacharyya (arccos)	0.19	0.22	0.34	0.12	0.09	0.14
Bhattacharyya (log)	0.02	0.02	0.06	0.01	0	0.01
City Block	0.2	0.4	0.67	0.2	0.14	0.29
Euclídea	0.14	0.24	0.47	0.14	0.09	0.2
J-Divergencia	0.37	0.43	0.68	0.24	0.18	0.29
Logarítmica	0.43	0.50	0.43	0.18	0.23	0.18
Mahalanobis (raw)	3	4.46	5.07	3	1.63	0.93
Mahalanobis (clr)	5.12	5.12	5.12	0.88	0.88	0.88
Matusita	0.19	0.22	0.34	0.12	0.09	0.14
Minkowski (q=3)	0.13	0.21	0.42	0.13	0.08	0.18

Este ejemplo pone en evidencia que la distancia euclídea no es una medida de diferencia adecuada entre dos datos composicionales. Ciertamente, la traslación $\mathbf{t} = (0.2, 0.2, -0.4)$ transforma la observación $\mathbf{x}_1$ en la $\mathbf{x}_3$, y la observación $\mathbf{x}_2$ en la $\mathbf{x}_4$; es decir, se cumple que $\mathbf{x}_1 + \mathbf{t} = \mathbf{x}_3$ y $\mathbf{x}_2 + \mathbf{t} = \mathbf{x}_4$. Este hecho implica que todas las distancias de la familia Minkowski –la City Block y la euclídea entre ellas– proporcionan la misma distancia entre $\mathbf{x}_1$ y $\mathbf{x}_2$ que entre $\mathbf{x}_3$ y $\mathbf{x}_4$, puesto que todas estas medidas de diferencia son invariantes por traslaciones. Sin embargo, desde el punto de vista de los datos composicionales, donde lo que importan son las proporciones entre las partes, la diferencia entre $\mathbf{x}_1$ y $\mathbf{x}_2$ debe ser mayor que la diferencia entre $\mathbf{x}_3$ y $\mathbf{x}_4$. Obsérvese que $\mathbf{x}_1$ y $\mathbf{x}_2$ sólo difieren en ± 0.1 en las dos primeras partes, y que lo mismo ocurre con $\mathbf{x}_3$ y $\mathbf{x}_4$. Sin embargo, en el primer caso la diferencia ± 0.1 se produce sobre un total de 0.3 $(= 1 - 0.7)$, mientras que en el segundo caso la misma diferencia ± 0.1 sobre un total de 0.7 $(= 1 - 0.3)$. Esta argumentación resulta mucho más clara a partir de la comparación de las correspondientes subcomposiciones

$$\mathbf{s}_1 = (\frac{1}{3}, \frac{2}{3}), \quad \mathbf{s}_2 = (\frac{2}{3}, \frac{1}{3}), \quad \mathbf{s}_3 = (\frac{3}{7}, \frac{4}{7}) \quad \text{y} \quad \mathbf{s}_4 = (\frac{4}{7}, \frac{3}{7}),$$

cuya representación también se muestra en la figura 2.18. En esta figura se observa claramente que la diferencia entre las subcomposiciones $\mathbf{s}_1$ y $\mathbf{s}_2$ es mayor que la diferencia entre las subcomposiciones $\mathbf{s}_3$ y $\mathbf{s}_4$. De los resultados de la tabla 2.2 se desprende que tampoco la distancia angular tiene un comportamiento coherente con el carácter composicional de los datos, puesto que la distancia angular entre $\mathbf{x}_3$ y $\mathbf{x}_4$ resulta ser mayor que la distancia entre $\mathbf{x}_1$ y $\mathbf{x}_2$.

Si se comparan entre si los resultados de las columnas $d(\mathbf{x}_1, \mathbf{x}_2)$, $d(\mathbf{x}_1^*, \mathbf{x}_2^*)$ y $d(\mathbf{s}_1, \mathbf{s}_2)$ de la tabla 2.2 se comprueba que únicamente las distancias Aitchison y Mahalanobis (clr) verifican simultáneamente todos los requisitos de la Propiedad 2.8 (invariación por permutaciones, invariación por perturbaciones y dominación por subcomposiciones). Ninguna de las demás medidas conserva la diferencia entre las dos observaciones $\mathbf{x}_1$ y $\mathbf{x}_2$ al aplicar a ambas una misma la perturbación $\mathbf{p}$. A su vez, a excepción de la distancia logarítmica, todas las otras medidas proporcionan una diferencia mayor entre las subcomposiciones $\mathbf{s}_1$ y $\mathbf{s}_2$ que entre las observaciones $\mathbf{x}_1$ y $\mathbf{x}_2$. Sin embargo, este hecho no significa que la distancia logarítmica verifique el requisito de la dominación por subcomposiciones para cualquier par de observaciones. Ciertamente, si calculamos la distancia logarítmica entre los elementos $\mathbf{x}_2 = (0.2, 0.1, 0.7)$ y $\mathbf{x}_3 = (0.3, 0.4, 0.3)$ se obtiene $d_{\text{Log}}(\mathbf{x}_2, \mathbf{x}_3) = 0.727$. En cambio, si calculamos la distancia entre sus correspondientes subcomposiciones $\mathbf{s}_2 = (1/8, 7/8)$ y $\mathbf{s}_3 = (4/7, 3/7)$ formadas por las dos últimas partes, se obtiene una distancia mayor $d_{\text{Log}}(\mathbf{s}_2, \mathbf{s}_3) = 0.729$.

Por otra parte, en la tabla 2.2 puede observarse que los valores correspondientes a las distancias Bhattacharyya (arccos) y Matusita coinciden totalmente. Este hecho es debido a la íntima relación entre ellas y al efecto del redondeo a la segunda cifra decimal. □

2.5 Medida de diferencia entre dos composiciones

En esta sección denominamos composición $\mathbf{X}$ a un vector aleatorio que toma valores en el símplex $\mathcal{S}^D$. Sin embargo, todos los conceptos son fácilmente generalizables al caso en que $\mathbf{X}$ representa un conjunto $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de datos composicionales. Es por esta razón que también nos referiremos a estas medidas de diferencia como medidas entre dos conjuntos de datos. Por otra parte, puesto que existe una íntima relación entre la disimilitud de dos composiciones y la disimilitud entre un dato composicional y una composición, hemos optado por presentarlas de manera indistinta.

2.5.1 Generalidades

En la Sección 1.4.4 *Medidas de diferencia entre dos conjuntos de datos* se han expuesto los aspectos básicos referentes a las medidas de diferencia entre dos conjuntos de datos. Recordemos que, en general, la definición de una medida de este tipo pasa por dos etapas: la primera, elegir una disimilitud entre dos observaciones –véase la tabla 2.1–; y la segunda, elegir entre definir la diferencia entre los dos conjuntos de datos como la disimilitud entre los representantes de cada uno de los dos grupos, o definirla como la disimilitud mínima, máxima o media de todas las diferencias interindividuales entre los elementos de los dos conjuntos. En la primera etapa de la definición de la disimilitud entre dos conjuntos de datos, parece lógico que, en todo caso, procure elegirse una disimilitud entre observaciones que sea coherente con el carácter composicional de los datos. Este hecho nos lleva a descartar de entrada las distancias de la familia Minkowski, la distancia angular y la distancia de Mahalonobis (raw) $d_{\text{M-r}}$. –véase el Ejemplo 2.1. También será deseable que la disimilitud definida entre dos composiciones cumpla los requisitos R1-R4 dados en la Propiedad 2.8. La propiedad de invariación por permutaciones la cumplen todas las disimilitudes de la tabla 2.1 y, por lo tanto, también la cumplirán las medidas de diferencia entre dos conjuntos basadas en ellas. Si consideramos dos conjuntos de datos $\mathbf{X}$ y $\mathbf{X}^*$ podemos expresar los requisitos R2 y R3 del siguiente modo:

R2. Dominación respecto de las subcomposiciones.

$$d(\mathbf{X}, \mathbf{X}^*) \geq d(\mathcal{C}(\mathbf{X}_s), \mathcal{C}(\mathbf{X}^*{}_s)), \quad \forall \mathbf{X}, \mathbf{X}^* \in \mathcal{S}^D,$$

sea cual sea la subcomposición s escogida.

R3. Invariación respecto de las perturbaciones.

$$d(\mathbf{p} \circ \mathbf{X}, \mathbf{p} \circ \mathbf{X}^*) = d(\mathbf{X}, \mathbf{X}^*), \quad \forall \mathbf{X}, \mathbf{X}^* \in \mathcal{S}^D, \ \forall \mathbf{p} \in \mathbb{R}_+^D.$$

□

En las expresiones anteriores se entiende por perturbación y por subcomposición de un conjunto $\mathbf{X}$ los conjuntos que se obtienen al aplicar la transformación correspondiente a cada uno de los elementos del conjunto.

Entre las disimilitudes de la tabla 2.1, las únicas que verifican los requisitos R2 y R3 son la distancia de Aitchison y la distancia de Mahalanobis (clr). Por lo tanto, si definimos la medida de diferencia entre dos conjuntos basándonos en ellas, tenemos la garantía de establecer una medida coherente con el carácter composicional de los datos.

En la segunda etapa de la definición de la disimilitud entre dos conjuntos de datos tenemos que escoger entre considerar la disimilitud interindividual (mínima, máxima, o media) o la disimilitud entre los representantes de los respectivos conjuntos. La primera opción no está afectada por la tipología composicional de los datos si no más bien por la estructura y la disposición relativa de los grupos de observaciones, tal y como se ha expuesto en el Capítulo 1. En cambio, si la elección pasa por establecer la diferencia entre los representantes de cada grupo, estamos obligados a establecer previamente una medida de tendencia central coherente con el carácter composicional de los datos con el fin de establecer un criterio de elección del representante de un grupo de observaciones. Esta cuestión la tratamos con detalle en la Sección 2.6.

A nuestro entender la distancia de Mahalanobis merece un tratamiento especial debido a su importancia en el análisis multivariante de datos. Por este motivo, a continuación tratamos en profundidad la distancia entre dos composiciones basada en la distancia de Mahalanobis.

2.5.2 Distancia de Mahalanobis entre un dato composicional y una composición

En las expresiones (1.7) y (1.9) se presentó la distancia de Mahalanobis entre dos individuos de $\mathbb{R}^D$. Podemos extender esta definición a una medida de diferencia entre una observación $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_D^*)$ y un vector aleatorio $\mathbf{X}$.

Definición 2.12 La distancia de Mahalanobis entre una observación $\mathbf{x}^*$ y un vector aleatorio $\mathbf{X}$ con vector de esperanzas $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, se define por

$$\Delta(\mathbf{x}^*, \mathbf{X}) = [(\mathbf{x}^* - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu})^t]^{1/2}. \tag{2.24}$$

□

Esta distancia cumple, entre otras, la propiedad de ser invariante respecto de las transformaciones lineales no singulares (en particular, respecto los cambios de escala). Es decir, se cumple que

$$\Delta(\mathbf{M}\mathbf{x}^*, \mathbf{M}\mathbf{X}) = \Delta(\mathbf{x}^*, \mathbf{X}),$$

para cualquier $D \times D$ matriz $\mathbf{M}$ no singular.

Por otra parte, en caso que la matriz de covarianzas $\boldsymbol{\Sigma}$ fuese singular, diversos autores (Cuadras, 1991; Krzanowski y Marriot, 1994) sugieren utilizar en (2.24) la matriz pseudo-inversa $\boldsymbol{\Sigma}^-$ en substitución de la matriz inversa de $\boldsymbol{\Sigma}$, teniendo en cuenta que se cumplen las mismas propiedades que en el caso no singular, verificándose además la invariación respecto de las transformaciones lineales que conservan el rango de $\boldsymbol{\Sigma}$, y la independencia respecto de la pseudo-inversa $\boldsymbol{\Sigma}^-$ utilizada.

Parece lógico exigir también a una medida de diferencia entre un dato composicional $\mathbf{x}^*$ y una composición $\mathbf{X}$ en $\mathcal{S}^D$ unas propiedades semejantes a las propiedades R1-R4 –véase la Propiedad 2.8– exigidas por Aitchison (1992) a la hora de definir la medida de diferencia entre dos D-partes. En el Ejemplo 2.1 hemos mostrado que la distancia de Mahalanobis (2.24) definida directamente sobre $\mathcal{S}^D$ no cumple la invariación respecto a las perturbaciones ni la dominación por subcomposiciones. Debemos, pues, utilizar la medida de Mahalanobis (clr) o, de manera equivalente, la Mahalanobis (alr).

Definición 2.13 Se define la medida de diferencia $d_{\text{M-a}}(\mathbf{x}^*, \mathbf{X})$ entre una D-parte $\mathbf{x}^*$ y una composición $\mathbf{X}$ en $\mathcal{S}^D$ como la distancia de Mahalanobis en $\mathbb{R}^{D-1}$ entre el vector $\text{alr}(\mathbf{x}^*)$ y la variable aleatoria transformada $\text{alr}(\mathbf{X})$:

$$d_{\text{M-a}}(\mathbf{x}^*, \mathbf{X}) = \Delta(\text{alr}(\mathbf{x}^*), \text{alr}(\mathbf{X})) = \left[(\text{alr}(\mathbf{x}^*) - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\text{alr}(\mathbf{x}^*) - \boldsymbol{\mu})^t\right]^{1/2}, \tag{2.25}$$

donde $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ representan el vector de esperanzas y la matriz de covarianzas de la distribución transformada $\mathbf{Y} = \text{alr}(\mathbf{X})$ de $\mathbb{R}^{D-1}$. □

La medida de diferencia $d_{\text{M-a}}(\mathbf{x}^*, \mathbf{X})$ que acabamos de definir se generaliza fácilmente al caso en que $\mathbf{X}$ representa un conjunto de datos composicionales $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de $\mathcal{S}^D$. Basta substituir en (2.25) el vector de esperanzas $\boldsymbol{\mu}$ y la matriz de covarianzas $\boldsymbol{\Sigma}$ por, respectivamente, el vector de medias $\mathbf{m}_Y$ y la matriz de covarianzas $\mathbf{S}$ de las observaciones transformadas $\mathbf{y}_1 = \text{alr}(\mathbf{x}_1), \mathbf{y}_2 = \text{alr}(\mathbf{x}_2), \ldots, \mathbf{y}_n = \text{alr}(\mathbf{x}_n)$:

$$\mathbf{m}_Y = \frac{1}{n}\sum_{i=1}^{n} \text{alr}(\mathbf{x}_i) \tag{2.26}$$

y

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\text{alr}(\mathbf{x}_i) - \mathbf{m}_Y)^t (\text{alr}(\mathbf{x}_i) - \mathbf{m}_Y). \tag{2.27}$$

Definición 2.14 Se define la medida de diferencia $d_{\text{M-a}}(\mathbf{x}^*, \mathbf{X})$ entre una D-parte $\mathbf{x}^*$ y un conjunto $\mathbf{X}$ de datos composicionales $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de $\mathcal{S}^D$ como:

$$d_{\text{M-a}}(\mathbf{x}^*, \mathbf{X}) = \left[(\text{alr}(\mathbf{x}^*) - \mathbf{m}_Y)\mathbf{S}^{-1}(\text{alr}(\mathbf{x}^*) - \mathbf{m}_Y)^t\right]^{1/2}, \tag{2.28}$$

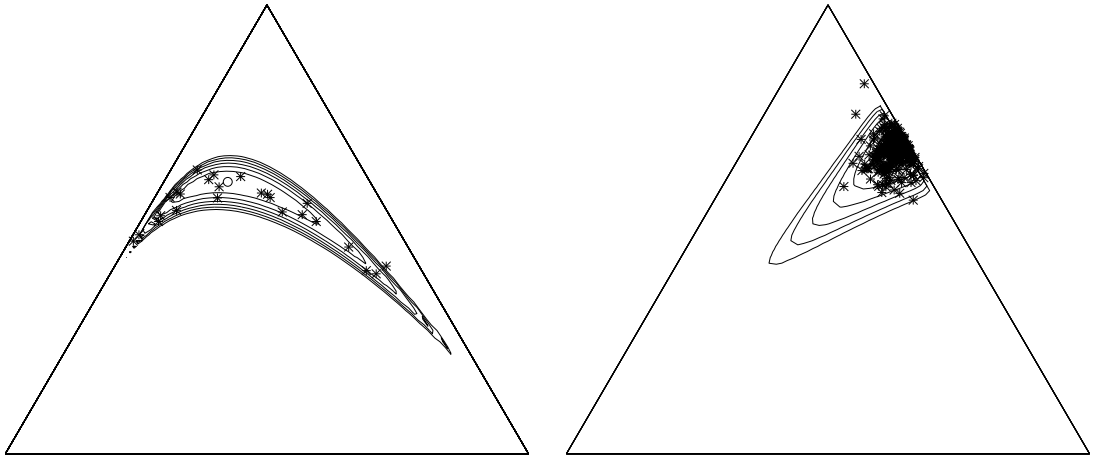
donde $\mathbf{m}$ y $\mathbf{S}$ vienen dados, respectivamente, por (2.26) y (2.27). □

Es inmediato comprobar que esta medida de diferencia verifica igualmente todas las propiedades R1-R4. Ciertamente, es importante resaltar que la distancia definida en (2.28) está íntimamente relacionada con la distancia de Mahalanobis $d_{\text{M-a}}$ entre dos datos composicionales presentada en (2.20). Así, la distancia definida en (2.28) no es más que la distancia $d_{\text{M-a}}$ entre $\mathbf{x}^*$ y el elemento $\text{agl}(\mathbf{m}_Y)$ de $\mathcal{S}^D$. Recordemos que $\text{agl}(\mathbf{m}_Y)$ es el elemento del símplex que se obtiene al aplicar a $\mathbf{m}_Y$ la transformación inversa de la transformación alr. Es importante resaltar que, este elemento $\text{agl}(\mathbf{m}_Y)$ coincide con el elemento $\text{ilc}(\mathbf{m}_Z) \in \mathcal{S}^D$ que se obtiene al aplicar al vector de medias $\mathbf{m}_Z$ del conjunto clr-transformado $\mathbf{Z} = \text{clr}(\mathbf{X})$ la transformación inversa de la transformación clr. En la Sección 2.6 se analiza en detalle este elemento y se demuestra que no es más que la media geométrica composicional del conjunto $\mathbf{X}$. Como veremos en las secciones siguientes, esta media geométrica composicional desempeña el papel de representante de un conjunto de observaciones composicionales, semejante al papel que juega el centroide o media aritmética de un conjunto de datos en el espacio $\mathbb{R}^D$.

Por otra parte, dada una composición o un conjunto $\mathbf{X}$ de datos composicionales sobre el símplex $\mathcal{S}^D$, tiene sentido considerar el entorno formado por aquellos puntos $\mathbf{x}^* \in \mathcal{S}^D$ cuya medida de diferencia a $\mathbf{X}$ se mantiene menor o igual que r ($r \in \mathbb{R}^+$): $\text{E}_{d_{\text{M-a}}}(\mathbf{X}; r) = \{\mathbf{x}^* \in \mathcal{S}^D : d_{\text{M-a}}(\mathbf{x}^*, \mathbf{X}) < r\}$.

En el caso $D = 3$ es posible visualizar estos entornos $\text{E}(\mathbf{X}; r)$ en $\mathcal{S}^3$. La figura 2.19 nos muestra los entornos que se obtienen cuando se utilizan los conjuntos de datos Hongita y Halimba. El primer conjunto se encuentra en Aitchison (1986) y el segundo en Mateu-Figueras et al. (1998).

En las figuras 2.20(a) y 2.20(b) se representan las observaciones del conjunto de datos Polen –que se encuentra en Aitchison (1986)– y los entornos obtenidos, respectivamente, con las distancias $d_{\text{M-r}}$ y $d_{\text{M-c}}$. Observemos que los entornos obtenidos mediante la distancia de Mahalanobis $d_{\text{M-r}}$ definida directamente en el símplex no son más que la intersección con el símplex de elipses de centro el punto $\text{agl}(\mathbf{m}_Y)$ representado en el gráfico con el símbolo '$\circ$'. Obsérvese en la figura 2.20(a) que los entornos resultantes no reconocen la forma del conjunto de datos Polen. Como puede apreciarse en la figura 2.20(b), sucede todo lo contrario con la distancia de Mahalanobis $d_{\text{M-c}}$. Los entornos calculados con esta distancia logran recoger de manera razonable la forma del conjunto de datos.



(a) (b)

Figura 2.19: *Entornos en $\mathcal{S}^3$ con la distancia $d_{\text{M-a}}$ para los conjuntos de datos composicionales:* (a) *Hongite;* (b) *Halimba.*

2.5.3 Distancia de Mahalanobis entre dos composiciones

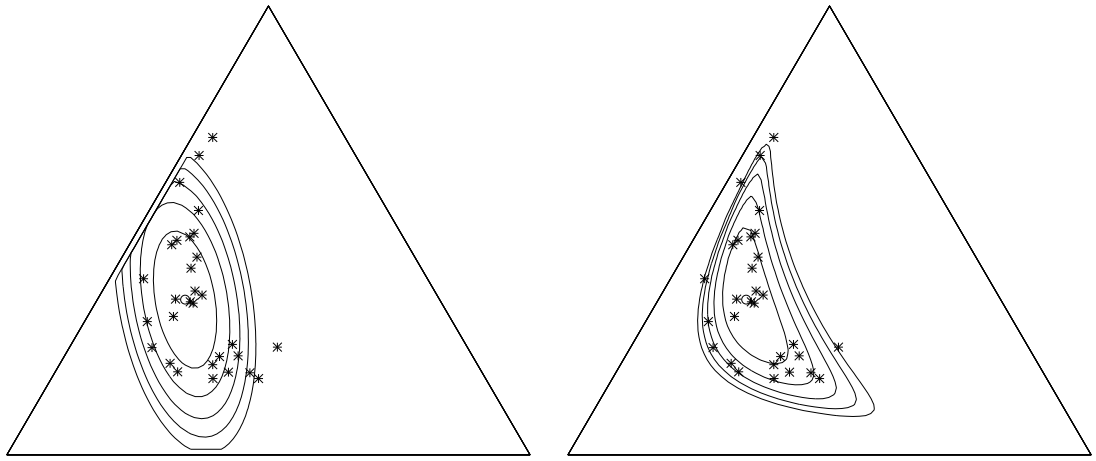
Las ideas anteriores respecto a una medida de diferencia entre un dato composicional y una composición pueden extenderse a una medida de diferencia entre dos composiciones.

Es sabido que si $\mathbf{X}_1$ y $\mathbf{X}_2$ son dos variables aleatorias homocedásticas definidas sobre $\mathbb{R}^D$, la medida de diferencia de Mahalanobis $\Delta(\mathbf{X}_1, \mathbf{X}_2)$ entre ambas se define como

$$\Delta(\mathbf{X}_1, \mathbf{X}_2) = \{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\}^{1/2},$$

siendo $\boldsymbol{\mu}_j = E(\mathbf{X}_j)$ $(j = 1, 2)$, y $\boldsymbol{\Sigma}$ la matriz de covarianzas común de ambas distribuciones.

Podemos definir una medida de diferencia entre dos composiciones como la distancia de Mahalanobis entre sus alr-transformadas.



(a) (b)

Figura 2.20: Entornos en el símplex $\mathcal{S}^3$ del conjunto de datos Polen (a) *con la distancia Mahalanobis (raw);* (b) *con la distancia Mahalanobis (*clr*).*

Definición 2.15 Si $\mathbf{X}_1$ y $\mathbf{X}_2$ son dos composiciones del símplex $\mathcal{S}^D$ tales que las distribuciones transformadas $\text{alr}(\mathbf{X}_1)$ y $\text{alr}(\mathbf{X}_2)$ tienen la misma matriz de covarianzas $\boldsymbol{\Sigma}$, se define la medida de diferencia $d_{\text{M-a}}(\mathbf{X}_1, \mathbf{X}_2)$ entre ambas como la medida de diferencia de Mahalanobis en $\mathbb{R}^{D-1}$ entre $\text{alr}(\mathbf{X}_1)$ y $\text{alr}(\mathbf{X}_2)$:

$$d_{\text{M-a}}(\mathbf{X}_1, \mathbf{X}_2) = \Delta(\text{alr}(\mathbf{X}_1), \text{alr}(\mathbf{X}_1)) = \{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\}^{1/2}, \tag{2.29}$$

siendo $\boldsymbol{\mu}_j = E(\text{alr}(\mathbf{X}_j))$ $(j = 1, 2)$. □

En Barceló (1996) se demuestra que esta distancia cumple todas las propiedades exigibles a una medida de diferencia composicional y que puede extenderse de manera simple al caso de la distancia entre dos conjuntos de datos composicionales. Ciertamente, podemos establecer la distancia entre dos conjuntos de observaciones composicionales de $\mathcal{S}^D$ con la definición siguiente:

Definición 2.16 Sean $\mathbf{X}_1 = \{\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}\}$ y $\mathbf{X}_2 = \{\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}\}$ dos conjuntos de composiciones del símplex $\mathcal{S}^D$. Se define la distancia $d_{\text{M-a}}(\mathbf{X}_1, \mathbf{X}_2)$ entre ambos como:

$$d_{\text{M-a}}(\mathbf{X}_1, \mathbf{X}_2) = \Delta(\text{alr}(\mathbf{X}_1), \text{alr}(\mathbf{X}_2)) = \{(\mathbf{m}_{Y1} - \mathbf{m}_{Y2})\mathbf{S}^{-1}(\mathbf{m}_{Y1} - \mathbf{m}_{Y2})^t\}^{1/2},$$

siendo

$$\mathbf{m}_{Yj} = \frac{1}{n_j}\sum_{i=1}^{n_j} \text{alr}(\mathbf{x}_{ji}) \qquad (j = 1, 2),$$

y

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2},$$

donde

$$\mathbf{S}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \{\text{alr}(\mathbf{x}_{ji}) - \mathbf{m}_{Yj}\}\{\text{alr}(\mathbf{x}_{ji}) - \mathbf{m}_{Yj}\}^t \qquad (j = 1, 2).$$

□

Esta distancia satisface las mismas propiedades que la distancia definida en (2.29) entre dos composiciones.

2.6 Medida de tendencia central de un conjunto de datos composicionales

2.6.1 La media geométrica composicional

Las medidas de tendencia central constituyen, junto con las medidas de diferencia y de dispersión, el elemento diferenciador clave entre las técnicas habituales de clasificación automática no paramétrica. Es bien sabido que la medida de tendencia central más utilizada para conjuntos de datos en el espacio real es la media aritmética o centroide del conjunto. En Aitchison (1997) se muestra que esta medida tan usualmente aplicada no es compatible con el carácter composicional de los datos. En el mismo trabajo Aitchison aboga por el uso de la media geométrica como medida representativa del centro de un conjunto de datos composicionales y demuestra que es compatible con el grupo de transformaciones definido en el símplex, e.g., con el grupo de las perturbaciones.

Definición 2.17 Sea $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ un conjunto de datos composicionales de $\mathcal{S}^D$. Se define la *media geométrica composicional* $g(\mathbf{X})$ del conjunto $\mathbf{X}$ como:

$$g(\mathbf{X}) = \mathcal{C}(g_1, g_2, \ldots, g_D) = \left(\frac{g_1}{\sum g_k}, \frac{g_2}{\sum g_k}, \ldots, \frac{g_D}{\sum g_k} \right), \tag{2.30}$$

donde $g_k = \left(\prod_{i=1}^{n} x_{ik} \right)^{1/n}$ representa la media geométrica de la k-ésima componente de los datos.

□

Observemos que de la definición de operación perturbación y de la definición de producto por escalar –véanse las Definiciones 2.7 y 2.8– puede establecerse que la media geométrica composicional de un conjunto $\mathbf{X}$ viene dada por la expresión

$$g(\mathbf{X}) = (\frac{1}{n} \cdot \mathbf{x}_1) \circ (\frac{1}{n} \cdot \mathbf{x}_2) \circ \ldots \circ (\frac{1}{n} \cdot \mathbf{x}_n) = \prod_{i=1}^{n} (\frac{1}{n} \cdot \mathbf{x}_i) = \frac{1}{n} \cdot \left(\prod_{i=1}^{n} \mathbf{x}_i \right).$$

Esta última expresión es análoga a la definición de la media aritmética de un conjunto de observaciones en el espacio real $\mathbb{R}^D$. Para ilustrar el comportamiento de la media geométrica composicional y de la media aritmética consideremos el conjunto de datos Hongite publicado en Aitchison (1986). La figura 2.21 muestra el conjunto de datos y sus dos medidas de tendencia central. Se observa que el centroide no es una medida representativa del centro de este conjunto de datos composicionales. En general esto ocurrirá siempre que el conjunto de datos composicionales de $\mathcal{S}^3$ tenga apariencia cóncava (de "media luna").

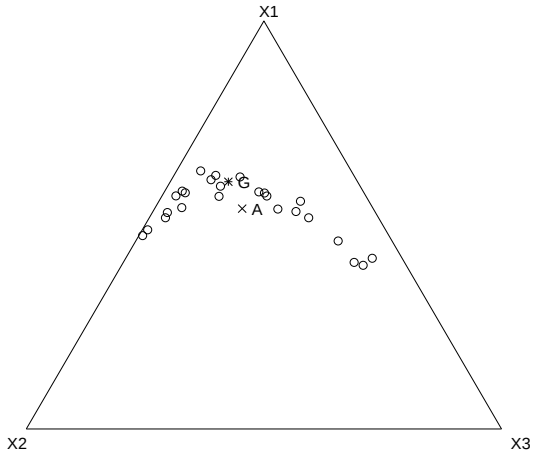


Figura 2.21: *Medidas de tendencia central para el conjunto de datos Hongite: (A) media aritmética; (G) media geométrica.*

Un estudio en profundidad de la media geométrica composicional nos conduce a constatar que esta medida de tendencia central posee propiedades que la convierten en una medida coherente con la naturaleza composicional de los datos. En la propiedad siguiente demostramos que la medida de tendencia central $g(\mathbf{X})$ es compatible con las operaciones básicas: perturbación, producto por escalar, y subcomposición.

Propiedad 2.12 Sea $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ un conjunto de datos composicionales de $\mathcal{S}^D$ y $g(\mathbf{X})$ su media geométrica composicional. Entonces, se cumple que

i. $g(\mathbf{p} \circ \mathbf{X}) = \mathbf{p} \circ g(\mathbf{X}),\ \forall \mathbf{p} \in \mathbb{R}^D_+$,

ii. $g(\alpha \cdot \mathbf{X}) = \alpha \cdot g(\mathbf{X}),\ \forall \alpha \in \mathbb{R}$,

iii. $g(\mathcal{C}(\mathbf{X}_s)) = \mathcal{C}(g(\mathbf{X})_s)$, sea cual sea la subcomposición s escogida,

donde las expresiones $\mathbf{p} \circ \mathbf{X}$, $\alpha \cdot \mathbf{X}$, y $\mathcal{C}(\mathbf{X}_s)$ simbolizan los conjuntos que se obtienen al aplicar la correspondiente operación al conjunto $\mathbf{X}$.

Demostración

La demostración de estas propiedades se basa en las definiciones de las operaciones perturbación, producto por escalar, y subcomposición, y en las propiedades del operador clausura $\mathcal{C}$.

i.

$$\begin{aligned} g(\mathbf{p}\circ\mathbf{X}) &= \mathcal{C}\left(\left[\prod_{i=1}^{n}\frac{p_1x_{i1}}{\sum p_kx_{ik}}\right]^{1/n},\left[\prod_{i=1}^{n}\frac{p_2x_{i2}}{\sum p_kx_{ik}}\right]^{1/n},\ldots,\left[\prod_{i=1}^{n}\frac{p_Dx_{iD}}{\sum p_kx_{ik}}\right]^{1/n}\right)\\ &= \mathcal{C}\left(\left[\prod_{i=1}^{n}p_1x_{i1}\right]^{1/n},\left[\prod_{i=1}^{n}p_2x_{i2}\right]^{1/n},\ldots,\left[\prod_{i=1}^{n}p_Dx_{iD}\right]^{1/n}\right)\\ &= \mathcal{C}\left(p_1\left[\prod_{i=1}^{n}x_{i1}\right]^{1/n},p_2\left[\prod_{i=1}^{n}x_{i2}\right]^{1/n},\ldots,p_D\left[\prod_{i=1}^{n}x_{iD}\right]^{1/n}\right)\\ &= \mathbf{p}\circ g(\mathbf{X}), \end{aligned}$$

ii.

$$\begin{aligned} g(\alpha\cdot\mathbf{X}) &= \mathcal{C}\left(\left[\prod_{i=1}^{n}\frac{x_{i1}^{\alpha}}{\sum x_{ik}^{\alpha}}\right]^{1/n},\left[\prod_{i=1}^{n}\frac{x_{i2}^{\alpha}}{\sum x_{ik}^{\alpha}}\right]^{1/n},\ldots,\left[\prod_{i=1}^{n}\frac{x_{iD}^{\alpha}}{\sum x_{ik}^{\alpha}}\right]^{1/n}\right)\\ &= \mathcal{C}\left(\left[\prod_{i=1}^{n}x_{i1}^{\alpha}\right]^{1/n},\left[\prod_{i=1}^{n}x_{i2}^{\alpha}\right]^{1/n},\ldots,\left[\prod_{i=1}^{n}x_{iD}^{\alpha}\right]^{1/n}\right)\\ &= \mathcal{C}\left(\left[\prod_{i=1}^{n}x_{i1}\right]^{\alpha/n},\left[\prod_{i=1}^{n}x_{i2}\right]^{\alpha/n},\ldots,\left[\prod_{i=1}^{n}x_{iD}\right]^{\alpha/n}\right)\\ &= \alpha\cdot g(\mathbf{X}), \end{aligned}$$

iii.

$$\begin{aligned} g(\mathcal{C}(\mathbf{X}_s)) &= \mathcal{C}\left(\left[\prod_{i=1}^{n}\frac{x_{i1}}{\sum_{k=1}^{s}x_{ik}}\right]^{1/n},\left[\prod_{i=1}^{n}\frac{x_{i2}}{\sum_{k=1}^{s}x_{ik}}\right]^{1/n},\ldots,\left[\prod_{i=1}^{n}\frac{x_{is}}{\sum_{k=1}^{s}x_{ik}}\right]^{1/n}\right)\\ &= \mathcal{C}\left(\left[\prod_{i=1}^{n}x_{i1}\right]^{1/n},\left[\prod_{i=1}^{n}x_{i2}\right]^{1/n},\ldots,\left[\prod_{i=1}^{n}x_{is}\right]^{1/n}\right)\\ &= \mathcal{C}(g(\mathbf{X})_s). \end{aligned}$$

□

Por otra parte, de acuerdo con lo establecido en Cuadras et al. (1997), la definición siguiente presenta de manera natural un concepto análogo al concepto de media aritmética de un conjunto de datos.

Definición 2.18 Sea $\mathbf{E}$ el espacio muestral o soporte de las observaciones, d una disimilitud definida sobre $\mathbf{E}$, y $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ un conjunto de datos de $\mathbf{E}$. Se define el *d-centro* $cen(\mathbf{X})$

del conjunto $\mathbf{X}$ como un elemento de $\mathbf{E}$ que minimiza la expresión $\sum_{i=1}^{n} d^2(\mathbf{x}, \mathbf{x}_i)$. Es decir, tal que

$$\sum_{i=1}^{n} d^2(cen(\mathbf{X}), \mathbf{x}_i) = \min\{\sum_{i=1}^{n} d^2(\mathbf{x}, \mathbf{x}_i) : \mathbf{x} \in \mathbf{E}\}. \tag{2.31}$$

□

Tal y como se expone en Cuadras et al. (1997), en general, el $cen(\mathbf{X})$ de un conjunto no es único. Sin embargo la siguiente propiedad muestra que en el caso de los datos composicionales y para la distancia de Aitchison, el *d-centro* es único y coincide con la media geométrica composicional.

Propiedad 2.13 Sea $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ un conjunto de datos composicionales del símplex $\mathcal{S}^D$ sobre el que consideramos definida la distancia de Aitchsion d_{Ait}. Sean $cen(\mathbf{X})$ el d_{Ait}-*centro* y $g(\mathbf{X})$ la media geométrica composicional del conjunto $\mathbf{X}$. Entonces, se cumple que $cen(\mathbf{X}) = g(\mathbf{X})$. □

La demostración de esta propiedad anterior se basa en otra propiedad que relaciona el concepto de media geométrica composicional y el concepto de media aritmética para datos en el espacio real $\mathbb{R}^D$. La propiedad siguiente muestra cómo estos conceptos están relacionados mediante la transformación clr.

Propiedad 2.14 Sea $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ un conjunto de datos composicionales de $\mathcal{S}^D$ y $g(\mathbf{X})$ su media geométrica composicional. Entonces se cumple que

$$\text{clr}(g(\mathbf{X})) = \overline{\text{clr}(\mathbf{X})}, \tag{2.32}$$

donde $\overline{\text{clr}(\mathbf{X})} = \frac{1}{n}\sum_{i=1}^{n} \text{clr}(\mathbf{x}_i)$ es la media aritmética del conjunto $\text{clr}(\mathbf{X})$ o clr-transformado. □

Observemos que, de manera análoga a la relación establecida en la Propiedad 2.8, en la propiedad anterior se establece una equivalencia entre los roles desempeñados por la media geométrica composicional y la media aritmética para datos en $\mathbb{R}^D$. Esta relación y el hecho que la media geométrica composicional sea una medida compatible con la operación perturbación, nos permite establecer (Martín-Fernández et al., 1999) una transformación de los datos que, por analogía al caso de datos en el espacio $\mathbb{R}^D$, llamaremos *centrado* de un conjunto de datos. Al aplicar esta transformación a un conjunto de datos composicionales de $\mathcal{S}^D$ se obtiene un conjuto de datos cuya media geométrica composicional es el baricentro del símplex $\mathcal{S}^D$, $\mathbf{e} = (1/D, 1/D, \ldots, 1/D)$. Esta composición es el elemento neutro por la operación perturbación –véase la Propiedad 2.3.

Propiedad 2.15 Sea $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ un conjunto de datos composicionales de $\mathcal{S}^D$ y $g(\mathbf{X})$ su media geométrica composicional. Entonces se cumple que $g\left(g(\mathbf{X})^{-1} \circ \mathbf{X}\right) = \mathbf{e}$, donde $\mathbf{e} = (\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D})$ representa el *baricentro* del símplex $\mathcal{S}^D$. □

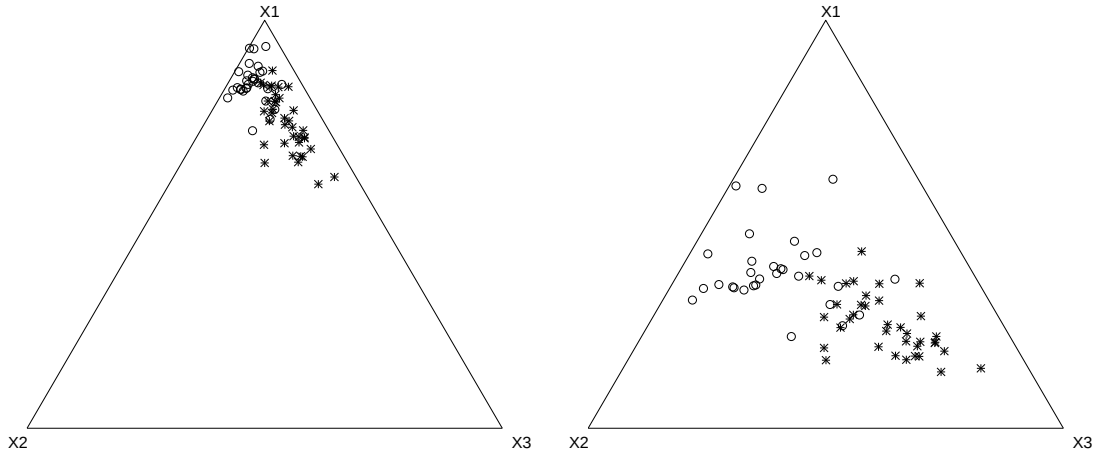
Con el propósito de ilustrar la utilidad de esta transformación de centrado de conjuntos de datos composicionales en el análisis estadístico de datos presentamos el ejemplo siguiente.

Ejemplo 2.2 Consideremos el conjunto de datos composicionales $\mathbf{X}$ llamado *Metabol* que aparece en Aitchison (1986). Los datos de este conjunto recogen la composición de las excreciones urinarias (mg/24 horas) de 37 adultos y de 30 niños. Cada observación está formada por las tres componentes siguientes:

1. x_1: total cortisol metabolitos;
2. x_2: total corticosterone metabolitos;
3. x_3: total pregnanetriol y Λ-5-pregnentriol.

En la figura 2.22(a) podemos observar que el conjunto de datos $\mathbf{X}$ se encuentra cercano al vértice x_1 debido a que esta componente adquiere valores cercanos a 1. En nuestro caso observamos que es muy difícil establecer visualmente en el gráfico si existen diferencias entre adultos y niños en relación al patrón de sus excreciones urinarias. En la figura 2.22(b) aparece el conjunto de datos $\mathbf{X}$ centrado. Este conjunto se ha obtenido calculando la media geométrica composicional del conjunto de datos $\mathbf{X}$ –$g(\mathbf{X}) = (0.8230, 0.0886, 0.0884)$–, y perturbando cada observación del conjunto de datos $\mathbf{X}$ por la perturbación $g(\mathbf{X})^{-1}$. De este modo, la media geométrica composicional del conjunto centrado es el baricentro del símplex $\mathcal{S}^3$. Podemos observar que en la figura 2.22(b) las diferencias entre los patrones relativos de las excreciones de adultos y niños aparecen con mayor claridad.

□



(a) (b)

Figura 2.22: Conjunto Metabol $\mathbf{X}$ en el diagrama ternario (símbolos: '$*$'-adulto y '$\circ$'-niño) donde: (a) *conjunto inicial;* (b) *conjunto centrado.*

2.7 Medida de dispersión de un conjunto de datos composicionales

2.7.1 La variabilidad composicional

Es bien sabido que una de las medidas de dispersión más utilizada para conjuntos de datos en el espacio real $\mathbb{R}^D$ es la traza de la matriz de covarianzas asociada al conjunto. Tal y como hemos citado en la Sección 2.3, diversos autores han hecho hincapié en la falta de interpretabilidad de la matriz de covarianzas directas de un conjunto de datos composicionales. Puesto que esta medida no es compatible con el carácter composicional de los datos, Aitchison (1997) define una medida de variabilidad $totvar(\mathbf{X})$ igual a la traza, traza($\mathbf{\Gamma}$), de la matriz de covarianzas del conjunto de datos clr-transformados. Siguiendo esta definición y usando la expresión (2.11) de la distancia de Aitchison que hemos establecido en la Sección 2.4.2, la siguiente propiedad, que aparece en Martín-Fernández et al. (1998b), muestra que esta medida de variabilidad $totvar(\mathbf{X})$ es compatible con la distancia d_{Ait}.

Propiedad 2.16 Sea $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ un conjunto de datos composicionales de $\mathcal{S}^D$ y $totvar(\mathbf{X}) = \text{traza}(\mathbf{\Gamma})$ su variabilidad total. Entonces se cumple que

$$totvar(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} d^2_{\text{Ait}}(\mathbf{x}_i, g(\mathbf{X})) = \frac{1}{2n^2}\sum_{i,j=1}^{n} d^2_{\text{Ait}}(\mathbf{x}_i, \mathbf{x}_k), \tag{2.33}$$

donde $g(\mathbf{X})$ es la media geométrica composicional del conjunto $\mathbf{X}$.

Demostración

Simbolizamos por $\mathbf{Z} = \text{clr}(\mathbf{X})$ al conjunto de datos clr-transformados y por $\mathbf{z}_i,\ i = 1, 2, \ldots, n$ a sus elementos. La siguiente secuencia de identidades nos demuestra la primera igualdad de la propiedad

$$\begin{aligned} totvar(\mathbf{X}) &= \text{traza}(\mathbf{\Gamma}) \\ &= \frac{1}{n}\sum_{k=1}^{D}\sum_{i=1}^{n}(z_{ik} - \overline{\mathbf{Z}}_k)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{D}(z_{ik} - \overline{\mathbf{Z}}_k)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} d^2_{\text{Euc}}(\mathbf{z}_i, \overline{\mathbf{Z}}) \\ &= \frac{1}{n}\sum_{i=1}^{n} d^2_{\text{Ait}}(\mathbf{x}_i, g(\mathbf{X})). \end{aligned}$$

La demostración de la segunda igualdad de la propiedad se obtiene mediante la siguiente secuencia de identidades

$$\begin{aligned} \frac{1}{2n^2}\sum_{i,j=1}^{n} d^2_{\text{Ait}}(\mathbf{x}_i, \mathbf{x}_j) &= \frac{1}{2n^2}\sum_{i,j=1}^{n}\sum_{k=1}^{D}(z_{ik} - z_{jk})^2 \\ &= \frac{1}{2n^2}\Big[\sum_{i,j=1}^{n}\sum_{k=1}^{D}(z_{ik} - \overline{\mathbf{Z}}_k)^2 + \sum_{i,j=1}^{n}\sum_{k=1}^{D}(z_{jk} - \overline{\mathbf{Z}}_k)^2 - \\ &\quad -2\sum_{i,j=1}^{n}\sum_{k=1}^{D}(z_{ik} - \overline{\mathbf{Z}}_k)(z_{jk} - \overline{\mathbf{Z}}_k)\Big] \\ &= \frac{1}{2n^2}\Big[2n\sum_{i=1}^{n}\sum_{k=1}^{D}(z_{ik} - \overline{\mathbf{Z}}_k)^2\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n} d^2_{\text{Ait}}(\mathbf{x}_i, g(\mathbf{X})). \end{aligned}$$

□

En el mismo trabajo de Cuadras et al. (1997) en que se define el concepto de *d-centro* de un conjunto de datos, también se define el concepto de *variabilidad geométrica* como una generalización de la medida de variabilidad total.

Definición 2.19 Sea $\mathbf{X}$ un vector aleatorio con función de densidad $f(\mathbf{X})$, respecto a una medida adecuada λ y espacio soporte $\mathbf{E}$. Sea $\delta(\mathbf{x}, \mathbf{x}^*)$ una distancia entre observaciones de $\mathbf{X}$. Se define la *variabilidad geométrica* de $\mathbf{X}$ respecto δ como

$$V_\delta(\mathbf{X}) = \frac{1}{2}\int_{\mathbf{E}\times\mathbf{E}} \delta^2(\mathbf{x}, \mathbf{x}^*) f(\mathbf{x}) f(\mathbf{x}^*) \lambda(d\mathbf{x}) \lambda(d\mathbf{x}^*). \tag{2.34}$$

□

La medida de variabilidad que acabamos de definir se generaliza fácilmente al caso en que $\mathbf{X}$ representa un conjunto de datos $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ como

$$V_\delta(\mathbf{X}) = \frac{1}{2n^2} \sum_{i,j}^{n} \delta^2(\mathbf{x}_i, \mathbf{x}_j). \tag{2.35}$$

Una simple comparación entre este concepto de variabilidad geométrica (2.35) y la expresión de la variabilidad total para datos composicionales (2.33) nos hace ver que, para el caso de datos composicionales y la distancia de Aitchison se trata del mismo concepto. En el mismo trabajo los autores muestran que para el caso de conjuntos de datos en el espacio $\mathbb{R}^D$, una medida de variabilidad total basada en la traza de la matriz de covarianzas está relacionada con la distancia euclídea de la misma manera que aparece en la expresión (2.33). Siguiendo este paralelismo entre los dos tipos de datos y las dos distancias, surge la idea de exigir que cualquier medida de variabilidad de un conjunto de datos composicionales sea coherente con las operaciones básicas perturbación, producto por escalar, y subcomposición. La propiedad siguiente establece esta coherencia.

Propiedad 2.17 Sea $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ un conjunto de datos composicionales de $\mathcal{S}^D$ y $totvar(\mathbf{X})$ su variabilidad total. Entonces se cumple que

i. $totvar(\mathbf{p} \circ \mathbf{X}) = totvar(\mathbf{X}),\ \forall \mathbf{p} \in \mathbb{R}_+^D$.

ii. $totvar(\alpha \cdot \mathbf{X}) = \alpha^2 totvar(\mathbf{X}),\ \forall \alpha \in \mathbb{R}$.

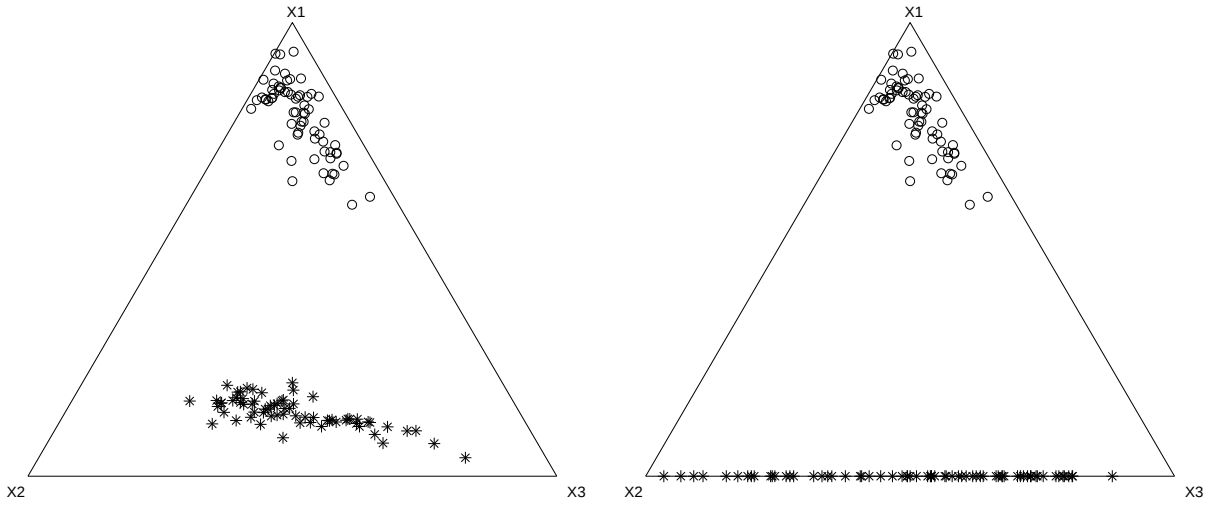
iii. $totvar(\mathcal{C}(\mathbf{X}_s)) \leq totvar(\mathbf{X})$, sea cual sea la subcomposición s escogida,

donde las expresiones $\mathbf{p} \circ \mathbf{X}$, $\alpha \cdot \mathbf{X}$, y $\mathcal{C}(\mathbf{X}_s)$ simbolizan los conjuntos que se obtienen al aplicar la correspondiente operación al conjunto $\mathbf{X}$. □

La demostración de estas propiedades se basa en los requerimientos, expuestos en la Propiedad 2.8, que satisface la distancia de Aitchison respecto de las operaciones básicas definidas en $\mathcal{S}^D$. Estos requerimientos, en combinación con el resultado (2.33), reducen la demostración de estas propiedades a unos sencillos cálculos.

La primera de las tres propiedades anteriores establece que la medida de variabilidad *totvar* es invariante por perturbaciones. En particular, al centrar cualquier conjunto $\mathbf{X}$ de datos composicionales, el valor de la medida de su variabilidad $totvar(\mathbf{X})$ se conservará. Este hecho se ilustra en la figura 2.22 para el centrado del conjunto de datos Metabol, donde $totvar(\mathbf{X}) = 0.8012$. La figura 2.23(a) ilustra la segunda propiedad: se muestra el conjunto de datos Metabol $\mathbf{X}$ y

el conjunto de datos $(-0.5) \cdot \mathbf{X}$ resultado de la operación producto por escalar para $\alpha = -0.5$. Finalmente, la figura 2.23(b) ilustra la última de las anteriores propiedades: puede observarse el conjunto resultado $\mathbf{X}_S$ de aplicar la operación subcomposición en las componentes $\mathbf{x}_2$ y $\mathbf{x}_3$ al conjunto Metabol $\mathbf{X}$.

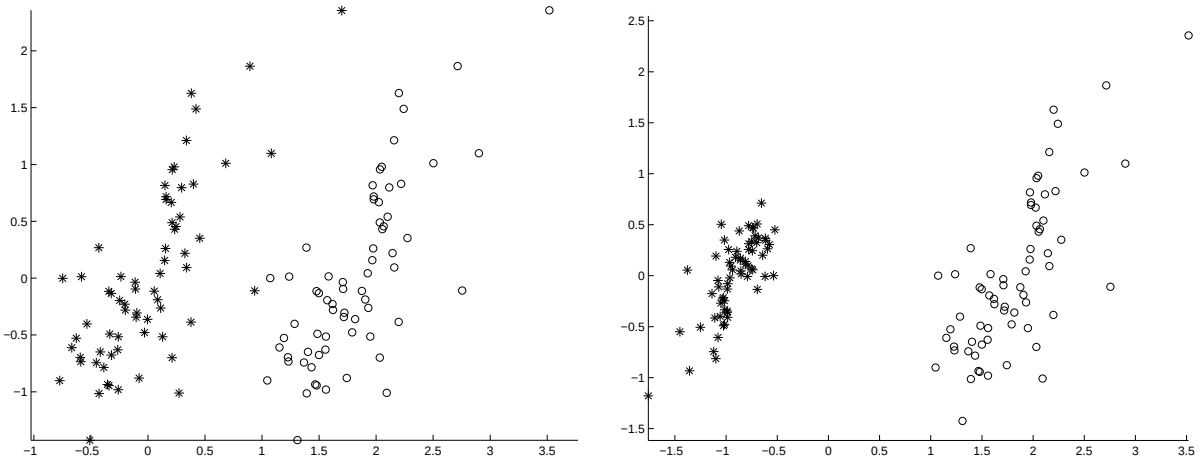


(a) (b)

Figura 2.23: Conjunto Metabol $\mathbf{X}$ en el diagrama ternario (símbolo '$\circ$') y el conjunto de datos que se obtiene al aplicar la operación: (a) *producto por escalar,* $(-0.5) \cdot \mathbf{X}$*;* (símbolo '$*$') (b) *subcomposición,* $\mathbf{X}_S$(símbolo '$*$')).

Es importante resaltar que, desde un punto de vista euclídeo somos incapaces de visualizar las propiedades en estos tres gráficos. Sin embargo, podemos visualizar estas propiedades en el espacio clr-transformado. En la figura 2.24(a) se representan los conjuntos que se obtienen al aplicar la transformación clr al conjunto Metabol y al conjunto Metabol centrado. En la figura 2.24(a), que se relaciona con la figura 2.22, podemos observar que una simple traslación nos transforma el conjunto $\mathrm{clr}(\mathbf{X})$ en el conjunto *centrado* $\mathrm{clr}(g(\mathbf{X})^{-1} \circ \mathbf{X})$. En consecuencia, los dos conjuntos tiene la misma variabilidad. En la figura 2.24(b) se representan los conjuntos clr-transformados de los conjuntos $\mathbf{X}$ y $(-0.5) \cdot \mathbf{X}$, respectivamente –véase el diagrama ternario de la figura 2.23(a). En la figura 2.24(b) se aprecia que la operación producto por escalar ($\alpha = -0.5$) ha producido una reducción de la variabilidad.

El hecho de disponer de una medida de variabilidad coherente con la estructura de espacio métrico del símplex $\mathcal{S}^D$ nos permite adaptar (Martín-Fernández et al., 1998b) los métodos de clasificación no paramétrica que se fundamentan en el concepto de variabilidad. Entre estos métodos se encuentran el método de Ward y los métodos de optimización.



(a) (b)

Figura 2.24: Conjunto Metabol transformado $\mathrm{clr}(\mathbf{X})$ (símbolo '$\circ$') y el conjunto transformado del conjunto que se obtiene a aplicar la operación: (a) *centrado*, $\mathrm{clr}(g(\mathbf{X})^{-1} \circ \mathbf{X})$ (símbolo '$*$'); (b) *producto por escalar*, $\mathrm{clr}((-0.5) \cdot \mathbf{X})$ (símbolo '$*$').

En resumen podemos afirmar que, la clasificación de un conjunto de datos composicionales utilizando la distancia de Aitchison (2.11), la media geométrica composicional (2.30) y la medida de variabilidad composicional (2.33), será coincidente con la clasificación de los datos clr-transformados utilizando la distancia euclídea, la media aritmética y la traza de la matriz de covarianzas.

Capítulo 3

Medidas de divergencia composicionales

3.1 Introducción

En su acepción clásica, se entiende por medidas de divergencias aquellas medidas que intentan reflejar la discrepancia o la diferencia entre dos distribuciones de probabilidad. Entre las diferentes familias de divergencias conocidas destacaremos en nuestro estudio las medidas conocidas como *Divergencias de Jeffreys* o $\mathcal{J}$*-divergencias*. Las medidas de divergencia han sido utilizadas en disciplinas muy diversas, pero quizás es en el campo de la teoría de la información donde las divergencias juegan su papel más importante (Cover, 1991). Las relaciones entre las medidas de divergencia y las medidas de diferencia, de disimilitud o de distancia han sido ampliamente estudiadas en los trabajos de Rao (1982), Burbea y Rao (1982), Burbea (1983), y Cuadras (1989). De la lectura de estos trabajos se ha extraído la información que se expone en la sección siguiente donde presentamos los conceptos más básicos en relación a las medidas de divergencia. Las demás secciones de este capítulo las dedicamos a estudiar la aplicación de las medidas de divergencia en problemas donde los datos son de tipología composicional. Es por este motivo que en este trabajo de investigación nos centraremos en las medidas de divergencia aplicadas a distribuciones de probabilidad multinomiales, siendo conscientes que todos los conceptos que se exponen pueden extenderse al caso de otras distribuciones de probabilidad, tanto discretas como continuas.

En los artículos de Medak y Cressie (1991), Rayens y Srinivasan (1991), y Aitchison (1997) se utilizan las medidas de divergencia en estudios de inferencia estadística para datos composicionales. Las divergencias tratadas como medidas de diferencia entre dos datos composicionales

aparecen en Martín (1996). En este capítulo proponemos una nueva medida de diferencia entre dos observaciones composicionales basada en la medida de Kullback-Leibler. La definición y primeras propiedades de esta nueva medida se presentan en Martín-Fernández et al. (1998c) y su análisis en profundidad y su generalización se aborda en Martín-Fernández et al. (1999).

3.2 Definiciones y propiedades básicas

En esta sección introducimos los aspectos más básicos de las medidas de divergencia entre dos distribuciones de probabilidad multinomiales. La mayoría de estos aspectos se presentan siguiendo la línea de exposición de Burbea (1983). En la mayoría de las definiciones y propiedades que se presentan aparecen los vectores de probabilidades $\mathbf{p}$ y $\mathbf{q}$ correspondientes a dos distribuciones multinomiales. Estos vectores de probabilidad pertenecen al espacio de parámetros

$$\mathcal{P}^D = \{\ \mathbf{p} = (p_1, p_2, \ldots, p_D) \in \mathbb{R}^D\ ;\ p_k \geq 0\ ;\ \sum p_k = 1\ \}.$$

Definición 3.1 Una aplicación

$$\begin{aligned} \mathcal{D} : \mathcal{P}^D \times \mathcal{P}^D &\rightarrow \mathbb{R} \\ \mathbf{p} \quad , \quad \mathbf{q} &\rightarrow \mathcal{D}(p_1, p_2, \ldots, p_D; q_1, q_2, \ldots, q_D), \end{aligned}$$

se denomina *medida de divergencia* si verifica las siguientes propiedades:

i. Es continua en todo punto de $\mathcal{P}^D \times \mathcal{P}^D$.

ii. Es invariante por permutaciones: $\mathcal{D}(\mathbf{P}\mathbf{p}^t, \mathbf{P}\mathbf{q}^t) = \mathcal{D}(\mathbf{p}, \mathbf{q}),\ \forall\ \mathbf{p}, \mathbf{q} \in \mathcal{P}^D$, sea cual sea la permutación $\mathbf{P}$ que se aplique.

iii. $\mathcal{D}(p_1, p_2, \ldots, p_D, 0; q_1, q_2, \ldots, q_D, 0) = \mathcal{D}(\mathbf{p}, \mathbf{q}),\ \forall\ \mathbf{p}, \mathbf{q} \in \mathcal{P}^D$.

iv. $\mathcal{D}(\mathbf{p}, \mathbf{q}) \geq 0,\ \forall\ \mathbf{p}, \mathbf{q} \in \mathcal{P}^D$.

v. $\mathcal{D}(\mathbf{p}, \mathbf{q}) = 0 \iff \mathbf{p} = \mathbf{q},\ \forall\ \mathbf{p}, \mathbf{q} \in \mathcal{P}^D$. □

Comparando la definición anterior con las definiciones de medida de diferencia y de disimilitud –véanse las Definiciones 1.1, 1.2 y 1.3– observamos que una medida de divergencia que sea simétrica la podemos catalogar como una disimilitud definida.

Las medidas de divergencia más utilizadas se basan en el concepto de *entropía*. La entropía es una medida de la incertidumbre o heterogeneidad de una variable aleatoria.

Definición 3.2 Una aplicación

$$\begin{aligned} \mathcal{H} : \mathcal{P}^D &\rightarrow \mathbb{R} \\ \mathbf{p} &\rightarrow \mathcal{H}(p_1, p_2, \ldots, p_D), \end{aligned}$$

se denomina *medida de entropía* si verifica las siguientes propiedades:

i. Es continua en todo punto de $\mathcal{P}^D$.

ii. Es invariante por permutaciones.

iii. $\mathcal{H}(p_1, p_2, \ldots, p_D, 0) = \mathcal{H}(\mathbf{p})$, $\forall\, \mathbf{p} \in \mathcal{P}^D$.

iv. $\mathcal{H}(\mathbf{p}) \geq 0$, $\forall\, \mathbf{p} \in \mathcal{P}^D$.

v. $\mathcal{H}(\mathbf{p}) = 0$ si, y solo si, existe un valor p_j tal que $p_j = 1$.

vi. $\mathcal{H}$ adquiere su valor máximo para $\mathbf{p} = (1/D, 1/D, \ldots, 1/D)$ y este valor máximo es una función creciente de la dimensión D. □

Entre las medidas de entropía destaca la *entropía de Shannon* definida como

$$\mathcal{H}(\mathbf{p}) = -\sum p_k \log(p_k), \tag{3.1}$$

donde si para algún valor se tiene que $p_k = 0$ se considera, por continuidad, que $p_k \log(p_k) = 0$.

Si consideramos que el vector de probabilidades $\mathbf{p}$ representa la distribución de probabilidad $\mathbf{f_X}$ de una variable aleatoria discreta $\mathbf{X}$, podemos interpretar la entropía de Shannon como la esperanza matemática de la variable $\log(\mathbf{f_X}(x))$

$$\mathcal{H}(\mathbf{p}) = -\mathbf{E_f}(\log(\mathbf{f_X}(x))) = -\mathbf{E_p}(\log(\mathbf{p})), \tag{3.2}$$

donde los subíndices de la esperanza matemática $\mathbf{E}$ indican la distribución de probabilidad a la que está asociada.

En Harris (1983) se propone que la medida de entropía pueda considerarse como un parámetro estadístico descriptivo de una variable aleatoria. En el caso de variables discretas se interpreta como una medida que nos informa si la probabilidad está concentrada en unos pocos puntos –*baja entropía*– o si está repartida en muchos puntos –*alta entropía*. En consecuencia, la entropía es una medida de dispersión como puede serlo la desviación estándar. En realidad una medida de dispersión basada en la entropía debe establecerse en términos de la diferencia

$\mathcal{H}(1/D, 1/D, \ldots, 1/D) - \mathcal{H}(\mathbf{p})$. Un aspecto remarcable del concepto de entropía es su invariación por transformaciones biyectivas de los valores de la variable aleatoria discreta. En la tabla 3.1 –extraída de Harris (1983), pág. 515– se muestran los valores x_k comunes a dos variables aleatorias discretas $\mathbf{X}_1$ y $\mathbf{X}_2$, y sus respectivas distribuciones de probabilidad $\mathbf{f}_{\mathbf{X}_1}$ y $\mathbf{f}_{\mathbf{X}_2}$. Puede comprobarse fácilmente que en ambos casos la entropía de Shannon es igual a 1.5 y la esperanza es igual a 0. Por el contrario, la varianza en el primer caso es igual a 0.5 y en el segundo toma el valor 2. Nótese que la entropía es una medida que, al contrario que la varianza, no depende de los valores de la variable aleatoria y únicamente basa su valor en la distribución de la probabilidad.

Tabla 3.1: Distribuciones de probabilidad de las variables aleatorias $\mathbf{X}_1$ y $\mathbf{X}_2$.

x_k	$\mathbf{f}_{\mathbf{X}_1}(x_k)$	$\mathbf{f}_{\mathbf{X}_2}(x_k)$
-2	0	$\frac{1}{4}$
-1	$\frac{1}{4}$	0
0	$\frac{1}{2}$	$\frac{1}{2}$
1	$\frac{1}{4}$	0
2	0	$\frac{1}{4}$

A la entropía de Shannon (3.1) se le asocia la divergencia

$$\mathcal{I}(\mathbf{p}, \mathbf{q}) = \mathcal{H}(\mathbf{q}) - \mathcal{H}(\mathbf{p}) = -\mathbf{E}_{\mathbf{p}}(\log(\mathbf{q})) + \mathbf{E}_{\mathbf{p}}(\log(\mathbf{p})). \tag{3.3}$$

Si desarrollamos la expresión (3.3) obtenemos la divergencia denominada *medida de información de Kullback-Leibler*:

$$\mathcal{I}(\mathbf{p}, \mathbf{q}) = \sum p_k \log\left(\frac{p_k}{q_k}\right). \tag{3.4}$$

En el cálculo de esta divergencia se sigue el convenio, que hemos introducido en (3.1) y que se hará extensivo a otras divergencias, de no considerar en (3.4) los sumandos en los que los elementos p_k y q_k tengan simultáneamente el valor cero. Por otra parte, obsérvese que en (3.4) hemos substituido la letra $\mathcal{D}$ por la $\mathcal{I}$. Substituciones similares se irán efectuando a medida que vayamos presentando las diferentes divergencias.

Con el propósito de poseer una divergencia que sea simétrica respecto las distribuciones $\mathbf{p}$ y $\mathbf{q}$, se define la medida

$$\mathcal{J}(\mathbf{p}, \mathbf{q}) = \mathcal{I}(\mathbf{p}, \mathbf{q}) + \mathcal{I}(\mathbf{q}, \mathbf{p}) = \sum (p_k - q_k) \log\left(\frac{p_k}{q_k}\right), \tag{3.5}$$

conocida como el *invariante de Jeffreys* o la *$\mathcal{J}$-divergencia* entre $\mathbf{p}$ y $\mathbf{q}$. La aplicación de esta medida a datos composicionales se propone en Martín (1996). Recordemos que en esta tesis

–véase la Sección 2.4.2– se ha expuesto que la $\mathcal{J}$-divergencia es una medida que no verifica los requerimientos de invariación por perturbaciones y de dominación por subcomposiciones introducidos en la Propiedad 2.8. Sin embargo, los entornos en $\mathcal{S}^3$ muestran que la $\mathcal{J}$-divergencia (3.5) tiene un comportamiento coherente con la naturaleza composicional. Este buen comportamiento se interpreta si expresamos la medida de información de Kullback-Leibler dada en (3.4) como el valor esperado de una función de los cocientes de los dos vectores de probabilidad

$$\mathcal{I}(\mathbf{p},\mathbf{q}) \;=\; \mathbf{E}_{\mathbf{p}}\left(\log\left(\frac{\mathbf{p}}{\mathbf{q}}\right)\right). \tag{3.6}$$

El hecho de medir la diferencia en términos de los cocientes de las componentes entronca de lleno con las ideas expuestas por Aitchison (1986) respecto a las características deseables para que una disimilitud sea adecuada a la naturaleza composicional. La entropía de Shannon y la medida de información de Kullback-Leibler son las medidas más utilizadas en la disciplina de la Teoría de la Información. A continuación presentamos una familia de medidas que generaliza las medidas $\mathcal{H}$ e $\mathcal{I}$.

3.3 Tipos de divergencias más usuales

La entropía de Shannon pertenece a una familia de entropías más general denominada *entropías de Havrda-Charvát de grado* α cuya expresión viene dada por

$$\mathcal{H}_\alpha(\mathbf{p}) = \frac{1}{\alpha-1}\left(1-\sum p_k^\alpha\right)\;; \qquad \alpha>0, \tag{3.7}$$

donde para $\alpha\to 1$ se obtiene la entropía de Shannon, $\mathcal{H}_1(\mathbf{p})=\mathcal{H}(\mathbf{p})$. Para $\alpha=2$ se obtiene el conocido *índice de diversidad de Gini-Simpson* $\mathbf{G}(\mathbf{p})=\mathcal{H}_2(\mathbf{p})=1-\sum p_k^2$.

De manera similar al caso de la entropía de Shannon –véase (3.2)– se considera, para cada $\alpha>0$, la función real de variable real $\boldsymbol{\phi}_\alpha(x)=(\alpha-1)^{-1}(x^{\alpha-1}-1)$ y se define la entropía de Havrda-Charvát de grado α como

$$\mathcal{H}_\alpha(\mathbf{p}) = -\mathbf{E}_{\mathbf{p}}\left(\boldsymbol{\phi}_\alpha(\mathbf{p})\right)\;; \qquad \alpha>0. \tag{3.8}$$

Ciertamente, existe una familia de entropías aún más general: las $\boldsymbol{\phi}$-entropías. Éstas se definen de manera similar a (3.8) considerando funciones reales de variable real $\boldsymbol{\phi}$ tales que: 1) sean convexas en el intervalo $[0,1]$; 2) sean de clase C^2; y 3) que $\boldsymbol{\phi}(0)=\boldsymbol{\phi}(1)=0$. Por ser las entropías de Havrda-Charvát las medidas más utilizadas, en este trabajo de investigación nos centramos en ellas y en sus correspondientes medidas de divergencia asociadas.

Para definir medidas de divergencia basadas en el concepto de entropía, se siguen diversas estrategias que dan lugar a las familias de divergencias siguientes: *$\mathcal{I}$-divergencias*, *$\mathcal{J}$-divergencias*, *$\mathcal{K}$-divergencias*, y *$\mathcal{L}$-divergencias*. A continuación se exponen las definiciones de cada una de estas familias, haciendo mención especial a la familia de divergencias denominadas *distancias de Hellinger de grado α*.

- **$\mathcal{I}$-divergencias de grado α**

  Una divergencia de esta familia la simbolizaremos por $\mathcal{I}_\alpha(\mathbf{p},\mathbf{q})$. Esta familia de divergencias, conocida también como *$\boldsymbol{\phi}$-divergencias de Csiszár*, se define a partir de las entropías de Havrda-Charvát de grado α.

  Definición 3.3 Si $\mathbf{p},\mathbf{q} \in \mathcal{P}^D$ son dos vectores de probabilidad, se define la *divergencia directa de grado α* entre $\mathbf{p}$ y $\mathbf{q}$ como

$$\mathcal{I}_\alpha(\mathbf{p},\mathbf{q}) = \frac{1}{\alpha-1}\left(\sum p_k^\alpha q_k^{1-\alpha} - 1\right); \quad \alpha > 0. \tag{3.9}$$

  □

  Obsérvese que para $\alpha \to 1$ se obtiene la medida de información de Kullback-Leibler – véase (3.4). De manera similar a (3.6) puede definirse la *divergencia directa de grado α* en términos de la esperanza del vector de los cocientes

$$\mathcal{I}_\alpha(\mathbf{p},\mathbf{q}) \;=\; \mathbf{E}_\mathbf{p}\left(\boldsymbol{\phi}_\alpha\left(\frac{\mathbf{p}}{\mathbf{q}}\right)\right), \tag{3.10}$$

  y, en consecuencia, parece lógico esperar que todas estas medidas de divergencia aplicadas a datos composicionales produzcan resultados razonables. Por otra parte, nótese que la función $\boldsymbol{\phi}_\alpha$ aplicada a cada componente del vector de cocientes $\mathbf{p}/\mathbf{q}$ viene dada por la expresión

$$\boldsymbol{\phi}_\alpha(p_k/q_k) = \frac{1}{\alpha-1}((\frac{p_k}{q_k})^{\alpha-1} - 1).$$

  Esta expresión recoge la divergencia entre la ratio $(p_k/q_k)^{\alpha-1}$ y el valor 1. Entonces podemos interpretar (3.10) como una medida del valor esperado de la diferencia entre el vector ratio $(\mathbf{p}/\mathbf{q})^{\alpha-1}$ y el vector $(1,1,\ldots,1)$. Por esta razón, a la familia de medidas $\mathcal{I}_\alpha$ también se la conoce por *Power divergence class*. Esta observación que acabamos de realizar nos será de gran utilidad cuando definamos una medida de divergencia para datos composicionales.

Entre las divergencias $\mathcal{I}_\alpha$ destaca la medida que se obtiene para $\alpha = \frac{1}{2}$:

$$\mathcal{I}_{1/2}(\mathbf{p},\mathbf{q}) \;=\; 2\left(1 - \sum p_k^{1/2} q_k^{1/2}\right) = \sum \left(p_k^{1/2} - q_k^{1/2}\right)^2, \tag{3.11}$$

que coincide con la distancia de Matusita presentada en la Sección 2.4.2 de esta tesis. Recordemos que esta medida está, además, íntimamente relacionada con las disimilitudes de Bhattacharyya que hemos presentado en la tabla 2.1. Todas estas medidas muestran un comportamiento razonable –véase la figura 2.17– cuando se aplican a datos composicionales a pesar de no verificar los requisitos de invariación por perturbaciones y de dominación por subcomposiciones.

Por otra parte, observemos que la medida $\mathcal{I}_{1/2}$ dada en (3.11) es la única de las divergencias directas de grado α que tiene la buena propiedad de ser simétrica, $\mathcal{I}_{1/2}(\mathbf{p},\mathbf{q}) = \mathcal{I}_{1/2}(\mathbf{q},\mathbf{p})$ y por lo tanto, es la única que es una disimilitud. Con el propósito de tener medidas de divergencia simétricas se definen las divergencias denominadas $\mathcal{J}$-divergencias.

- **$\mathcal{J}$-divergencias de grado α**

  En general una $\mathcal{I}$-divergencia no es una medida simétrica. Si se desea obtener a partir de $\mathcal{I}_\alpha$ una divergencia simétrica puede definirse la medida siguiente:

  $$\mathcal{J}_\alpha(\mathbf{p},\mathbf{q}) = \mathcal{I}_\alpha(\mathbf{p},\mathbf{q}) + \mathcal{I}_\alpha(\mathbf{q},\mathbf{p}). \tag{3.12}$$

  Este tipo de medida se denomina $\mathcal{J}$-divergencia de grado α. Destacamos que para el caso $\alpha \to 1$ se obtiene el invariante de Jeffreys –véase (3.5)– y que para el valor $\alpha = 1/2$ se tiene que $\mathcal{J}_{1/2} = 2\,\mathcal{I}_{1/2}$, es decir, es una medida proporcional a la distancia de Matusita.

  A partir de su definición, es inmediato comprobar que cualquier $\mathcal{J}$-divergencia aplicada a datos composicionales hereda de la $\mathcal{I}$-divergencia dos características esenciales: el comportamiento razonable y el incumplimiento de los requisitos expuestos en la Propiedad 2.8.

- **$\mathcal{K}$-divergencias de grado α**

  Una medida perteneciente a esta familia de divergencias se simboliza por $\mathcal{K}_\alpha^\lambda(\mathbf{p},\mathbf{q})$ y se define como

  $$\mathcal{K}_\alpha^\lambda(\mathbf{p},\mathbf{q}) = \mathcal{H}_\alpha(\lambda\mathbf{p} + (1-\lambda)\mathbf{q}) - [\lambda\mathcal{H}_\alpha(\mathbf{p}) + (1-\lambda)\mathcal{H}_\alpha(\mathbf{q})] \quad 0 < \lambda < 1,\ \alpha > 0, \tag{3.13}$$

  donde $\mathcal{H}_\alpha$ es la entropía Havrda-Charvát de grado α definida en (3.8). Obsérvese que la definición de la familia de $\mathcal{K}$-divergencias se basa en la desigualdad de Jensen. Por este

motivo a esta familia de divergencias se la conoce también por el nombre de *diferencias de Jensen de grado* α. Entre las medidas que pertenecen a esta familia destacamos la que se obtiene para $\alpha = 2$. Esta medida, basada en el índice de diversidad de Gini-Simpson $\mathbf{G}(\mathbf{p}) = \mathcal{H}_2(\mathbf{p}) = 1 - \sum p_k^2$, resulta ser una medida proporcional al cuadrado de la distancia euclídea:

$$\mathcal{K}_2^\lambda(\mathbf{p}, \mathbf{q}) = \lambda(1-\lambda) \sum_{k=1}^{D} (p_k - q_k)^2 = \lambda(1-\lambda)\, d^2_{\text{Euc}}(\mathbf{p}, \mathbf{q}).$$

El hecho que la medida $\mathcal{K}_2^\lambda$ y la distancia euclídea estén relacionadas por una transformación monótona las hace equivalentes cuando se aplica cualquier método de clasificación automática que sea invariante por transformaciones monótonas de la matriz de disimilitudes.

Cualquier medida de la familia de las $\mathcal{K}$-divergencias no presentará un comportamiento adecuado para los datos composicionales puesto que como puede observarse en su definición (3.13), estas medidas no se expresan en términos de los cocientes de las composiciones $\mathbf{p}$ y $\mathbf{q}$.

- **$\mathcal{L}$-divergencias de grado α**

  Una medida de la familia de $\mathcal{L}$-divergencias se define como

$$\mathcal{L}_\alpha(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{D} (p_k - q_k) \left(\boldsymbol{\phi}_\alpha(p_k) - \boldsymbol{\phi}_\alpha(q_k)\right), \quad \alpha > 0, \tag{3.14}$$

  donde $\boldsymbol{\phi}_\alpha(x) = (\alpha-1)^{-1}(x^{\alpha-1} - 1)$ es la función que hemos usado en (3.8) para definir la entropía de Havrda-Charvát de grado α. Si se desarrolla la expresión (3.14) se obtiene la expresión

$$\mathcal{L}_\alpha(\mathbf{p}, \mathbf{q}) = \frac{1}{\alpha - 1} \sum_{k=1}^{D} (p_k - q_k) \left(p_k^{\alpha-1} - q_k^{\alpha-1}\right). \tag{3.15}$$

  Obsérvese que para $\alpha = 2$ obtenemos que $\mathcal{L}_2(\mathbf{p}, \mathbf{q}) = d^2_{\text{Euc}}(\mathbf{p}, \mathbf{q})$. Para $\alpha \to 1$ la $\mathcal{L}$-divergencia se basa en la entropía de Shannon y tiene la expresión

$$\mathcal{L}_1(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{D} (p_k - q_k) \left(\log(p_k) - \log(q_k)\right), \tag{3.16}$$

  que coincide con la definición del invariante de Jeffreys –véase (3.5). Es éste el único caso en que una medida de la familia de las $\mathcal{L}$-divergencias presenta un comportamiento adecuado para los datos composicionales. Sin embargo, los requisitos de invariación por perturbaciones y de dominación por subcomposiciones no se satisfacen para ningún valor de α.

- **Distancias de Hellinger de grado α**

  En conexión con las familias de divergencias anteriores se definen las distancias de Hellinger de grado α

$$\mathcal{M}_\alpha = 2\,|\alpha|^{-1}\left[\sum_{k=1}^{D}(p_k^{\alpha/2} - q_k^{\alpha/2})^2\right]^{1/2}, \tag{3.17}$$

  donde para $\alpha \to 0$

$$\mathcal{M}_0 = \left[\sum_{k=1}^{D}(\log(p_k) - \log(q_k))^2\right]^{1/2}. \tag{3.18}$$

  La familia $\mathcal{M}_\alpha$ no proviene como las otras familias de la entropía de Havrda-Charvát de grado α. Sin embargo, está fuertemente relacionada con ellas. Por ejemplo, para el caso $\alpha = 2$ resulta que $\mathcal{M}_2$ es la distancia euclídea. Si consideramos el caso $\alpha = 1$ se observa que la medida $\mathcal{M}_1$ coincide con la distancia de Matusita presentada en la Sección 2.4.2. Por lo tanto, tal y como hemos expuesto en (3.11), la distancia $\mathcal{M}_1$ puede considerarse una $\mathcal{I}$-divergencia. Por otra parte, queremos destacar que para el caso especial de $\alpha \to 0$ la medida $\mathcal{M}_0$ coincide con la distancia logarítmica presentada en la Sección 2.4.2. En consecuencia, puede afirmarse que el comportamiento de las medidas $\mathcal{M}_0$ y $\mathcal{M}_1$ para los datos composicionales será razonable.

Dado el enorme abanico de medidas que se abre al considerar las familias de divergencias anteriores, se hace preciso establecer algún criterio de selección que nos permita identificar las que puedan ser las más adecuadas para los datos composicionales. En Burbea y Rao (1982) se expone que la familia de distancias de Hellinger de grado α son distancias geodésicas para el espacio de parámetros de las distribuciones de probabilidad multinomiales. En Martín (1996) se expone que entre las $\mathcal{J}$-divergencias la más adecuada para los datos composicionales es el invariante de Jeffreys que hemos definido en (3.5). En consecuencia, centraremos nuestro estudio en las medidas $\mathcal{M}_\alpha$ y las $\mathcal{J}$-divergencias.

3.4 Divergencias composicionales

Realmente, de todas las medidas de diferencia que hemos presentado en este trabajo de investigación, la distancia de Aitchison y la distancia de Mahalanobis (clr) son las únicas medidas que muestran un comportamiento adecuado para los datos composicionales y que satisfacen los requerimientos de invariación por perturbaciones y dominación por subcomposiciones –véase la Propiedad 2.8. En su uso práctico, el cálculo de la distancia de Mahalanobis necesita de la existencia de un conjunto de datos. Cuando se realiza una clasificación automática se desconoce

la estructura de grupos existente y, en consecuencia, los métodos de clasificación más habituales no utilizan esta distancia. Por este motivo, y a la luz de los objetivos que nos hemos marcado, en esta sección solo estudiaremos la distancia de Aitchison.

En la sección anterior hemos expuesto que entre las familias de divergencias se encuentran medidas que muestran un comportamiento adecuado para los datos composicionales. La razón expuesta ha sido que son medidas que se expresan en términos de los cocientes de las dos observaciones composicionales. Sin embargo, ninguna de estas medidas satisfacen los requerimientos de la Propiedad 2.8. La pregunta que surge inmediatamente es: ¿es posible definir una divergencia que mantenga el comportamiento adecuado y que, además, satisfaga los requerimientos de la Propiedad 2.8? Este problema ha sido parcialmente estudiado en Martín-Fernández et al. (1998c, 1999). Para resolver esta cuestión hemos optado por profundizar en el estudio de la distancia de Aitchison. Este estudio nos plantea nuevas cuestiones:

- ¿Cómo puede expresarse la distancia de Aitchison en términos de los cocientes?
- ¿Puede interpretarse como una divergencia?
- ¿Existen características especiales que impliquen que esta distancia satisfaga los requisitos?

Recordemos que la distancia de Aitchison (al cuadrado) de define por

$$d^2_{\mathrm{Ait}}(\mathbf{x},\mathbf{x}^*) = \sum_{k=1}^{D}\left[\log\left(\frac{x_k}{g(\mathbf{x})}\right) - \log\left(\frac{x^*_k}{g(\mathbf{x}^*)}\right)\right]^2 = d^2_{\mathrm{Euc}}(\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{x}^*)), \tag{3.19}$$

donde $g(\mathbf{x})$ es la media geométrica de la observación $\mathbf{x}$, d_{Euc} representa la distancia euclídea, y clr es la transformación logratio centrada. Las siguientes manipulaciones algebraicas nos muestran la distancia de Aitchison expresada en términos de los cocientes de las observaciones:

$$\begin{aligned} d^2_{\mathrm{Ait}}(\mathbf{x},\mathbf{x}^*) &= \sum_{k=1}^{D}\left[\log\left(\frac{x_k}{g(\mathbf{x})}\right) - \log\left(\frac{x^*_k}{g(\mathbf{x}^*)}\right)\right]^2 \\ &= \sum_{k=1}^{D}\left[\log\left(\frac{x_k}{x^*_k}\right) - \log\left(g(\frac{\mathbf{x}}{\mathbf{x}^*})\right)\right]^2 \\ &= \sum_{k=1}^{D}\left[\log\left(\frac{\frac{x_k}{x^*_k}}{g(\frac{\mathbf{x}}{\mathbf{x}^*})}\right)\right]^2. \end{aligned} \tag{3.20}$$

Por otra parte, el hecho que la distancia d_{Ait} sea invariante por perturbaciones nos permite expresarla en los siguientes términos:

$$d_{\mathrm{Ait}}(\mathbf{x},\mathbf{x}^*) = d_{\mathrm{Ait}}(\mathbf{e},\mathbf{x}\circ\mathbf{x}^{*-1}), \tag{3.21}$$

donde el centro del símplex $\mathbf{e} = (1/D, 1/D, \ldots, 1/D)$ es el elemento neutro de la operación perturbación. Obsérvese que hemos expresado la distancia de Aitchison como una medida de

diferencia entre el elemento $\mathbf{e}$ y la perturbación diferencia $\mathbf{x} \circ \mathbf{x}^{*-1}$. Una propiedad análoga a la expresada en (3.21) se satisface para datos en el espacio $\mathbb{R}^D$ si consideramos la distancia euclídea y la operación suma:

$$d_{\text{Euc}}(\mathbf{x}, \mathbf{x}^*) = d_{\text{Euc}}(\mathbf{0}, \mathbf{x} - \mathbf{x}^*). \tag{3.22}$$

Las expresiones (3.20) y (3.21) nos llevan a plantearnos la búsqueda de una expresión única de la d_{Ait} que fusione los dos planteamientos. Una solución a esta cuestión surge de la siguiente cadena de igualdades:

$$\begin{aligned} d^2_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) &= d^2_{\text{Ait}}(\mathbf{e}, \mathbf{x} \circ \mathbf{x}^{*-1}) \\ &= d^2_{\text{Euc}}(\text{clr}(\mathbf{e}), \text{clr}(\mathbf{x} \circ \mathbf{x}^{*-1})) \\ &= d^2_{\text{Euc}}(\mathbf{0}, \text{clr}(\mathbf{x} \circ \mathbf{x}^{*-1})) \\ &= \sum_{k=1}^{D} \left[\log \left(\frac{(\mathbf{x} \circ \mathbf{x}^{*-1})_k}{g(\mathbf{x} \circ \mathbf{x}^{*-1})} \right) \right]^2. \end{aligned} \tag{3.23}$$

Esta última expresión muestra la d_{Ait} en términos de la perturbación diferencia. Si volvemos a fijarnos en la expresión (3.23) vemos que el hecho de definir la distancia en términos de $\mathbf{x} \circ \mathbf{x}^{*-1}$ nos asegura que la distancia satisface el requisito de la invariación por perturbaciones. Ciertamente, puesto que $\mathbf{x} \circ \mathbf{x}^{*-1} = (\mathbf{p} \circ \mathbf{x}) \circ (\mathbf{p} \circ \mathbf{x}^*)^{-1} \quad \forall\, \mathbf{p}, \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$, entonces es obvio que $d_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) = d_{\text{Ait}}(\mathbf{p} \circ \mathbf{x}, \mathbf{p} \circ \mathbf{x}^*)$. Esta observación nos servirá en el momento de definir una nueva medida de diferencia, puesto que una medida de diferencia será invariante por perturbaciones si, y solo si, puede expresarse en términos de la perturbación diferencia $\mathbf{x} \circ \mathbf{x}^{*-1}$.

Por otra parte desarrollando la expresión (3.23) se obtiene

$$\begin{aligned} d^2_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) &= \sum_{k=1}^{D} \left[\log \left(\frac{(\mathbf{x} \circ \mathbf{x}^{*-1})_k}{g(\mathbf{x} \circ \mathbf{x}^{*-1})} \right) \right]^2 \\ &= \sum_{k=1}^{D} \left[\log((\mathbf{x} \circ \mathbf{x}^{*-1})_k) - \log(g(\mathbf{x} \circ \mathbf{x}^{*-1})) \right]^2 \\ &= \sum_{k=1}^{D} \left[\log((\mathbf{x} \circ \mathbf{x}^{*-1})_k) - \overline{\log(\mathbf{x} \circ \mathbf{x}^{*-1})} \right]^2 \\ &= D \cdot \frac{1}{D} \sum_{k=1}^{D} \left[\log((\mathbf{x} \circ \mathbf{x}^{*-1})_k) - \overline{\log(\mathbf{x} \circ \mathbf{x}^{*-1})} \right]^2, \end{aligned} \tag{3.24}$$

donde $\overline{\log(\mathbf{x} \circ \mathbf{x}^{*-1})}$ representa la media aritmética de las componentes del vector $\log(\mathbf{x} \circ \mathbf{x}^{*-1})$. La expresión (3.24) nos define la distancia de Aitchison en términos de la varianza del vector $\log(\mathbf{x} \circ \mathbf{x}^{*-1})$. Esta varianza es mínima –igual a zero– cuando el vector es constante, lo que ocurre sólo para el caso $\log(\mathbf{x} \circ \mathbf{x}^{*-1}) = \log(\mathbf{e})$, es decir, para $\mathbf{x} = \mathbf{x}^*$. Recordemos que, tal y

como se ha expuesto en la Sección 3.2, para medir el grado de heterogeneidad de un vector $\mathbf{p}$ de probabilidades pueden usarse medidas de dispersión basadas en una medida de entropía, por ejemplo $\mathcal{H}_\alpha(\mathbf{p})$ –véase (3.7). Puesto que el valor máximo de la entropía se obtiene para $\mathbf{p} = \mathbf{e}$, puede definirse una medida de dispersión basada en el concepto de entropía en términos de la diferencia $\mathcal{H}_\alpha(\mathbf{e}) - \mathcal{H}_\alpha(\mathbf{p})$. La combinación de estas ideas con la expresión (3.24) nos ofrece la posibilidad de enunciar la siguiente definición:

Definición 3.4 Sea $\mathcal{H}_\alpha$ una medida de entropía de Havrda-Charvát de grado α. Definimos la medida de diferencia $d_{\mathcal{H}_\alpha}$ (al cuadrado) entre las composiciones $\mathbf{x}$ y $\mathbf{x}^*$ del símplex $\mathcal{S}^D$ como

$$d^2_{\mathcal{H}_\alpha}(\mathbf{x}, \mathbf{x}^*) = D \cdot \frac{1}{2}\left(\mathcal{H}_\alpha(\mathbf{e}) - \mathcal{H}_\alpha(\mathbf{x} \circ \mathbf{x}^{*-1}) + \mathcal{H}_\alpha(\mathbf{e}) - \mathcal{H}_\alpha(\mathbf{x}^* \circ \mathbf{x}^{-1})\right). \tag{3.25}$$

□

Cualquier medida de diferencia así definida satisface la propiedad siguiente:

Propiedad 3.1 Sea $d_{\mathcal{H}_\alpha}$ una medida de diferencia establecida en la definición anterior, entonces se cumplen las siguientes propiedades:

1. $d_{\mathcal{H}_\alpha}(\mathbf{x}, \mathbf{x}^*) \geq 0$, $\forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$.
2. $d_{\mathcal{H}_\alpha}(\mathbf{x}, \mathbf{x}^*) = 0 \Longleftrightarrow \mathbf{x} = \mathbf{x}^*$, $\forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$.
3. $d_{\mathcal{H}_\alpha}(\mathbf{x}, \mathbf{x}^*) = d_{\mathcal{H}_\alpha}(\mathbf{x}^*, \mathbf{x})$, $\forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$.
4. $d_{\mathcal{H}_\alpha}(\mathbf{P}\mathbf{x}^t, \mathbf{P}\mathbf{x}^{*t}) = d_{\mathcal{H}_\alpha}(\mathbf{x}, \mathbf{x}^*)$, $\forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$, sea cual sea la permutación $\mathbf{P}$ que se aplique a las componentes de una D-parte.
5. $d_{\mathcal{H}_\alpha}(\mathbf{p} \circ \mathbf{x}, \mathbf{p} \circ \mathbf{x}^*) = d_{\mathcal{H}_\alpha}(\mathbf{x}, \mathbf{x}^*)$, $\forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D, \forall \mathbf{p} \in \mathbb{R}^D_+$. □

Omitimos las demostraciones de estas propiedades debido a su extrema sencillez.

Nótese que estas propiedades nos permiten catalogar a cualquier medida del tipo $d_{\mathcal{H}_\alpha}$ como una disimilitud definida que es invariante por permutaciones y por perturbaciones.

Existe un camino alternativo que también nos conduce a la expresión presentada en la Definición 3.4. Esta nueva vía consiste en considerar la medida de similitud definida en el símplex por la expresión

$$s_{\mathcal{H}_\alpha}(\mathbf{x}, \mathbf{x}^*) = \frac{1}{2}\left(\mathcal{H}_\alpha(\mathbf{x} \circ \mathbf{x}^{*-1}) + \mathcal{H}_\alpha(\mathbf{x}^* \circ \mathbf{x}^{-1})\right).$$

La demostración de que la expresión anterior es una similitud es extremadamente simple y hemos optado por omitirla. Entonces, a partir de esta similitud $s_{\mathcal{H}_\alpha}$ podemos establecer una medida de diferencia $d_{\mathcal{H}_\alpha}$ mediante la transformación presentada en la Sección 1.4.2

$$d_{\mathcal{H}_\alpha}(\mathbf{x},\mathbf{x}^*) = \sqrt{s_{\mathcal{H}_\alpha}(\mathbf{x},\mathbf{x}) + s_{\mathcal{H}_\alpha}(\mathbf{x}^*,\mathbf{x}^*) - 2s_{\mathcal{H}_\alpha}(\mathbf{x},\mathbf{x}^*)}. \tag{3.26}$$

Si en esta expresión se substituyen los valores de $s_{\mathcal{H}_\alpha}$ se obtiene

$$d_{\mathcal{H}_\alpha}(\mathbf{x},\mathbf{x}^*) = \sqrt{\mathcal{H}_\alpha(\mathbf{e}) + \mathcal{H}_\alpha(\mathbf{e}) - \mathcal{H}_\alpha(\mathbf{x}\circ\mathbf{x}^{*-1}) - \mathcal{H}_\alpha(\mathbf{x}^*\circ\mathbf{x}^{-1})}. \tag{3.27}$$

Ciertamente, esta medida coincide con la dada en la Definición 3.4 salvo el factor constante $\sqrt{D/2}$.

Observemos que nos aparece como línea de investigación futura realizar un desarrollo de estos conceptos con el objetivo de analizar para qué valores de α una medida del tipo $d_{\mathcal{H}_\alpha}$ cumple las propiedades de la dominación por subcomposiciones y la desigualdad triangular. Obsérvese que si para algún valor de α se satisfacen estas dos propiedades, entonces se dispondrá de una métrica composicional. En relación con este objetivo, en la siguiente sección presentamos una medida de diferencia basada en la entropía de Shannon.

3.5 Medida de Kullback-Leibler composicional

En la Definición 3.4 se ha establecido una familia de medidas de diferencia basadas en el concepto de entropía. Estas medidas poseen un conjunto de propiedades que las convierten en medidas adecuadas para los datos composicionales. En esta sección vamos a desarrollar el estudio de las medidas $d_{\mathcal{H}_\alpha}$ para el caso que $\alpha \to 1$, es decir la disimilitud basada en la entropía de Shannon.

La entropía de Shannon, que simbolizamos simplemente por $\mathcal{H}$, se define –véase (3.1)– mediante la expresión

$$\mathcal{H}(\mathbf{p}) = -\sum p_k \log(p_k). \tag{3.28}$$

La disimilitud $d_{\mathcal{H}}$ (al cuadrado) asociada a esta entropía resulta ser

$$\begin{aligned}
d^2_{\mathcal{H}}(\mathbf{x},\mathbf{x}^*) &= \frac{D}{2}\Big(\mathcal{H}(\mathbf{e}) - \mathcal{H}(\mathbf{x}\circ\mathbf{x}^{*-1}) + \mathcal{H}(\mathbf{e}) - \mathcal{H}(\mathbf{x}^*\circ\mathbf{x}^{-1})\Big) \\
&= \frac{D}{2}\left(\log(D) + \sum_{k=1}^{D}(\mathbf{x}\circ\mathbf{x}^{*-1})_k \log(\mathbf{x}\circ\mathbf{x}^{*-1})_k + \right. \\
&\qquad \left. + \log(D) + \sum_{k=1}^{D}(\mathbf{x}^*\circ\mathbf{x}^{-1})_k \log(\mathbf{x}^*\circ\mathbf{x}^{-1})_k\right) \\
&= \frac{D}{2}\left(\sum_{k=1}^{D}(\mathbf{x}\circ\mathbf{x}^{*-1})_k \log\left(\frac{(\mathbf{x}\circ\mathbf{x}^{*-1})_k}{1/D}\right) + \sum_{k=1}^{D}(\mathbf{x}^*\circ\mathbf{x}^{-1})_k \log\left(\frac{(\mathbf{x}\circ\mathbf{x}^{*-1})_k}{1/D}\right)\right) \\
&= \frac{D}{2}\Big(\mathcal{I}(\mathbf{x}\circ\mathbf{x}^{*-1},\mathbf{e}) + \mathcal{I}(\mathbf{x}^*\circ\mathbf{x}^{-1},\mathbf{e})\Big),
\end{aligned} \tag{3.29}$$

donde $\mathcal{I}$ representa la divergencia denominada medida de información de Kullback-Leibler presentada en (3.3).

Nótese que hemos expresado $d_{\mathcal{H}}$ en términos de las divergencias entre el elemento $\mathbf{e}$ (neutro de la operación perturbación) y las perturbaciones diferencia $\mathbf{x} \circ \mathbf{x}^{*-1}$ y $\mathbf{x}^* \circ \mathbf{x}^{-1}$. La distancia de Aitchison puede definirse mediante una expresión análoga –véase (3.21). Se obtiene que

$$\begin{aligned} d^2_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) &= \frac{1}{2}[d^2_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) + d^2_{\text{Ait}}(\mathbf{x}^*, \mathbf{x})] \\ &= \frac{1}{2}[d^2_{\text{Ait}}(\mathbf{e}, \mathbf{x} \circ \mathbf{x}^{*-1}) + d^2_{\text{Ait}}(\mathbf{e}, \mathbf{x}^* \circ \mathbf{x}^{-1})] \\ &= \frac{1}{2}[d^2_{\text{Euc}}(\text{clr}(\mathbf{e}), \text{clr}(\mathbf{x} \circ \mathbf{x}^{*-1})) + d^2_{\text{Euc}}(\text{clr}(\mathbf{e}), \text{clr}(\mathbf{x}^* \circ \mathbf{x}^{-1}))] \\ &= \frac{1}{2}[d^2_{\text{Euc}}(\mathbf{0}, \text{clr}(\mathbf{x}) - \text{clr}(\mathbf{x}^*)) + d^2_{\text{Euc}}(\mathbf{0}, \text{clr}(\mathbf{x}^*) - \text{clr}(\mathbf{x}))]. \end{aligned} \tag{3.30}$$

La constatación de este hecho nos abre toda una nueva vía de desarrollo de medidas de diferencia basadas en las divergencias. Llamaremos a estas nuevas medidas *divergencias composicionales*. En Martín-Fernández et al. (1998c) se encuentra definida una divergencia composicional (al cuadrado), que simbolizamos por $d_{\mathcal{KL}}$, según la siguiente expresión:

$$d^2_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) = \frac{D}{2}\left(\mathcal{I}(\mathbf{e}, \mathbf{x} \circ \mathbf{x}^{*-1}) + \mathcal{I}(\mathbf{e}, \mathbf{x}^* \circ \mathbf{x}^{-1})\right). \tag{3.31}$$

Nótese que la diferencia entre las expresiones (3.29) y (3.31) es únicamente el orden en que se establece la divergencia entre $\mathbf{e}$ y las perturbaciones diferencia. Por lo tanto, es inmediato comprobar que la medida de diferencia $d_{\mathcal{KL}}$ tiene todas las buenas propiedades de la medida $d_{\mathcal{H}}$, es decir es una disimilitud definida que es invariante por permutaciones y por perturbaciones. Además de estas propiedades, una medida de diferencia para datos composicionales debe satisfacer el requisito de la dominación por subcomposiciones. Previamente a la demostración de este requerimiento, presentamos una propiedad que nos será de gran utilidad.

Propiedad 3.2 Sean dos composiciones $\mathbf{x}$, $\mathbf{x}^* \in \mathbf{S}^D$, entonces se cumple que

$$d^2_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) = \frac{D}{2} \log\left(A(\frac{\mathbf{x}}{\mathbf{x}^*}) \cdot A(\frac{\mathbf{x}^*}{\mathbf{x}})\right), \tag{3.32}$$

donde $A(\mathbf{x}/\mathbf{x}^*)$ representa la media aritmética del vector de ratios $\mathbf{x}/\mathbf{x}^*$.

Demostración

De la expresión (3.31) se obtiene que

$$d^2_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) = \frac{D}{2}\left[\sum_{k=1}^{D} \frac{1}{D} \log\left(\frac{1/D}{(\mathbf{x} \circ \mathbf{x}^{*-1})_k}\right) + \sum_{k=1}^{D} \frac{1}{D} \log\left(\frac{1/D}{(\mathbf{x}^* \circ \mathbf{x}^{-1})_k}\right)\right],$$

donde

$$(\mathbf{x} \circ \mathbf{x}^{*-1})_k = \frac{\frac{x_k}{x_k^*}}{\sum_{c=1}^{D} \frac{x_c}{x_c^*}} \quad \text{y} \quad (\mathbf{x}^* \circ \mathbf{x}^{-1})_k = \frac{\frac{x_k^*}{x_k}}{\sum_{c=1}^{D} \frac{x_c^*}{x_c}}.$$

Entonces, si formamos un único sumatorio, aplicamos las propiedades de la función log y realizamos unas simples manipulaciones algebraicas obtenemos el resultado

$$d^2_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) = \frac{D}{2} \log \left(\frac{\sum_{k=1}^{D} x_k/x_k^*}{D} \frac{\sum_{k=1}^{D} x_k^*/x_k}{D} \right) = \frac{D}{2} \log \left(A\left(\frac{\mathbf{x}}{\mathbf{x}^*}\right) \; A\left(\frac{\mathbf{x}^*}{\mathbf{x}}\right) \right),$$

con lo que se demuestra la propiedad. □

Nótese que la formulación (3.32) de la disimilitud $d_{\mathcal{KL}}$ es más sencilla que la de su propia definición. Esta nueva expresión de la disimilitud nos simplifica la demostración del requisito de dominación por subcomposiciones.

Propiedad 3.3 La disimilitud $d_{\mathcal{KL}}$ definida en (3.31) es dominante por subcomposiciones, es decir,

$$d_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) \geq d_{\mathcal{KL}}(\mathbf{x}_s, \mathbf{x}_s^*), \quad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D,$$

para toda subcomposición de **s** componentes que se considere.

Demostración

El requerimiento de invariación por permutaciones nos asegura que podemos suponer, sin pérdida de generalidad, que las subcomposiciones de $\mathcal{S}^s$ que consideremos son el resultado de prescindir de las $D - s$ últimas componentes de las composiciones de $\mathcal{S}^D$. Por otro lado, las propiedades de las subcomposiciones nos hacen ver que para demostrar la propiedad es suficiente probar que la dominación se verifica cuando consideramos las subcomposiciones resultado de prescindir únicamente del último componente. En consecuencia, es suficiente demostrar que se satisface la siguiente desigualdad

$$d_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) \geq d_{\mathcal{KL}}(\mathbf{x}_{D-1}, \mathbf{x}_{D-1}^*) \quad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D,$$

donde $\mathbf{x}_{D-1}$, $\mathbf{x}_{D-1}^*$ son las correspondientes subcomposiciones que se obtienen al prescindir de la D-ésima componente:

$$\mathbf{x}_{D-1} = \left(\frac{x_1}{\sum_{k=1}^{D-1} x_k}, \frac{x_2}{\sum_{k=1}^{D-1} x_k}, \ldots, \frac{x_{D-1}}{\sum_{k=1}^{D-1} x_k} \right) \text{ y } \mathbf{x}_{D-1}^* = \left(\frac{x_1^*}{\sum_{k=1}^{D-1} x_k^*}, \frac{x_2^*}{\sum_{k=1}^{D-1} x_k^*}, \ldots, \frac{x_{D-1}^*}{\sum_{k=1}^{D-1} x_k^*} \right).$$

Demostrar la desigualdad anterior utilizando la expresión (3.32) equivale a probar que

$$D \cdot \log \left[A(\mathbf{x}/\mathbf{x}^*) \cdot A(\mathbf{x}^*/\mathbf{x}) \right] \geq (D-1) \cdot \log \left[A(\mathbf{x}_{D-1}/\mathbf{x}_{D-1}^*) \cdot A(\mathbf{x}_{D-1}^*/\mathbf{x}_{D-1}) \right], \tag{3.33}$$

donde $A(\mathbf{x}/\mathbf{x}^*)$ representa la medida aritmética del vector de cocientes $\mathbf{x}/\mathbf{x}^*$.

Veamos primero que

$$A(\mathbf{x}_{D-1}/\mathbf{x}^*_{D-1}) \cdot A(\mathbf{x}^*_{D-1}/\mathbf{x}_{D-1}) = A\left(\frac{x_1}{x_1^*}, \frac{x_2}{x_2^*}, \ldots, \frac{x_{D-1}}{x_{D-1}^*}\right) \cdot A\left(\frac{x_1^*}{x_1}, \frac{x_2^*}{x_2}, \ldots, \frac{x_{D-1}^*}{x_{D-1}}\right). \quad (3.34)$$

En efecto,

$$\begin{aligned} A\left(\frac{\mathbf{x}_{D-1}}{\mathbf{x}^*_{D-1}}\right) \cdot A\left(\frac{\mathbf{x}^*_{D-1}}{\mathbf{x}_{D-1}}\right) &= A\left(\frac{\sum_{k=1}^{D-1} x_k^*}{\sum_{i=1}^{D-1} x_k} \left(\frac{x_1}{x_1^*}, \frac{x_2}{x_2^*}, \ldots, \frac{x_{D-1}}{x_{D-1}^*}\right)\right) \cdot \\ &\quad \cdot A\left(\frac{\sum_{k=1}^{D-1} x_k}{\sum_{i=1}^{D-1} x_k^*} \left(\frac{x_1^*}{x_1}, \frac{x_2^*}{x_2}, \ldots, \frac{x_{D-1}^*}{x_{D-1}}\right)\right) \\ &= A\left(\frac{x_1}{x_1^*}, \frac{x_2}{x_2^*}, \ldots, \frac{x_{D-1}}{x_{D-1}^*}\right) \cdot A\left(\frac{x_1^*}{x_1}, \frac{x_2^*}{x_2}, \ldots, \frac{x_{D-1}^*}{x_{D-1}}\right), \end{aligned}$$

que es la igualdad (3.34) que se deseaba probar.

El resultado que acabamos de obtener implica que la desigualdad (3.33) que debemos demostrar es equivalente a la desigualdad

$$D \cdot \log\left(\frac{\sum_{k=1}^{D} \frac{x_k}{x_k^*}}{D} \cdot \frac{\sum_{k=1}^{D} \frac{x_k^*}{x_k}}{D}\right) \geq (D-1) \cdot \log\left(\frac{\sum_{k=1}^{D-1} \frac{x_k}{x_k^*}}{D-1} \cdot \frac{\sum_{k=1}^{D-1} \frac{x_k^*}{x_k}}{D-1}\right). \quad (3.35)$$

La demostración de esta desigualdad se consigue operando su primer miembro y aplicando el hecho que la función logarítmo es una función cóncava:

$$\begin{aligned} D \cdot \log\left(\frac{\sum_{k=1}^{D} \frac{x_k}{x_k^*}}{D} \cdot \frac{\sum_{k=1}^{D} \frac{x_k^*}{x_k}}{D}\right) &= D \cdot \log\left(\frac{\sum_{k=1}^{D} \frac{x_k}{x_k^*}}{D}\right) + D \cdot \log\left(\frac{\sum_{k=1}^{D} \frac{x_k^*}{x_k}}{D}\right) \\ &= D \cdot \log\left(\frac{D-1}{D}\left(\frac{\sum_{k=1}^{D-1} \frac{x_k}{x_k^*}}{D-1}\right) + \frac{1}{D}\left(\frac{x_D}{x_D^*}\right)\right) \\ &\quad + D \cdot \log\left(\frac{D-1}{D}\left(\frac{\sum_{k=1}^{D-1} \frac{x_k^*}{x_k}}{D-1}\right) + \frac{1}{D}\left(\frac{x_D^*}{x_D}\right)\right) \\ &\geq D \cdot \left[\frac{D-1}{D} \log\left(\frac{\sum_{k=1}^{D-1} \frac{x_k}{x_k^*}}{D-1}\right) + \frac{1}{D} \log\left(\frac{x_D}{x_D^*}\right)\right] \\ &\quad + D \cdot \left[\frac{D-1}{D} \log\left(\frac{\sum_{k=1}^{D-1} \frac{x_k^*}{x_k}}{D-1}\right) + \frac{1}{D} \log\left(\frac{x_D^*}{x_D}\right)\right]. \end{aligned}$$

Nótese que los términos $\frac{1}{D}\log\left(\frac{x_D}{x_D^*}\right)$ y $\frac{1}{D}\log\left(\frac{x_D^*}{x_D}\right)$ se cancelan y que el factor D se simplifica con los denominadores D. Entonces obtenemos la expresión siguiente:

$$\begin{aligned} D \cdot \log\left(\frac{\sum_{k=1}^{D} \frac{x_k}{x_k^*}}{D} \cdot \frac{\sum_{k=1}^{D} \frac{x_k^*}{x_k}}{D}\right) &\geq (D-1) \cdot \left[\log\left(\frac{\sum_{k=1}^{D-1} \frac{x_k}{x_k^*}}{D-1}\right) + \log\left(\frac{\sum_{k=1}^{D-1} \frac{x_k^*}{x_k}}{D-1}\right)\right] \\ &= (D-1) \cdot \log\left(\frac{\sum_{k=1}^{D-1} \frac{x_k}{x_k^*}}{D-1} \cdot \frac{\sum_{k=1}^{D-1} \frac{x_k^*}{x_k}}{D-1}\right), \end{aligned}$$

con lo que hemos demostrado que la disimilitud propuesta cumple el requisito de la dominación por subcomposiciones. □

El hecho que la disimilitud $d_{\mathcal{KL}}$ propuesta en (3.31) satisfaga tres de los requisitos para una medida de diferencia entre datos composicionales convierte a esta medida en una disimilitud compatible con la naturaleza composicional de los datos. Sin embargo, la plenitud de esta compatibilidad no se consigue puesto que la medida $d_{\mathcal{KL}}$ no es coherente respecto el producto por un escalar –véase la Propiedad 2.8. Para ilustrarlo podemos considerar los datos que hemos utilizado en la figura 2.12. Sean las composiciones $\mathbf{x} = (0.5, 0.15, 0.35)$, $\mathbf{x}^* = (0.5, 0.35, 0.15)$ y sus correspondientes $2 \cdot \mathbf{x}$, $2 \cdot \mathbf{x}^*$. Entonces, unos sencillos cálculos nos proporcionan que $2d_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) = 1.648$ y que $d_{\mathcal{KL}}(2 \cdot \mathbf{x}, 2 \cdot \mathbf{x}^*) = 1.542$.

En la figura 3.1 podemos observar la forma de los entornos en el diagrama ternario que se obtienen con la disimilitud $d_{\mathcal{KL}}$. A pesar de que la medida $d_{\mathcal{KL}}$ no es plenamente compatible con la estructura de espacio vectorial definida en el símplex, estas formas nos indican que la disimilitud tiene un comportamiento coherente con la naturaleza de los datos composicionales. Nótese que los entornos cuyo centro está situado cerca de los lados o vértices del símplex son entornos que aparecen más apretados entre si.

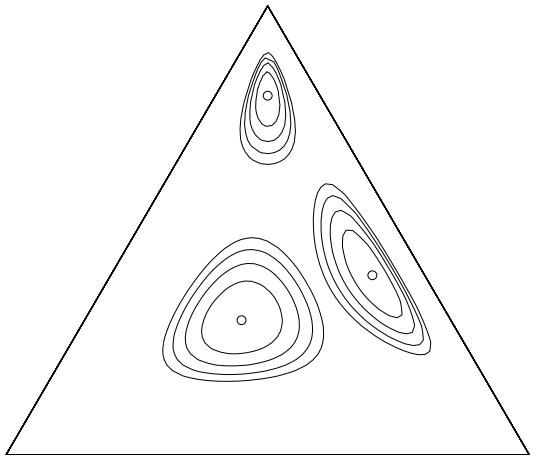


Figura 3.1: *Entornos en $\mathcal{S}^3$ con la disimilitud $d_{\mathcal{KL}}$.*

Queremos resaltar la extrema semejanza entre los entornos que se obtienen para la distancia de Aitchison –véase la figura 2.13– y los entornos de la figura 3.1. Este hecho induce a pensar que existe una estrecha conexión entre ambas medidas de diferencia. En la propiedad siguiente se establece la expresión que relaciona la distancia de Aitchison d_{Ait} con la disimilitud $d_{\mathcal{KL}}$.

Propiedad 3.4 Sean dos composiciones $\mathbf{x}$, $\mathbf{x}^* \in \mathbf{S}^D$. Entonces se cumple que

$$d^2_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) \approx 2d^2_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*), \tag{3.36}$$

o, de manera equivalente,

$$d_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) \approx \sqrt{2}\, d_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*). \tag{3.37}$$

Demostración

De la expresión (3.32) que aparece en la Propiedad 3.2 se tiene que

$$d^2_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) = \frac{D}{2} \log\left(A\left(\frac{\mathbf{x}}{\mathbf{x}^*}\right) A\left(\frac{\mathbf{x}^*}{\mathbf{x}}\right)\right),$$

donde $A(\mathbf{x}/\mathbf{x}^*)$ representa la media aritmética del vector de ratios $\mathbf{x}/\mathbf{x}^*$. Las siguientes manipulaciones algebraicas nos conducen a obtener una expresión de la disimilitud $d_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*)$ en términos de la covarianza muestral de los dos vectores de ratios $\widehat{cov}\left(\frac{\mathbf{x}}{\mathbf{x}^*}, \frac{\mathbf{x}^*}{\mathbf{x}}\right)$:

$$\begin{aligned} d^2_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) &= \frac{D}{2} \log\left(A\left(\frac{\mathbf{x}}{\mathbf{x}^*}\right) A\left(\frac{\mathbf{x}^*}{\mathbf{x}}\right)\right) \\ &= \frac{D}{2} \log\left(A\left(\frac{\mathbf{x}}{\mathbf{x}^*}\right) A\left(\frac{\mathbf{x}^*}{\mathbf{x}}\right) - 1 + 1\right) \\ &= \frac{D}{2} \log\left(-\widehat{cov}\left(\frac{\mathbf{x}}{\mathbf{x}^*}, \frac{\mathbf{x}^*}{\mathbf{x}}\right) + 1\right). \end{aligned}$$

Recordemos que el desarrollo de Taylor de primer orden de la función $\log(1-x)$ para valores de x cercanos a cero puede expresarse como $\log(1-x) = -x + \theta(\epsilon)$, donde el término $\theta(\epsilon)$ representa el error de segundo orden en la aproximación. Es decir, se cumple que $\lim_{\epsilon \to 0} \frac{\theta(\epsilon)}{\epsilon} = 0$.

Aplicando esta aproximación a nuestra última expresión se obtiene que

$$\begin{aligned} d^2_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) &= \frac{D}{2} \log\left(-\widehat{cov}\left(\frac{\mathbf{x}}{\mathbf{x}^*}, \frac{\mathbf{x}^*}{\mathbf{x}}\right) + 1\right) \\ &\approx -\frac{D}{2}\, \widehat{cov}\left(\frac{\mathbf{x}}{\mathbf{x}^*}, \frac{\mathbf{x}^*}{\mathbf{x}}\right). \end{aligned} \tag{3.38}$$

Por otra parte, recordando la definición de la distancia de Aitchison y realizando unas manipulaciones algebraicas se obtiene que la distancia d_{Ait} puede expresarse en términos de la covarianza muestral de los vectores ratios. En nuestro desarrollo aparece el término $\widehat{var}$ que utilizamos para identificar la varianza muestral.

$$\begin{aligned} d^2_{\text{Ait}}(\mathbf{x}, \mathbf{x}^*) &= \sum_{k=1}^{D} \left(\log\left(\frac{x_k}{g(\mathbf{x})}\right) - \left(\frac{x^*_k}{g(\mathbf{x}^*)}\right)\right)^2 \\ &= \frac{1}{2}\left[\sum_{k=1}^{D} \left(\log\left(\frac{x_k/x^*_k}{g(\mathbf{x}/\mathbf{x}^*)}\right)\right)^2 + \sum_{k=1}^{D} \left(\log\left(\frac{x^*_k/x_k}{g(\mathbf{x}^*/\mathbf{x})}\right)\right)^2\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\left[\sum_{k=1}^{D}\left(\log\left(\frac{(\mathbf{x}\circ\mathbf{x}^{*-1})_k}{g(\mathbf{x}\circ\mathbf{x}^{*-1})}\right)\right)^2+\sum_{k=1}^{D}\left(\log\left(\frac{(\mathbf{x}^*\circ\mathbf{x}^{-1})_k}{g(\mathbf{x}^*\circ\mathbf{x}^{-1})}\right)\right)^2\right]\\
&= \frac{D}{2}\left[\frac{1}{D}\sum_{k=1}^{D}\left(\log((\mathbf{x}\circ\mathbf{x}^{*-1})_k)-\frac{1}{D}\sum_{l}\log((\mathbf{x}\circ\mathbf{x}^{*-1})_l)\right)^2\right]+\\
&\quad+\frac{D}{2}\left[\frac{1}{D}\sum_{k=1}^{D}\left(\log((\mathbf{x}^*\circ\mathbf{x}^{-1})_k)-\frac{1}{D}\sum_{l}\log((\mathbf{x}^*\circ\mathbf{x}^{-1})_l)\right)^2\right]\\
&= \frac{D}{2}\left[\widehat{var}(\log(\mathbf{x}\circ\mathbf{x}^{*-1}))+\widehat{var}(\log(\mathbf{x}^*\circ\mathbf{x}^{-1}))\right]\\
&= \frac{D}{2}\left[\widehat{var}\left(\log(\mathbf{x}\circ\mathbf{x}^{*-1})+\log(\mathbf{x}^*\circ\mathbf{x}^{-1})\right)-2\widehat{cov}(\log(\mathbf{x}\circ\mathbf{x}^{*-1}),\log(\mathbf{x}^*\circ\mathbf{x}^{-1}))\right]\\
&= \frac{D}{2}\left[\widehat{var}\left(\log((\mathbf{x}\circ\mathbf{x}^{*-1}).(\mathbf{x}^*\circ\mathbf{x}^{-1}))\right)-2\widehat{cov}\Big(\log\left(\frac{\mathbf{x}}{\mathbf{x}^*}\right),\log\left(\frac{\mathbf{x}^*}{\mathbf{x}}\right)\Big)\right]\\
&= -D\widehat{cov}\left(\log\left(\frac{\mathbf{x}}{\mathbf{x}^*}\right),\log\left(\frac{\mathbf{x}^*}{\mathbf{x}}\right)\right).
\end{aligned}
$$

Recordemos que el desarrollo de Taylor de primer orden de la función $\log(x)$ para valores de x cercanos a uno puede expresarse como $\log(x) = x-1+\theta(\delta)$, donde el término $\theta(\delta)$ representa el error de segundo orden en la aproximación. Aplicando esta aproximación a nuestra última expresión se obtiene que

$$
\begin{aligned}
d^2_{\mathrm{Ait}}(\mathbf{x},\mathbf{x}^*) &= -D\widehat{cov}\left(\log\left(\frac{\mathbf{x}}{\mathbf{x}^*}\right),\log\left(\frac{\mathbf{x}^*}{\mathbf{x}}\right)\right)\\
&\approx -D\widehat{cov}\left(\frac{\mathbf{x}}{\mathbf{x}^*}-1,\frac{\mathbf{x}^*}{\mathbf{x}}-1\right)\\
&= -D\widehat{cov}\left(\frac{\mathbf{x}}{\mathbf{x}^*},\frac{\mathbf{x}^*}{\mathbf{x}}\right), \qquad (3.39)
\end{aligned}
$$

con lo que se demuestra la propiedad.

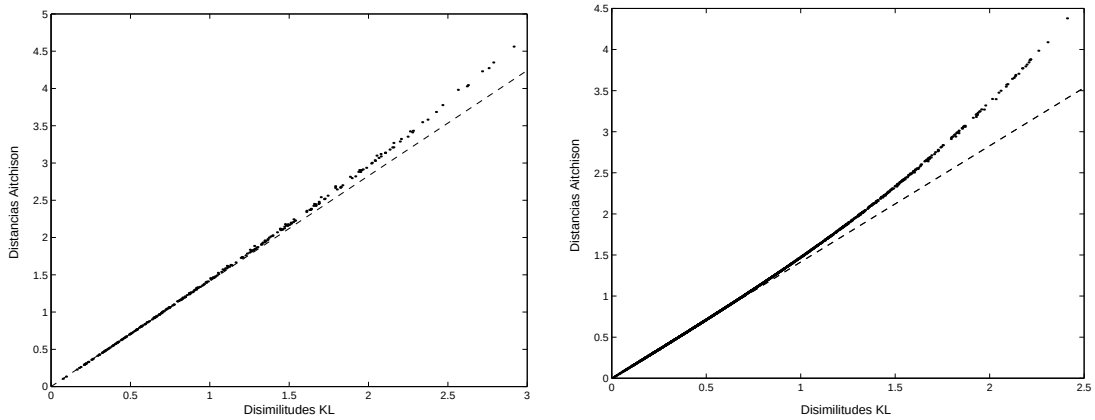
□

Es importante resaltar que, cuando las dos composiciones $\mathbf{x},\mathbf{x}^*$ son *cercanas* puede considerarse que los vectores ratios $\frac{\mathbf{x}}{\mathbf{x}^*}$ y $\frac{\mathbf{x}^*}{\mathbf{x}}$ son ambos cercanos al vector unidad $(1,1,\ldots,1)\in\mathbb{R}^D$. En este caso, los errores en las aproximaciones expresadas en (3.38) y (3.39) serán razonablemente pequeños. Por lo tanto, a medida que las composiciones consideradas $\mathbf{x},\mathbf{x}^*$ sean más *distantes* el error en la aproximación $d_{\mathrm{Ait}}(\mathbf{x},\mathbf{x}^*)\approx\sqrt{2}\,d_{\mathcal{KL}}(\mathbf{x},\mathbf{x}^*)$ empeorará.

Para ilustrar la relación (3.36) entre la distancia de Aitchison y la disimilitud $d_{\mathcal{KL}}$ presentamos el siguiente ejemplo basado en los conjuntos de datos Hongite y Metabol. Estos conjuntos ya han sido utilizados en diversos ejemplos en el Capítulo 2 de esta tesis.

Ejemplo 3.1 Consideramos los conjuntos de datos Hongite y Metabol. Para cada conjunto por separado, calculamos la distancia de Aitchison para cualquier par de observaciones distintas del

conjunto y llamamos $\mathbf{v}_{\mathrm{Ait}}$ al vector que contiene estas distancias. Repetimos el cálculo, para cada uno de los dos conjuntos de datos, usando ahora la medida $d_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*)$, obteniendo el vector $\mathbf{v}_{\mathcal{KL}}$. La figura 3.2(a) muestra la nube de puntos que se obtiene al representar el diagrama de dispersión de los dos vectores para el caso del conjunto de datos Hongite. En el eje de abscisas se han situado los valores del vector $\mathbf{v}_{\mathcal{KL}}$ y en el eje de ordenadas los valores del vector $\mathbf{v}_{\mathrm{Ait}}$. En la misma figura y con trazado discontinuo se ha representado la recta $y = \sqrt{2}\, x$, con el propósito de facilitar la interpretación del diagrama. En la figura 3.2(b) se muestra la nube de puntos que se obtiene cuando se trata del conjunto de datos Metabol. Observemos que, para valores de $d_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*) \leq 1$ se percibe una relación muy estrecha entre las dos medidas de diferencia y que la relación es del tipo $y = \sqrt{2}\, x$.



(a) (b)

Figura 3.2: Diagrama de dispersión de las medidas $d_{\mathcal{KL}}$ y d_{Ait} para los conjuntos de datos: (a) *Hongite;* (b) *Metabol. La línea de trazo discontínuo representa la recta* $y = \sqrt{2}\, x$.

Recordemos que en la Sección 1.6.2 del Capítulo 1 de esta tesis se ha expuesto que la mayoría de los métodos de clasificación automática jerárquicos son invariantes por transformaciones monótonas de la matriz de distancias. El hecho que la relación entre la distancia de Aitchison y la medida $d_{\mathcal{KL}}$ consista en una transformación *aproximadamente* monótona justifica que las clasificaciones obtenidas al aplicar estos tipos de métodos con estas dos medidas sean clasificaciones extremadamente coincidentes. En el capítulo siguiente mostramos algunos ejemplos de esta coincidencia de resultados.

Por otra parte, podemos observar en los gráficos de la figura 3.2 que la relación $d_{\mathrm{Ait}}(\mathbf{x}, \mathbf{x}^*) \approx \sqrt{2}\, d_{\mathcal{KL}}(\mathbf{x}, \mathbf{x}^*)$ implica que la distancia de Aitchison siempre es mayor que la disimilitud $d_{\mathcal{KL}}$ y que este hecho se acentúa para los valores grandes de la distancia. Es importante tener en cuenta esta propiedad cuando en el conjunto de datos a estudiar existan individuos anormalmente

alejados del resto de observaciones (en inglés, *outliers*). Si se usan criterios de detección de *outliers* basados en medidas de diferencia, entonces resultará que la distancia de Aitchison puede ser de mayor utilidad que la medida $d_{\mathcal{KL}}$. Por el contrario, cuando al realizar una clasificación automática no paramétrica utilicemos una técnica de agrupación sensible a la presencia de *outliers*, como por ejemplo el método del máximo, la elección de la disimilitud $d_{\mathcal{KL}}$ producirá una menor distorsión de los resultados finales debido a la presencia de este tipo de observaciones.

Capítulo 4

Clasificación automática no paramétrica de datos composicionales

4.1 Introducción

En el Prólogo de esta tesis hemos constatado la inexistencia de un cuerpo teórico y metodológico apropiado que permita desarrollar pautas y recomendaciones a seguir en el momento de realizar una clasificación no paramétrica de datos composicionales. En este capítulo presentamos la metodología a aplicar en la realización de una clasificación automática de datos composicionales. Esta metodología se fundamenta en aspectos de carácter general expuestos en el Capítulo 1, *Introducción a la clasificación automática no paramétrica*, y en el cuerpo teórico presentado en el Capítulo 2, *Datos composicionales*.

Préviamanete a esta tesis ha habido otros trabajos donde el núcleo de la investigación ha sido la clasificación de datos composicionales. Entre los trabajos cuyo desarrollo consiste en un estudio de un conjunto de datos concreto desde un enfoque predominantemente aplicado, queremos resaltar los trabajos de Zhou et al. (1991), Davis et al. (1995), Bohling et al. (1996), Martín-Fernández et al. (1997), Zhou (1997) y Bren y Martín-Fernández (1999). Entre los estudios donde se abordan conceptos teóricos relacionados con la clasificación de datos composicionales destacamos los trabajos de Martín (1996), Pawlowsky et al. (1997), Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (1998a, 1998b, 1999) y Tauber (1999).

4.2 Metodología propuesta

En el Capítulo 1 de esta tesis se ha expuesto a grandes rasgos la metodología a aplicar en la realización de una clasificación automática no paramétrica. En su desarrollo se ha insistido en la gran importancia que tienen la fase de elección de la medida de disimilitud y la fase de elección del método de clasificación. Por lo que se refiere a la elección de la medida de disimilitud, la idea clave a tener en cuenta es que una disimilitud puede ser adecuada o no dependiendo de la tipología de los datos a clasificar. No existe una disimilitud adecuada para todos los tipos de datos, y, en general, para cualquier tipo de dato puede encontrarse más de una medida de disimilitud que sea adecuada. En la fase de elección del método de clasificación debe decidirse en primer lugar qué tipo de técnica no paramétrica vamos a utilizar: jerárquica o no jerárquica. La decisión se toma fundamentalmente en base a si se conoce o no el número de grupos a construir. De nuestra experiencia en la realización de clasificaciones, se desprende que en la gran mayoría de los estudios se desconoce a priori el número de grupos a considerar, y en consecuencia, las técnicas más utilizadas son las jerárquicas.

En la figura 4.1 presentamos de manera esquemática las fases a seguir en la realización de una clasificación automática no paramétrica de datos composicionales mediante un método jerárquico. Si el método de clasificación no fuese jerárquico el esquema seguiría siendo válido suprimiendo la etapa intermedia de elección del número de grupos a considerar. Como puede apreciarse, este esquema no es únicamente válido para datos de tipo composicional. Las particularidades a tener en cuenta para el caso de datos composicionales las exponemos en las secciones siguientes donde desarrollamos cada una de las etapas del esquema propuesto. En la figura 4.1 se observa que la realización de una clasificación se basa en un proceso de naturaleza inductiva-deductiva. La naturaleza de este proceso es común a la gran mayoría de técnicas estadísticas y está en el fundamento del propio método estadístico. En la realización de una clasificación, la etapa de diagnosis o crítica de resultados consiste en analizar si la agrupación obtenida puede considerarse razonable. En este contexto, entendemos que una clasificación razonable es aquella agrupación de los datos en la que observaciones que pertenezcan a grupos diferentes muestren un patrón claramente diferenciado en el valor que toman en las diferentes variables. Este patrón diferenciador de los grupos obtenidos deberá ser interpretable en relación al contexto o población de la que haya sido extraído el conjunto de los datos. Si la clasificación no se considera razonable el proceso iterativo-deductivo contempla la posibilidad de modificar la elección de la medida de disimilitud, la elección del método de clasificación o, en su caso, la elección del número de grupos a considerar.

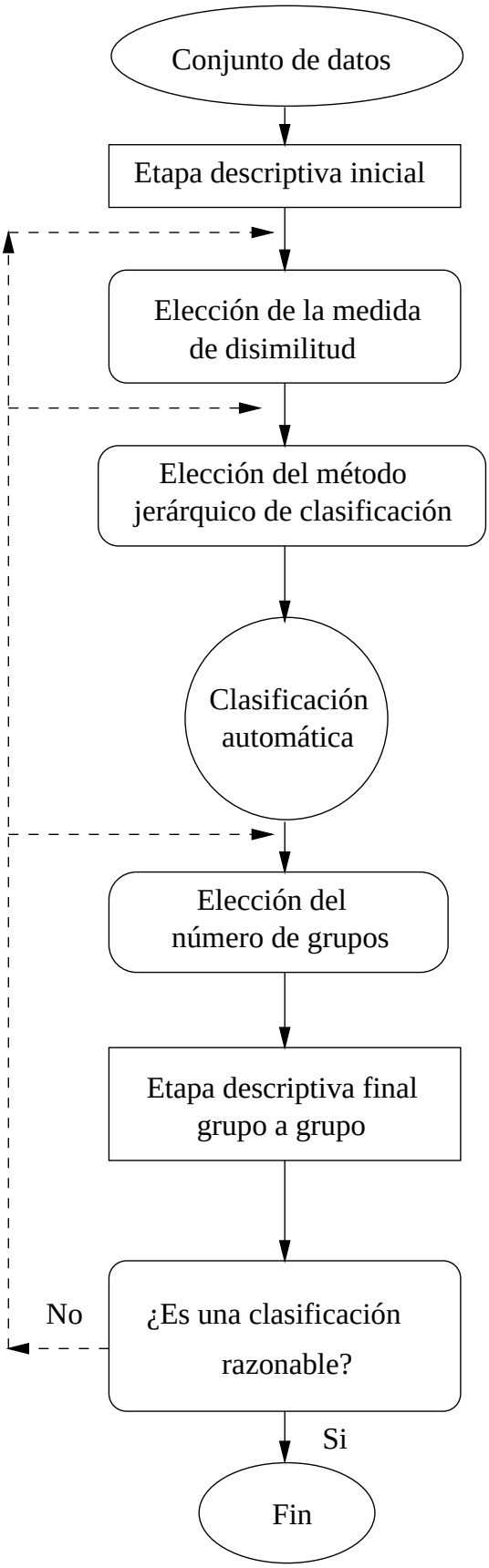


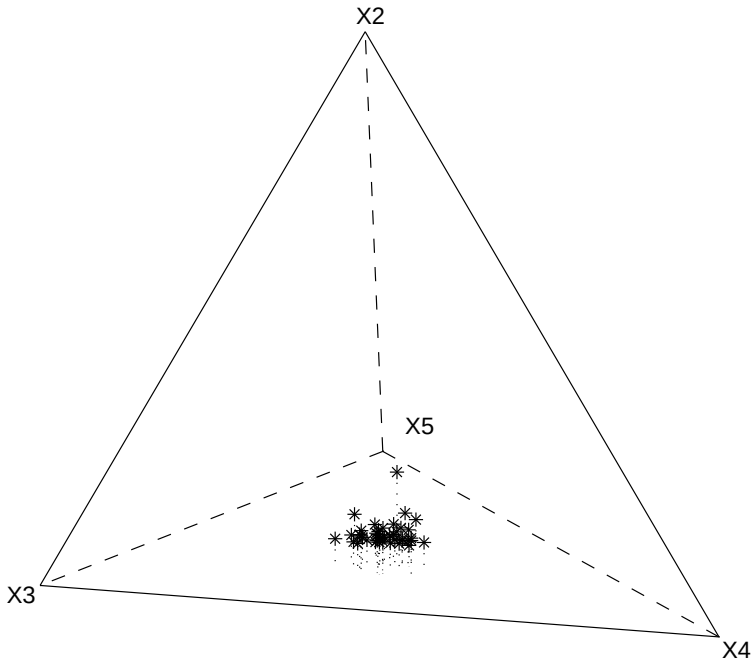
Figura 4.1: Esquema de las fases a seguir en la realización de una clasificación automática no paramétrica de datos composicionales.

4.2.1 Etapa descriptiva inicial

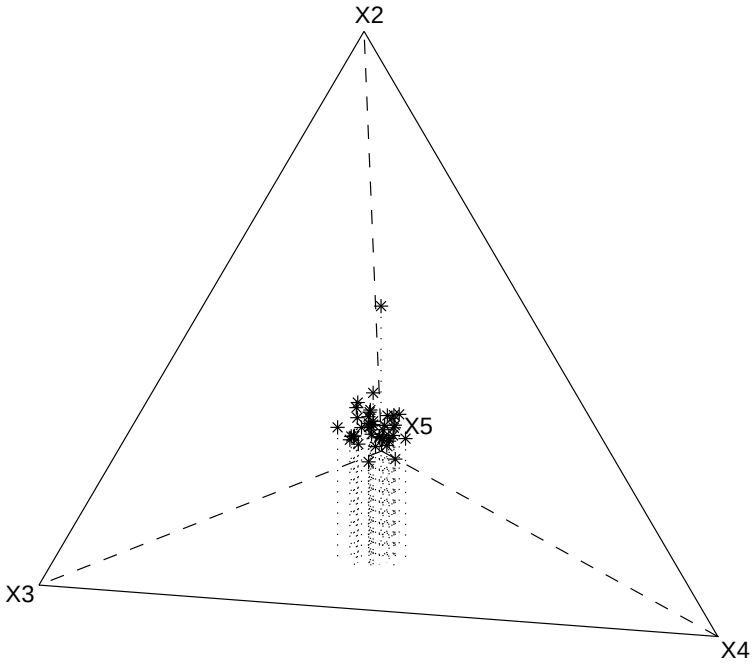
Esta primera fase consiste en una etapa donde se realiza un primer acercamiento al conjunto de datos. En esta fase se verifica que en el conjunto no existen observaciones que toman el valor cero en alguna componente. En el caso de detectar observaciones con valores nulos se deberá aplicar, previamente a la clasificación, la metodología que se propone en el Capítulo 5 de esta tesis. En esta etapa descriptiva inicial se calculan los estadísticos básicos expuestos en el Capítulo 2: la media geométrica composicional y la variabilidad total del conjunto de datos. Estos estadísticos son de gran utilidad cuando se usan conjuntamente con otras herramientas descriptivas, como por ejemplo los diagramas de caja. Los diagramas de caja de cada una de las componentes pueden permitir detectar componentes con un rango de valores muy grande que nos sugiera la existencia de heterogeneidad en el conjunto de datos. Sin embargo, son las técnicas que permiten representar gráficamente el propio conjunto de datos las que con mayor claridad nos aportan evidencias de que en el conjunto de datos existe heterogeneidad.

Los conjuntos de datos del símplex $\mathcal{S}^3$ pueden representarse en el diagrama ternario –véase la figura 2.22 del Capítulo 2. En el caso de conjuntos de datos del símplex $\mathcal{S}^4$ puede utilizarse la representación en el diagrama cuaternario. Es importante resaltar que en estos diagramas los conjuntos de datos que contienen observaciones con valores muy próximos a cero aparecen situados cerca de una cara, de una arista o de un vértice. En consecuencia, la visualización de la estructura del conjunto de datos es muy pobre. En estas situaciones (Martín-Fernández et al., 1999) debemos aplicar la transformación denominada *centrado* del conjunto de datos expuesta en el Capítulo 2.

La figura 4.2(a) muestra el conjunto de datos resultante de considerar la subcomposición en las componentes X_2, X_3, X_4, y X_5 de las observaciones del conjunto *Población ocupada*. Este conjunto de datos –véase el Apéndice A.1– será objeto de un estudio más profundo en este capítulo. En la figura puede apreciarse que el hecho que las observaciones tomen en la componente X_2 valores relativamente cercanos a cero produce que el conjunto aparezca situado muy cercano a la cara inferior del diagrama cuaternario. En la figura 4.2(b) se muestra el conjunto de datos que se obtiene al centrar el conjunto de la figura 4.2(a). En esta figura la visualización del conjunto de datos ha mejorado ostensiblemente. Por lo que se refiere a la existencia de heterogeneidad en el conjunto de datos, en la representación de esta subcomposición no se aprecian evidencias de la existencia de grupos en el conjunto de datos *Población ocupada*.



(a)



(b)

Figura 4.2: Dos perspectivas diferentes en $\mathcal{S}^4$ del conjunto resultante de considerar la subcomposición en las componentes X_2, X_3, X_4, y X_5 de las observaciones del conjunto *Población ocupada*: (a) *Sin centrar;* (b) *Conjunto centrado.*

Los conjuntos de datos pertenecientes a espacios símplex de mayor dimensión, $\mathcal{S}^D$, $D > 4$, deben representarse utilizando otras estrategias. Una primera estrategia consiste en representar los conjuntos de datos resultado de considerar todas sus subcomposiciones posibles. Si se utiliza el diagrama ternario será necesario representar tantas subcomposiciones como indica el número combinatorio C_3^D. Por el contrario, si se utiliza el diagrama cuaternario se representarán C_4^D diagramas cuaternarios. Cuando se analicen los diagramas de las subcomposiciones debe tenerse en cuenta que el hecho de hallar evidencias de heterogeneidad en una subcomposición nos sugiere la existencia de grupos en el conjunto total. Sin embargo, la afirmación recíproca no es cierta. Es decir, puede suceder que no existan evidencias claras de heterogeneidad en ninguna subcomposición, pero que el conjunto total contenga grupos naturales de observaciones. Una segunda estrategia consiste en utilizar herramientas que permiten representar conjuntos de datos multivariantes en el plano. En el trabajo de Aitchison (1997) se desarrolla para datos composicionales la técnica de representación conocida como *biplot*. Fundamentalmente la técnica expuesta en Aitchison (1997) se basa en la representación mediante un diagrama *biplot* para datos en el espacio real multivariante (Gower y Hand, 1996) del conjunto de datos resultante de aplicar la transformación clr al conjunto de datos originales a clasificar.

4.2.2 Elección de la medida de disimilitud

En el Capítulo 1 hemos presentado en detalle la importancia que tiene la medida de disimilitud en el resultado de la clasificación. En ese mismo capítulo hemos expuesto una idea clave: *la medida de disimilitud debe ser coherente con el tipo de datos a clasificar*. Hemos presentado algunas de las medidas más usuales para diferentes tipos de datos. Somos conscientes que el hecho de utilizar una medida coherente con la tipología de los datos no es garantía *sine qua non* para obtener una clasificación razonable. También somos conscientes que una medida de disimilitud incoherente con la tipología de los datos puede proporcionar en ciertos casos una clasificación razonable. Sin embargo, pensamos que este caso será inusual, y que el hecho de realizar una agrupación usando una medida inadecuada nos llevará a obtener resultados erróneos y clasificaciones poco verosímiles. En el Capítulo 2 de esta tesis hemos desarrollado en detalle las propiedades de los datos de tipo composicional. Hemos mostrado (Martín-Fernández et al., 1998a) que las medidas de disimilitud más usuales no son coherentes con la naturaleza composicional de los datos. En ese mismo capítulo hemos presentado la distancia de Aitchison d_{Ait} (al cuadrado):

$$d^2_{\mathrm{Ait}}(\mathbf{x},\mathbf{x}^*) = \sum_{k=1}^{D} (\log(\frac{x_k}{g(\mathbf{x})}) - \log(\frac{x^*_k}{g(\mathbf{x}^*)}))^2, \tag{4.1}$$

como una medida de disimilitud adecuada por ser compatible con las operaciones básicas del símplex. Es sencillo comprender que esta compatibilidad viene inducida por la transformación clr si recordamos que la distancia de Aitchison entre dos datos composicionales es igual a la distancia euclídea entre sus clr transformados: $d_{\mathrm{Ait}}(\mathbf{x},\mathbf{x}^*) = d_{\mathrm{Euc}}(\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{x}^*))$. En el Capítulo 3 de esta tesis hemos desarrollado un estudio de las medidas de divergencia más usuales, y hemos propuesto una medida de disimilitud (Martín-Fernández et al., 1998c) coherente con los datos composicionales: la medida de Kullback-Leibler composicional $d_{\mathcal{KL}}$ (al cuadrado)

$$d^2_{\mathcal{KL}}(\mathbf{x},\mathbf{x}^*) = \frac{D}{2}\log\left(A(\frac{\mathbf{x}}{\mathbf{x}^*})\cdot A(\frac{\mathbf{x}^*}{\mathbf{x}})\right), \tag{4.2}$$

donde $A(\mathbf{x}/\mathbf{x}^*)$ representa la media aritmética del vector de ratios $\mathbf{x}/\mathbf{x}^*$. En ese mismo capítulo se ha mostrado que las dos medidas, d_{Ait} y $d_{\mathcal{KL}}$, están fuertemente relacionadas. En consecuencia, en esta fase de la clasificación, la elección de la medida de disimilitud consistirá en decidir si utilizamos la distancia d_{Ait} o la disimilitud $d_{\mathcal{KL}}$. En el caso práctico que presentamos en este capítulo utilizaremos las dos medidas y analizaremos si existe relación entre los resultados obtenidos en las correspondientes clasificaciones.

4.2.3 Elección del método de clasificación

En el Capítulo 1 de esta tesis hemos expuesto con detalle las características más relevantes de las dos grandes familias que se consideran dentro de los métodos de clasificación: las *técnicas jerárquicas* y las *técnicas no jerárquicas*. Para aplicar alguna de las técnicas no jerárquicas expuestas en el Capítulo 1 se debe conocer a priori el número de grupos a construir en la clasificación. En consecuencia, el conocimiento o no del número de grupos a construir es un factor determinante en el momento de elegir a que familia pertenecerá el método de clasificación que aplicaremos. Sin embargo, tal y como se expone en el Capítulo 1, para el caso de conjuntos de datos con un elevado número de observaciones, existen estrategias *mixtas* que combinan la aplicación de métodos jerárquicos y no jerárquicos, para salvar la dificultad de desconocer a priori el número de grupos a obtener. Dentro de los métodos jerárquicos, los métodos aglomerativos son los que habitualmente se encuentran implementados en los paquetes de *software* estadístico. En el momento de elegir entre un método aglomerativo u otro, debe tenerse en cuenta que un factor muy influyente en la clasificación resultante es la posición relativa de los grupos que existan en el conjunto de datos a clasificar. En el Capítulo 1 se han descrito en detalle las particularidades que distinguen a cada uno de los métodos aglomerativos. A continuación exponemos de una manera muy resumida estas particularidades:

- Método del mínimo: tiene tendencia a construir grupos con forma alargada.

- Método del máximo: tiende a formar clases muy compactas.
- Método de la media: opción intermedia entre los dos métodos anteriores.
- Método del centroide: tiene el defecto de producir inversiones.
- Método de Ward: tiende a crear grupos de forma esférica y con el mismo número de observaciones.

De estas propiedades se deduce otra razón más para considerar la etapa descriptiva inicial como una fase que juega un papel fundamental en la clasificación no paramétrica. Una vez se ha elegido un método de clasificación y se ha obtenido una agrupación, ésta puede considerarse como razonable o no en la etapa de diagnosis. Si la respuesta es negativa, puede retrocederse en el esquema de la clasificación hasta la fase de elección del método de clasificación. Entonces se elegirá otro de los métodos y se reiniciará la clasificación.

Sin pretender dar unas pautas a seguir, creemos importante resaltar que en nuestra experiencia en la realización de clasificaciones hemos constatado que hemos obtenido los resultados más razonables cuando se han utilizado el método de la media y el método de Ward. El hecho de haber obtenido inversiones en las clasificaciones realizadas usando método del centroide nos ha llevado a considerarlo como uno de los métodos más inadecuados. También se ha constatado en nuestros estudios que los métodos del máximo y de la media proporcionan clasificaciones globalmente coincidentes, y que por el contrario, utilizando el método del mínimo, a menudo se obtienen clasificaciones marcadamente diferentes.

4.2.4 Clasificación automática

Una vez se ha elegido el método de clasificación se llevará a cabo la agrupación automática propiamente dicha. En el Capítulo 1 se ha presentado una descripción somera de algunas de las aplicaciones informáticas existentes en el mercado comercial.

El tratamiento específico de los datos composicionales lo realizamos mediante funciones programadas en lenguaje del paquete Matlab. Cuando se trata de aplicar técnicas estadísticas básicas usamos indistintamente el paquete Minitab y el paquete Matlab. El paquete S-plus lo utilizamos cuando el tratamiento de los conjuntos de datos contempla la aplicación de técnicas estadísticas avanzadas. Por lo que se refiere a los métodos de clasificación automática, unos paquetes incluyen más facilidades que otros. El paquete Matlab únicamente permite realizar los cinco métodos aglomerativos antes mencionados. En nuestra experiencia con el uso del Matlab queremos destacar que cuando el conjunto de datos a clasificar contiene un número elevado

de observaciones –del orden de un millar– el programa no puede realizar la agrupación por problemas de gestión de memoria. Estos métodos aglomerativos también se encuentran en el paquete Minitab. Este paquete incluye además el método no jerárquico por grupos disjuntos conocido como *k-means*. El paquete S-plus es la aplicación informática más completa en cuanto a métodos de clasificación. Este paquete incluye métodos aglomerativos, métodos divisivos, métodos de grupos disjuntos, e incluso métodos de clasificación estocásticos o paramétricos. En nuestra experiencia con el uso del S-plus no nos ha surgido ningún problema relacionado con la dimensión del conjunto de datos a clasificar.

4.2.5 Elección del número de grupos

En el caso de que no se conozca a priori el número de grupos a considerar en la agrupación y de que se haya elegido un método de clasificación jerárquico deberá decidirse de antemano el número de grupos a contemplar. En el Capítulo 1 de esta tesis se ha explicado que esta decisión equivale a decidir a qué nivel de fusión se *corta* el dendrograma. Esta decisión puede llevarse a cabo de manera totalmente subjetiva por simple prospección del dendrograma. Si se desea utilizar alguna herramienta objetiva, puede escogerse alguno de los índices expuestos en el Capítulo 1. Recordemos que el índice de Mojena –véase la expresión (1.11)– se basa en el intento de detectar de manera automática un *salto grande* en la estructura del dendrograma. El índice de Calinski –véase la expresión (1.12)– intenta detectar cual es el número de grupos que proporciona una mayor homogeneidad dentro de cada grupo y, simultáneamente, proporciona una mayor heterogeneidad entre los grupos. En nuestra experiencia en la realización de clasificaciones hemos constatado la gran utilidad de este tipo de índices. En nuestra opinión es muy destacable el hecho que estos índices aportan una información que se obtiene de manera automática y objetiva. En consecuencia, valoramos muy positivamente el cálculo de estos índices ya que nos proporcionan un primer acercamiento a la decisión sobre número de grupos a considerar.

4.2.6 Etapa descriptiva final grupo a grupo

Para juzgar si la clasificación resultante es razonable es necesario disponer de información sobre los grupos obtenidos. Por lo tanto, realizamos un estudio descriptivo sobre cada grupo, similar al realizado en la *Etapa descriptiva inicial*. En esta fase descriptiva *grupo a grupo* se calcularán diversos estadísticos básicos y se realizarán diferentes representaciones gráficas. Empezaremos, por ejemplo, calculando la media geométrica composicional –véase la Definición 2.30– de cada grupo. De esta manera podremos comparar si, por lo que se refiere a la medida de tendencia

central, los grupos tienen un patrón diferenciado. Una representación gráfica de mucha utilidad se obtiene a partir de los diagramas de caja de cada componente. Es decir, para cada grupo representamos un diagrama de caja múltiple donde cada diagrama corresponde valores tomados en una componente. De la observación de este diagrama se obtiene una comparación grupo a grupo de las medidas de posición y del rango de cada componente. Si en estos diagramas se detecta un rango muy grande en los valores que toman en alguna de las componentes las observaciones de determinado grupo podemos analizar la conveniencia de subdividir este grupo. Si la dimensión de los datos lo permite realizaremos una representación en el diagrama ternario o en el diagrama cuaternario identificando cada observación con un símbolo que identifique el grupo al que ha sido asignada. Si la dimensión no lo permite representaremos las diferentes subcomposiciones de $\mathcal{S}^3$ y de $\mathcal{S}^4$ en los diagramas correspondientes. En cualquier caso, siempre tendremos presente la operación centrado de los datos para conjuntos en los que alguna componente tome valores muy próximos a cero.

Otras representaciones gráficas imprescindibles son las que permiten representar datos multivariantes en el plano. Así, repetiremos el diagrama *biplot* de la etapa descriptiva inicial pero representando cada observación mediante un símbolo según el grupo al que haya sido asignada. Es importante resaltar que la distancia euclídea entre dos puntos cualesquiera que aparecen en el diagrama *biplot* utilizado está relacionada (Krzanowski, 1988b) con la distancia euclídea entre las correspondientes observaciones clr-transformadas, y por lo tanto, está relacionada con la distancia de Aitchison entre las correspondientes observaciones del conjunto.

4.2.7 ¿Es una clasificación razonable?

En el Capítulo 1 de esta tesis se ha expuesto una serie de ayudas y estrategias que serán útiles en esta fase de crítica de la clasificación obtenida. Por ejemplo, en el caso de haber realizado la agrupación mediante un método jerárquico, el cálculo del coeficiente de correlación cofenética nos proporciona una medida de la calidad de la clasificación obtenida. A la vista de la información grupo a grupo elaborada en la etapa anterior y de los resultados de las ayudas a la clasificación mencionadas se decidirá si la clasificación obtenida es o no razonable. Si la respuesta es negativa debe retrocederse en el esquema de la clasificación y considerar, bien otra medida de disimilitud, bien otro método de agrupación, o en su caso, un número diferente de grupos.

4.3 Aplicación a un caso práctico: *Población ocupada por grupos profesionales*

4.3.1 El conjunto de datos

El conjunto de datos *Población ocupada por grupos profesionales* –véase el Apéndice A.1– ha sido motivo de un estudio detallado en el trabajo de Vives y Villarroya (1996). Este conjunto de datos, que simbolizamos por $\mathbf{W}$, está formado por la observación sobre 41 unidades muestrales. Cada fila corresponde a una de las 41 comarcas en que se encuentra dividida Catalunya en el censo del año 1991 –véase la figura 4.3.

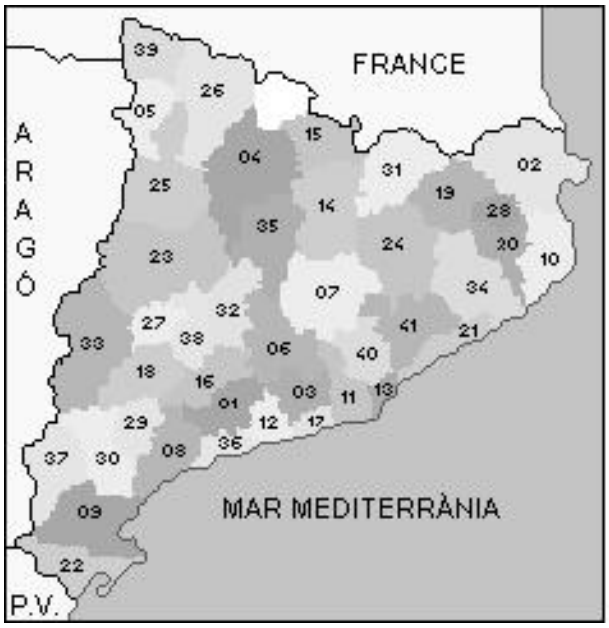


Figura 4.3: Mapa de las comarcas de Catalunya. Cada comarca se indica con el número que ocupa en la tabla del Apéndice A.1. (Fuente: *Generalitat de Catalunya*)

De cada una de las comarcas se observó el reparto de la población activa en los 8 grupos profesionales siguientes:

$\mathbf{W}_1$: *Profesionales y técnicos*; $\mathbf{W}_2$: *Personal directivo*;
$\mathbf{W}_3$: *Servicios administrativos*; $\mathbf{W}_4$: *Comerciantes y vendedores*;
$\mathbf{W}_5$: *Hostelería y otros*; $\mathbf{W}_6$: *Agricultura y pesca*;
$\mathbf{W}_7$: *Industria*; $\mathbf{W}_8$: *Fuerzas armadas.*

Cada uno de estos grupos profesionales es una variable o columna del conjunto de datos. Por lo tanto, el conjunto $\mathbf{W}$ está formado por 41 observaciones, $\mathbf{w}_i$, $i = 1, 2, \ldots, 41$, en el espacio

$\mathbb{R}^8_+$. Nótese que cada una de las 41 observaciones puede considerarse como la realización de un vector aleatorio $\mathbf{W}_i = (w_{i1}, w_{i2}, \ldots, w_{i8})$, $i = 1, 2, \ldots, 41$ distribuido según una ley multinomial multivariante de parámetros $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{i8})$.

El objetivo del estudio consiste en realizar una clasificación automática no paramétrica que permita analizar la existencia de grupos de comarcas que sean similares por lo que se refiere a la distribución de la población activa.

Préviamente a la aplicación de la metodología propuesta en esta tesis presentamos un breve resumen de los resultados expuestos en el trabajo de Vives y Villarroya (1996).

4.3.2 Resumen del estudio realizado por otros investigadores

En el trabajo de Vives y Villarroya (1996) los autores realizaron una clasificación automática no paramétrica del conjunto de datos *Población ocupada por grupos profesionales.* En la realización de la clasificación se utilizó como medida de diferencia la disimilitud de Bhattacharyya (arccos) –véase la tabla 2.1– y como método de agrupación se eligió el método jerárquico aglomerativo de la media o *average linkeage method* –véase la Sección 1.6.2. Según los autores, la elección de la disimilitud de Bhattacharyya se basó en dos motivos diferentes. Un primer motivo consistió en que esta medida es, en realidad, una medida de divergencia entre distribuciones de probabilidad multinomiales. La justificación de que esta medida es aplicable al conjunto de datos $\mathbf{W}$ se basa en la construcción del vector de proporciones $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{i8})$ para cada una de las 41 comarcas. Este vector de proporciones se calcula dividiendo el correspondiente vector de observaciones $\mathbf{w}_i$ por su total, es decir mediante la expresión $\mathbf{p}_i = \mathcal{C}(\mathbf{w}_i)$, donde la función $\mathcal{C}$ es el operador clausura expuesto en la Definición 2.4 de esta tesis. Asumiendo que estos vectores de proporciones son estimaciones muestrales del vector de parámetros de la distribución multinomial correspondiente queda justificada la elección de una medida de divergencia. La segunda razón por la que los autores eligieron la disimilitud de Bhattacharyya es que esta medida es la medida de diferencia que utiliza el método de representación de datos multivariantes conocido como método *IDA*. El método IDA, *Intrinsic Data Analysis*, es un método de representación gráfica en un espacio de dimensión reducida. Este método, según el enfoque que los autores Vives y Villarroya (1996) sintetizan en su trabajo, está basado en una métrica riemanniana.

En los resultados que presentan los autores en su trabajo las 41 comarcas de Catalunya se clasifican en los 9 grupos que se muestran en la tabla 4.1.

En la exposición de sus resultados los autores resaltan que la variable $\mathbf{W}_6$: *Agricultura y pesca* es el factor diferenciador más importante en el momento de construir los grupos. La

Tabla 4.1: Clasificación de las comarcas de Catalunya en 4 bloques y 9 grupos según los autores Vives y Villarroya (1996). Entre paréntesis se muestra el número de comarca según el Apéndice A.1.

Bloque 1 o *Agrícola*		*Bloque 3* o *Administrativo*	
	Grupo 1		**Grupo 6**
	Garrigues (18)		Barcelonès (13)
	Priorat (29)	*Bloque 4* o *Industrial*	
	Terra Alta (37)		**Grupo 7**
			Anoia (6)
	Grupo 2		Bages (7)
	Baix Ebre (9)		Baix Llobregat (11)
	Conca de Barberà (16)		Vallès Occidental (40)
	Montsià (22)		Vallés Oriental (41)
	Noguera (23)		
	Pla d'Urgell (27)		**Grupo 8**
	Ribera d'Ebre (30)		Baix Camp (8)
	Segarra (32)		Baix Empordà (10)
	Solsonès (35)		Baix Penedès (12)
	Urgell (38)		Garraf (17)
			Gironès (20)
	Grupo 3		Maresme (21)
	Pallars Jussà (25)		Selva (34)
	Pallars Sobirà (26)		Tarragonès (36)
	Grupo 4		**Grupo 9**
	Alt Empordà (2)		Alt Camp (1)
	Alta Ribagorça (5)		Alt Penedès (3)
	Alt Urgell (4)		Berguedà (14)
	Cerdanya (15)		Garrotxa (19)
	Segrià (33)		Osona (24)
Bloque 2 o *Turístico*			Pla de l'Estany (28)
	Grupo 5		Ripollès (31)
	Val d'Aran (39)		

segunda variable en importancia es la $\mathbf{W}_7$: *Industria*. Respecto la variable $\mathbf{W}_8$: *Fuerzas armadas* los autores destacan que cualquier interpretación sería cuestionable por la escasa importancia relativa de esta componente debido a ser una componente que presenta valores muy próximos a cero. El resto de variables $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_5$, que corresponden a grupos profesionales de tipo *Servicios* los autores resaltan que no son de utilidad en el momento de diferenciar los grupos de comarcas. Finalmente, y después de analizar otras características de cada comarca como son, entre otras, la renta per cápita, la densidad de población, y la situación geográfica, los autores agrupan los 9 grupos que aparecen en la tabla 4.1 en los 4 bloques siguientes:

- Bloque 1 o *Agrícola*: formado por las comarcas que aparecen en los grupos 1, 2, 3, y, 4 de la tabla;
- Bloque 2 o *Turístico*: formado por la comarca de la Val d'Aran (grupo 5);
- Bloque 3 o *Administrativo*: formado por la comarca del Barcelonès (grupo 6);
- Bloque 4 o *Industrial*: formado por las comarcas que aparecen en los grupos 7, 8, y 9 de la tabla.

Es importante resaltar que estos 4 bloques corresponden a la clasificación que se obtendría si se cortase el dendrograma presentado en el trabajo de Vives y Villarroya (1996) por un nivel de jerarquía superior al nivel con el que los autores obtuvieron los 9 grupos de la tabla 4.1.

4.3.3 Clasificación utilizando la metodología propuesta

Los autores Vives y Villarroya (1996) destacan en su trabajo que han querido dar el mismo peso a todas las comarcas. Esta decisión responde a su interés en estudiar las relaciones y las características de las comarcas, independientemiente del total de población activa de cada comarca. En este sentido los autores resaltan que una alternativa habría sido la aplicación de las técnicas del Análisis de Correspondencias (AC). Sin embargo, es bien conocido que los resultados obtenidos mediante el AC se ven afectados por los tamaños muestrales como consecuencia de que las técnicas del AC se basan en la distancia χ^2 –véase la expresión 1.5 en Capítulo 1. Este planteamiento fue el que nos indujo a realizar el estudio del conjunto de datos *Población ocupada por grupos profesionales* utilizando la metodología propuesta en esta tesis para datos composicionales.

Consideremos el conjunto de datos composicionales $\mathbf{X}$ formado por las 41 observaciones $\mathbf{x}_i,\ i = 1, 2, \ldots, 41$ del símplex $\mathcal{S}^8$. Cada observación $\mathbf{x}_i$ se obtiene dividiendo el vector $\mathbf{w}_i$

correspondiente por su total. Es decir, se cumple que $\mathbf{x}_i = \mathcal{C}(\mathbf{w}_i),\ i = 1, 2, \ldots, 41$, donde $\mathcal{C}$ es el operador clausura. Nótese que las observaciones composicionales $\mathbf{x}_i$ coinciden con los vectores de proporciones $\mathbf{p}_i$ que construyeron los autores Vives y Villarroya en su estudio. Por lo tanto, en cada composición $\mathbf{x}_i \in \mathcal{S}^8$ se recogen las proporciones de población activa para cada comarca según los 8 grupos profesionales.

Dentro de la etapa descriptiva inicial calculamos algunos estadísticos básicos del conjunto $\mathbf{X}$. En la tabla 4.2 se muestran la media geométrica composicional del conjunto, y la mediana, mínimo, y máximo de cada componente por separado. Téngase en cuenta que la mayoría de resultados de la tabla aparecen redondeados a su segunda cifra decimal, pero que los valores cercanos a cero se muestran con tres cifras decimales para diferenciarlos del valor nulo. De la observación de los valores de la tabla destacamos los aspectos siguientes:

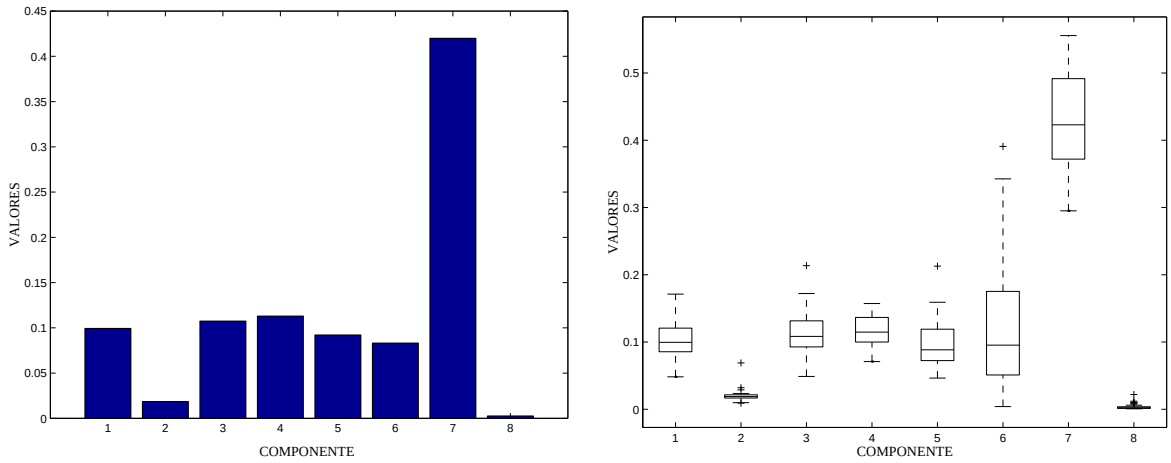
- la componente $\mathbf{X}_6$ es la que posee un rango de variación mayor;
- en la componente $\mathbf{X}_7$ se toman los valores más altos y en la componente $\mathbf{X}_8$ los más próximos a cero;
- las componentes $\mathbf{X}_1$, $\mathbf{X}_3$, $\mathbf{X}_4$, y $\mathbf{X}_5$ toman valores parecidos en su tendencia central y su rango de variación;
- en la componente $\mathbf{X}_2$ se toman valores mayores y con menor rango de variación relativo que en la componente $\mathbf{X}_8$.

Tabla 4.2: Estadísticos básicos (*media geométrica composicional, mediana, mínimo y máximo*) del conjunto de datos composicionales *Población ocupada por grupos profesionales.*

Componente	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$
Media geom.	0.11	0.02	0.11	0.12	0.1	0.09	0.45	0.003
Mediana	0.1	0.02	0.11	0.12	0.09	0.1	0.42	0.002
Mín.-Máx.	0.05-0.17	0.01-0.07	0.05-0.21	0.07-0.16	0.05-0.21	0.004-0.39	0.3-0.56	0.001-0.02

Todos estos aspectos pueden apreciarse de manera gráfica en la figura 4.4. En el diagrama de barras de la figura 4.4(a) se han representado los valores de la media geométrica composicional del conjunto $\mathbf{X}$. En esta figura destacan los valores altos de la componente $\mathbf{X}_7$ y los valores cercanos a cero de la componente $\mathbf{X}_8$. En la figura 4.4(b) se han representado los diagramas de caja de cada una de las 8 componentes por separado. En ellos se aprecia el comportamiento

similar de las componentes $\mathbf{X}_1$, $\mathbf{X}_3$, $\mathbf{X}_4$, y $\mathbf{X}_5$, y se observa el gran rango de variación de la componente $\mathbf{X}_6$.



(a) (b)

Figura 4.4: Gráficos descriptivos del conjunto *Población ocupada por grupos profesionales*: (a) *diagrama de barras de la media geométrica composicional;* (b) *diagrama de caja múltiple.*

La figura 4.5 muestra el diagrama ternario de algunas de las $C_3^8 = 56$ subcomposiciones de $\mathcal{S}^3$ que se pueden considerar para el conjunto de datos $\mathbf{X} \in \mathcal{S}^8$. En estos diagramas, las 41 observaciones se han representado mediante el número correspondiente a la posición que ocupan en la tabla del Apéndice A.1. En las figuras 4.5(a) y 4.5(b) se aprecia que en la componente $\mathbf{X}_2$ se toman valores muy bajos pero con poco rango de variación en relación a las componentes $\mathbf{X}_1$, $\mathbf{X}_3$, y $\mathbf{X}_4$. Obsérvese que la composición número 39 correspondiente a la comarca de la Val d'Aran aparece alejada del resto de observaciones del conjunto $\mathbf{X}$. En la figura 4.5(c) se observa que la nube de puntos tiene forma circular y que está situada en el centro del diagrama ternario. Este comportamiento está directamente relacionado con el hecho que las componentes $\mathbf{X}_3$, $\mathbf{X}_4$, y $\mathbf{X}_5$ tienen valores similares en su tendencia central y en su rango de variación. Por el contrario, en la figura 4.5(d) se aprecia en la forma alargada de la nube de puntos que la componente $\mathbf{X}_6$ posee un gran rango de variación relativo en relación a las componentes $\mathbf{X}_4$ y $\mathbf{X}_5$. La figura 4.5(e) muestra una nube de puntos donde se observa que de las componentes $\mathbf{X}_5$, $\mathbf{X}_6$, y $\mathbf{X}_7$, la que tiene menor rango de variación es la componente $\mathbf{X}_5$. Tanto en la figura 4.5(e) como en la 4.5(f) se observa que es la componente $\mathbf{X}_7$ la que toma valores más altos. En la figura 4.5(f) se aprecia además que la componente $\mathbf{X}_8$ toma valores extremadamente cercanos a cero y que, sin embargo, esta componente tiene un rango de variación relativo grande.

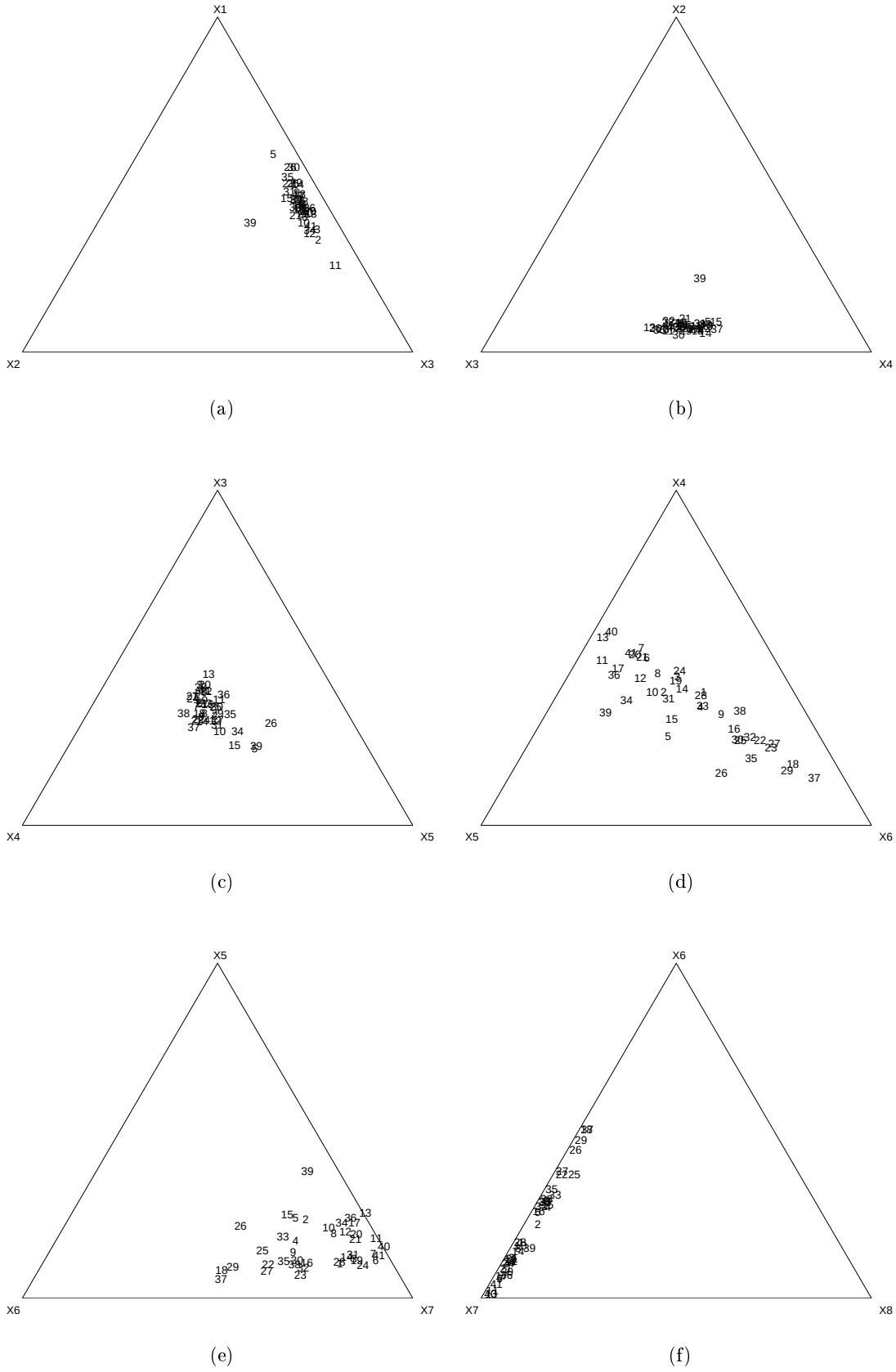


Figura 4.5: Diversas subcomposiciones del conjunto *Población ocupada por grupos profesionales.* En las componentes: (a) X_1, X_2, *y* X_3; (b) X_2, X_3, *y* X_4; (c) X_3, X_4, *y* X_5; (d) X_4, X_5, *y* X_6; (e) X_5, X_6, *y* X_7; (f) X_6, X_7, *y* X_8. Las 41 comarcas de Catalunya se simbolizan con el número que ocupan en la tabla del Apéndice A.1.

Con el propósito de completar esta etapa descriptiva inicial, utilizamos la técnica *biplot* de representación en espacios de baja dimensión. En la figura 4.6 se muestra el diagrama *biplot* del conjunto $\mathbf{X}$ construido en el espacio clr siguiendo el trabajo de Aitchison (1997). Nótese que el diagrama representa las coordenadas en el espacio de los dos primeros ejes y que estos dos ejes recogen un 90% de la variabilidad total.

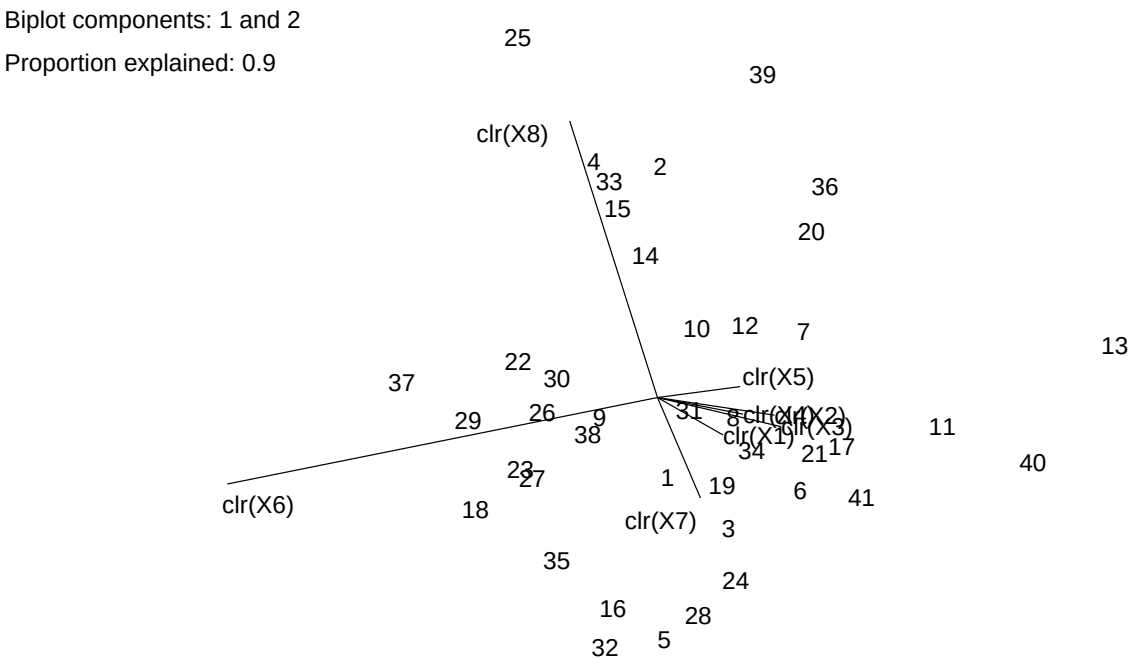


Figura 4.6: Diagrama *biplot* en el espacio clr del conjunto *Población ocupada por grupos profesionales.* Las 41 comarcas de Catalunya se simbolizan con el número que ocupan en la tabla del Apéndice A.1.

En el diagrama se observa que la componente $\mathrm{clr}(\mathbf{X}_6)$, y por ende la componente $\mathbf{X}_6$, es la que manifiesta una mayor variabilidad. También se aprecia que la dirección del eje de esta componente no dista excesivamente de la dirección del primer eje *biplot*. Este hecho puede comprobarse observando en la tabla 4.3 el gran peso de esta componente en el primer eje. La segunda componente que manifiesta mayor variabilidad es la componente $\mathrm{clr}(\mathbf{X}_8)$. La dirección de esta componente no dista excesivamente de la dirección del segundo eje *biplot*, tal y como puede confirmarse en los valores de la tabla 4.3. Las otras componentes manifiestan, con diferencia, una menor variabilidad. Observemos que la componente $\mathrm{clr}(\mathbf{X}_5)$ aparece contrapuesta a la componente $\mathrm{clr}(\mathbf{X}_6)$, y que lo mismo sucede entre las componentes $\mathrm{clr}(\mathbf{X}_7)$ y $\mathrm{clr}(\mathbf{X}_8)$. Las direcciones de las componentes $\mathrm{clr}(\mathbf{X}_1)$, $\mathrm{clr}(\mathbf{X}_2)$, $\mathrm{clr}(\mathbf{X}_3)$ y $\mathrm{clr}(\mathbf{X}_4)$, aparecen muy próximas y se situan entre las direcciones de las componentes $\mathrm{clr}(\mathbf{X}_5)$ y $\mathrm{clr}(\mathbf{X}_7)$.

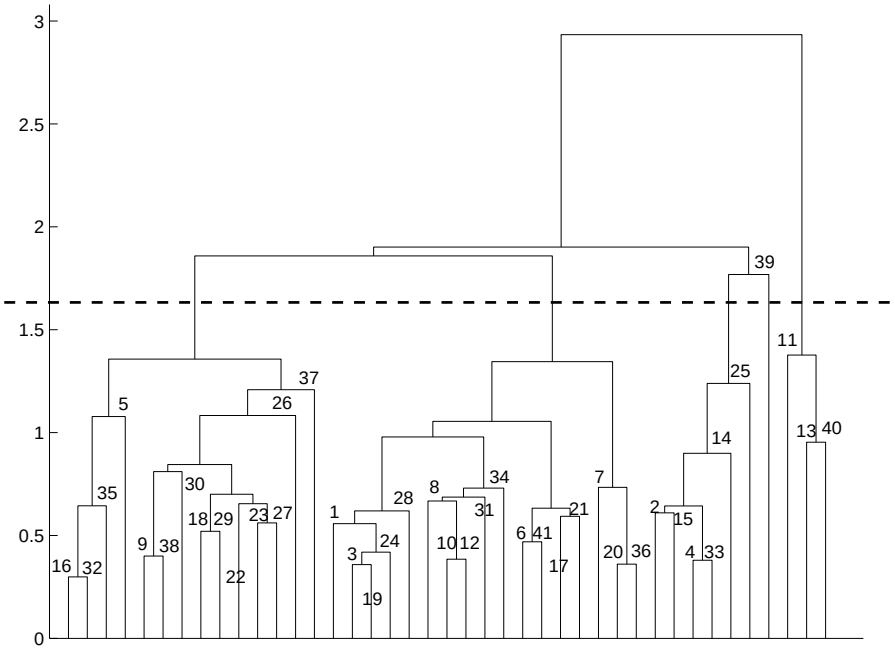
Tabla 4.3: Componentes de los dos primeros ejes del *biplot* del conjunto de datos composicionales *Población ocupada por grupos profesionales.*

Componente	$\mathrm{clr}(\mathbf{X}_1)$	$\mathrm{clr}(\mathbf{X}_2)$	$\mathrm{clr}(\mathbf{X}_3)$	$\mathrm{clr}(\mathbf{X}_4)$	$\mathrm{clr}(\mathbf{X}_5)$	$\mathrm{clr}(\mathbf{X}_6)$	$\mathrm{clr}(\mathbf{X}_7)$	$\mathrm{clr}(\mathbf{X}_8)$
Primer eje	0.9	1.7	1.8	1.2	1.2	-6.1	0.6	-1.2
Segundo eje	-0.5	-0.3	-0.4	-0.2	0.2	-1.2	-1.4	3.9

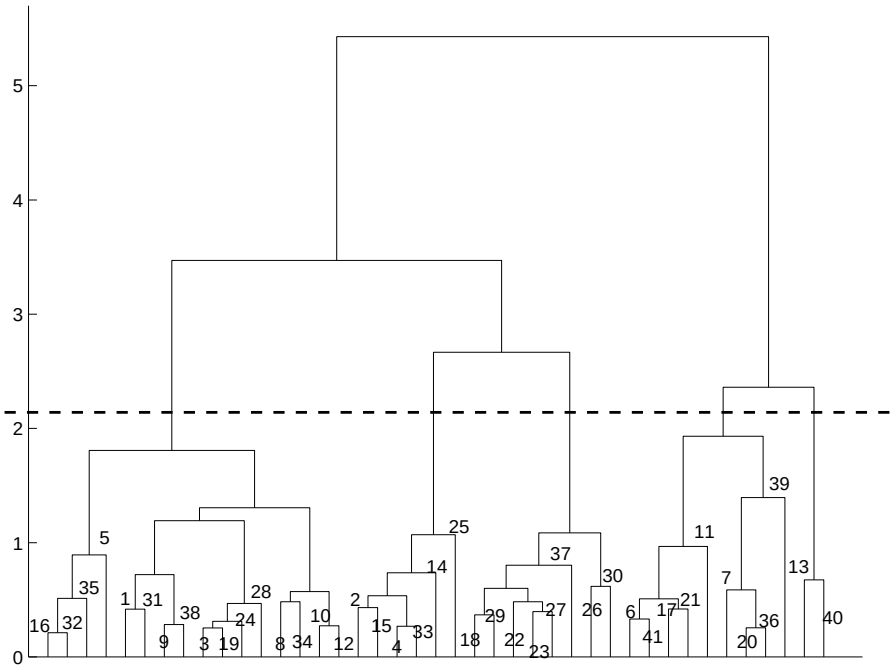
Dentro de la etapa de elección de la medida de diferencia a utilizar en la clasificación, recordamos que la disimilitud de Bhattacharyya, utilizada en el trabajo de Vives y Villarroya (1996), ha sido analizada en el Capítulo 2 de esta tesis –véase la tabla 2.1– y en el Capítulo 3. En este análisis se ha puesto de manifiesto que la medida de Bhattacharyya tiene un comportamiento coherente con la naturaleza de los datos composicionales. Sin embargo, esta disimilitud no es una medida compatible con la estructura de espacio vectorial definida en el símplex puesto que no es una medida invariante por perturbaciones. En consecuencia decidimos no utilizar esta medida en la realización de nuestra clasificación. Con el propósito de comparar los resultados, decidimos utilizar dos medidas compatibles con la operación perturbación: la distancia de Aitchison –véase la expresión (4.1)– y la disimilitud $d_{\mathcal{KL}}$ –véase la fórmula (4.2).

Entre los diferentes métodos de clasificación automática no paramétrica escogemos para nuestro estudio los jerárquicos aglomerativos. Esta elección se basa en la intención de usar métodos de la misma familia que el método jerárquico de la media, utilizado en el trabajo de Vives y Villarroya (1996).

Después de realizar diferentes clasificaciones utilizando la distancia de Aitchison y la disimilitud $d_{\mathcal{KL}}$, se ha puesto de manifiesto la coincidencia total de resultados. Este hecho puede apreciarse en la observación de los dendrogramas representados en las figuras 4.7 y 4.8. En estos dendrogramas las observaciones se representan por el número que ocupan en la tabla de datos del Apéndice A.1. La observación de estas figuras pone de manifiesto la estrecha relación entre las medidas d_{Ait} y $d_{\mathcal{KL}}$. Esta relación ha sido analizada en la Sección 3.5 de esta tesis. Por este motivo, decidimos limitar nuestro estudio a las clasificaciones realizadas usando la distancia de Aitchison.

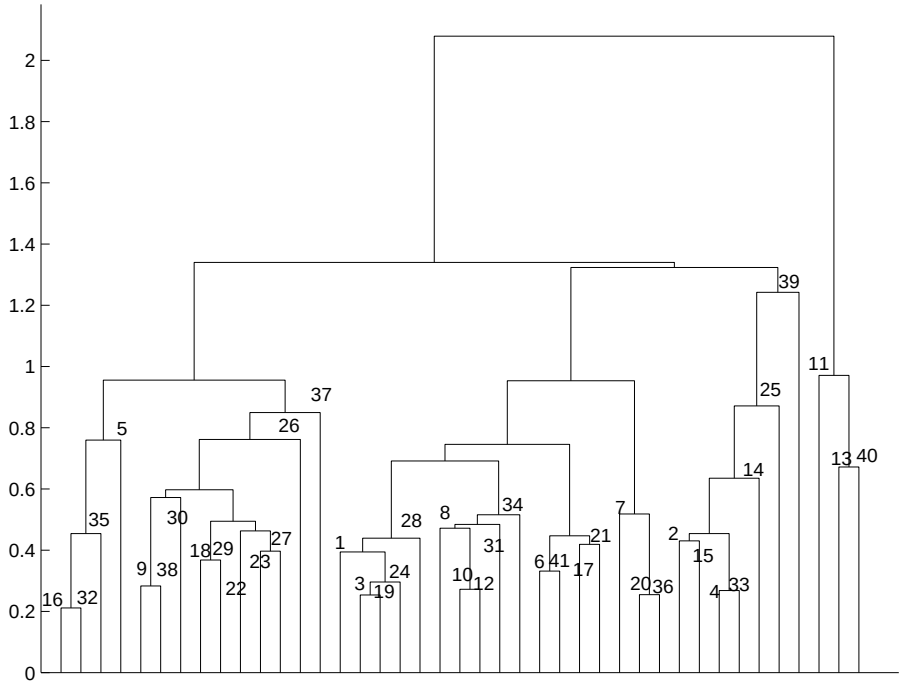


(a)

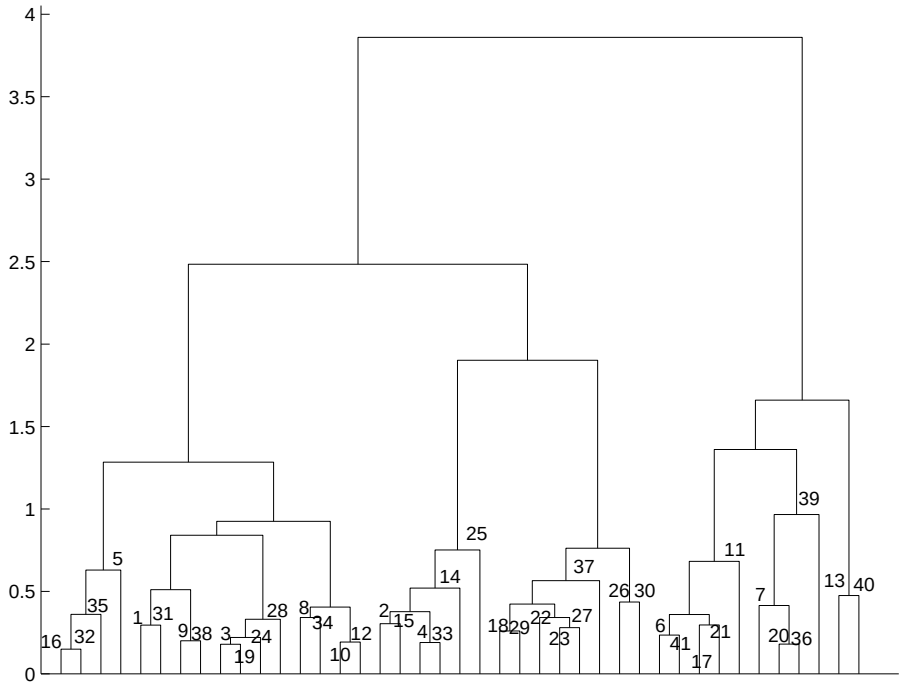


(b)

Figura 4.7: Dendrogramas obtenidos de la clasificación del conjunto *Población ocupada por grupos profesionales* utilizando la distancia de Aitchison y el método de agrupación: (a) *de la media;* (b) *de Ward.* La línea discontinua indica el nivel de corte del árbol para obtener 5 grupos.



(a)



(b)

Figura 4.8: Dendrogramas obtenidos de la clasificación del conjunto *Población ocupada por grupos profesionales* utilizando la disimilitud $d_{\mathcal{KL}}$ y el método de agrupación: (a) *de la media;* (b) *de Ward.*

En la tabla 4.4 se muestran los valores de tres coeficientes que han sido presentados en la Sección 1.8 de esta tesis: el coeficiente de correlación cofenética, el índice de Mojena, y el índice de Calinski. Estos tres índices han sido calculados para cada uno de los cinco métodos aglomerativos de clasificación. Recordemos que el coeficiente de correlación cofenética mide el grado de relación entre el índice de jerarquía resultado de la estructura jerárquica y la medida de diferencia. Observamos en la tabla que los valores más altos en este índice se manifiestan para los métodos del centroide y de la media. El índice de Mojena informa sobre el número de grupos que refleja la estructura jerárquica resultado de la clasificación. Recordemos que este índice se calcula en base a la búsqueda de "saltos grandes" en los niveles de fusión del dendrograma. Los valores del índice de Mojena sugieren para todos los métodos, excepto para el método del centroide, la existencia de 5 grupos en el conjunto de datos $\mathbf{X}$. En el índice de Calinski o índice C se aprecia una mayor divergencia entre los resultados para los diferentes métodos de clasificación. Recordemos que este índice se basa en la comparación entre la variabilidad dentro de los grupos y la variabilidad entre los grupos. El índice calcula el número de grupos en que debe dividirse el conjunto a clasificar de manera que los grupos resulten ser lo más homogéneos dentro de si y lo más heterogéneos entre ellos. La tabla 4.4 muestra que el único método que sigue indicando la existencia de 5 grupos es el método de la media. A la vista de los valores de la tabla 4.4 y con el objetivo de realizar una comparación de resultados con la clasificación del trabajo de Vives y Villarroya (1996), decidimos analizar únicamente la clasificación obtenida con el método de la media.

Tabla 4.4: Índices de correlación cofenética, de Mojena, y de Calinski para las clasificaciones del conjunto *Población ocupada por grupos profesionales* según los diferentes métodos de agrupación.

Índice	Ward	Centroide	Mínimo	Máximo	Media
Correlación cofenética	0.59	0.74	0.55	0.52	0.74
Mojena	5	6	5	5	5
Calinski	2	6	3	6	5

En la figura 4.7(a) se ha representado el dendrograma obtenido al aplicar el método de la media usando la distancia de Aitchison al conjunto $\mathbf{X}$. En esta misma figura se muestra un nivel de corte del árbol que da lugar a 5 grupos. Es importante resaltar que, si bien en una primera opción se ha analizado la clasificación resultante de considerar los 5 grupos que sugieren los índices de Mojena y de Calinski, se han analizado también los grupos resultantes al considerar un menor o un mayor número de grupos. Sin embargo, a la vista de los resultados obtenidos,

se ha decidido que la clasificación en 5 grupos es la agrupación más razonable puesto que esta clasificación es la que manifiesta un patrón diferenciador entre grupos más acusado.
En la tabla 4.5 se muestran los cinco grupos de comarcas resultantes de aplicar el método de media usando la distancia d_{Ait}. En la tabla, los grupos están ordenados, en sentido decreciente, según el valor de la componente $\mathbf{X}_6$ o *Agricultura y pesca* de la media geométrica composicional del grupo.

Tabla 4.5: Clasificación de las comarcas de Catalunya en 5 grupos resultantes de aplicar el método de la media con la distancia de Aitchison al conjunto *Población ocupada por grupos profesionales*. Entre paréntesis se muestra el número de comarca según el Apéndice A.1.

Grupo 1 o *Agrícola*	
	Alta Ribagorça (5)
	Baix Ebre (9)
	Conca de Barberà (16)
	Garrigues (18)
	Montsià (22)
	Noguera (23)
	Pallars Sobirà (26)
	Pla d'Urgell (27)
	Priorat (29)
	Ribera d'Ebre (30)
	Segarra (32)
	Solsonès (35)
	Terra Alta (37)
	Urgell (38)
Grupo 2 o *Militar*	
	Alt Empordà (2)
	Alt Urgell (4)
	Berguedà (14)
	Cerdanya (15)
	Pallars Jussà (25)
	Segrià (33)
Grupo 3 o *Turístico*	
	Val d'Aran (39)
Grupo 4 o *Industrial-Medio*	
	Alt Camp (1)
	Alt Penedès (3)
	Anoia (6)
	Bages (7)
	Baix Camp (8)
	Baix Empordà (10)
	Baix Penedès (12)
	Garraf (17)
	Garrotxa (19)
	Gironès (20)
	Maresme (21)
	Osona (24)
	Pla de l'Estany (28)
	Ripollès (31)
	Selva (34)
	Tarragonès (36)
	Vallés Oriental (41)
Grupo 5 o *Industria-Servicios*	
	Baix Llobregat (11)
	Barcelonès (13)
	Vallès Occidental (40)

De estos grupos se destacan las siguientes características:

- Grupo 1 o *Agrícola*. Las 14 comarcas que pertenecen a este grupo tienen como característica principal la de tomar valores altos en la componente $\mathbf{X}_6$. Esta característica se

aprecia en la tabla 4.6 tanto en los valores de la media geométrica y mediana como en los valores mínimo y máximo de la componente $\mathbf{X}_6$. En la figura 4.9 se observa que es el grupo con mayor valor en la componente $\mathbf{X}_6$ de la media geométrica composicional. En los diagramas de caja de la figura 4.10(a) puede apreciarse que, en relación a los otros grupos, el Grupo 1 es el que tiene situada la caja de la componente $\mathbf{X}_6$ en la zona de valores más altos. En todas las subcomposiciones representadas en la figura 4.11 se observa que el Grupo 1 aparece situado como el grupo más cercano al vértice de la componente $\mathbf{X}_6$. En los diagramas *biplot* de la figura 4.12 el Grupo 1 se encuentra situado alrededor del eje de la componente $\mathrm{clr}(\mathbf{X}_6)$ y en posiciones alejadas del centro del diagrama manifestando el hecho que las comarcas pertenecientes a este grupo poseen una elevada proporción de población activa dedicada a la *Agricultura y pesca*.

- Grupo 2 o *Militar*. Este grupo está formado por 6 comarcas que tienen un elevado porcentaje de personal perteneciente a las actividades correspondientes al grupo *Fuerzas armadas*. Son comarcas fronterizas o con un número elevado de instalaciones militares que provoca que sean observaciones que toman valores relativamente altos en la componente $\mathbf{X}_8$. En la tabla 4.6 se observa que este grupo toma un valor medio de un 1% de la población activa dedicada a las *Fuerzas armadas* cuando la proporción media a nivel de Catalunya se sitúa entorno al 0.3% –véase tabla 4.2. Observando la figura 4.9 puede apreciarse que, exceptuando el Grupo 3, el Grupo 2 es el grupo con mayor valor en la componente $\mathbf{X}_8$ de la media geométrica composicional. Por lo que se refiere a las componentes correspondientes a los otros grupos profesionales, el Grupo 2 no destaca por tomar valores alejados de la media del total de comarcas de Catalunya. Los diagramas de caja de la figura 4.10(b) reflejan las principales características del Grupo 2, destacándose los valores relativamente bajos en la componente industrial $\mathbf{X}_7$. Las figuras 4.11(a) y 4.11(b) muestran los diagramas ternarios de subcomposiciones en las que no interviene la componente $\mathbf{X}_8$. Nótese que en estos diagramas las comarcas del Grupo 2 aparecen en la parte central de la nube de puntos. Por el contrario, en los diagramas que muestran las figuras 4.11(c), 4.11(d), 4.11(e) y 4.11(f), las comarcas del Grupo 2 aparecen entre las comarcas más cercanas al vértice de la componente $\mathbf{X}_8$. En el diagrama *biplot* que muestra la figura 4.12 se aprecia que las comarcas del Grupo 2 aparecen situadas muy cercanas al eje de la variable $\mathrm{clr}(\mathbf{X}_8)$ y alejadas del origen de coordenadas. Este gráfico pone de manifiesto que el Grupo 2 está formado por comarcas que toman valores relativamente altos en la componente $\mathbf{X}_8$ y toman valores medios en las otras componentes.

- Grupo 3 o *Turístico*. Este grupo está formado únicamente por la comarca de la Val d'Aran. Esta comarca se distingue por tomar valores altos conjuntamente en las componentes $\mathbf{X}_2$, $\mathbf{X}_5$ y $\mathbf{X}_8$. Recordemos que la componente $\mathbf{X}_2$ recoge la proporción de población activa considerada como *Personal directivo* y la componente $\mathbf{X}_5$ la proporción de población activa dedicada a la actividad *Hostelería y otros*. Estas características ponen de manifiesto la existencia de una fuerte industria turística de ámbito local. Recordemos que en los diagramas ternarios de las figuras 4.5(a), 4.5(b), y 4.5(e), la comarca de la Val d'Aran (39) aparece claramente separada de la nube de puntos. En los diagramas de barras de la figura 4.9 se aprecia que la barra de las componentes $\mathbf{X}_2$ y $\mathbf{X}_5$ de la media geométrica composicional es más alta en este grupo que en el resto de grupos. Nótese que la barra de la componente $\mathbf{X}_8$ también es alta, mostrando la existencia de numerosas instalaciones militares en la comarca de la Val d'Aran. En la figura 4.10(c) se han representado mediante un diagrama de puntos los valores de la tabla 4.6 para el Grupo 3. En este diagrama se aprecia que en la comarca de la Val d'Aran se toma un valor bajo en la componente agrícola $\mathbf{X}_6$. En los diagramas ternarios y cuaternarios de la figura 4.11 la comarca de la Val d'Aran se ha representado mediante un triángulo invertido. Observemos que en las figuras 4.11(a) y 4.11(b) esta comarca es la más cercana al vértice de la componente $\mathbf{X}_5$ de la actividad *Hostelería y otros*. En la figuras 4.11(c) y 4.11(d), en las que no interviene la componente $\mathbf{X}_5$, la comarca de la Val d'Aran aparece entre las comarcas más cercanas al vértice de la componente $\mathbf{X}_8$. En los diagramas cuaternarios 4.11(e) y 4.11(f) se observa que la comarca de la Val d'Aran aparece entre las comarcas más cercanas a la arista determinada por las componentes $\mathbf{X}_5$ y $\mathbf{X}_8$. En el diagrama *biplot* de la figura 4.12 se aprecia que la comarca de la Val d'Aran aparece situada en el semiplano determinado por las variables $\mathrm{clr}(\mathbf{X}_5)$ y $\mathrm{clr}(\mathbf{X}_8)$, y muy alejada del origen de coordenadas. Nótese que el símbolo de esta comarca aparece dentro de un círculo indicando que la comarca de la Val d'Aran puede ser catalogada como una observación atípica.

- Grupo 4 o *Industrial-Medio*. Las 17 comarcas pertenecientes a este grupo se caracterizan por tomar valores medios en las componentes $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, $\mathbf{X}_5$, y $\mathbf{X}_8$. Por lo tanto son comarcas cuya distribución de la población activa en estas variables se asemeja a la distribución de la población activa en toda Catalunya. Sin embargo, en contraposición a las comarcas pertenecientes al Grupo 1, las comarcas de este Grupo 4 toman valores más altos en la componente $\mathbf{X}_7$: *Industria* y más bajos en la componente agrícola $\mathbf{X}_6$. Estas características de la tendencia central del Grupo 4 pueden apreciarse de manera gráfica

en el diagrama de barras de la figura 4.9 y en los diagramas de caja de la figura 4.10(d). En los diagramas ternarios y cuaternarios de la figura 4.11 se observa que las comarcas del Grupo 4 aparecen en la zona central de la nube de puntos pero siempre situadas más cercanas al vértice de la componente industrial $\mathbf{X}_7$ que del vértice de la componente agrícola $\mathbf{X}_6$. En los diagramas *biplot* de la figura 4.12 puede apreciarse en su zona central las comarcas de esta Grupo 4. Nótese que la mayoría de las comarcas del grupo aparecen en la zona del semieje negativo de la variable $\text{clr}(\mathbf{X}_6)$ con lo que se manifiesta que la comarcas pertenecientes al Grupo 4 toman valores bajos en la componente agrícola.

- Grupo 5 o *Industria-Servicios*. Este grupo está formado por tres comarcas cuya característica principal es la de tomar valores relativamente altos conjuntamente en las cinco primeras componentes y en la componente $\mathbf{X}_7$ de actividades industriales. En este grupo destacan el valor alto en la componente $\mathbf{X}_3$ de *Servicios Administrativos* aportado por la comarca del Barcelonès (13) y el valor alto en la componente industrial $\mathbf{X}_7$ aportado por las comarcas del Baix LLobregat (11) y la comarca del Vallès Occidental (40). Tanto en los diagramas de barras de la figura 4.9 como en los diagramas de caja de la figura 4.10(e) puede apreciarse que las comarcas del Grupo 5 toman valores altos en todas las componentes excepto en la componente agrícola y en la componente militar $\mathbf{X}_8$. En los diagramas ternarios y cuaternarios de la figura 4.11 en los que interviene la componente $\mathbf{X}_6$ las tres observaciones perteneciente a este Grupo 5 aparecen como las más alejadas del vértice de la componente $\mathbf{X}_6$. En los diagramas donde se representan subcomposiciones en las que interviene la componente $\mathbf{X}_7$ las tres observaciones del Grupo 5 aparecen entre las más cercanas a su vértice. Como puede apreciarse en los diagramas *biplot* de la figura 4.12 la componente agrícola de estas comarcas es muy pequeña. Nótese que dos de estas tres comarcas, el Baix LLobregat (11) y el Barcelonès (13), pueden ser catalogadas como observaciones atípicas.

Las características que acabamos de exponer de cada uno de los 5 grupos resultantes de la clasificación ponen de manifiesto que las observaciones que pertenecen a grupos diferentes muestran un patrón claramente diferenciado en el valor que toman en las diferentes variables. En consecuencia, consideramos que la agrupación obtenida es una clasificación razonable. Somos conscientes que al habernos limitado a estudiar los resultados que proporciona el método de la media no hemos completado el estudio del conjunto *Población ocupada por grupos profesionales*. Pueden obtenerse otras clasificaciones razonables mediante la aplicación de otros métodos de clasificación automática no paramétrica. Sin embargo, con el ánimo de no extender en demasía

el estudio del caso práctico que nos ocupa y recordando que uno de los centros de interés del estudio es la comparación de resultados con la clasificación presentada en Vives y Villarroya (1996), decidimos no desarrollar otras clasificaciones resultantes de aplicar métodos diferentes al de la media.

Tabla 4.6: Estadísticos básicos (*Número de observaciones, Media geométrica composicional, Mediana, Mínimo y Máximo*) de cada uno de los 5 grupos resultantes de aplicar el método de la media al conjunto de datos composicionales *Población ocupada por grupos profesionales.*

	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$
Grupo 1								
(14 obs.)								
Med. geo.	0.09	0.02	0.08	0.09	0.08	0.22	0.41	0.002
Mediana	0.097	0.02	0.09	0.09	0.07	0.21	0.40	0.003
Mín.-Máx.	0.05-0.14	0.01-0.02	0.05-0.1	0.07-0.13	0.05-0.16	0.13-0.39	0.3-0.51	0.001-0.004
Grupo 2								
(6 obs.)								
Med. geo.	0.11	0.02	0.11	0.13	0.11	0.13	0.37	0.01
Mediana	0.1	0.02	0.11	0.14	0.11	0.14	0.34	0.009
Mín.-Máx.	0.08-0.13	0.01-0.02	0.09-0.14	0.1-0.16	0.08-0.16	0.08-0.21	0.32-0.51	0.006-0.02
Grupo 3								
(1 obs.)								
Val d'Aran	0.11	0.07	0.11	0.14	0.21	0.05	0.3	0.01
Grupo 4								
(17 obs.)								
Med. geo.	0.10	0.02	0.12	0.13	0.1	0.05	0.47	0.002
Mediana	0.1	0.02	0.12	0.12	0.09	0.06	0.48	0.002
Mín.-Máx.	0.07-0.14	0.02-0.03	0.08-0.17	0.1-0.15	0.07-0.15	0.02-0.1	0.38-0.56	0.001-0.006
Grupo 5								
(3 obs.)								
Med. geo.	0.11	0.02	0.17	0.14	0.11	0.01	0.44	0.001
Mediana	0.12	0.02	0.15	0.13	0.11	0.01	0.48	0.001
Mín.-Máx.	0.06-0.17	0.02-0.03	0.15-0.21	0.13-0.15	0.09-0.12	0.004-0.01	0.32-0.52	0.001-0.002

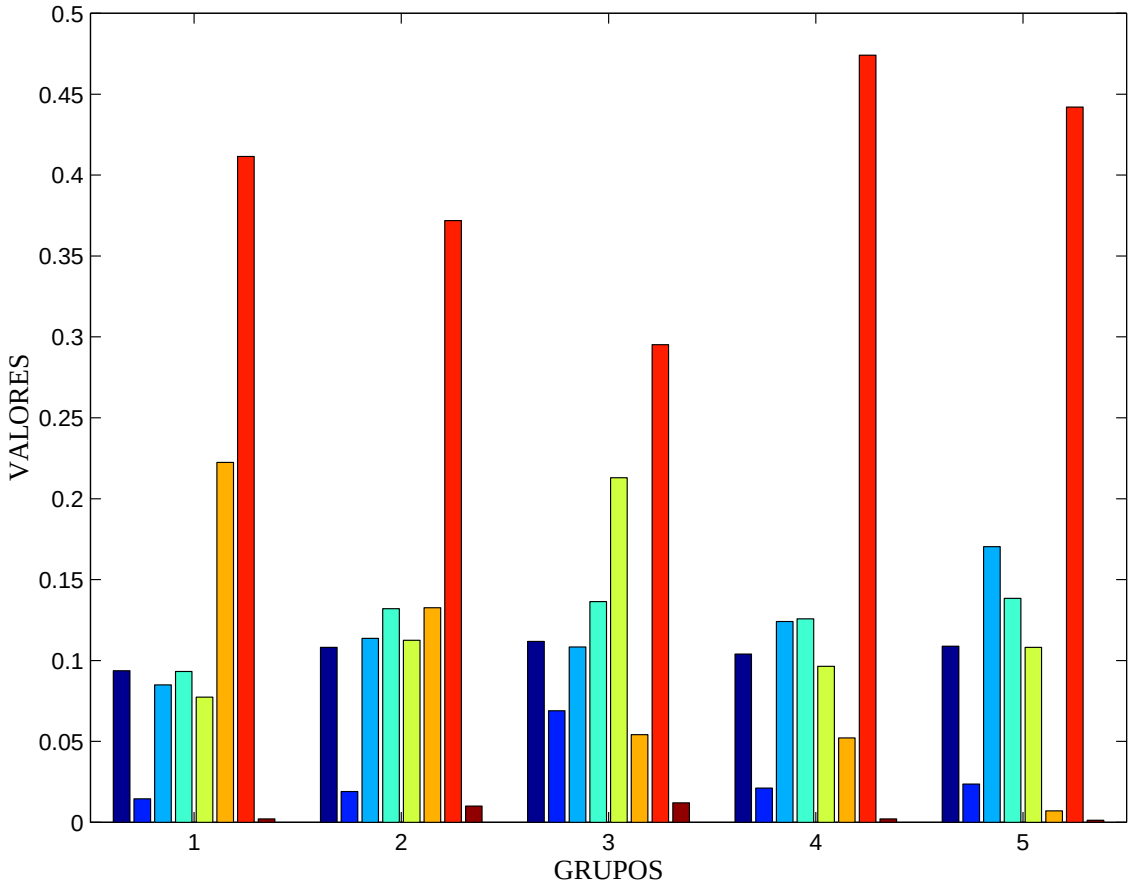


Figura 4.9: Diagrama de barras de las medias geométricas composicionales de los 5 grupos determinados por el método de la media aplicado al conjunto *Población ocupada por grupos profesionales.*

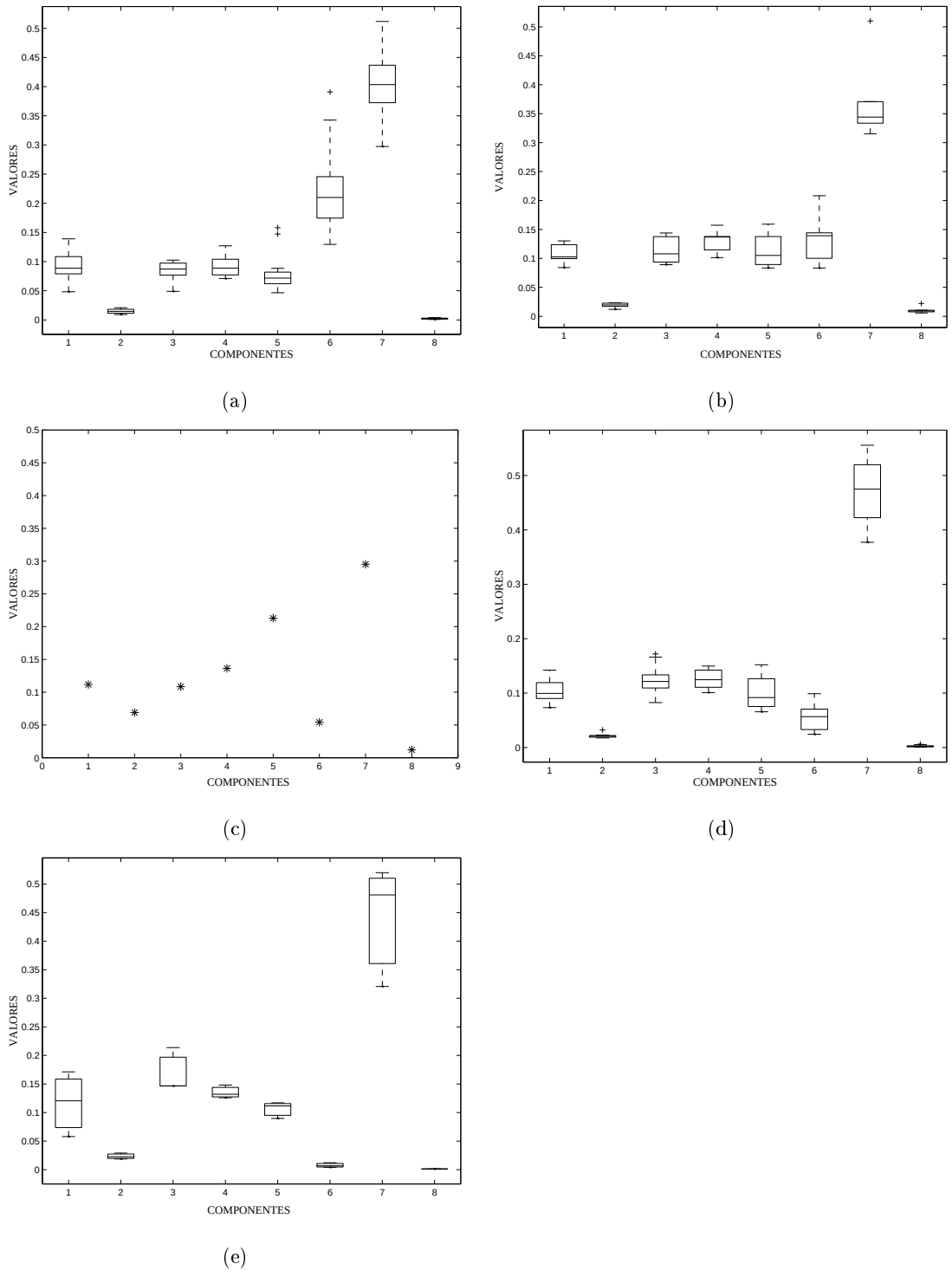


Figura 4.10: Diagramas de caja de los 5 grupos determinados por el método de la media aplicado al conjunto *Población ocupada por grupos profesionales*: (a) *grupo 1;* (b) *grupo 2;* (c) *grupo 3; comarca de la Val d'Aran* (d) *grupo 4;* (e) *grupo 5.*

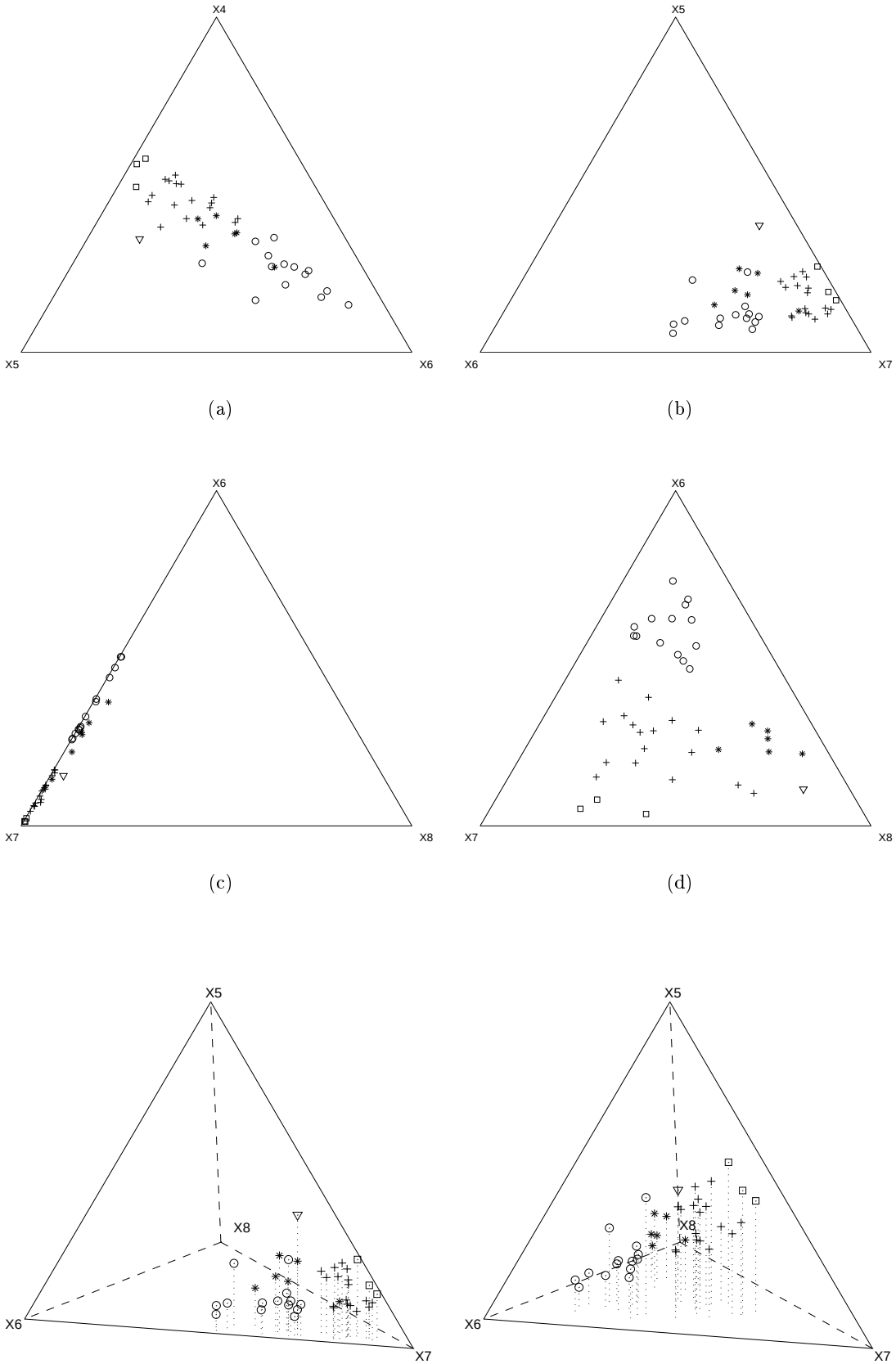
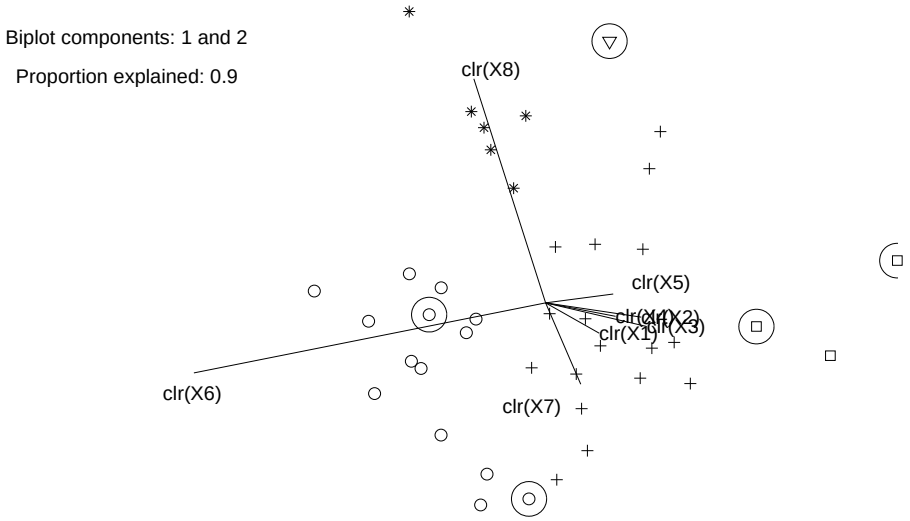
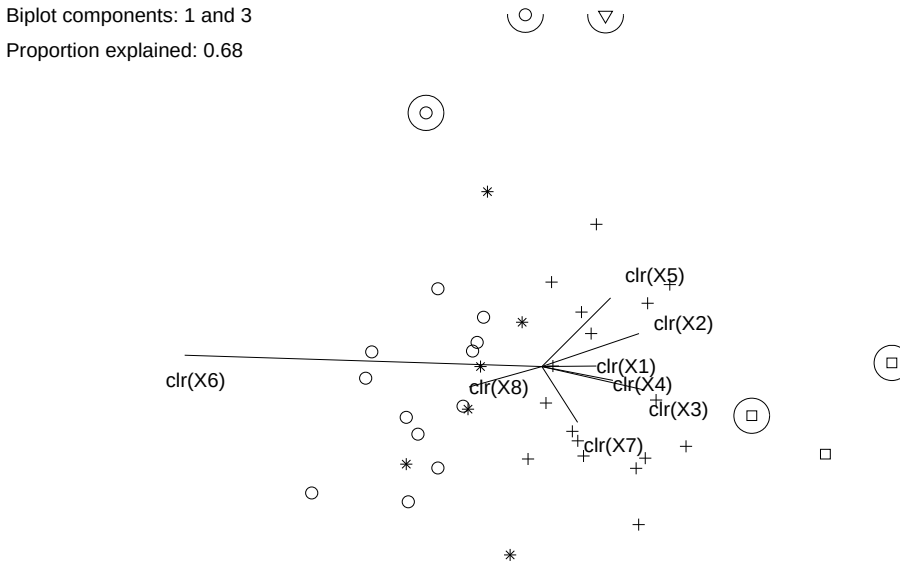


Figura 4.11: Diversas subcomposiciones del conjunto *Población ocupada por grupos profesionales.* En las componentes: (a) X_4, X_5, *y* X_6; (b) X_5, X_6, *y* X_7; (c) X_6, X_7, *y* X_8; (d) X_6, X_7, *y* X_8, *con los datos centrados;* (e) X_5, X_6, X_7, *y* X_8 ; (f) X_5, X_6, X_7, *y* X_8, *con los datos centrados.* Se muestran los 5 grupos determinados por el método de la media. (Grupo 1:'$\circ$'; Grupo 2: '$*$'; Grupo 3: '∇'; Grupo 4:'$+$'; Grupo 5:'$\square$').



(a)



(b)

Figura 4.12: Diagramas biplot en el espacio clr del conjunto *Población ocupada por grupos profesionales*: (a) *primer y segundo ejes;* (b) *primer y tercer ejes.* Se muestran los 5 grupos determinados por el método de la media. (Grupo 1:'$\circ$'; Grupo 2: '$*$'; Grupo 3: '∇'; Grupo 4:'$+$'; Grupo 5:'$\square$'). Las observaciones dentro de un círculo pueden ser catalogables como atípicas.

4.3.4 Comparación de resultados

En este apartado comparamos los resultados expuestos en el trabajo de Vives y Villarroya (1996) y los resultados obtenidos utilizando la metodología propuesta para datos composicionales. En una primera lectura puede observarse que existe una notable similitud entre los resultados de las dos clasificaciones. De esta similitud destacan los dos aspectos siguientes:

- Gran coincidencia en las comarcas que son calificadas como agrícolas y como industriales en las dos clasificaciones.
- En las dos agrupaciones la comarca de la Val d'Aran constituye, por si sola, un grupo cuya característica diferenciadora es una alta proporción en la componente turística.

Esta notable similitud de resultados pone de manifiesto que la medida de diferencia de Bhattacharyya tiene un comportamiento coherente con la naturaleza de los datos composicionales, tal y como se ha expuesto en el Capítulo 3 de esta tesis. Sin embargo, una comparación más pausada nos lleva a detectar la existencia de diferencias entre las dos clasificaciones. De las diferencias detectadas destacamos los dos aspectos siguientes:

- En nuestra clasificación se consideran 5 grupos de comarcas. Un grupo más que en la clasificación de Vives y Villarroya (1996) que contemplaba 4 bloques diferentes. Las comarcas que pertenecen a este nuevo grupo –véase el Grupo 2 de la tabla 4.5– se distinguen por su alta proporción relativa de población activa dedicada a actividades englobadas bajo el nombre de *Fuerzas armadas*. De las 6 comarcas pertenecientes a este grupo la comarca del Berguedà (14) era asignada en el trabajo de Vives y Villarroya (1996) al Bloque Industrial –véase la tabla 4.1. Las otras 5 comarcas pertenecientes al Grupo 2 de nuestra clasificación formaban el Grupo 4 del Bloque Agrícola en la clasificación de Vives y Villarroya (1996).
- En nuestra agrupación la comarca del Barcelonès (13) no constituye un grupo por si sola. El Grupo 5 o *Industria-Servicios* de nuestra agrupación está formado, además de la comarca del Barcelonès, por las comarcas del Baix Llobregat (11) y del Vallès Occidental. Este grupo pone de manifiesto la existencia en Cataluña de una área geográfica, que engloba la ciudad de Barcelona y sus alrededores, donde la proporción de servicios administrativos y de tejido industrial es muy elevada. En la clasificación de Vives y Villarroya (1996), la comarca del Barcelonès constituía por si sola el Bloque Administrativo y las otras dos comarcas eran asignadas al Bloque Industrial.

Sin estar en nuestro ánimo el calificar como mejor o peor una de los dos medidas de diferencia, creemos importante destacar que la distancia de Aitchison es mucho más sensible a las variaciones relativas en las proporciones que la disimilitud utilizada en el trabajo de Vives y Villarroya (1996). Esta característica se pone especialmente de manifiesto en el hecho que nuestra clasificación contempla la existencia de un grupo de comarcas cuya distinción principal se basa en la alta proporción relativa de la componente *Fuerzas armadas* cuyo rango de variación en toda Catalunya abarca desde un mínimo del 0.1% hasta un máximo del 2% –véanse los valores de la tabla 4.2. Consideramos que esta mayor sensibilidad de la distancia de Aitchison delante de valores cercanos a cero es una virtud que convierte a esta distancia en una medida muy útil en el estudio de conjuntos de datos que contengan componentes con valores casi nulos. En el estudio de conjuntos de datos sin componentes con valores cercanos a cero las divergencias de resultados obtenidos mediante la distancia de Aitchison y la disimilitud de Bhattacharyya serían menores, tal y como se ha expuesto en el Capítulo 2 de esta tesis.

Capítulo 5

El problema de los ceros

5.1 Introdución

La mayor parte de las técnicas utilizadas en un análisis estadístico de datos composicionales se fundamenta en la transformación de los datos mediante las aplicaciones alr y clr. Lamentablemente estas transformaciones no son aplicables a observaciones que tengan nula alguna de sus componentes. La realización de una clasificación automática no paramétrica de un conjunto de datos composicionales tiene como elemento clave la distancia entre dos observaciones. Tanto la distancia de Aitchison –véase (2.11) en el Capítulo 2– como la disimilitud de Kullback-Leibler –véase (3.32) en el Capítulo 3– tienen el handicap de no poder aplicarse a observaciones que contengan ceros en sus componentes. En este capítulo nos ocuparemos únicamente de la distancia de Aitchison por ser la que posee mejores propiedades respecto de las operaciones básicas definidas en el símplex. Debido a que el tratamiento del problema de los ceros no depende de la medida de diferencia escogida, para el caso de la disimilitud de Kullback-Leibler podrán aplicarse todas las consideraciones que desarrollamos en este capítulo para la distancia de Aitchison.

En muchas situaciones prácticas podemos encontrar conjuntos de datos con observaciones cuyas componentes contengan valores nulos. Por ejemplo, en el estudio de datos referentes al reparto de las diferentes partidas del presupuesto familiar podemos encontrar familias en las que la componente correspondiente a la partida de *tabaco y bebidas alcohólicas* sea nula. Otro caso diferente de valor nulo puede aparecer en el estudio de la composición mineral de diferentes rocas en el que podemos encontrar componentes nulas debido a que un particular mineral no ha sido detectado.

En el análisis estadístico de datos composicionales se distinguen dos tipos de valores nulos o ceros: *ceros esenciales* y *ceros por redondeo*. El valor nulo que aparece en un estudio de los

presupuestos de las familias es un cero de tipo esencial o absoluto. Este tipo de valor nulo aparecerá mayoritariamente en estudios cuyos datos composicionales puedan entenderse como realizaciones de variables aleatorias multinomiales. Por el contrario, el valor nulo que aparece en el estudio de la composición mineral es habitualmente considerado un cero por redondeo, es decir, es un valor nulo que indica que no se ha registrado la presencia del mineral en cuestión puesto que no se ha superado el umbral de detección inherente al proceso de medida.

En la realización de una clasificación automática no paramétrica de un conjunto de datos composicionales, el tratamiento de datos que contengan uno u otro tipo de cero es diferente. En este capítulo se expone el tratamiento adecuado para el caso de ceros esenciales haciendo mención de los principales aspectos a tener en cuenta y presentando una técnica de clasificación automática. Por lo que se refiere al caso de ceros por redondeo, en la monografía de Aitchison (1986) se propone un reemplazamiento de estas observaciones con componentes nulas por otras observaciones sin ceros. Este reemplazamiento ha sido analizado recientemente por Tauber (1999) desde un punto de vista descriptivo en el contexto de las clasificaciones automáticas de datos composicionales. En ese trabajo el autor argumenta que la substitución propuesta por Aitchison no es satisfactoria porque provoca la aparición de grupos espúreos de observaciones. En esta tesis se propone una nueva fórmula de reemplazamiento de ceros por redondeo introducida por primera vez en Martín-Fernández et al. (2000).

5.2 La operación amalgama y el problema de los ceros

En la mayor parte de los estudios de conjuntos de datos composicionales sería factible inducir la presencia de componentes con valores nulos simplemente aumentando el número de componentes a considerar en las observaciones. Por ejemplo, si en un estudio de la composición de los presupuestos de las familias, subdividimos la componente *Vestido y Calzado* en las componentes: *Camisas, Pantalones, Faldas, Suéteres, Chaquetas, Abrigos, Ropa interior, Botas, Zapatos, y Calzado Deportivo*, nos aparecerán componentes con valores nulos. En consecuencia, una primera cuestión que debemos resolver en un estudio de datos con ceros es si estos ceros son o no producto de una subdivisión excesiva de las componentes que estamos observando. En el caso que la respuesta sea afirmativa, es necesario realizar una *amalgama* (Aitchison, 1986) de algunas de las componentes de las observaciones. Recordemos que en el Capítulo 2 de esta tesis, donde hemos expuesto los aspectos básicos de los datos composicionales, hemos presentado una primera definición de la operación amalgama. La definición que damos a continuación nos presenta la operación amalgama en términos matriciales.

Definición 5.1 Sea $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ una composición de $\mathcal{S}^D$. Consideremos una matriz $\mathbf{A}$ de orden $C \times D$, donde $C \leq D$ y todos los elementos de la matriz son iguales a cero excepto D elementos iguales a 1 que aparecen uno en cada columna y como mínimo uno en cada fila. Entonces, la observación $\mathbf{t}^t = \mathbf{A}\mathbf{x}^t \in \mathcal{S}^C$ recibe el nombre de *amalgama* en C componentes de $\mathbf{x}$. □

El ejemplo siguiente ilustra la realización práctica de una amalgama de una observación de $\mathcal{S}^8$ que contiene componentes con ceros para obtener como resultado una observación de $\mathcal{S}^5$ sin valores nulos.

Ejemplo 5.1 Supongamos que se ha decidido aplicar una amalgama a la observación $\mathbf{x} = (0.2, 0.1, 0, 0, 0.3, 0.1, 0.1, 0.2) \in \mathcal{S}^8$ para obtener una observación de $\mathcal{S}^5$ a partir de la matriz

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

es decir aglutinamos la segunda y la tercera componentes, la cuarta y la quinta componentes, y la sexta y la séptima componentes. La observación resultante $\mathbf{t}^t = \mathbf{A}\mathbf{x}^t$ es la observación sin ceros $\mathbf{t} = (0.2, 0.1, 0.3, 0.2, 0.2)$ perteneciente a $\mathcal{S}^5$. □

Observemos que uno de los efectos de la operación amalgama es eliminar la presencia de valores nulos en las componentes. En consecuencia, la amalgama debe considerarse como una fase previa a la realización de una clasificación. Esta operación debe realizarse teniendo siempre muy presente la propia naturaleza de las componentes a aglutinar.

En Martín-Fernández et al. (1997) la estrategia seguida para realizar una clasificación incorporaba la operación amalgama. En ese trabajo los autores analizan datos de naturaleza *granulométrica*. Mediante el trabajo de campo se habían recogido 1281 muestras de sedimentos marinos en diferentes puntos geográficos del fondo del Mar Báltico de una área conocida como *Darss Sill*. Los componentes arenosos de estos sedimentos se separaron según el tamaño del grano en 8 componentes ordenadas de mayor a menor tamaño. Entonces, los datos composicionales a considerar consistían en el porcentaje en peso de cada tamaño de grano respecto del peso total de la muestra recogida. El propósito del estudio consistió en agrupar las muestras según su composición granulométrica para facilitar la confección de un mapa del fondo marino

donde se reflejaran las diferentes zonas según el tipo de sedimento. Cuando se empezó a analizar el conjunto de datos $\mathbf{X}$ pertenecientes a $\mathcal{S}^8$ se observó que el 90.8% de las 1281 observaciones contenían algún valor nulo y que de los 1281×8 elementos de la matriz de datos un 27.8% eran ceros. Estos valores nulos estaban concentrados mayoritariamente en las 4 primeras componentes, es decir en los 4 tamaños mayores de grano. En consecuencia se decidió aglutinar las cuatro primeras componentes y se obtuvo un conjunto de observaciones de $\mathcal{S}^5$ de las cuales un 12.5% contenían algún valor nulo y, únicamente, el 2.7% de los elementos de la matriz de datos amalgamada eran nulos. Superada esta fase previa, los autores asumieron que los ceros restantes eran ceros por redondeo y aplicaron un tratamiento de reemplazamiento de estos ceros por valores relacionados con el umbral de detección del proceso de medida, para después realizar la clasificación automática propiamente dicha.

En general, una vez se ha superado la fase de amalgama de los datos deberá decidirse si se asumen los valores nulos como ceros esenciales o como ceros por redondeo. En la aplicación de la mayor parte de técnicas estadísticas, el tratamiento de los datos composicionales depende del tipo de cero que contengan. En la sección siguiente presentamos una estrategia a seguir en la aplicación de una técnica de clasificación en el caso de un conjunto de datos composicionales con ceros esenciales.

5.3 Ceros esenciales

En el contexto de las clasificaciones automáticas la presencia de un cero esencial en una componente de una observación nos informa que, en relación a otra observación que en la misma componente contenga un valor no nulo, estas dos observaciones deben pertenecer a grupos diferentes. Esta idea se cita en Doveton (1998) en referencia al *Problema del Martini Perfecto*, del inglés *Perfect Martini Problem*. Se considera que un *Martini* es una bebida consistente en una mezcla, en diferentes proporciones, de *ginebra*, de *vermut seco*, y de *vermut dulce*. La idea fundamental que aparece en este ejemplo es que una observación –*bebida*– que contenga un cero esencial en una componente –*le falta un ingrediente*– no es un Martini, sino una bebida diferente. En consecuencia, cuando se está interesado en realizar una clasificación automática de un conjunto de datos con observaciones que contienen ceros esenciales, estos valores nulos juegan un papel de atributos que separan a las observaciones entre si, según el número y la disposición de sus ceros. De esta manera, dos observaciones $\mathbf{x}$ y $\mathbf{x}^*$ inicialmente pertenecen al mismo grupo si son observaciones con *ceros comunes*, es decir con el mismo número y disposición de los valores nulos. A partir de esta *preclasificación* inicial, y dentro de cada grupo, aplicaremos un método

de clasificación automática para analizar la existencia de agrupaciones de observaciones. Este método de clasificación utilizará una medida de diferencia adecuada para datos composicionales y, dentro de cada grupo, tendrá en cuenta únicamente las componentes no nulas.

5.3.1 Preclasificación binaria de datos con ceros esenciales

Recordemos que el objetivo final del análisis que nos ocupa es obtener una clasificación de los datos de un conjunto $\mathbf{X}$ mediante técnicas automáticas. Por este motivo es importante desarrollar un método automático para obtener la preclasificación. En esta tesis proponemos una técnica muy simple consistente en una *codificación* binaria de las observaciones del conjunto $\mathbf{X}$. Esta codificación de cada observación $\mathbf{x}_i \in \mathbf{X}$, $i = 1, 2, \ldots, n$, consiste en obtener un vector binario $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iD}) \in \{0, 1\}^D$ mediante la siguiente expresión:

$$v_{ik} = \begin{cases} 0, & \text{si } x_{ik} = 0, \\ 1, & \text{si } x_{ik} > 0. \end{cases} \tag{5.1}$$

Resulta obvio que la codificación anterior satisface la propiedad inyectiva respecto al número y la disposición de los valores nulos. En consecuencia, dos observaciones con ceros comunes se codificarán mediante el mismo vector binario. Adicionalmente, es factible convertir cada vector binario $\mathbf{v}_i$ en un número entero c_i mediante la expresión

$$c_i = \sum_{k=1}^{D} v_{ik} 2^k. \tag{5.2}$$

Es sencillo demostrar que se cumple que $2 < c_i < 2^{D+1} - 2$, $i = 1, 2, \ldots, n$. Como resultado del proceso de combinar las expresiones (5.1) y (5.2), obtendremos un vector de números enteros que será el que indica a qué grupo pertenece cada observación. Este vector tiene tantos números enteros diferentes como grupos distintos de observaciones con ceros comunes tengamos en el conjunto $\mathbf{X}$.

La técnica automática de clasificación prosigue aplicando a cada grupo por separado el método de clasificación automático que se haya elegido. La disimilitud que se utilice deberá aplicarse sobre las componentes no nulas de las observaciones. En particular, si $\mathbf{x}$ y $\mathbf{x}^*$ son dos observaciones con C ceros comunes se asumirá que $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^{D-C}$ y se aplicará la expresión de la disimilitud para observaciones del símplex $\mathcal{S}^{D-C}$.

5.3.2 Preclasificación divisiva de datos con ceros esenciales

De manera natural surge otra estrategia para realizar de forma automática la preclasificación de un conjunto de datos composicionales $\mathbf{X} \in \mathcal{S}^D$ con ceros esenciales. Esta estrategia se inspira

en los algoritmos jerárquicos *divisivos* presentados en la Sección 1.6.1 del Capítulo 1 de esta tesis. Inicialmente se considera que únicamente existe un grupo y éste es el conjunto total de observaciones $\mathbf{X}$. En una primera fase del algoritmo, el conjunto $\mathbf{X}$ se divide en dos grupos. Esta división se realiza escogiendo una de las D componentes y asignando cada observación del conjunto $\mathbf{X}$ a uno de los dos subconjuntos según si en la componente escogida la observación tiene o no el valor nulo. En una segunda etapa del algoritmo, se escoge otra componente, diferente de la primera, y se subdivide cada uno de los dos subconjuntos en dos grupos según si las observaciones tienen o no el valor cero en la componente escogida. El final del algoritmo se alcanza cuando se han usado las D componentes para subdividir los grupos existentes. En consecuencia, la preclasificación resultante estará formada por un máximo de 2^D grupos diferentes. La estructura jerárquica divisiva construida podrá representarse mediante un dendrograma de $D+1$ niveles de división. Existe la posibilidad de disminuir el tiempo de ejecución del algoritmo si únicamente subdividimos a partir de aquellas componentes que contienen algún valor nulo. En particular, si un conjunto $\mathbf{X}$ tiene los ceros esenciales concentrados en C componentes, entonces el algoritmo constará de C etapas divisivas y la estructura jerárquica estará formada por $C+1$ niveles de fusión. De manera análoga a la preclasificación binaria, a los grupos resultantes de esta preclasificación divisiva se les aplicará por separado el método de clasificación automática escogido.

Queremos resaltar que, en realidad, estas dos técnicas de preclasificación son totalmente equivalentes. Observemos que en la técnica de preclasificación divisiva podemos asociar inicialmente a cada observación $\mathbf{x}_i$ del conjunto $\mathbf{X}$ el vector binario unitario $\mathbf{v}_i = (1, 1, \ldots, 1) \in \{0, 1\}^D$. A continuación, en los niveles de división correspondientes a la k-ésima componente, conservamos el 1 en la k-ésima componente del vector $\mathbf{v}_i$ si la componente en cuestión de la observación $\mathbf{x}_i$ no es nula, o cambiamos el 1 por un 0 si la componente de la observación es un cero. Al final del proceso divisivo cada observación $\mathbf{x}_i$ tendrá asociado el vector binario $\mathbf{v}_i$ cuya expresión aparece en la fórmula (5.1).

5.4 Ceros por redondeo

En esta sección nos centramos en el problema de los ceros por redondeo en el contexto de las clasificaciones automáticas no paramétricas. Este tipo de valor nulo que aparece en una componente es un dato que se ha traducido por un cero debido a que corresponde a valores que no han sido registrados o detectados por ser valores extremadamente pequeños. Es decir, en una componente de la observación aparece un cero que proviene de un dato censurado por tener

un valor inferior al umbral de detección de la variable en cuestión. Este umbral de detección se deriva de la precisión con la que se trabaja en el proceso de medida. Primeramente presentaremos el tratamiento propuesto en Aitchison (1986) para este tipo de cero y analizaremos los inconvenientes que presenta. Realizaremos una breve presentación de los aspectos básicos del tratamiento de datos censurados en el espacio $\mathbb{R}^D$ y presentaremos un nuevo método de reemplazamiento (Martín-Fernández et al., 2000) para el caso de datos composicionales.

5.4.1 Reemplazamiento de tipo aditivo

En Aitchison (1986) se propone reemplazar un cero no esencial de una observación $\mathbf{x} \in \mathcal{S}^D$ y construir un dato composicional $\mathbf{r} = (r_1, r_2, \ldots, r_D)$ utilizando la fórmula siguiente:

$$r_k = \begin{cases} \frac{\delta(C+1)(D-C)}{D^2}, & \text{si } x_k = 0, \\ x_k - \frac{\delta(C+1)C}{D^2}, & \text{si } x_k > 0, \end{cases} \tag{5.3}$$

donde C es el número de ceros presentes en la observación $\mathbf{x}$ y δ es un número real positivo inferior al valor del umbral de detección del proceso de medida.

En la bibliografía se encuentran, entre otros, los trabajos de Martín-Fernández et al. (1997), Zhou (1997) y Tauber (1999), donde se utiliza y analiza este tipo de reemplazamiento. En los trabajos de Tauber y de Zhou Di se pone especial énfasis, desde un punto de vista descriptivo, en la aparición de grupos espúreos debido a la extrema sensibilidad de la distancia de Aitchison en función del valor δ que se considere.

Observemos que la restricción de que la suma de las componentes sea igual a uno obliga en el reemplazamiento (5.3) a modificar las componentes no nulas de la observación $\mathbf{x}$. El hecho que la modificación de las componentes no nulas sea *aditiva* es el principal inconveniente para que este reemplazamiento sea coherente con las operaciones básicas definidas en el símplex. Desde un punto de vista teórico, cabe resaltar que el reemplazamiento (5.3) presenta los siguientes inconvenientes:

i. No parece razonable que el valor de la substitución $\frac{\delta(C+1)(D-C)}{D^2}$, por el que se substituye un cero presente en la observación $\mathbf{x}$, dependa de la cantidad total de ceros C presentes en $\mathbf{x}$.

ii. Consideremos el caso de dos observaciones $\mathbf{x}$ y $\mathbf{x}^*$ con ceros comunes, es decir con valores nulos en las mismas componentes. En este caso, el reemplazamiento (5.3) no satisface que $d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*) = d_{\text{Ait}}(\mathbf{x}_s, \mathbf{x}_s^*)$, donde $\mathbf{x}_s$ y $\mathbf{x}_s^*$ representan, respectivamente, las *máximas subcomposiciones no nulas comunes*. Es sencillo comprobar que, en este caso, el valor de

$d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*)$ depende del valor δ y que se cumple que:

$$\lim_{\delta \to 0^+} d^2_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*) = d^2_{\text{Ait}}(\mathbf{x}_s, \mathbf{x}^*_s) + \frac{C}{D(D-C)} \left[\sum_{x_l \neq 0} \log \left(\frac{x_l}{x^*_l} \right) \right]^2 .$$

iii. En el caso de dos observaciones $\mathbf{x}$ y $\mathbf{x}^*$ con ceros no comunes la substitución (5.3) satisface que:

- $\lim_{\delta \to 0^+} d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*) = +\infty$,
- $\lim_{\delta \to 1^-} d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*) = +\infty$.

iv. El reemplazamiento (5.3) no es compatible con la operación de formación de subcomposiciones en el sentido que no conserva las proporciones entre datos observados no nulos. Más concretamente:

- $\frac{r_k}{r_l} \neq \frac{x_k}{x_l}$, si $x_k > 0, x_l > 0$,
- $\frac{r_k}{r^*_k} \neq \frac{x_k}{x^*_k}$, si $x_k > 0, x^*_k > 0$.

Queremos poner énfasis que este inconveniente implica que para conjuntos de datos en los que existan componentes que sean no nulas para todas las observaciones, la estructura de covarianza de la subcomposición formada por estas componentes se distorsiona al aplicar el reemplazamiento (5.3) propuesto por Aitchison.

En el contexto de las clasificaciones automáticas se observa que, si se consideran valores δ muy pequeños las observaciones que contienen valores nulos tienden, como consecuencia de estas propiedades, a formar grupos separados de las observaciones que no contienen ceros. Análogamente, las observaciones que contengan ceros tienden a subdividirse en grupos según el número de ceros que contienen y según las componentes que los contienen. Por lo tanto, si los valores δ tienden a cero la clasificación resultante tenderá a coincidir con la clasificación que se obtendría si los valores nulos se considerasen ceros esenciales.

Con el propósito de formular un nuevo método de reemplazamiento de ceros por redondeo que sea coherente con el carácter composicional de los datos, presentamos en la sección siguiente los aspectos básicos que aparecen en el análisis estadístico de conjuntos de datos en $\mathbb{R}^D$ que contengan datos censurados.

5.4.2 Datos ausentes en conjuntos de datos de $\mathbb{R}^D$

En esta sección tratamos los aspectos básicos que hacen referencia a conjuntos de datos en el espacio $\mathbb{R}^D$ en los que aparecen observaciones con datos censurados. Es importante precisar

que por dato censurado entendemos aquel dato que no ha podido observarse por tener un valor inferior al umbral de detección de la variable en cuestión. Los aspectos que expondremos aparecen en la bibliografía como características comunes a otro tipo de dato: el dato ausente o, en inglés, *missing data*. Este último es un tipo de dato que generaliza el tipo de dato censurado. Por este motivo en nuestro desarrollo, cuando se considere que no es relevante, se usará indistintamente el término dato ausente o dato censurado. Nuestro interés se centra en aquellos aspectos de los datos censurados que están relacionados con las técnicas de clasificación automática no paramétrica. En particular, observemos que los algoritmos jerárquicos de clasificación producen una clasificación de los datos basándose en la matriz de distancias entre observaciones. En consecuencia, cuando el conjunto de datos a clasificar contenga datos ausentes será necesario *completar* la matriz de distancias entre individuos. A medida que presentamos una breve descripción de algunos de los métodos más usuales en el tratamiento de datos ausentes, analizaremos su posible aplicación en el contexto de los datos composicionales:

Métodos de tratamiento de datos ausentes en $\mathbb{R}^D$

M1. Método de ponderación de las distancias.

Este método ha sido analizado y utilizado en el contexto de las clasificaciones automáticas no paramétricas. En los trabajos de Murtagh y Heck (1987), Krzanowski (1988b) y Murtagh y Hernández-Pajares (1995), los autores exponen que no es recomendable hacer un reemplazamiento de los datos ausentes por la misma cantidad en todas las observaciones porque ello comporta añadir semejanza entre las observaciones y distorsiona la matriz de distancias. Los mismos autores sugieren que una alternativa adecuada consiste en adjudicar como valor de disimilitud entre dos observaciones $\mathbf{x}$, $\mathbf{x}^* \in \mathbb{R}^D$ con datos ausentes la distancia entre los dos *máximos subvectores observados comunes* –$\mathbf{x}_s, \mathbf{x}_s^*$– ponderada por un factor que depende del número de datos no observados según la expresión siguiente:

$$d(\mathbf{x}, \mathbf{x}^*) = \frac{D}{D-C}\, d(\mathbf{x}_s, \mathbf{x}_s^*), \tag{5.4}$$

donde $\mathbf{x}_s, \mathbf{x}_s^* \in \mathbb{R}^{D-C}$ y $D - C$ es el número de datos no ausentes o datos observados conjuntamente en $\mathbf{x}$ y $\mathbf{x}^*$.

Si se desea utilizar este método como aproximación al problema de los ceros en los datos composicionales debe utilizarse una medida de diferencia coherente con la naturaleza de los datos, y el concepto de máximo subvector observado común debe reemplazarse por el concepto de *máxima subcomposición no nula común*. Sin embargo, la utilización de este método posee los siguientes inconvenientes:

- Si las dos observaciones no tienen subvector común la distancia no queda definida.
- Es posible encontrar situaciones incongruentes.

Consideremos, por ejemplo, las observaciones $\mathbf{x} = (0,\ 0.8,\ 0.2)$, $\mathbf{x}^* = (0.95,\ 0.04,\ 0.01)$, y $\mathbf{x}' = (0.06,\ 0.76,\ 0.18)$ que aparecen en la figura 5.1. Adaptando para datos composicionales la estrategia propuesta por Krzanowski (1988b) para datos en $\mathbb{R}^D$, formamos las subcomposiciones respecto a la segunda y tercera componente y se obtienen las observaciones:

$$\mathbf{x}_s = (0.8, 0.2),\ \mathbf{x}_s^* = (0.8, 0.2),\ \text{y}\ \mathbf{x}_s' = (0.81, 0.19).$$

Suponiendo que el cero en la observación $\mathbf{x}$ es realmente un cero por redondeo, se espera que las observaciones $\mathbf{x}$ y $\mathbf{x}'$ sean más similares que las observaciones $\mathbf{x}$ y $\mathbf{x}^*$. Sin embargo se obtiene que $d_{\text{Ait}}(\mathbf{x}_s, \mathbf{x}_s^*) = 0$ y que $d_{\text{Ait}}(\mathbf{x}_s, \mathbf{x}_s') = 0.07$.

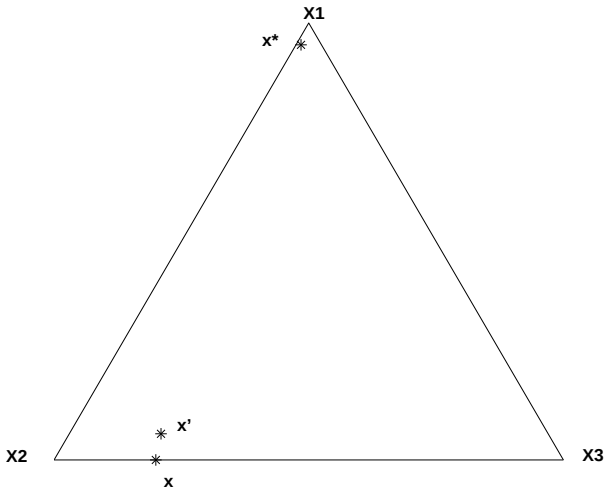


Figura 5.1: *Diagrama ternario con las observaciones* $\mathbf{x} = (0, 0.8, 0.2)$, $\mathbf{x}^* = (0.95, 0.04, 0.01)$ *y* $\mathbf{x}' = (0.06, 0.76, 0.18)$.

M2. Método de reemplazamiento de los datos ausentes basado en la descomposición en valores singulares de la matriz de datos.

Este método no paramétrico, que se propone en Krzanowski (1988a), consiste en el reemplazamiento de los datos ausentes x_{ik} de un conjunto de datos $\mathbf{X}$ por valores r_{ik} obtenidos mediante un algoritmo iterativo basado en la descomposición en valores singulares de la matriz de datos $\mathbf{X}$. El algoritmo consta de las siguientes etapas:

1. Se realiza un reemplazamiento inicial de todos los datos no observados x_{ik} por un mismo valor $r_{ik} = r_0$. En general, los datos ausentes se substituyen por la media

aritmética de los datos observados en la componente o, en el caso más concreto de los datos censurados, se substituyen por un valor relacionado con el valor umbral de detección derivado del proceso de medida.

2. Para cada dato no observado x_{ik} de la matriz de datos $\mathbf{X}$ se deben realizar dos descomposiciones en valores singulares: una, la de la matriz $\mathbf{X}$ sin la fila i, $\mathbf{X}_{-i}$, y otra, la de la matriz $\mathbf{X}$ sin la columna k, $\mathbf{X}^{-k}$. Las siguientes expresiones representan las dos descomposiciones:

$$\begin{cases} \mathbf{X}_{-i} = & U_i D_i V_i, \\ \mathbf{X}^{-k} = & U^k D^k V^k. \end{cases}$$

Entonces, el valor r_{ik} que substituye al dato no observado x_{ik} se obtiene a partir de la siguiente expresión:

$$r_{ik} = \sum_{t=1}^{D-1} [(u^k)_{it} \sqrt{(d^k)_t}][[(v_i)_{tk} \sqrt{(d_i)_t}].$$

3. Volver al paso anterior hasta conseguir estabilidad en la solución.

Si se desea aplicar este reemplazamiento en el contexto de las clasificaciones automáticas no paramétricas aparece un inconveniente que el mismo autor menciona en su artículo: cuando se sabe a priori que en el conjunto de datos existen grupos de individuos entonces debe aplicarse el algoritmo para cada grupo de observaciones por separado. Por lo tanto, en el contexto de las clasificaciones automáticas, donde a priori no se conocen los grupos, el algoritmo es manifiestamente insuficiente. En el mismo trabajo Krzanowski (1988a) expone que en este caso es conveniente basarse en un algoritmo de descomposición en valores singulares (Rao y Mitra, 1971) más general que el expuesto en este apartado. Una línea de investigación futura que nos proponemos es la adaptación de este método de substitución de datos ausentes al contexto de los datos composicionales. Una posible estrategia pasará por adaptar el algoritmo para trabajar con los datos en el símplex, de manera que se incorpore en el algoritmo la restricción de suma de las componentes igual a uno. En los casos de conjuntos de datos composicionales que contengan una componente con valores no nulos en todas las observaciones nos aparece otra posible estrategia: aplicar la transformación alr con denominador la componente no nula y aplicar el algoritmo a los datos alr-transformados. Para poder desarrollar esta segunda estrategia, deberemos analizar si los resultados dependen o no del denominador escogido en la transformación alr.

M3. Método de reemplazamiento de los datos ausentes basado en técnicas de regresión lineal múltiple.

Este método, basado en técnicas de regresión lineal múltiple, se encuentra desarrollado en profundidad en Little y Rubin (1987). Las técnicas de regresión múltiple se aplican sobre las componentes que no contienen datos ausentes, para inferir a partir de ellas el valor de los datos no observados. El método consiste en un algoritmo iterativo de reemplazamiento y inferencia que se aplica hasta conseguir convergencia o estabilidad en la solución. Según se expone en Little y Rubin (1987), este método es útil cuando se asume que el valor no observado en una componente depende únicamente de los valores observados en la observación y no depende del dato ausente en si mismo.

Si se desea adaptar este método al caso de los ceros en datos composicionales nos encontramos con las siguientes dificultades:

- En nuestro caso los ceros son datos de tipo censurado. Por lo tanto, el valor no observado también depende de la componente en cuestión.
- En el contexto de las clasificaciones automáticas de conjuntos de datos este método puede ser ineficaz. El hecho de existir diferentes grupos de observaciones puede ser debido precisamente a que las relaciones entre las componentes son diferentes para grupos diferentes. En este caso carece de sentido encontrar una relación *global* entre las componentes que contienen datos observados y las que contienen datos ausentes.

M4. Método de reemplazamiento de los datos ausentes basado en técnicas paramétricas.

El desarrollo de estos métodos paramétricos de reemplazamiento se basa fundamentalmente en la maximización de la función de verosimilitud o en el algoritmo *EM*. En ambos casos estas técnicas se aplican bajo hipótesis de normalidad multivariante de los datos. En estos métodos se asume otra hipótesis esencial: todos las observaciones son *iid*. En particular, se asume que las observaciones con datos ausentes siguen el mismo modelo de distribución que las observaciones con datos observados. Estos métodos son útiles (Little y Rubin, 1987) tanto para el caso de datos ausentes que dependen únicamente de los datos observados, como para el caso de datos no observados totalmente aleatorios. En este último caso se calcula en cada componente el valor esperado de los datos no observados. Entonces se procede a reemplazar los datos ausentes por el valor esperado obtenido. Esta estrategia ha sido utilizada en el trabajo de Sandford et al. (1993) para conjuntos de datos composicionales con ceros.

Sin embargo, al aplicar estos métodos sobre datos composicionales aparecen las siguientes dificultades:

- Si en el conjunto de datos existen grupos diferenciados de observaciones, la hipótesis de que las observaciones son *iid* no puede aceptarse. En este caso, se hace necesario reformular la función de máxima verosimilitud con lo que la complejidad del método aumenta.
- Este método es aplicable a los datos composicionales una vez se han transformado mediante la aplicación alr puesto que entonces puede asumirse la hipótesis de normalidad multivariante de los datos. Entonces, se hace necesario disponer de una componente que no contenga ningún valor nulo para poder considerarla como el denominador de la transformación alr.

A la vista de lo expuesto para cada uno de los diferentes métodos de tratamiento de datos ausentes, en esta tesis nos limitamos a estudiar en profundidad un tratamiento consistente en el reemplazamiento de un cero en una observación composicional mediante una estrategia similar a la que propone Aitchison (1986) y que hemos presentado en este capítulo mediante la expresión (5.3). En el trabajo de Sandford et al. (1993) esta estrategia se conoce como *simple-substitución*. En el mismo trabajo el autor analiza y compara, para datos en $\mathbb{R}^D$, dos diferentes posibilidades para una simple-substitución de datos censurados. Las dos posibilidades consisten en reemplazar el dato censurado x_{ik} por un valor r_{ik} igual a un porcentaje del valor del umbral de detección. En las conclusiones de su estudio, Sandford et al. (1993) defiende como mejor alternativa la simple-substitución del dato censurado x_{ik} por el valor r_{ik} igual al 55% del umbral de detección, frente a la otra posibilidad, más habitualmente utilizada, en la que r_{ik} es igual al 75% del umbral. Sin embargo, es necesario aclarar que en el mismo trabajo, Sandford et al. (1993) analizan el reemplazamiento basado en la maximización de la función de máxima verosimilitud y que en las conclusiones de su estudio, consideran este método como el que proporciona mejores resultados. Debido a que nuestro trabajo de investigación está centrado en las técnicas de clasificación automática no paramétrica, decidimos limitar nuestro análisis del reemplazamiento de ceros a los métodos no paramétricos y dejamos como línea de investigación futura el trabajo de adaptar el resto de métodos de tratamiento de datos ausentes al caso de los ceros en datos composicionales.

En la siguiente sección presentamos con detalle la relación existente entre una estrategia de reemplazamiento del tipo simple-substitución para datos ausentes en el espacio $\mathbb{R}^D$ y la distancia euclídea. El propósito es analizar con detalle esta relación con el objetivo de adaptar

sus características principales al caso de los datos composicionales y la distancia de Aitchison.

5.4.3 Simple-substitución de datos ausentes en $\mathbb{R}^D$ y la distancia euclídea

En el contexto de las clasificaciones no paramétricas, la medida de diferencia entre dos observaciones de $\mathbb{R}^D$ es un elemento fundamental del análisis. Por ser la medida más habitual, nos centraremos en el estudio de las clasificaciones que utilizan la distancia euclídea. En este apartado exponemos algunas consideraciones a tener en cuenta cuando se pretende reemplazar datos no observados en una componente por un mismo valor fijo para todos los individuos. Este tipo de substitución es el tipo de reemplazamiento propuesto por Aitchison (1986) en su monografía.

Consideremos el caso de una observación $\mathbf{x}$ cuya k-ésima componente ha sido censurada por ser un valor extremadamente pequeño. La simple-substitución consiste en reemplazar estas componentes por valores $r_k = \delta_k$, siendo δ_k un valor que se deriva del umbral de detección ligado al proceso de medida en la k-ésima componente. La simple-substitución de la observación $\mathbf{x}$ por la observación $\mathbf{r} = (r_1, r_2, \ldots, r_D)$ se puede formular mediante la siguiente expresión:

$$r_k = \begin{cases} \delta_k, & \text{si } x_k \text{ es censurado;} \\ x_k, & \text{si } x_k \text{ no es censurado;} \end{cases} \qquad \forall k = 1, 2, \ldots, D. \tag{5.5}$$

Obsérvese que este planteamiento puede extenderse de manera sencilla a datos censurados por ser valores extremadamente grandes.

Al analizar las principales características de la relación entre una simple-substitución y la distancia euclídea se han observado los siguientes propiedades:

p1. Si $\mathbf{x},\ \mathbf{x}^* \in \mathbb{R}^D$ son dos observaciones con *datos censurados comunes*, es decir, con la misma cantidad de datos censurados y situados en las mismas componentes, se obtiene que

$$d_{\text{Euc}}(\mathbf{r}, \mathbf{r}^*) = d_{\text{Euc}}(\Pi(\mathbf{x}), \Pi(\mathbf{x}^*)) = d_{\text{Euc}}(\mathbf{x}_s, \mathbf{x}_s^*),$$

donde Π representa la operación proyección canónica en las componentes sin datos ausentes, y $\mathbf{x}_s, \mathbf{x}_s^*$ los dos máximos subvectores observados comunes. Por lo tanto, la distancia euclídea entre observaciones reemplazadas $\mathbf{r}, \mathbf{r}^*$ no depende ni del valor δ_k ni de la cantidad de valores ausentes.

p2. En el caso de observaciones con datos censurados comunes las diferencias entre componentes se conservan. Es decir:

- $x_k - x_l = r_k - r_l$, si x_k, x_l son datos no censurados,

- $x_k - x_k^* = r_k - r_k^*$, si x_k, x_l son datos no censurados.

p3. Si $\mathbf{x},\ \mathbf{x}^* \in \mathbb{R}^D$ son dos observaciones con datos censurados no comunes se obtendrá que

$$\lim_{\delta_k \to \pm\infty} d_{\text{Euc}}(\mathbf{r}, \mathbf{r}^*) = +\infty.$$

p4. Si $\mathbf{x},\ \mathbf{x}^* \in \mathbb{R}_+^D$ son dos observaciones con datos censurados no comunes y con componentes no negativas se tiene que:

- $\lim_{\delta_k \to +\infty} d_{\text{Euc}}(\mathbf{r}, \mathbf{r}^*) = +\infty$,
- $\lim_{\delta_k \to 0^+} d_{\text{Euc}}(\mathbf{r}, \mathbf{r}^*) < +\infty$.

Podemos generalizar esta propiedad para el caso de una componente que contenga valores censurados y que esté acotada, $m \leq \mathbf{x}_k \leq M$, mediante las expresiones

- $\lim_{\delta_k \to m^+} d_{\text{Euc}}(\mathbf{r}, \mathbf{r}^*) < +\infty$,
- $\lim_{\delta_k \to M^-} d_{\text{Euc}}(\mathbf{r}, \mathbf{r}^*) < +\infty$.

Teniendo presente en todo momento estas propiedades, vamos a analizar el problema de los ceros por redondeo en los datos composicionales. Para este tipo de datos el espacio muestral es el símplex $\mathcal{S}^D$ y una medida de diferencia adecuada es la distancia de Aitchison. En consecuencia, en la propiedad p1 los conceptos proyección y máximo subvector común deben substituirse por los conceptos formación de subcomposición y máxima subcomposición común. Análogamente, en la propiedad p2 la conservación de la diferencia entre componentes se traducirá por la conservación de la proporción entre componentes, y en la propiedad p3 los límites para $\delta_k \to \pm\infty$ deben ser substituidos por los límites para $\delta_k \to 0^+$ y $\delta_k \to 1^-$, respectivamente. Observemos que, a diferencia de los casos de observaciones en $\mathbb{R}^D$, cualquier reemplazamiento de un dato composicional censurado implica que los demás valores en la observación también deben ser modificados. Ello es debido a la restricción de la suma de componentes igual a uno. Esta particularidad hace imposible conseguir un reemplazamiento que, para observaciones con ceros comunes, conserve la distancia de Aitchison entre las máximas subcomposiciones comunes. Sin embargo, la propiedad p1 nos sugiere que parece lógico exigir que la distancia de Aitchison entre observaciones reemplazadas no dependa del valor δ_k utilizado en el reemplazamiento.

5.4.4 Reemplazamiento de tipo multiplicativo

Las dificultades que presenta el reemplazamiento (5.3) propuesto por Aitchison y el interés por buscar una substitución coherente con el carácter composicional de los datos nos ha motivado (Martín-Fernández et al., 2000) a buscar otra fórmula de reemplazamiento de los ceros por

redondeo. En esta tesis proponemos una nueva aproximación al problema basada en el reemplazamiento de los ceros por redondeo mediante una fórmula que tenga buenas propiedades respecto de las operaciones perturbación y formación de subcomposiciones. Consideremos δ_k el valor del reemplazamiento derivado del umbral de detección para la k-ésima componente. Sea $\mathbf{x}$ una observación que contenga ceros por redondeo. Entonces, construimos la observación $\mathbf{r} = (r_1, r_2, \ldots, r_D)$ substituyendo los ceros de $\mathbf{x}$ mediante la expresión siguiente:

$$r_k = \begin{cases} \delta_k & \text{si } \; x_k = 0, \\ x_k \left(1 - \sum_{\{x_l \,=0\}} \delta_l\right) & \text{si } \; x_k > 0. \end{cases} \tag{5.6}$$

Obsérvese que, a diferencia del reemplazamiento (5.3), en la fórmula anterior la modificación de las componentes no nulas de la observación $\mathbf{x}$ es una modificación de tipo *multiplicativo*. De esta característica se deriva que el reemplazamiento (5.6) posee las siguientes propiedades:

P1. El reemplazamiento (5.6) es una substitución que surge de manera natural si pensamos en el proceso de censura de los datos que no superan un umbral establecido para la k-ésima componente. Sin pérdida de generalidad podemos considerar que la censura se ha producido sobre las observaciones de $\mathbb{R}_+^D$ que dan lugar a la composición y sólo en la k-ésima componente. Denominamos $\mathbf{w} = (w_1, w_2, \ldots, w_D)$ a la observación de $\mathbb{R}_+^D$ que, una vez censurada y normalizada, toma el valor:

$$\mathbf{x} = (x_1, x_2, \ldots, x_D) = \left(\frac{w_1}{\sum_{l\neq k} w_l}, \frac{w_2}{\sum_{l\neq k} w_l}, \ldots, \frac{w_{k-1}}{\sum_{l\neq k} w_l}, 0, \frac{w_{k+1}}{\sum_{l\neq k} w_l}, \ldots, \frac{w_D}{\sum_{l\neq k} w_l}\right).$$

Una vez realizado el reemplazamiento obtenemos:

$$\mathbf{r} = (x_1(1-\delta_k), \ldots, x_{k-1}(1-\delta_k), \delta_k, x_{k+1}(1-\delta_k), \ldots, x_D(1-\delta_k)).$$

Obsérvese que si en la substitución se utiliza el *verdadero valor* $\delta_k = \frac{w_k}{\sum w_l}$ entonces, se obtiene el *verdadero valor* de la observación:

$$\mathbf{r} = \left(\frac{w_1}{\sum w_l}, \frac{w_2}{\sum w_l}, \ldots, \frac{w_{k-1}}{\sum w_l}, \frac{w_k}{\sum w_l}, \frac{w_{k+1}}{\sum w_l}, \ldots, \frac{w_D}{\sum w_l}\right).$$

Si la observación tiene más de un cero o si la censura se realiza sobre las observaciones $\mathbf{x}$ en el símplex, puede razonarse de manera análoga.

P2. La simple-substitución (5.6) es más general que el reemplazamiento (5.3) propuesto por Aitchison puesto que el valor δ_k no tiene porqué depender ni del número de ceros C presentes en la observación $\mathbf{x}$ ni de la dimensión D del símplex. Siguiendo lo propuesto en Sandford et al. (1993) puede optarse por considerar δ_k igual al 55% del valor del umbral del proceso de medida.

P3. El reemplazamiento (5.6) tiene propiedades razonables respecto la operación de formación de subcomposiciones en el sentido que conserva las proporciones entre datos observados no nulos. Más concretamente si $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$ son dos observaciones con ceros por redondeo, entonces se cumple que:

- $\frac{r_k}{r_l} = \frac{x_k}{x_l}$, si $x_k > 0, x_l > 0$,
- $\frac{r_k}{r_k^*} = \frac{x_k}{x_k^*}$, si $x_k > 0, x_k^* > 0$,

donde la segunda igualdad solo es cierta para observaciones $\mathbf{x}$ y $\mathbf{x}^*$ con ceros comunes. Para el caso particular en que $\delta_k = \delta$, $\forall\, k = 1, 2, \ldots, D$, esta propiedad también se cumple para observaciones $\mathbf{x}$ y $\mathbf{x}^*$ con la misma cantidad de ceros por redondeo aunque no sean comunes. A partir de esta propiedad es fácil demostrar que se verifica la igualdad siguiente: $\mathbf{r}_s = \mathbf{x}_s$, donde s es cualquier subcomposición formada por componentes no nulas de $\mathbf{x}$. La figura 5.2 muestra un ejemplo donde se visualiza esta propiedad. Obsérvese que esta igualdad implica a su vez que la estructura de covarianzas de cualquier subcomposición formada por componentes no nulas se conserva después de realizar el reemplazamiento (5.6).

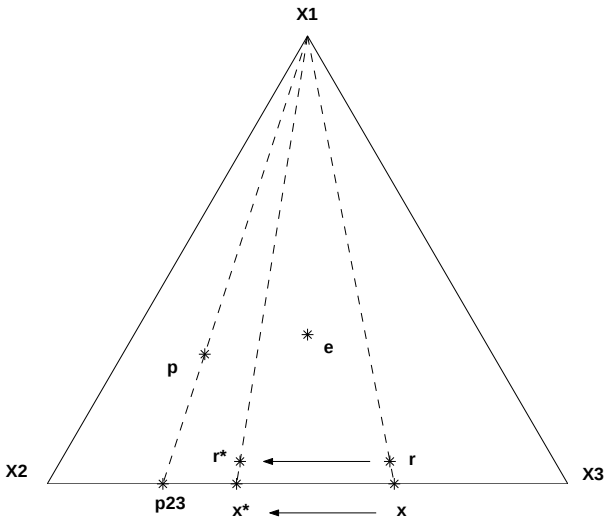


Figura 5.2: *Diagrama ternario con las observaciones* $\mathbf{x} = (0, 1/3, 2/3)$, $\mathbf{x}^* = (0, 0.64, 0.36)$. *Donde* $\mathbf{r} = (0.05, 0.32, 0.63)$ *y* $\mathbf{r}^* = (0.05, 0.6, 0.35)$ *son, respectivamente, las observaciones resultantes del reemplazamiento del cero de* $\mathbf{x}$ *y* $\mathbf{x}^*$ *por* $\delta = 0.05$. *Las observaciones* $\mathbf{p}23 = \mathbf{x}^* \circ \mathbf{x}^{-1}$ *y* $\mathbf{p} = \mathbf{r}^* \circ \mathbf{r}^{-1}$ *son las perturbaciones que nos transforman, respectivamente,* $\mathbf{x}$ *y* $\mathbf{r}$ *en* $\mathbf{x}^*$ *y* $\mathbf{r}^*$. *Las líneas discontinuas representan la operación subcomposición en las componentes* $\mathbf{X}_2$ *y* $\mathbf{X}_3$.

P4. El reemplazamiento (5.6) es compatible con las perturbaciones en el sentido que si se

perturban los valores no nulos de la observación $\mathbf{x}$ o se perturban los de la observación asociada $\mathbf{r}$, las proporciones entre las componentes no nulas se conservan. Es decir, se verifica la igualdad siguiente:

$$(\mathbf{p} \circ \mathbf{x})_s = (\mathbf{p} \circ \mathbf{r})_s,$$

donde s es cualquier subcomposición formada por variables no nulas de $\mathbf{x}$, y $\mathbf{p} \in \mathcal{S}^D$ es cualquier perturbación. En la figura 5.2 puede observarse la compatibilidad entre la operación perturbación y el reemplazamiento multiplicativo.

P5. En el caso de dos observaciones $\mathbf{x}$, $\mathbf{x}^* \in \mathcal{S}^D$ con ceros comunes el reemplazamiento (5.6) no satisface que $d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*) = d_{\text{Ait}}(\mathbf{x}_s, \mathbf{x}_s^*)$, donde s representa la máxima subcomposición común no nula. Sin embargo, a diferencia de la substitución (5.3), usando la substitución (5.6) la distancia $d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*)$ no depende del valor δ_k. Adicionalmente puede demostrarse que para $\delta \to 0^+$ la distancia de Aitchison entre $\mathbf{r}$ y $\mathbf{r}^*$, calculada con el reemplazamiento (5.3), tiende a la distancia $d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*)$ usando la substitución (5.6). Más concretamente, con la substitución (5.6) que se propone en este trabajo, se verifica que:

$$d_{Ait}^2(\mathbf{r}, \mathbf{r}^*) = d_{Ait}^2(\mathbf{x}_s, \mathbf{x}_s^*) + \frac{C}{D(D-C)} \left[\sum_{x_l \neq 0} \log \left(\frac{x_l}{x_l^*} \right) \right]^2,$$

donde s representa la máxima subcomposición común no nula, y C el número de ceros.

P6. En el caso de dos observaciones $\mathbf{x}$, $\mathbf{x}^* \in \mathcal{S}^D$ con ceros no comunes, la substitución (5.6) verifica que:

- $\lim_{\delta_k \to 0^+} d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*) = +\infty$,
- $\lim_{\delta_k \to 1^-} d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*) = +\infty$.

Obsérvese que, según hemos analizado en este capítulo, estas dos propiedades no son exclusivas ni de las substituciones (5.3) y (5.6), ni de los datos composicionales, puesto que aparecen los mismos problemas si se trabaja con observaciones en $\mathbb{R}^D$ y con la distancia euclídea.

P7. El reemplazamiento (5.6) es compatible con la transformación alr en el sentido que a las componentes no nulas de una observación $\mathbf{x}$ se les aplica el mismo tipo de modificación, tanto si realizamos un reemplazamiento según la expresión (5.6) como si se realiza una simple-substitución al vector $\text{alr}(\mathbf{x}) \in \mathbb{R}^{D-1}$. Para ilustrar esta propiedad podemos considerar, sin pérdida de generalidad, que la observación $\mathbf{x} \in \mathcal{S}^D$ contiene C ceros por redondeo en las C primeras componentes, y el resto son componentes no nulas. Es decir,

$\mathbf{x} = (0, 0, \ldots, 0, x_{C+1}, x_{C+2}, \ldots, x_D)$. Al aplicar la observación $\mathbf{x}$ el reemplazamiento (5.6) obtenemos la composición

$$\mathbf{r} = (\delta_1, \delta_2, \ldots, \delta_C, x_{C+1}(1 - \sum \delta_k), x_{C+2}(1 - \sum \delta_k), \ldots, x_D(1 - \sum \delta_k)). \tag{5.7}$$

Si transformamos la observación $\mathbf{r}$ mediante la aplicación alr resulta el vector $\mathbf{y} = \text{alr}(\mathbf{r}) \in \mathbb{R}^{D-1}$

$$y_k = \begin{cases} \log\left(\frac{\delta_k}{x_D(1-\sum \delta_k)}\right) & \text{si } \; k = 1, 2, \ldots, C, \\ \log\left(\frac{x_k}{x_D}\right) & \text{si } \; k = C+1, C+2, \ldots, D-1. \end{cases} \tag{5.8}$$

Obsérvese que la expresión anterior es equivalente a aplicar una simple-substitución al vector alr($\mathbf{x}$), al considerar las componentes nulas como datos ausentes. Es decir, se considera la observación $\mathbf{x} = (0, 0, \ldots, 0, x_{C+1}, x_{C+2}, \ldots, x_D)$ como una observación con datos ausentes "$*$", que podemos expresar como

$$\mathbf{x}^* = (*, *, \ldots, *, x_{C+1}, x_{C+2}, \ldots, x_D).$$

La observación $\mathbf{x}^*$ la transformamos mediante la aplicación alr en sus componentes observadas y obtenemos el vector $\mathbf{y}^* \in \mathbb{R}^{D-1}$,

$$\mathbf{y}^* = \text{alr}(\mathbf{x}^*) = (*, *, \ldots, *, \log(x_{C+1}/x_D), \log(x_{C+2}/x_D), \ldots, \log(x_{D-1}/x_D)).$$

A este vector $\mathbf{y}^*$ con datos ausentes le aplicamos una simple-substitución siguiendo la expresión (5.5) y obtenemos el vector sin datos ausentes $\mathbf{q}^* = (q_1^*, q_2^*, \ldots, q_{D-1}^*) \in \mathbb{R}^{D-1}$ que podemos expresar como

$$q_k^* = \begin{cases} l_k & \text{si } k = 1, 2, \ldots, C \\ \log(x_k/x_D) & \text{si } \; k = C+1, C+2, \ldots, D-1. \end{cases} \qquad \forall k = 1, 2, \ldots, D-1. \tag{5.9}$$

Un procedimiento análogo se encuentra en Sandford et al. (1993) pero con la sustancial diferencia que estos autores utilizan la transformación logarítmica log en vez de la transformación alr. Recordemos que en el Capítulo 2 presentamos en (2.6) la aplicación agl como la transformación inversa de la aplicación alr. Si aplicamos la transformación agl al vector $\mathbf{q}^*$ obtendremos la observación $\mathbf{r}^* = \text{agl}(\mathbf{q}^*) \in \mathcal{S}^D$, que podemos expresar como

$$r_k^* = \begin{cases} \frac{e^{l_k}}{1+\sum e^{q_m^*}} & \text{si } \; k = 1, 2, \ldots, C, \\ \frac{x_k/x_D}{1+\sum e^{q_m^*}} & \text{si } \; k = C+1, C+2, \ldots, D-1, \\ \frac{1}{1+\sum e^{q_m^*}} & \text{si } \; k = D. \end{cases} \tag{5.10}$$

Tomando $\gamma_k \in \mathbb{R}$ tal que $l_k = \log(\gamma_k / x_D)$ la expresión anterior es equivalente a

$$r_k^* = \begin{cases} \frac{\gamma_k}{1+\sum_{m=1}^{C} \gamma_m} & \text{si } \; k = 1, 2, \ldots, C, \\ \frac{x_k}{1+\sum_{m=1}^{C} \gamma_m} & \text{si } \; k = C+1, C+2, \ldots, D. \end{cases} \tag{5.11}$$

Si en la expresión anterior llamamos $\delta_k^* = \frac{\gamma_k}{1+\sum_{m=1}^{C} \gamma_m}$, resulta que podemos expresar la observación $\mathbf{r}^*$ como

$$r_k^* = \begin{cases} \delta_k^* & \text{si } \; k = 1, 2, \ldots, C, \\ x_k(1 - \sum_{m=1}^{C} \delta_k^*) & \text{si } \; k = C+1, C+2, \ldots, D. \end{cases} \tag{5.12}$$

Comparando esta expresión con la que hemos formulado en (5.7) se observa que la transformación alr y el reemplazamiento (5.6) son compatibles.

Una vez realizado el reemplazamiento de los ceros por redondeo podremos aplicar el método jerárquico de clasificación que creamos conveniente. Una vez obtenida una agrupación de las observaciones surgirá de manera natural la necesidad de realizar un *análisis de sensibilidad*. El problema que se nos plantea en el análisis de sensibilidad de los resultados es estudiar el grado de dependencia de la clasificación obtenida con respecto de los valores δ_k utilizados en el reemplazamiento. Este problema, por lo que se refiere al reemplazamiento (5.3) propuesto por Aitchison, ha sido estudiado desde un punto de vista descriptivo en los trabajos de Tauber (1999) y de Zhou (1997). Recordemos que los valores δ_k se derivan del valor del umbral de detección y que por lo tanto podemos hacer que $\delta_k \to 0^+$. Sin embargo, la propiedad P6 nos muestra que cuando dos observaciones $\mathbf{x}, \mathbf{x}^*$ tienen ceros no comunes, la distancia de Aitchison entre sus respectivas observaciones reemplazadas $\mathbf{r}, \mathbf{r}^*$ tiende a infinito: $\lim_{\delta_k \to 0^+} d_{\text{Ait}}(\mathbf{r}, \mathbf{r}^*) = +\infty$. En consecuencia, las clasificaciones que se obtengan al hacer que $\delta_k \to 0^+$, irán tendiendo a una clasificación en la que las observaciones se agruparán según el número y la disposición de sus ceros. En definitiva, se obtendrá (Martín-Fernández et al., 2000) la clasificación que resultaría si en el inicio del estudio los ceros hubieran sido considerados como ceros de tipo esencial. Esta dificultad no es, a nuestro parecer un defecto inherente al reemplazamiento (5.6) sino que es un comportamiento que es inherente a cualquier tipo de simple-substitución. Pensemos que en el caso de la simple-substitución de datos censurados en $\mathbb{R}^D$ con la distancia euclídea, nos aparece la misma situación al hacer tender $\delta_k \to \pm\infty$. Una estrategia adecuada para realizar un análisis de sensibilidad consiste en hacer variar el valor δ_k en un rango ligado al umbral de detección o al máximo error de redondeo. En particular, si denominamos δ_r al máximo error por redondeo, un rango adecuado (Aitchison, 1986) de variación de los valores δ_k consiste en

$$\frac{\delta_r}{5} \;\leq\; \delta_k \;\leq\; 2\delta_r. \tag{5.13}$$

De acuerdo con el trabajo de Sandford et al. (1993) –el cual propone como valor adecuado para la simple-substitución δ_k el 55% del valor del umbral δ_r– el rango de variación (5.13) es un rango suficientemente amplio.

En la sección siguiente presentamos un caso práctico de clasificación de un conjunto de datos con ceros por redondeo donde aplicamos el tratamiento que acabamos de exponer.

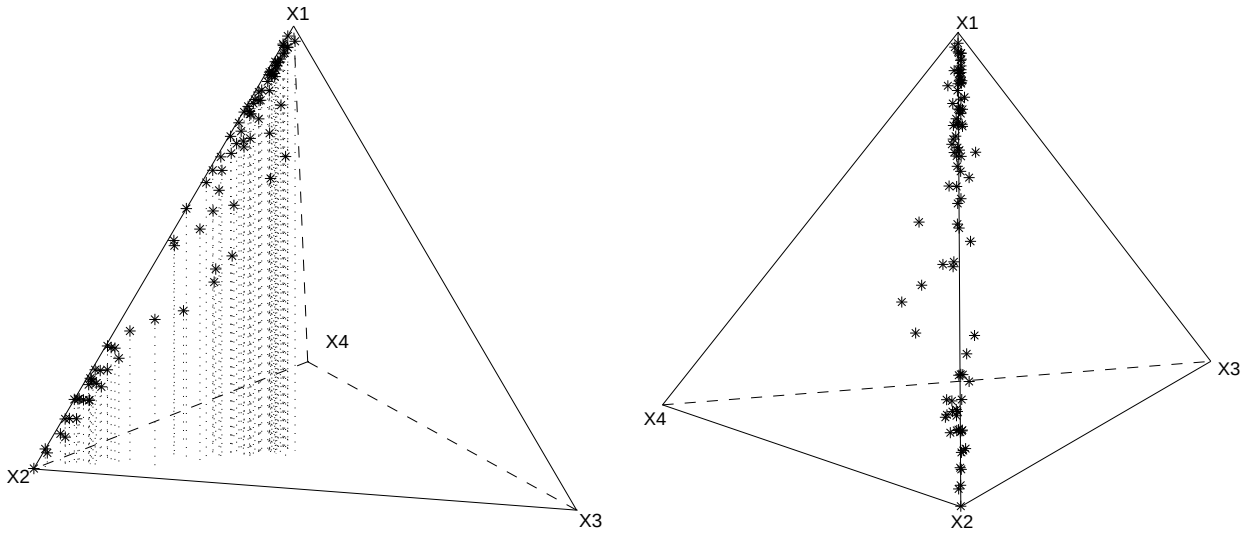
5.5 Los ceros por redondeo y la clasificación automática: dos casos prácticos

En esta sección presentamos el estudio de dos conjuntos de datos diferentes con el objetivo de ilustrar cómo influyen la cantidad de observaciones y la cantidad de ceros presentes en el conjunto de datos a clasificar. Los dos conjuntos de datos que vamos a estudiar contienen entre sus valores un número suficientemente alto de ceros de manera que es de esperar *a priori* que el valor utilizado en el reemplazamiento puede influir en el resultado de la clasificación automática. El primer conjunto de datos que vamos a estudiar aparece en Aitchison (1986) con el nombre *Glacial data set* y ha sido motivo de estudio en Martín-Fernández et al. (2000). Este conjunto de datos contiene 92 observaciones de $\mathcal{S}^4$ de las cuales 41 contienen algún valor nulo. El segundo conjunto de datos es el conjunto *Dars Sill* que hemos descrito en la Sección 5.2 de este mismo capítulo. Recordemos que en este conjunto de datos de $\mathcal{S}^8$, el 90.8% de las 1281 observaciones contiene algún valor nulo.

5.5.1 Estudio del conjunto *Glacial data set*

El conjunto de datos *Glacial* –véase el Apéndice A.2– está formado por 92 muestras recogidas en glaciares. De cada muestra se separaron los guijarros que contenían, según las cuatro categorías siguientes: *red sandstone*, *gray sandstone*, *crystalline*, y *miscellaneous*. Para cada muestra se consideró el peso total de los guijarros, y para cada una de las cuatro categorías se consideró una variable, $\mathbf{X}_i$ con $i = 1, 2, 3, 4$, que recoge el tanto por ciento en peso que representan los guijarros de la i-ésima categoría respecto del peso total de los guijarros de la muestra. De esta manera, cada una de las 92 muestras se expresa como una observación del símplex $\mathcal{S}^4$.

Las figuras 5.3(a) y 5.3(b) muestran dos perspectivas diferentes del conjunto de datos *Glacial* en el espacio $\mathcal{S}^4$. Las líneas discontinuas que aparecen en la figura 5.3(a) representan la proyección de cada observación sobre la cara inferior del símplex.



(a) (b)

Figura 5.3: Dos perspectivas diferentes del conjunto *Glacial* en $\mathcal{S}^4$: (a) *con proyecciones;* (b) *sin proyecciones.*

En las dos figuras se observa que la mayoría de las observaciones aparecen situadas cerca de la arista determinada por las variables $\mathbf{X}_1$ y $\mathbf{X}_2$. Este hecho es consecuencia de que las observaciones toman valores altos en estas variables y toman valores bajos en las variables $\mathbf{X}_3$ y $\mathbf{X}_4$.

Esta característica se aprecia con mayor claridad en los cuatro diagramas ternarios de la figura 5.4. Cada diagrama ternario representa una de las subcomposiciones de $\mathcal{S}^3$ que puede considerarse en el conjunto *Glacial*. En las figuras 5.4(a) y 5.4(b) se observa que la mayor parte de las observaciones se encuentran próximas a la arista determinada por las dos primeras variables. En las figuras 5.4(c) y 5.4(d) se observa que la mayoría de las observaciones aparecen cercanas al vértice opuesto a la arista formada por las variables $\mathbf{X}_3$ y $\mathbf{X}_4$.

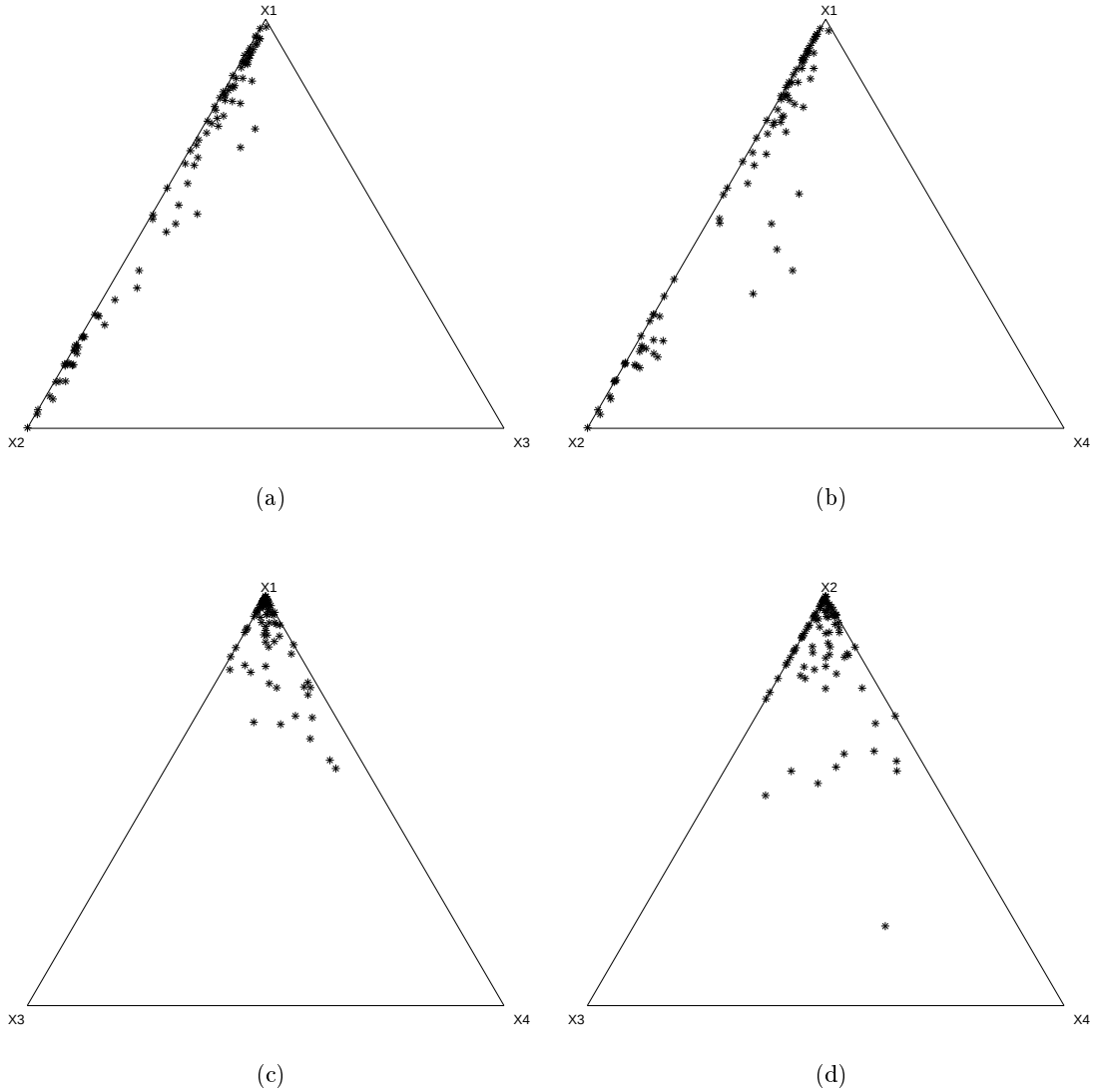


Figura 5.4: Diagramas ternarios de las subcomposiciones del conjunto *Glacial*: (a) *subcomposición* $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$; (b) *subcomposición* $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_4$; (c) *subcomposición* $\mathbf{X}_1, \mathbf{X}_3, \mathbf{X}_4$; (d) *subcomposición* $\mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$.

La tabla 5.1 recoge el valor de algunos estadísticos básicos de cada una de las cuatro variables del conjunto *Glacial*.

Tabla 5.1: Estadísticos básicos del conjunto *Glacial*.

Componente	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
Mediana	0.7081	0.2410	0.0095	0.0075
Mínimo-Máximo	0.0010-0.9770	0.0075-0.9990	0.0000-0.1082	0.0000-0.2282

El valor de la mediana de cada componente refleja la característica que las observaciones toman valores bajos en las componentes $\mathbf{X}_3$ y $\mathbf{X}_4$. Los valores del mínimo y máximo nos informan que la variabilidad relativa es alta en todas las componentes. Obsérvese que el mínimo de las variables $\mathbf{X}_3$ y $\mathbf{X}_4$ es cero y, por lo tanto, refleja la existencia de valores nulos en estas dos variables. Las componentes $\mathbf{X}_1$ y $\mathbf{X}_2$ toman valores estrictamente positivos en todas las observaciones. Estas características se muestran de una manera gráfica en la figura 5.5.

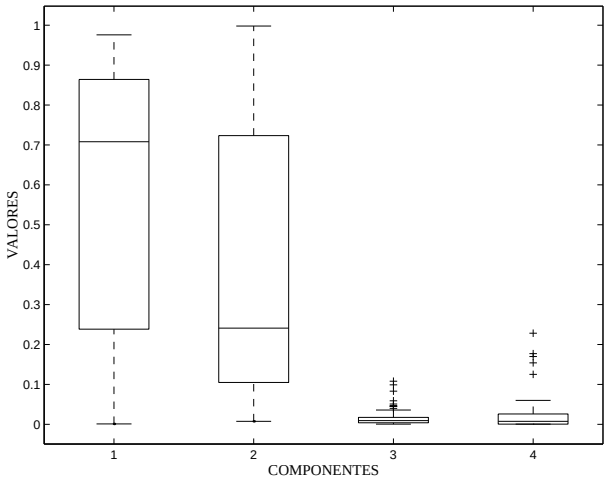


Figura 5.5: Diagrama de caja múltiple del conjunto *Glacial*.

En esta figura se ha representado el diagrama de caja de cada una de las cuatro componentes del conjunto. Obsérvese que las cajas de las componentes $\mathbf{X}_1$ y $\mathbf{X}_2$ muestran que estas dos componentes poseen asimetrías opuestas; y que las cajas de las componentes $\mathbf{X}_3$ y $\mathbf{X}_4$ muestran la presencia de datos atípicos de valores relativamente altos.

La observación de las figuras 5.3, 5.4, 5.5 nos sugiere la existencia de grupos diferenciados de observaciones dentro del conjunto *Glacial*. Previamente a realizar una clasificación automática debemos plantearnos resolver el problema de las observaciones que contienen ceros. En este

estudio suponemos que los valores nulos son ceros por redondeo, y llevamos a cabo su reemplazamiento siguiendo la fórmula (5.6) presentada en este capítulo. Sin embargo, antes de proceder a este reemplazamiento, estudiamos la cantidad y la disposición de los valores nulos en el conjunto de datos. Si observamos los valores recogidos en el conjunto de datos –véase el Apéndice A.2– detectamos que de las 92 observaciones, 42 de ellas poseen algún valor nulo. De estas 42 observaciones, 6 de ellas contienen el valor cero en las dos componentes $\mathbf{X}_3$ y $\mathbf{X}_4$; 6 de ellas toman el valor cero únicamente en la componente $\mathbf{X}_3$; y 30 de ellas toman el valor cero únicamente en la componente $\mathbf{X}_4$. Por otra parte, si observamos los valores no nulos del conjunto de datos vemos que el valor mínimo es 0.001. Por lo tanto, de acuerdo con Sandford et al. (1993) un valor adecuado para reemplazamiento de los ceros puede ser el 55% de esta cantidad, es decir 0.00055. Mediante el reemplazamiento (5.6) sustituimos los ceros que contengan las observaciones en las componentes $\mathbf{X}_3$ y $\mathbf{X}_4$ por el valor $\delta = 0.00055$, y modificamos convenientemente las componentes no nulas. Una vez reemplazados los ceros podemos realizar la clasificación automática escogiendo una medida de diferencia adecuada –d_{Ait} o $d_{\mathcal{KL}}$– y podemos aplicar un método de clasificación jerárquico. Las figuras 5.6(a) y 5.6(b) muestran los dendrogramas que se obtienen al aplicar, respectivamente, el método de Ward y el método de la media, usando la distancia de Aitchison d_{Ait}.

Previamente a analizar los resultados de las clasificaciones, debemos analizar si el valor $\delta = 0.00055$ usado en la imputación afecta el resultado de la clasificación. Por ello, hemos llevado a cabo un análisis de sensibilidad en función del valor δ. En Aitchison (1986) se sugiere que un análisis de sensibilidad adecuado de los resultados obtenidos al aplicar una técnica multivariante a datos composicionales con ceros por redondeo debe contemplar un rango de valores de imputación que vayan desde el 10% al 100% del valor mínimo no nulo observado. Entonces, en nuestro caso debemos analizar el resultado de la clasificación para valores de imputación que vayan desde 0.0001 hasta 0.001. El análisis de sensibilidad que realizamos se basa en el estudio de la variación del valor de tres coeficientes que han sido presentados en la Sección 1.8 de esta tesis: el coeficiente de correlación cofenética, el índice de Mojena, y el índice de Canlinski. El primero de estos tres coeficientes mide el grado de relación entre el índice de jerarquía resultado de la estructura jerárquica y la medida de diferencia. Los resultados obtenidos en el análisis del coeficiente de correlación cofenética se muestran en la tabla 5.2. En esta tabla también se muestra el valor medio y el coeficiente de variación de los resultados obtenidos al variar el valor de δ. De la observación de estos resultados destacamos que para todos los métodos los valores del coeficiente reflejan estabilidad en la variación de δ. Adicionalmente,

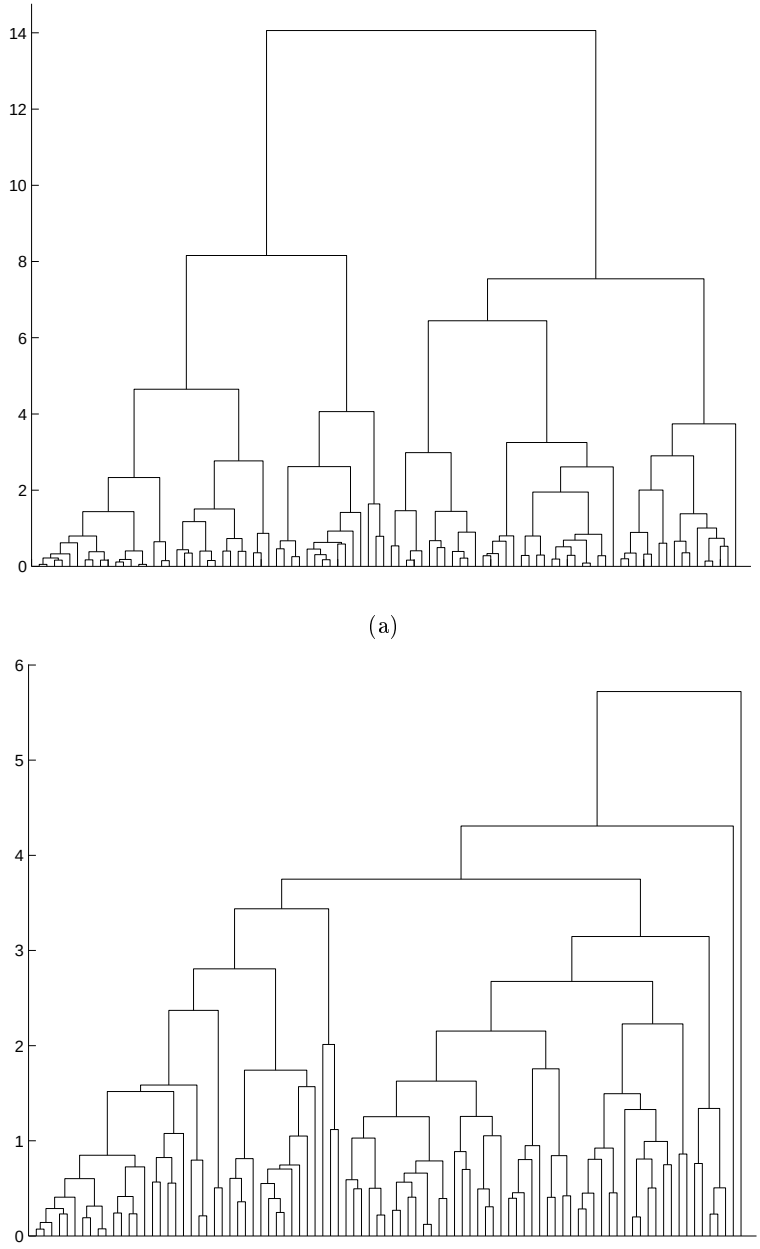


Figura 5.6: Dendrogramas obtenidos de la clasificación del conjunto resultante de reemplazar los ceros del conjunto *Glacial* por $\delta = 0.00055$. La distancia usada ha sido la distancia de Aitchison d_{Ait}: (a) *clasificación con el método de Ward;* (b) *clasificación con el método de la media.*

en la tabla 5.2 se observa que el método de la media es el que produce mejores resultados.

Tabla 5.2: Análisis de sensibilidad según δ para el coeficiente de correlación cofenética en el conjunto *Glacial*.

δ	Ward	Centroide	Mínimo	Máximo	Media
0.0001	0.71	0.83	0.78	0.81	0.86
0.0002	0.65	0.79	0.63	0.66	0.83
0.0003	0.62	0.78	0.64	0.63	0.80
0.0004	0.61	0.74	0.62	0.60	0.78
0.0005	0.61	0.73	0.62	0.62	0.78
0.0006	0.61	0.73	0.63	0.62	0.76
0.0007	0.60	0.73	0.64	0.62	0.76
0.0008	0.60	0.73	0.65	0.62	0.76
0.0009	0.59	0.72	0.66	0.62	0.76
0.001	0.60	0.72	0.66	0.62	0.76
Media	0.62	0.75	0.65	0.64	0.79
C.V.	0.06	0.05	0.07	0.09	0.05

El índice de Mojena informa sobre el número de grupos que refleja la estructura jerárquica resultado de la clasificación. Este índice se calcula en base a la búsqueda de "saltos grandes" de nivel de fusión del dendrograma. De entre los saltos de fusión que se consideran "grandes" puede escogerse el último salto, el primer salto o el salto máximo. Los dos primeros saltos nos proporcionan, respectivamente, el número mínimo y el máximo de grupos en que el índice de Mojena considera que puede dividirse el conjunto de datos. La tabla 5.3 muestra los análisis de sensibilidad de, respectivamente, el último, el primer, y el salto máximo. De la observación de la tabla 5.3 se destaca una gran estabilidad de los resultados para el índice mínimo y salto máximo. En estos dos casos el índice de Mojena sugiere para todos los métodos la existencia de 2 grupos en el conjunto de datos *Glacial*, siendo los más estables el método de Ward y el de la media. En los resultados del análisis de sensibilidad del número máximo de la misma tabla 5.3, se observa que el único método que muestra una gran estabilidad es el método de Ward, indicando la existencia de 2 grupos. El resto de métodos muestra una gran sensibilidad a la variación del valor de imputación δ. En este caso, y tal como muestra el método del centroide, el número de grupos a considerar oscila entre 2 y 8 grupos.

Tabla 5.3: Análisis de sensibilidad según δ para el número de grupos indicado por el índice de Mojena (mínimo/máximo/salto máximo) en el conjunto *Glacial*.

δ	Ward	Centroide	Mínimo	Máximo	Media
0.0001	2/ 2/ 2	3/ 5/ 5	2/ 2/ 2	2/ 6/ 2	2/ 5/ 2
0.0002	2/ 2/ 2	2/ 7/ 2	2/ 2/ 2	2/ 5/ 3	2/ 7/ 2
0.0003	2/ 2/ 2	2/ 6/ 2	2/ 2/ 2	2/ 5/ 2	2/ 7/ 2
0.0004	2/ 2/ 2	2/ 6/ 2	2/ 3/ 2	2/ 5/ 2	2/ 2/ 2
0.0005	2/ 2/ 2	2/ 8/ 2	2/ 3/ 2	2/ 3/ 2	2/ 2/ 2
0.0006	2/ 2/ 2	2/ 2/ 2	2/ 3/ 2	2/ 3/ 2	2/ 3/ 2
0.0007	2/ 2/ 2	2/ 2/ 2	2/ 3/ 2	2/ 3/ 2	2/ 3/ 2
0.0008	2/ 2/ 2	2/ 2/ 2	2/ 3/ 3	2/ 3/ 2	2/ 3/ 2
0.0009	2/ 2/ 2	2/ 4/ 2	2/ 3/ 3	2/ 2/ 2	2/ 3/ 2
0.001	2/ 2/ 2	2/ 4/ 2	2/ 3/ 3	2/ 2/ 2	2/ 3/ 2
Media	2/ 2/ 2	2.1/4.60/2.30	2/2.70/2.30	2/3.70/2.10	2/3.80/2
C.V.	0/ 0/ 0	0.15/0.47/0.41	0/0.18/0.21	0/0.38/0.15	0/0.49/0

Sin embargo, es en el índice de Calinski o índice C donde se aprecia una mayor divergencia entre los resultados para los diferentes métodos de clasificación. Recordemos que este índice se basa en la comparación entre la variabilidad dentro de los grupos y la variabilidad entre los grupos. El índice calcula el número de grupos en que debe dividirse el conjunto a clasificar, de manera que los grupos resulten ser lo más homogéneos dentro de si y lo más heterogéneos entre ellos. La tabla 5.4 muestra que el único método que sigue indicando la existencia de 2 grupos es el método de Ward. Sin embargo, para este método y para valores altos de δ, el índice C sugiere que 5 es el número óptimo de grupos. En la tabla 5.4 se observa que todos los métodos son muy sensibles al cambio en el valor de δ y que el índice C sugiere, en general, un número de grupos más elevado que el índice de Mojena. Esta diferencia entre los dos índices es atribuible a que el número de observaciones a clasificar 92 no es muy elevado. En consecuencia, si se considera un número alto de grupos, se obtendrán grupos con pocos individuos y, por lo tanto, muy homogéneos dentro de si. Como veremos en el segundo caso práctico que presentaremos, esta diferencia entre los dos índices no es tan acusada cuando el conjunto a clasificar es más numeroso.

A la vista de los resultados del análisis de sensibilidad y de los dendrogramas de la figura 5.6 decidimos seguir analizando la clasificación obtenida para $\delta = 0.00055$ mediante los métodos de Ward y de la media. Escogemos el método de Ward por ser el más estable de todos ellos y escogemos el método de la media por ser el que ha proporcionado mejores resultados para el

Tabla 5.4: Análisis de sensibilidad según δ para el número de grupos indicado por el índice de Calinski en el conjunto *Glacial*.

δ	Ward	Centroide	Mínimo	Máximo	Media
0.0001	2	6	9	2	10
0.0002	2	10	10	9	10
0.0003	2	10	10	5	9
0.0004	2	10	10	5	10
0.0005	2	9	7	3	8
0.0006	2	6	10	3	10
0.0007	2	6	10	3	8
0.0008	5	6	10	3	8
0.0009	5	7	10	4	7
0.001	5	7	8	6	9
Media	2.90	7.70	9.40	4.30	8.90
C.V.	0.50	0.24	0.11	0.48	0.12

coeficiente de correlación cofenética. En el dendrograma de la figura 5.6(a) se aprecia que para el método de Ward, una agrupación razonable estaría formada por 2 grupos. En el dendrograma de la figura 5.6(b) se observa que para el método de la media, una agrupación razonable contemplaría 4 grupos. De estos 4 grupos dos de ellos estarían formados por dos observaciones aisladas cada uno: la observación número 5 y la observación número 64. Si se consulta el Apéndice A.2 se puede observar que son dos observaciones que, de manera anómala, toman valores muy bajos en, respectivamente, las dos primeras componentes. En la primera componente la observación número 5 toma el valor $\mathbf{x}_{5,1} = 0.001$ y en la segunda componente la observación 64 toma el valor $\mathbf{x}_{64,2} = 0.0075$. Estos valores son además los mínimos que estas variables –véase la tabla 5.1– toman en todo el conjunto *Glacial*.

En las tablas 5.5 y 5.6 se encuentran los valores de algunos estadísticos básicos de los diferentes grupos que se obtienen mediante el método de Ward y el método de la media. Comparando los resultados de las dos tablas se observa que, exceptuando las observaciones número 64 y número 5, los valores de los estadísticos básicos presentan el mismo patrón para los dos métodos de clasificación. En las dos tablas los valores de los estadísticos del primer grupo indican que las observaciones que pertenecen a este grupo toman los valores más altos en la componente $\mathbf{X}_1$ y los valores más bajos en la componente $\mathbf{X}_3$. Los estadísticos del segundo grupo indican que las observaciones que pertenecen a este grupo toman los valores más altos en la componente $\mathbf{X}_1$

pero con más diferencia respecto la componente $\mathbf{X}_2$ que en el primer grupo. Los valores más bajos de las observaciones del segundo grupo se toman, a diferencia de lo que sucede en el primer grupo, en la componente $\mathbf{X}_4$. Si en estas dos tablas se observan los valores mínimo y máximo, se aprecia que los dos grupos manifiestan una variabilidad relativa alta en cada variable. Esta característica, unida a la observación de los dendrogramas de la figura 5.6, nos sugiere que el conjunto *Glacial* puede ser clasificado en un mayor número de grupos.

Tabla 5.5: Estadísticos básicos de los dos grupos del conjunto *Glacial* determinados por el método de Ward.

	Componente	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
Grupo 1 (46 obs.)					
	Mediana	0.5685	0.3261	0.0094	0.0260
	Mín.-Máx.	0.0010-0.9613	0.0075-0.9979	0.0006-0.1082	0.0006-0.2282
Grupo 2 (46 obs.)					
	Mediana	0.8335	0.1480	0.0095	0
	Mín.-Máx.	0.0452-0.9759	0.0230-0.9537	0.0006-0.0590	0.0006-0.0100

Tabla 5.6: Estadísticos básicos de los cuatro grupos del conjunto *Glacial* determinados por el método de la media.

	Componente	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
Grupo 1 (50 obs.)					
	Mediana	0.6590	0.2750	0.0101	0.0225
	Mín.-Máx.	0.0340-0.9150	0.0660-0.9520	0.0006-0.1082	0.0040-0.2282
Grupo 2 (40 obs.)					
	Mediana	0.8295	0.1610	0.0095	0
	Mín.-Máx.	0.0452-0.9759	0.0230-0.9537	0.0006-0.0590	0.0006-0.0030
Grupo 3 (1 obs.)					
	Observ. 64	0.9613	0.0075	0.0108	0.0204
Grupo 4 (1 obs.)					
	Observ. 5	0.0010	0.9979	0.0006	0.0006

En los dendrogramas de la figura 5.6 observamos que el dendrograma obtenido por el método de Ward muestra de una manera más nítida que los 2 grupos considerados hasta ahora pueden subdividirse a su vez en 2 subgrupos. De esta manera se obtiene una clasificación del conjunto *Glacial* en 4 grupos. Los estadísticos básicos de estos 4 grupos se muestran en la tabla 5.7.

Tabla 5.7: Estadísticos básicos de los cuatro grupos del conjunto *Glacial* determinados por el método de Ward.

	Componente	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
Grupo 1 (30 obs.)					
	Mediana	0.7281	0.2133	0.0170	0.0280
	Mín.-Máx.	0.3175-0.9613	0.0075-0.4709	0.0006-0.1082	0.0050-0.2282
Grupo 2 (16 obs.)					
	Mediana	0.1770	0.7884	0.0060	0.0190
	Mín.-Máx.	0.0010-0.2700	0.7020-0.9979	0.0006-0.0190	0.0006-0.0600
Grupo 3 (31 obs.)					
	Mediana	0.8940	0.0931	0.0100	0
	Mín.-Máx.	0.6466-0.9759	0.0230-0.3448	0.0006-0.0240	0.0006-0.0100
Grupo 4 (15 obs.)					
	Mediana	0.2242	0.7698	0.0080	0
	Mín.-Máx.	0.0452-0.5869	0.4092-0.9537	0.0006-0.0590	0.0006-0.0020

Los grupos primero y segundo mantienen la característica de tomar el valor mínimo en la componente $\mathbf{X}_3$ y se diferencian entre si por tomar el valor máximo en, respectivamente, la primera o segunda componente. Los grupos tercero y cuarto mantienen la característica de tomar el valor mínimo en la componente $\mathbf{X}_4$ y se diferencian entre si por tomar el valor máximo en, respectivamente, la primera o segunda componente. En consecuencia podemos resumir la clasificación en 4 grupos mediante dos criterios globales de clasificación:

- Las observaciones del primer y segundo grupos coinciden por el hecho de tomar valores mayores en la componente $\mathbf{X}_4$ que en la componente $\mathbf{X}_3$. Las observaciones del tercer y cuarto grupos coinciden por el hecho de tomar valores mayores en la componente $\mathbf{X}_3$ que en la componente $\mathbf{X}_4$.

- El primer y segundo grupos se diferencian entre si por tomar valores superiores en la componente $\mathbf{X}_1$ o $\mathbf{X}_2$. El tercer y cuarto grupos se diferencian entre si a partir del mismo criterio.

Los valores que toman estos estadísticos básicos de los cuatro grupos del conjunto *Glacial* tienen su traducción gráfica en los diagramas de caja de la figura 5.7.

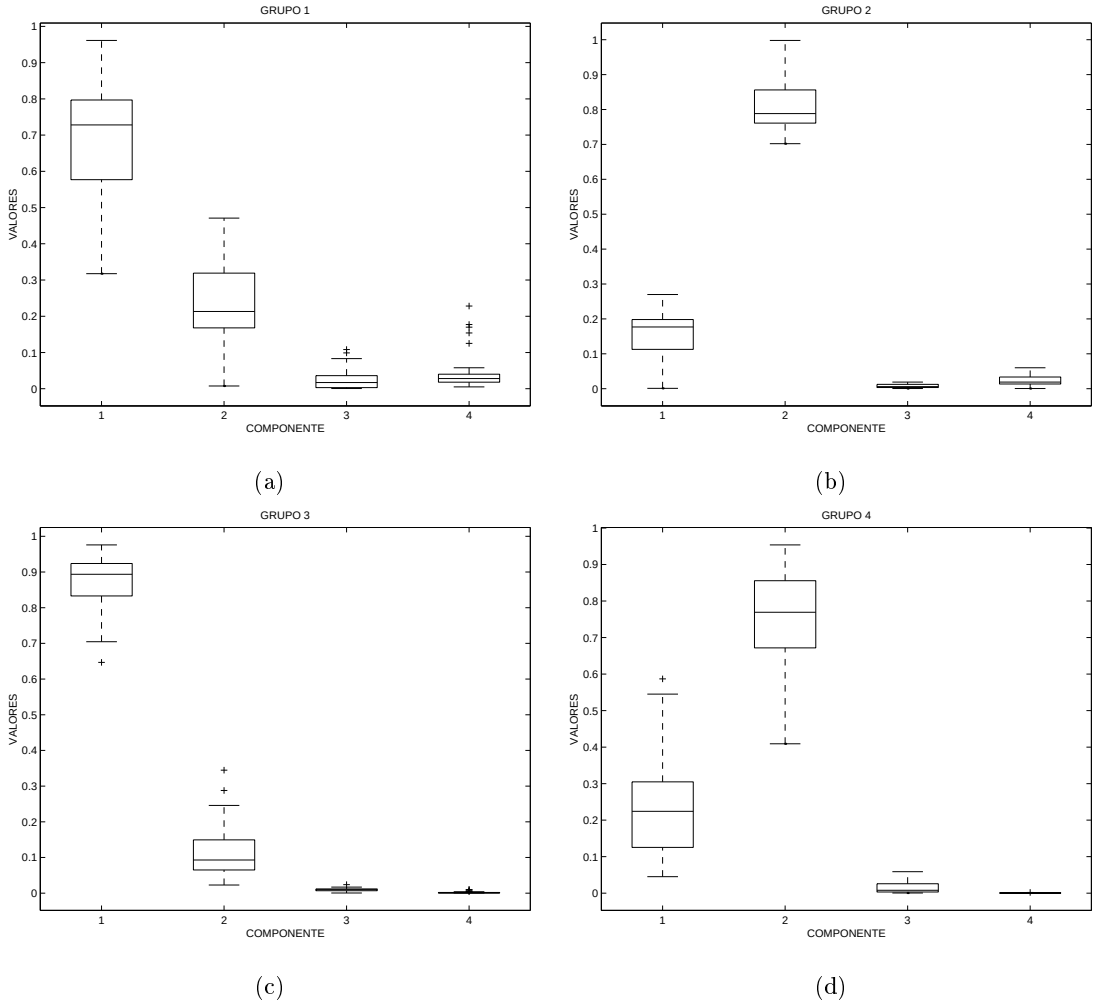


Figura 5.7: Diagramas de caja de los cuatro grupos determinados por el método de Ward aplicado al conjunto resultante de reemplazar los ceros del conjunto *Glacial* por $\delta = 0.00055$: (a) *grupo 1;* (b) *grupo 2;* (c) *grupo 3;* (d) *grupo 4.*

En esta figura puede apreciarse con mayor claridad cómo se contraponen los grupos primero y segundo, y, respectivamente, tercero y cuarto, en relación a los valores que toman las componentes $\mathbf{X}_1$ y $\mathbf{X}_2$.

La figura 5.8 muestra los valores de la media geométrica composicional para cada uno de los 4 grupos. Puede observarse que en los grupos primero y segundo, el valor de la componente $\mathbf{X}_4$ es superior al valor de $\mathbf{X}_3$. Sucede lo contrario entre los grupos tercero y cuarto. Asimismo puede observarse que en los grupos primero y tercero el valor de la componente $\mathbf{X}_1$ es superior al valor de $\mathbf{X}_2$ y que sucede lo contrario entre los grupos segundo y cuarto. Esta contraposición de valores altos/bajos en las variables puede apreciarse de una manera más clara en el diagrama biplot de la figura 5.9.

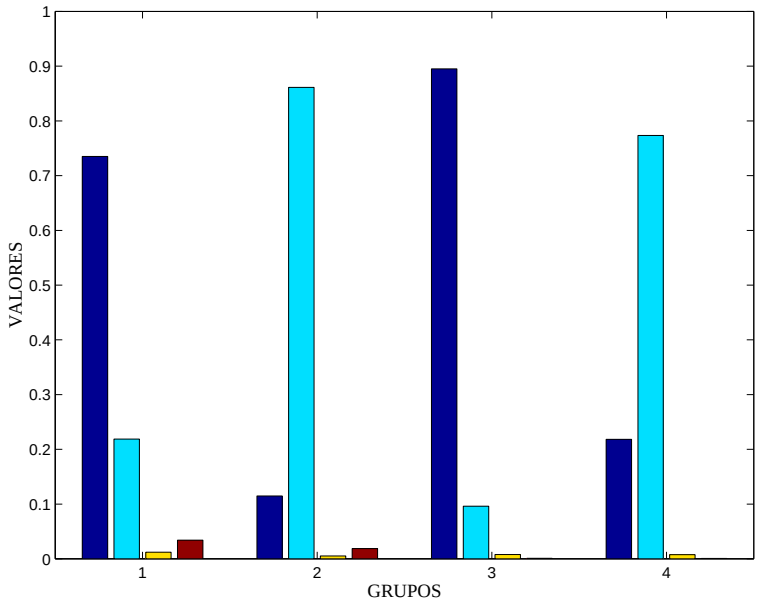


Figura 5.8: Diagramas de barras de las medias geométricas composicionales de los cuatro grupos determinados por el método de Ward aplicado al conjunto resultante de reemplazar los ceros del conjunto *Glacial* por $\delta = 0.00055$.

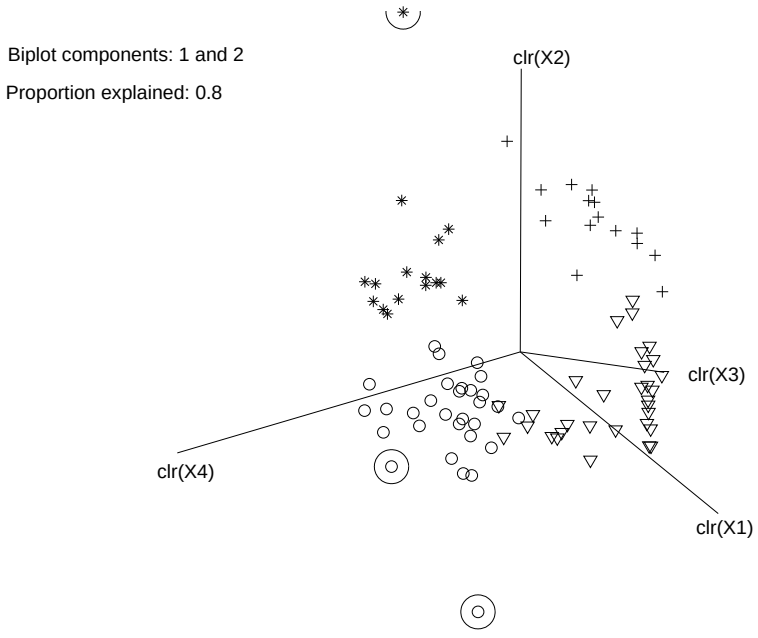
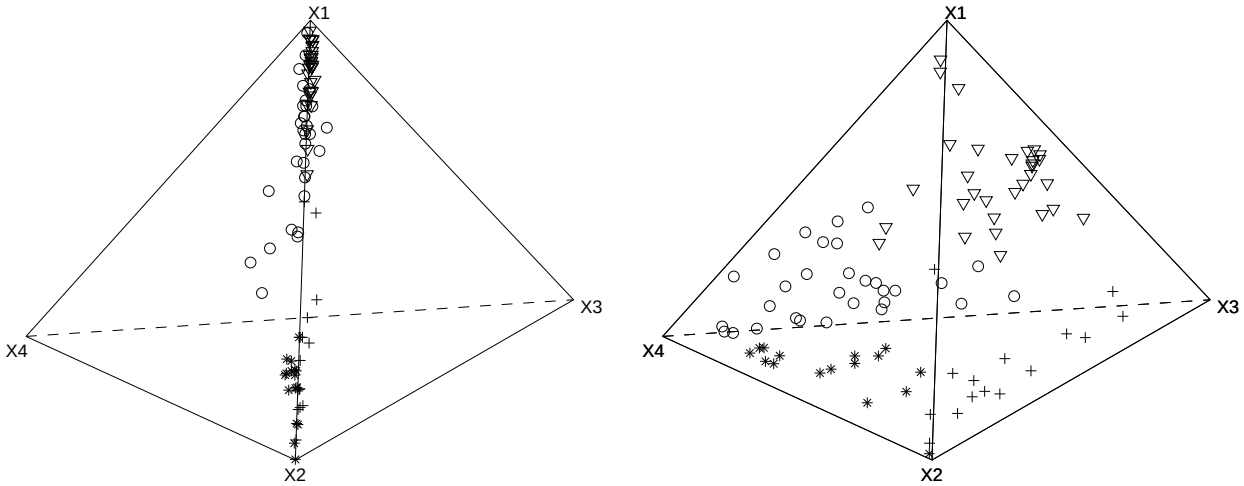


Figura 5.9: Diagrama biplot en el espacio clr del conjunto resultante de reemplazar los ceros del conjunto *Glacial* por $\delta = 0.00055$. Se muestran los cuatro grupos determinados por el método de Ward. (Grupo 1: '$\circ$'; Grupo 2: '$*$'; Grupo 3: '∇'; Grupo 4:'$+$'). Las observaciones dentro de un círculo pueden ser catalogables como atípicas.

Puede observarse que mayoritariamente las observaciones de los grupos primero y segundo se encuentran en el semieje positivo de la variable $\mathrm{clr}(\mathbf{X}_4)$ y en el semieje negativo de la variable $\mathrm{clr}(\mathbf{X}_3)$. Por el contrario los grupos tercero y cuarto aparecen colocados en disposición inversa. Se observa también que mayoritariamente las observaciones de los grupos segundo y cuarto aparecen en el semieje positivo de la variable $\mathrm{clr}(\mathbf{X}_2)$, y, por el contrario, los grupos primero y tercero se colocan en el semieje positivo de la variable $\mathrm{clr}(\mathbf{X}_1)$. En la misma figura se distinguen con un círculo tres observaciones. Esta distinción se debe a que estas observaciones pueden catalogarse como observaciones atípicas. Es importante resaltar que la observación atípica del primer grupo, que aparece en la parte inferior del diagrama más alejada del origen de coordenadas, es la observación número 64; y que la observación atípica del segundo grupo, que aparece en la parte superior del diagrama, es la observación número 5. Esta dos observaciones son las mismas que, por el método de la media, constituían cada una de ellas un grupo por separado –véase la tabla 5.6.

Finalmente, podemos observar en las figuras 5.10(a) y 5.10(b) dos perspectivas diferentes en el espacio $\mathcal{S}^4$ del conjunto de datos *Glacial*, donde cada observación se la ha simbolizado con el número del grupo al que pertenece según la clasificación obtenida mediante el método de Ward.



(a) (b)

Figura 5.10: Dos perspectivas diferentes en $\mathcal{S}^4$ del conjunto resultante de reemplazar los ceros del conjunto *Glacial* por $\delta = 0.00055$: (a) *sin centrar;* (b) *conjunto centrado.* Se muestran los cuatro grupos determinados por el método de Ward (Grupo 1: '$\circ$'; Grupo 2: '$*$'; Grupo 3: '∇'; Grupo 4:'$+$').

5.5.2 Estudio del conjunto *Darss Sill*

En la Sección 5.2 de este mismo capítulo hemos hecho referencia al conjunto de datos *Darss Sill*. Este tipo de datos se conoce en la literatura como datos de naturaleza *granulométrica*. Este nombre expresa la idea que el interés central radica en el estudio del tamaño del grano de los sedimentos recogidos en las muestras. En el caso del conjunto *Darss Sill* se recogieron 1281 muestras de sedimentos marinos en diferentes puntos geográficos del fondo del Mar Báltico. El área geográfica de muestreo es conocida con el nombre de *Darss Sill*. Los componentes arenosos de estos sedimentos se separaron según el tamaño del grano en 8 componentes que fueron ordenadas de mayor a menor tamaño. Los datos composicionales a considerar consisten en el porcentaje en peso de cada tamaño de grano respecto del peso total de la muestra recogida. Los nombres e intervalos de medida de estas 8 componentes o diferentes tamaños se muestran en la tabla 5.8. Debido al elevado número de observaciones que conforman el conjunto de datos, se ha optado por no incluirlo en los apéndices de esta tesis. Sin embargo, pueden obtenerse vía *FTP* al servidor *ftp.iamg.org* que mantiene la *International Association for Mathematical Geology*. Dentro del directorio *darssil* de este servidor se encuentra el fichero *darssil.txt* que contiene los datos del conjunto en formato *ascii*.

Tabla 5.8: Tamaños de grano del conjunto *Darss Sill*: nombre, intervalo de medida y número de observaciones que contienen el valor cero en la componente.

Componente	Nombre	Intervalo de medidas (mm)	Número de ceros
$\mathbf{X}_1$	*Gravel*	> 2.0	1156
$\mathbf{X}_2$	*Very coarse*	$2.0 - 1.0$	925
$\mathbf{X}_3$	*Coarse sand*	$1.0 - 0.63$	599
$\mathbf{X}_4$	*Medium sand*	$0.63 - 0.4$	95
$\mathbf{X}_5$	*Medium fine*	$0.4 - 0.2$	12
$\mathbf{X}_6$	*Fine sand*	$0.2 - 0.1$	0
$\mathbf{X}_7$	*Very fine sand*	$0.1 - 0.063$	8
$\mathbf{X}_8$	*Silt*	< 0.063	57

El interés en clasificar este conjunto de datos surge de la intención de realizar un mapa del fondo marino donde se reflejaran las diferentes zonas del fondo según el tipo de sedimento. Cuando se empezó a analizar el conjunto de datos composicionales $\mathbf{X} \in \mathcal{S}^8$ se observó que

contenía un número muy elevado de ceros –véase la tabla 5.8. Se apreció que el 90.8% de las 1281 observaciones contenían algún valor nulo y que de los 1281×8 elementos de la matriz de datos, el 27.8% eran ceros. Estos valores nulos estaban concentrados mayoritariamente en las 4 primeras componentes, es decir en los 4 tamaños mayores de grano. La distribución de los ceros en las diferentes componentes se muestra en la tabla 5.9. Esta tabla está inspirada en la información contenida en el trabajo de Tauber (1999).

Tabla 5.9: Distribución de los ceros del conjunto *Darss Sill*. Los valores nulos se representan con un cero y los valores no nulos con un punto.

Tipo de combinación	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$	Número de observaciones
1	0	0	0	0	0	·	·	·	12
2	0	0	0	0	·	·	·	·	83
3	0	0	0	·	·	·	·	0	23
4	0	0	0	·	·	·	·	·	480
5	0	0	0	·	·	·	0	·	2
6	0	0	·	·	·	·	·	0	17
7	0	0	·	·	·	·	·	·	308
8	0	·	0	·	·	·	·	·	1
9	0	·	·	·	·	·	0	·	3
10	0	·	·	·	·	·	·	0	13
11	0	·	·	·	·	·	·	·	214
12	·	·	·	·	·	·	0	·	3
13	·	·	·	·	·	·	·	0	4
14	·	·	·	·	·	·	·	·	118

En la tabla puede apreciarse que en el conjunto de datos aparecen 14 combinaciones distintas de valores nulos en las componentes. También se observa que la mayoría de los ceros aparecen en las primeras componentes y que la componente $\mathbf{X}_6$ es la única que no contiene valores nulos.

En la tabla 5.10 se muestran los estadísticos básicos del conjunto *Darss Sill*. Si se analizan los datos de la tabla puede apreciarse que en todas las componentes, excepto en la $\mathbf{X}_6$, el valor mínimo registrado es cero, y que en la componente $\mathbf{X}_6$ el valor mínimo es 0.002. Sin embargo, si se busca en el conjunto de datos, el valor mínimo registrado es 0.001. En consecuencia, el valor utilizado en el reemplazamiento de los ceros deberá ser inferior a 0.001. Otra característica de este conjunto de datos es la alta variabilidad dentro de los valores no nulos de las componentes. Este hecho se aprecia en los valores mínimo y máximo de la tabla 5.10, y en los diagramas de

caja que muestra la figura 5.11. En esta figura se aprecia que, exceptuando las componentes $\mathbf{X}_5$ y $\mathbf{X}_6$, en las demás componentes la mayor parte de los valores no nulos son catalogados como valores atípicos debido al elevado número de ceros presentes en las componentes.

Tabla 5.10: Estadísticos básicos del conjunto *Darss Sill*.

Componente	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$
Mediana	0	0	0.002	0.02	0.328	0.422	0.03	0.004
Mín.-Máx.	0-0.428	0-0.388	0-0.436	0-0.695	0-0.916	0.002-0.975	0-0.6	0-0.706

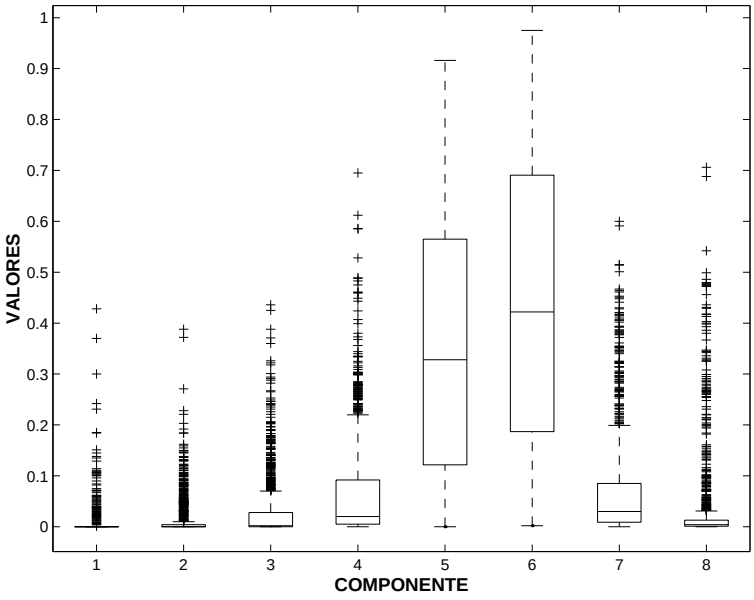


Figura 5.11: Diagrama de caja múltiple del conjunto *Darss Sill*.

En este estudio suponemos que los valores nulos son ceros por redondeo. En consecuencia, antes de realizar la clasificación automática reemplazamos los valores nulos utilizando la fórmula de reemplazamiento (5.6), y teniendo presente que el valor mínimo registrado es igual a 0.001. De manera análoga al procedimiento que hemos seguido en el estudio del conjunto de datos *Glacial*, aplicamos la fórmula del reemplazamiento (5.6) para un valor de $\delta = 0.00055$. Una vez sustituidos los ceros de las componentes nulas y modificadas convenientemente las componentes no nulas, podemos aplicar cualquier método jerárquico aglomerativo.

La figura 5.12 muestra el dendrograma resultado de aplicar el método de Ward utilizando la distancia de Aitchison al cuadrado.

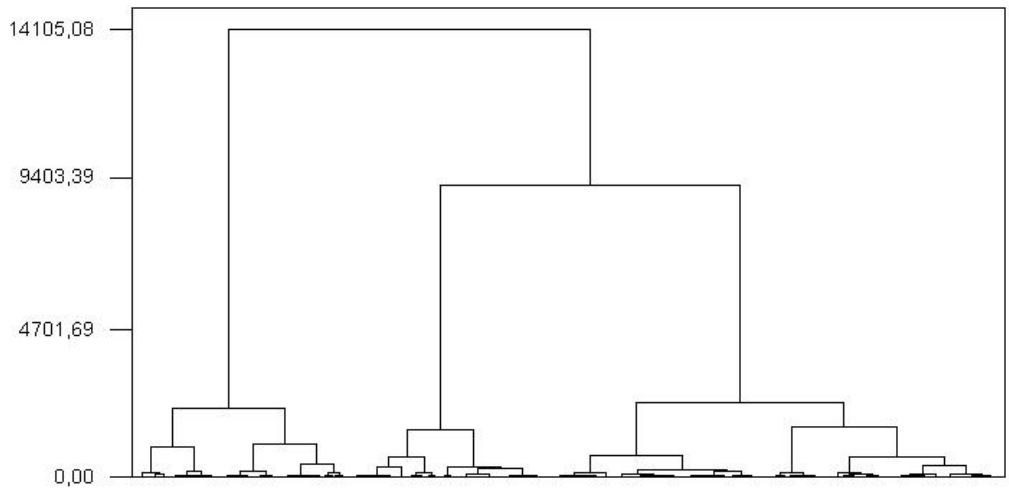


Figura 5.12: Dendrograma resultado de aplicar el método de Ward al conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$. La distancia usada ha sido la distancia d_{Ait} al cuadrado.

Puede observarse que si bien la jerarquía del dendrograma sugiere inicialmente una clasificación en 2 grupos, también pueden considerarse clasificaciones que contemplen 3, 4, e incluso 5 grupos razonablemente diferenciados. La cuestión que surge de manera inmediata es: ¿hasta qué punto el valor $\delta = 0.00055$ usado en la sustitución puede afectar al resultado de la clasificación? Para contestar esta cuestión llevamos a cabo el análisis de sensibilidad de los índices: de correlación cofenética, de Mojena y de Calinski. Realizamos el análisis de sensibilidad para valores de $\delta = 0.0002, 0.0004, \ldots, 0.001$. Para cada valor diferente de δ aplicamos el reemplazamiento (5.6) de los ceros y, a continuación, aplicamos los métodos de clasificación más usuales al conjunto de datos resultante. Para cada clasificación obtenida, calculamos los índices antes mencionados.

En la tabla 5.11 se muestran los valores obtenidos para el índice de correlación cofenética. Puede observarse que para todos los métodos de clasificación el coeficiente de variación –C.V.– toma valores que pueden ser considerados bajos y, por lo tanto se puede concluir que este índice no es sensible a los cambios efectuados en el valor de δ. Queremos resaltar que, si bien el método que toma valores más altos es el método de la media, es el método de Ward el menos sensible a los cambios en δ.

Tabla 5.11: Análisis de sensibilidad según δ para el coeficiente de correlación cofenética en el conjunto *Darss Sill*.

δ	Ward	Centroide	Mínimo	Máximo	Media
0.0002	0.66	0.67	0.57	0.70	0.72
0.0004	0.64	0.69	0.51	0.70	0.70
0.0006	0.64	0.59	0.49	0.68	0.68
0.0008	0.64	0.58	0.47	0.61	0.70
0.001	0.64	0.65	0.46	0.68	0.67
Media	0.64	0.64	0.50	0.68	0.70
C.V.	0.01	0.08	0.09	0.05	0.03

La tabla 5.12 muestra los valores *mínimo/máximo/salto máximo* para el índice de Mojena de las clasificaciones obtenidas al aplicar los diferentes métodos jerárquicos.

Tabla 5.12: Análisis de sensibilidad según δ para el número de grupos indicado por el índice de Mojena (mínimo/máximo/salto máximo) en el conjunto *Darss Sill*.

δ	Ward	Centroide	Mínimo	Máximo	Media
0.0002	2/ 5/ 2	2/ 175/ 47	2/ 1280/ 3	2/ 13/ 3	2/ 34/ 2
0.0004	2/ 5/ 2	3/ 193/ 13	2/ 1280/ 2	2/ 8/ 3	2/ 28/ 3
0.0006	2/ 4/ 3	2/ 141/ 23	2/ 1280/ 2	2/ 24/ 3	2/ 58/ 2
0.0008	2/ 4/ 3	13/ 173/ 13	2/ 1280/ 2	2/ 21/ 3	2/ 35/ 3
0.001	2/ 5/ 3	3/ 101/ 51	2/ 1280/ 2	2/ 9/ 3	2/ 21/ 2
Media	2/ 4.6/ 2.6	4.6/156.60/29.40	2/1280/2.2	2/15/3	2/35.2/2.4
C.V.	0/ 0.12/ 0.21	1.03/0.23/0.63	0/0/0.20	0/0.48/0	0/0.40/0.23

Puede apreciarse que, exceptuando el método del centroide, para todos los métodos los valores *mínimo* y *salto máximo* son poco sensibles a los cambios en el valor de δ. Por lo que se refiere a los valores *máximo* del índice, puede observarse que los métodos tienen entre ellos una gran disparidad de resultados, y que, exceptuando el método de Ward, todos los métodos toman unos valores poco razonables en el valor *máximo* del índice en comparación con los valores del *mínimo* y del *salto máximo*.

Los valores de la tabla 5.13 son los valores obtenidos para el índice de Calinski según el valor δ utilizado en la fórmula del remplazamiento (5.6).

Tabla 5.13: Análisis de sensibilidad según δ para el número de grupos indicado por el índice de Calinski en el conjunto *Darss Sill*.

δ	Ward	Centroide	Mínimo	Máximo	Media
0.0002	3	2	18	3	2
0.0004	3	3	19	4	2
0.0006	3	2	18	4	2
0.0008	3	2	3	4	4
0.001	3	3	3	3	2
Media	3	2.4	12.20	3.6	2.4
C.V.	0	0.23	0.69	0.15	0.37

Puede apreciarse que, excepto para el método jerárquico del *mínimo*, el índice de Calinski toma valores parecidos a los valores *mínimo* y *salto máximo* obtenidos con el método de Mojena. En la tabla 5.13 se observa que el índice de Calinski sugiere que en la clasificación pueden considerarse 2, 3 o 4 grupos, como una primera aproximación a una agrupación razonable. Queremos resaltar que, en general, el índice de Calinski muestra mayor sensibilidad que el índice de Mojena en los valores *mínimo* y *salto máximo*. Como excepción importante a este comportamiento, encontramos en la tabla 5.13 que el método de Ward aparece como un método muy robusto para el índice de Calinski con un coeficiente de variación igual a cero.

A la vista de los resultados obtenidos en el análisis de sensibilidad de los diferentes índices, escogemos el método de Ward para llevar a cabo la clasificación del conjunto de datos. La elección del método de Ward se ha basado en que es el método menos sensible a los cambios en el valor δ utilizado en el reemplazamiento. A tenor de los resultados de los índices de las tablas 5.13 y 5.12, la primera clasificación que aparece como razonable es la que contempla tres grupos –véase la figura 5.12. Recordemos que como clasificación razonable entendemos aquélla en la que las observaciones pertenecientes a los grupos obtenidos se diferencian por tener un patrón distinto en el valor relativo de las componentes. Al realizar la agrupación se obtienen tres grupos formados por, respectivamente, 209, 756, y 316 observaciones.

En la tabla 5.14 se muestran algunos de los estadísticos básicos de las observaciones que forman cada uno de los tres grupos.

Tabla 5.14: Estadísticos básicos de los tres grupos del conjunto *Darss Sill* determinados por el método de Ward.

	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$
Grupo 1								
(209 obs.)								
Mediana	0	0	0	0.001	0.01	0.744	0.155	0.02
Mín.-Máx.	0-0	0-0.005	0-0.007	0-0.01	0-0.313	0.102-0.971	0.012-0.6	0.002-0.688
Grupo 2								
(756 obs.)								
Mediana	0	0	0	0.015	0.399	0.489	0.024	0.002
Mín.-Máx.	0-0	0-0.024	0-0.436	0.001-0.695	0.006-0.916	0.01-0.975	0-0.43	0-0.472
Grupo 3								
(316 obs.)								
Mediana	0	0.041	0.073	0.162	0.412	0.118	0.014	0.003
Mín.-Máx.	0-0.428	0-0.388	0-0.425	0.006-0.586	0.018-0.822	0.002-0.729	0-0.274	0-0.706

En la figura 5.13 se aprecia como las observaciones que pertenecen al Grupo 1 toman en las componentes $\mathbf{X}_6$, $\mathbf{X}_7$, y $\mathbf{X}_8$ valores relativamente más altos que las observaciones pertenecientes a los otros grupos. En este Grupo 1 las observaciones toman en el resto de componentes valores relativamente más bajos que las observaciones de los otros grupos. Por el contrario, el Grupo 3 está formado por observaciones que toman valores relativamente altos en las cinco primeras componentes y valores relativamente bajos en las componentes $\mathbf{X}_6$, $\mathbf{X}_7$, y $\mathbf{X}_8$. Los valores que toman las observaciones pertenecientes al Grupo 2 convierten a este grupo, el más numeroso con diferencia, en un grupo *transición* entre el Grupo 1 y el Grupo 3.

En la figura 5.14 se han representado mediante diagramas de barras las medias geométricas composicionales de cada uno de los tres grupos. Para cada grupo, cada una de las barras representa el valor que toma la media geométrica composicional en cada una de las 8 componentes. En esta figura se observa el mismo patrón que hemos descrito mediante los diagramas de caja de la figura 5.13: para las variables $\mathbf{X}_6$, $\mathbf{X}_7$, y $\mathbf{X}_8$, los valores más altos aparecen en el Grupo 1 y los valores más bajos en el Grupo 3; y para las cinco primeras variables el comportamiento es el contrario.

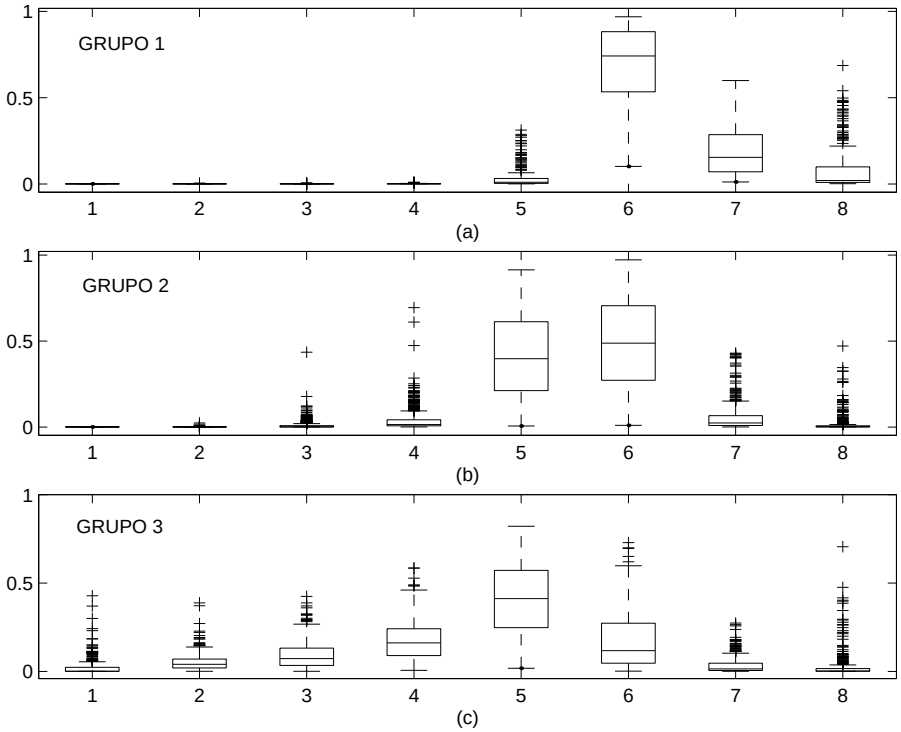


Figura 5.13: Diagramas de caja de los tres grupos determinados por el método de Ward aplicado al conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$: (a) *grupo 1;* (b) *grupo 2;* (c) *grupo 3.*

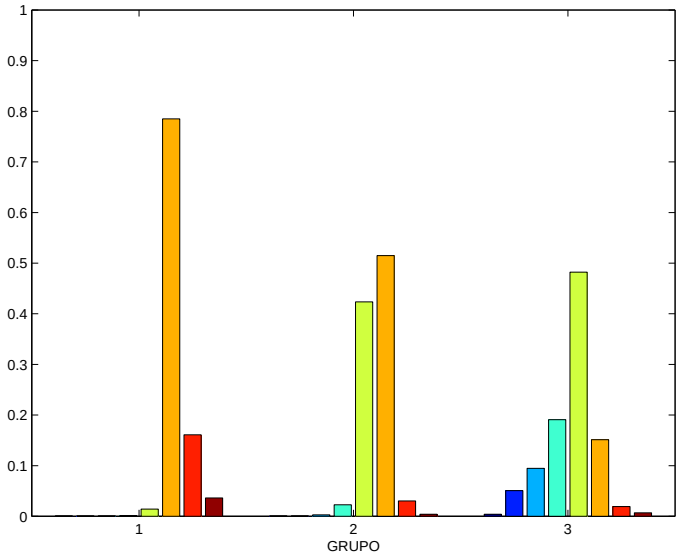


Figura 5.14: Diagramas de barras de las medias geométricas composicionales de los tres grupos determinados por el método de Ward aplicado al conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$. '1': *Grupo 1;* '2': *Grupo 2;* '3': *Grupo 3.*

La figura 5.15 muestra el diagrama *biplot* en el espacio clr-transformado, del conjunto de datos resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$.

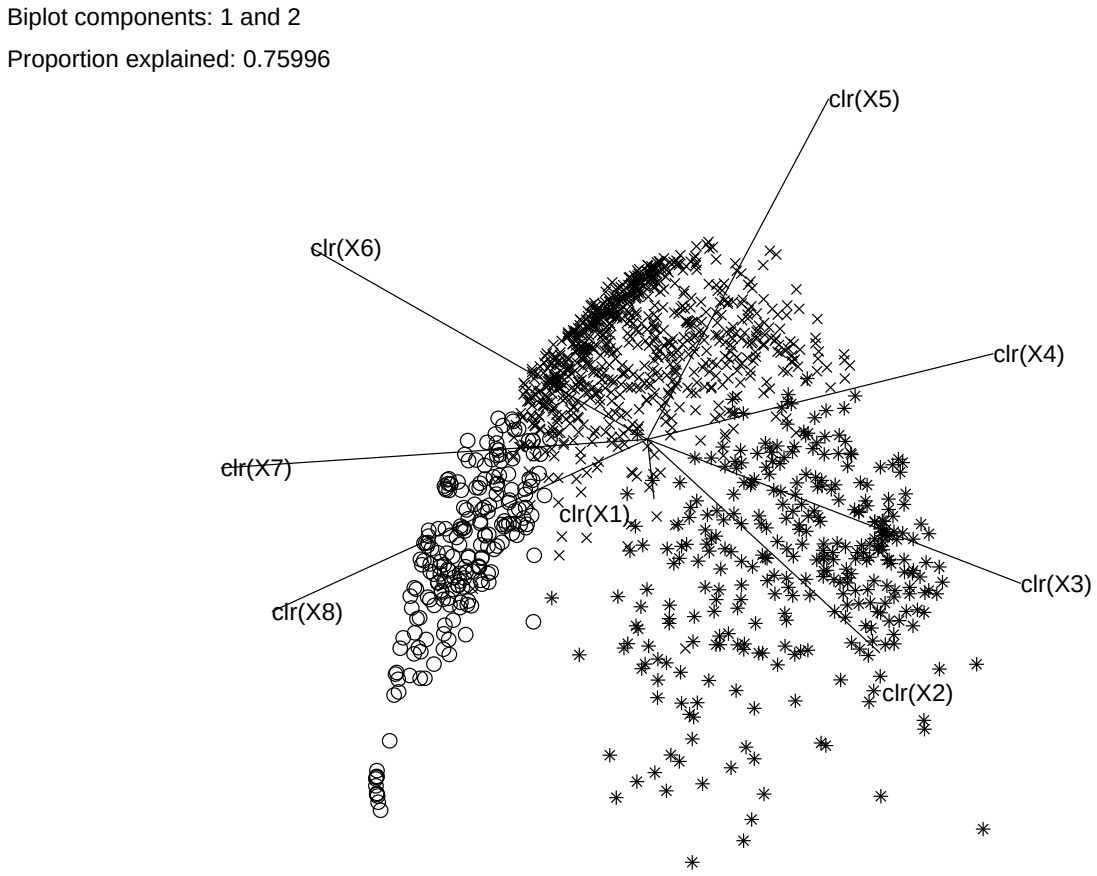


Figura 5.15: Diagrama biplot en el espacio clr del conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$. Se muestran los tres grupos determinados por el método de Ward. ('$\circ$': Grupo 1; '$\times$': Grupo 2; '$*$': Grupo 3).

Este diagrama *biplot*, que recoge un 76% de la variabilidad total, nos confirma el patrón descrito a partir de las otras figuras. Observemos que si se recorre el diagrama girando en el sentido de las agujas del reloj, se marca un movimiento desde las observaciones pertenecientes al Grupo 1 hasta las observaciones del Grupo 3. Se aprecia que las observaciones del Grupo 1 aparecen próximas a los ejes de las componentes clr($\mathbf{X}6$), clr($\mathbf{X}7$), y clr($\mathbf{X}8$), y que las observaciones pertenecientes al Grupo 3 aparecen situadas en posiciones cercanas al resto de los ejes. El Grupo 2 sigue mostrándose como un grupo transición entre el Grupo 1 y el Grupo 3.

Recordemos que el objetivo de este estudio era realizar un mapa del fondo marino del *Darss Sill* donde se distinguieran las zonas según el tipo de sedimento. La figura 5.16 muestra un mapa muy rudimentario donde se han situado las observaciones representándolas mediante símbolos diferentes según el grupo al que pertenecen. En el mapa de la figura se observa que las observa-

ciones del Grupo 1 aparecen situadas mayoritariamente en la zona central del mapa. También aparecen observaciones pertenecientes al Grupo 1 en el borde norte de la zona situada al suroeste del mapa. Esta zona es una área donde la erosión, el transporte y la deposición de los sedimentos es más cambiante. Entre los geólogos expertos esta zona se conoce como la zona del *canal*. Las observaciones del grupo 3 aparecen predominantemente en la zona situada más al noreste del mapa. Sin embargo, también se encuentran observaciones del Grupo 3 en la zona del *canal*. Las observaciones del Grupo 2, que se ha descrito como un grupo transición, aparecen mayoritariamente en la zona del *canal*.

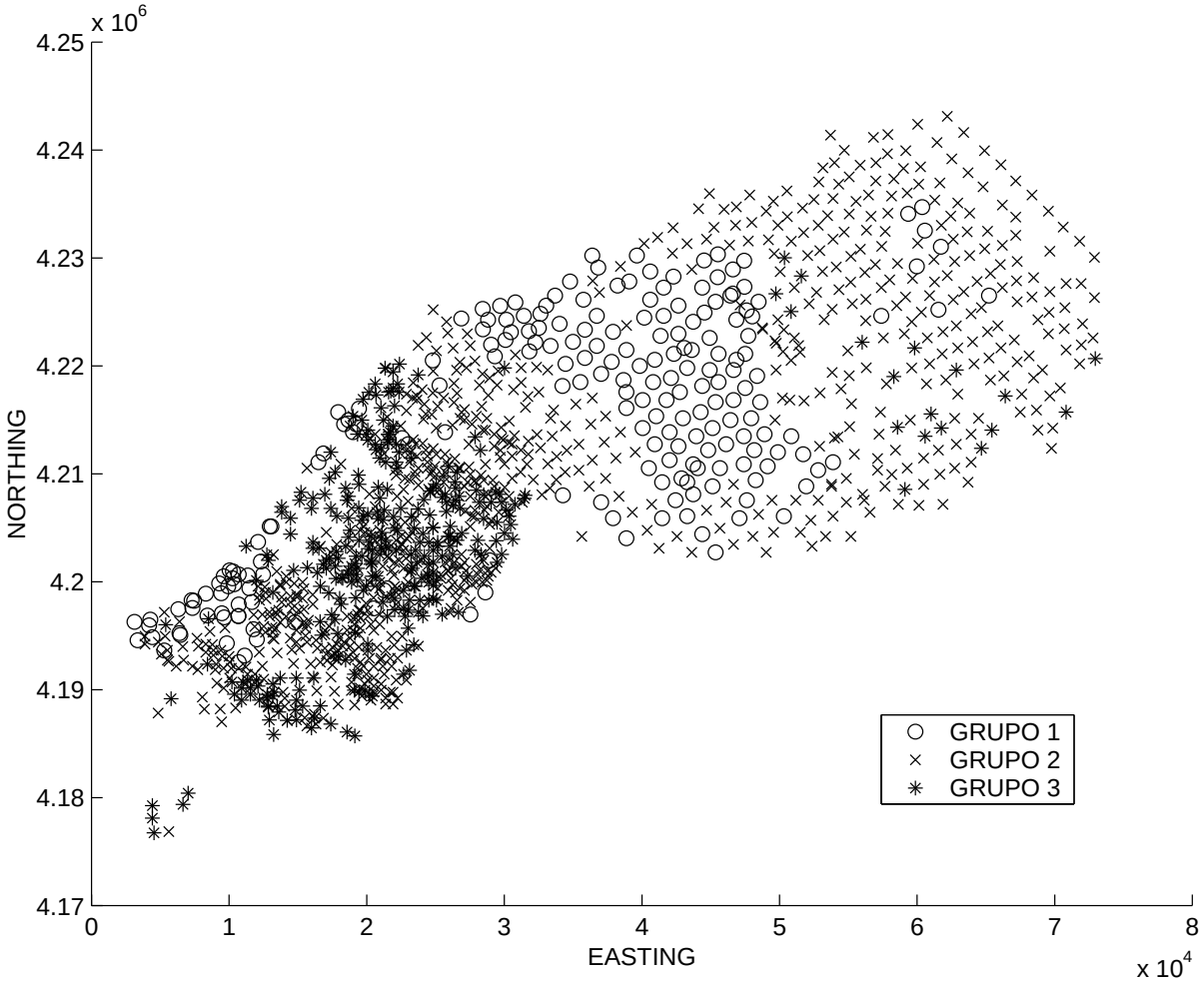


Figura 5.16: Mapa de la clasificación en tres grupos obtenida al aplicar el método de Ward al conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$. Los tres grupos se representan mediante diferentes símbolos ('$\circ$': Grupo 1; '$\times$': Grupo 2; '$*$': Grupo 3).

Al someter a comparación los resultados de nuestra clasificación en tres grupos con los resultados obtenidos en los principales trabajos que han tratado la clasificación del conjunto de datos *Darss Sill* (Davis et al., 1995; Martín-Fernández et al., 1997; Pawlowsky et al., 1997) se aprecia que

en estos trabajos se considera un mayor número de grupos. Con el propósito de averiguar si considerando un mayor número de grupos en nuestra clasificación, se obtienen agrupaciones razonables llevamos a cabo clasificaciones en más de los tres grupos iniciales. Siguiendo el criterio de buscar clasificaciones razonables, analizamos las agrupaciones que contemplan 4 o más grupos, siguiendo el método de clasificación de Ward –véase la figura 5.12. A la vista de los resultados de este análisis consideramos que la agrupación en seis grupos es una clasificación razonable. En la tabla 5.15 se muestra el cardinal de cada uno de los grupos y algunos de sus estadísticos básicos. Recordemos que una de las características del método de Ward es la tendencia a formar grupos con un número similar de observaciones. En la tabla 5.15 se aprecia que en nuestro caso es así, puesto que las 1281 observaciones del conjunto se han repartido en una cantidad relativamente similar entre los seis grupos.

Tabla 5.15: Estadísticos básicos de los seis grupos del conjunto *Darss Sill* determinados por el método de Ward. Se muestran los resultados redondeados a la segunda cifra decimal.

Componente	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$
Grupo 1 (209 obs.)								
Mediana	0	0	0	0	0.01	0.74	0.16	0.02
Mín.-Máx.	0-0	0-0.01	0-0.01	0-0.01	0-0.31	0.1-0.97	0.01-0.6	0-0.69
Grupo 2 (217 obs.)								
Mediana	0	0	0	0.01	0.24	0.6	0.08	0.01
Mín.-Máx.	0-0	0-0	0-0.03	0-0.2	0.01-0.76	0.01-0.95	0.01-0.43	0-0.47
Grupo 3 (280 obs.)								
Mediana	0	0	0	0.01	0.4	0.58	0.01	0
Mín.-Máx.	0-0	0-0.02	0-0.01	0-0.48	0-0.92	0.02-0.96	0-0.14	0-0.02
Grupo 4 (259 obs.)								
Mediana	0	0	0.01	0.05	0.59	0.28	0.02	0
Mín.-Máx.	0-0	0-0.01	0-0.47	0-0.7	0.07-0.88	0.01-0.84	0-0.26	0-0.28
Grupo 5 (164 obs.)								
Mediana	0	0.05	0.11	0.21	0.47	0.06	0.01	0
Mín.-Máx.	0-0.19	0-0.39	0.01-0.43	0.03-0.59	0.02-0.78	0-0.7	0-0.15	0-0.03
Grupo 6 (152 obs.)								
Mediana	0	0.03	0.04	0.1	0.33	0.22	0.05	0.02
Mín.-Máx.	0-0.43	0-0.18	0-0.29	0-0.37	0.04-0.82	0.01-0.73	0-0.27	0-0.71

En la figura 5.17 se observa como los valores tomados en las componentes siguen un patrón diferenciado según el grupo que consideremos. Los grupos se han ordenado de manera que este patrón quedara lo más patente posible.

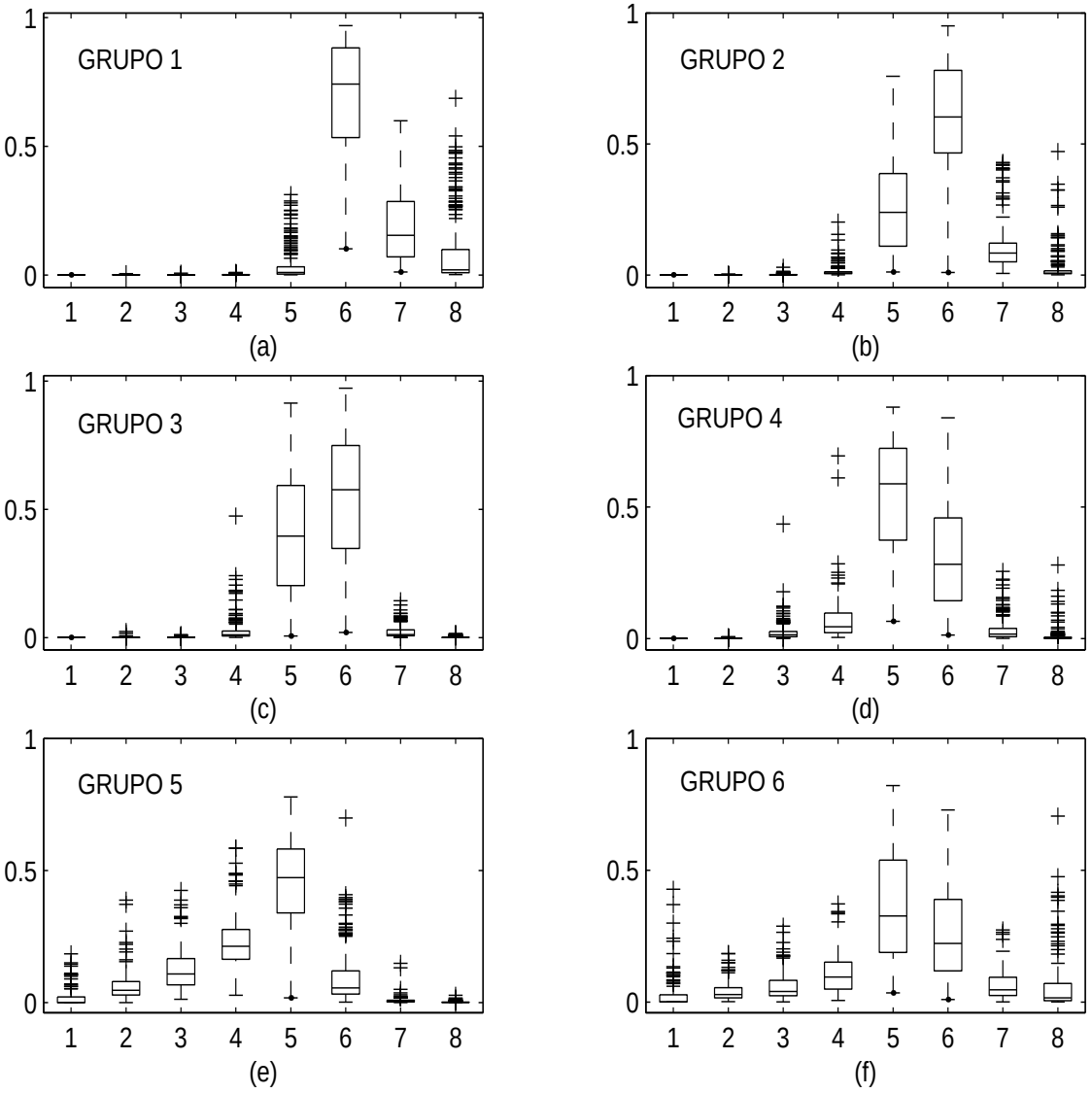


Figura 5.17: Diagramas de caja de los seis grupos determinados por el método de Ward aplicado al conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$: (a) *grupo 1;* (b) *grupo 2;* (c) *grupo 3;* (d) *grupo 4;* (e) *grupo 5;* (f) *grupo 6.*

En esta figura 5.17 se aprecia que los valores tomados en las componentes $\mathbf{X}_6$, $\mathbf{X}_7$, y $\mathbf{X}_8$ van disminuyendo a medida que pasamos del Grupo 1 al Grupo 5, y que al llegar al Grupo 6 estos valores, invirtiendo su tendencia, aumentan. Los valores tomados en las cinco primeras componentes siguen el patrón contrario: van aumentando según pasamos del Grupo 1 al Grupo 5, y en el Grupo 6 invierten su tendencia y disminuyen.

Este mismo patrón de comportamiento del valor de las componentes puede apreciarse en la figura 5.18 que muestra las medias geométricas composicionales de cada grupo.

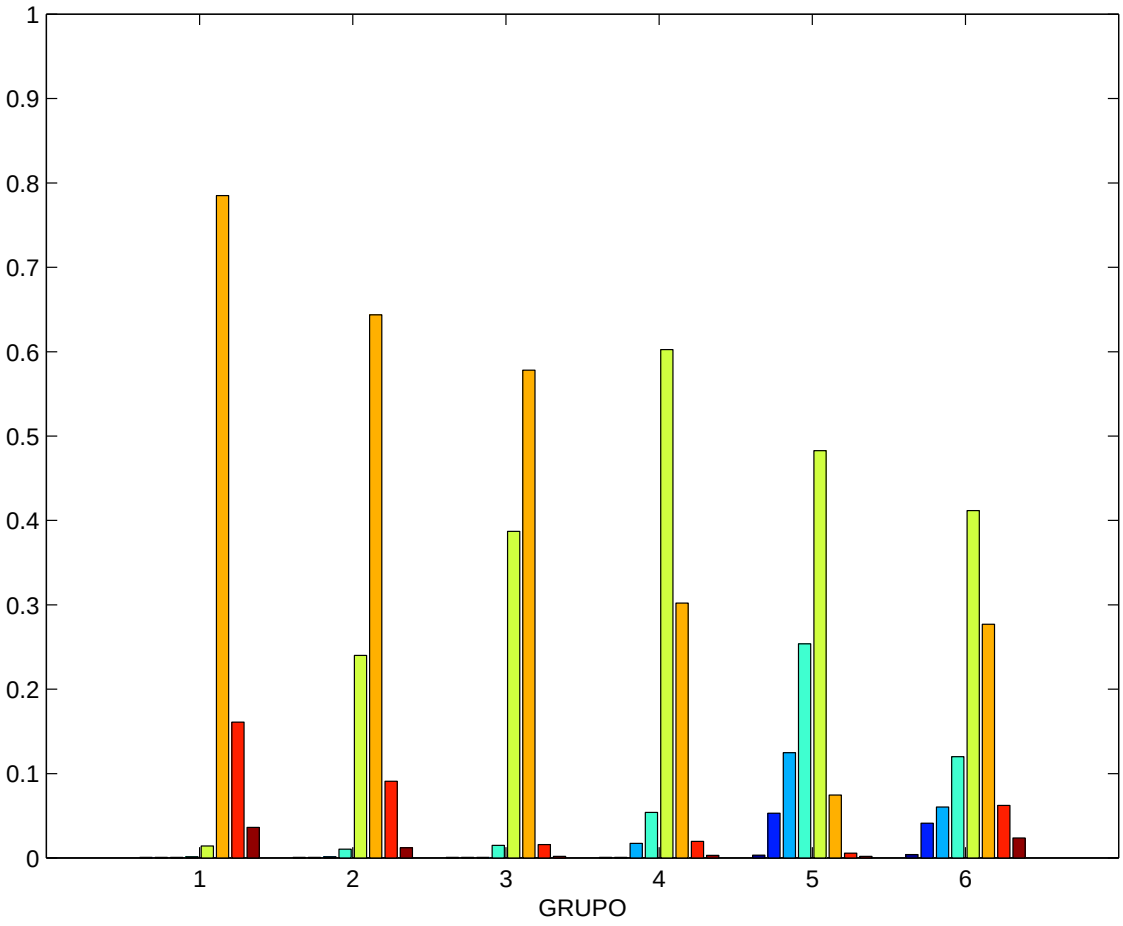


Figura 5.18: Diagramas de barras de las medias geométricas composicionales de los seis grupos determinados por el método de Ward aplicado al conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$ ('1': *Grupo 1;* '2': *Grupo 2;* '3': *Grupo 3;* '4': *Grupo 4;* '5': *Grupo 5;* '6': *Grupo 6).*

Si se representa, en el espacio clr-transformado, el diagrama *biplot* del conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$, se obtiene la figura 5.19. En esta figura se observa que la ordenación de los grupos se corresponde con un giro en el sentido de las agujas del reloj. La transición de valores altos a bajos en las variables $\mathbf{X}_6$, $\mathbf{X}_7$, y $\mathbf{X}_8$ se traduce en el *biplot* en el hecho que las observaciones aparecen situadas más o menos próximas a los ejes clr($\mathbf{X}6$), clr($\mathbf{X}7$), y clr($\mathbf{X}8$). En esta misma figura puede observarse la relación de los grupos resultado de la clasificación en seis grupos con los grupos resultado de la clasificación en tres grupos.

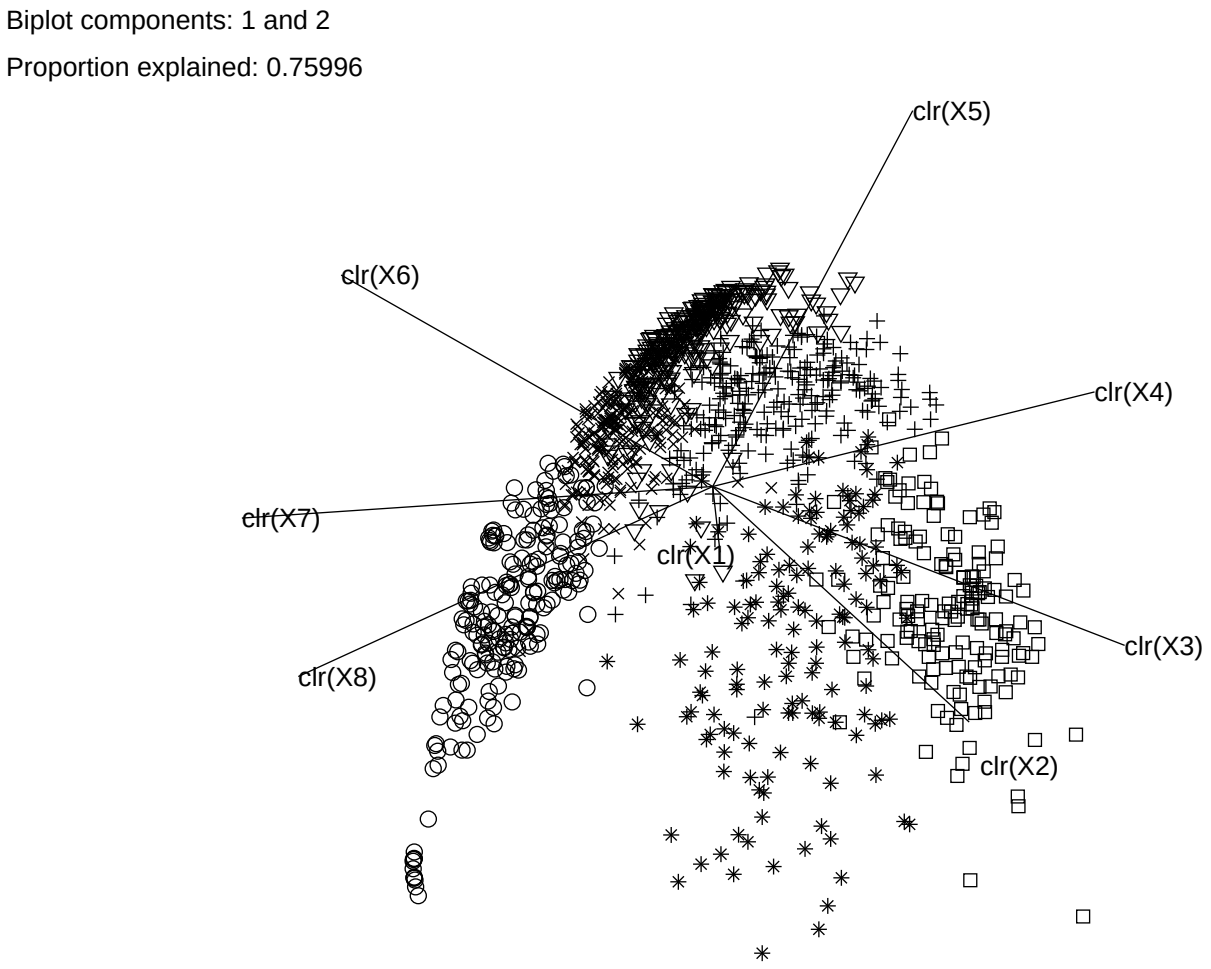


Figura 5.19: Diagrama biplot en el espacio clr del conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$. Se muestran los seis grupos determinados por el método de Ward. ('$\circ$': Grupo 1; '$\times$': Grupo 2; '∇': Grupo 3; '$+$': Grupo 4; '$\square$': Grupo 5; '$*$': Grupo 6).

Si se compara esta figura con la figura 5.15 se observa que el *antiguo* Grupo 1 o Grupo 1 de la clasificación en tres grupos aparece inalterado en la clasificación en seis grupos. El *antiguo* Grupo 2 se ha dividido en los *nuevos* Grupo 2, Grupo 3, y Grupo 4; y que las observaciones pertenecientes al *antiguo* Grupo 3 se han repartido entre los *nuevos* Grupo 5 y Grupo 6.

Esta relación entre las dos clasificaciones también se constata al comparar los dos mapas respectivos –véanse las figuras 5.16 y 5.20. En los mapas se aprecia que las observaciones pertenecientes al Grupo 1 siguen apareciendo situadas en la zona central y borde del *canal*. Las observaciones de los *nuevos* Grupo 2 y Grupo 3 aparecen predominantemente en la zona más noreste del mapa. Sin embargo, también aparecen en una franja de la zona central y en la zona del *canal*. Finalmente, en el mapa se aprecia que las observaciones de los *nuevos* Grupo 4, Grupo 5 y Grupo 6 aparecen mayoritariamente en la zona del *canal*.

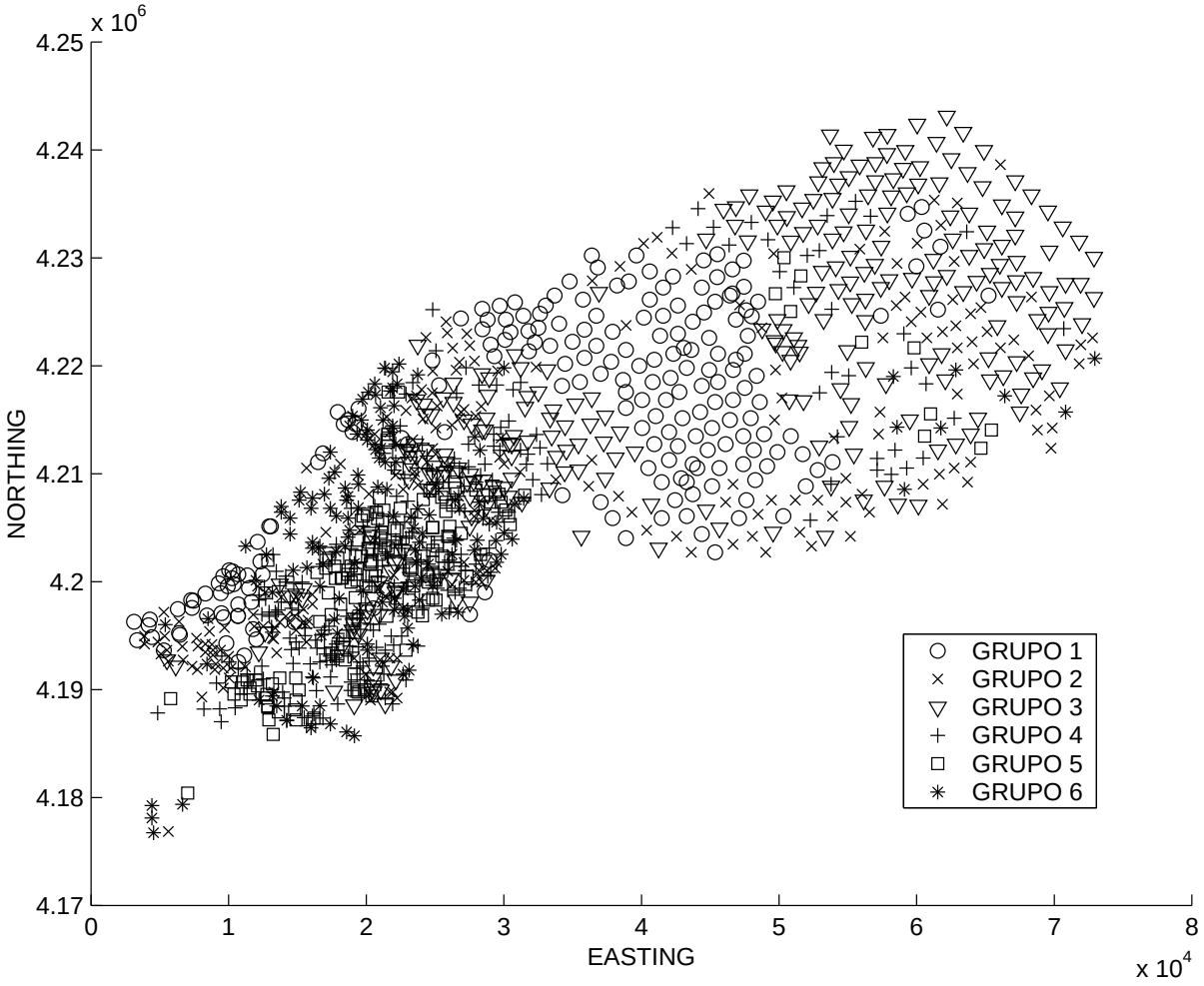


Figura 5.20: Mapa de la clasificación en seis grupos obtenida al aplicar el método de Ward al conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$. Los seis grupos se representan mediante diferentes símbolos ('$\circ$': Grupo 1; '$\times$': Grupo 2; '∇': Grupo 3; '$+$': Grupo 4; '$\square$': Grupo 5; '$*$': Grupo 6).

Con el objeto de ilustrar lo que sucede al considerar un mayor número de grupos, presentamos en la figura 5.21 el diagrama *biplot* considerando los siete grupos resultado de la aplicación del método de Ward. En esta figura se observa que las observaciones pertenecientes al Grupo 6 de la clasificación en seis grupos, al considerar siete grupos, se reparten entre el Grupo 6 y el Grupo 7. Si se representan los diagramas de caja correspondientes o se construyen los diagramas de barras de las medias geométricas composicionales, se concluye que esta agrupación en siete grupos no es una clasificación razonable. En los trabajos consultados (Davis et al., 1995; Martín-Fernández et al., 1997) los autores consideran siete grupos. Sin embargo, los propios autores reconocen que seis de los siete grupos aparecen bien diferenciados, y que uno de los grupos no posee un patrón claro de las variables que permita distinguirlo de los demás grupos.

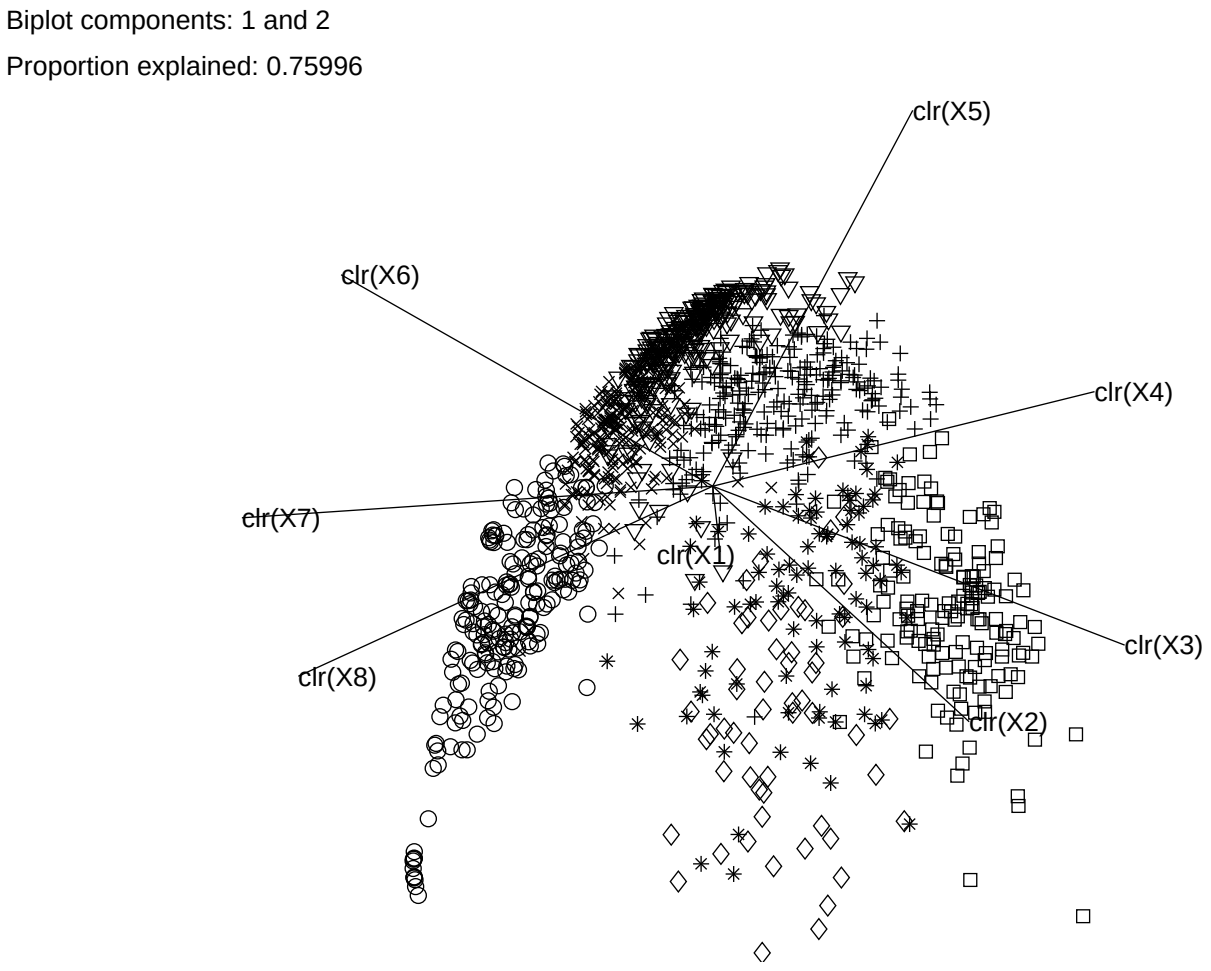


Figura 5.21: Diagrama biplot en el espacio clr del conjunto resultante de reemplazar los ceros del conjunto *Darss Sill* por $\delta = 0.00055$. Se muestran los siete grupos determinados por el método de Ward. ('$\circ$': Grupo 1; '$\times$': Grupo 2; '∇': Grupo 3; '$+$': Grupo 4; '$\square$': Grupo 5; '$*$': Grupo 6; '$\diamond$': Grupo 7.).

Capítulo 6

Epílogo

Durante el desarrollo de esta tesis se ha hecho patente que difícilmente un tema de investigación queda completamente cerrado después de su estudio. Más bien al contrario: a medida que va profundizándose en un tema de investigación, van surgiendo nuevos problemas y nuevas vías de desarrollo. En esta sección presentamos de manera somera las conclusiones relacionadas con los objetivos marcados al inicio del trabajo de investigación y exponemos una lista de ideas, problemas y demás aportaciones que han ido surgiendo a medida que se desarrollaba el trabajo de investigación relacionado con la clasificación de conjuntos de datos composicionales. Somos conscientes que el estudio en profundidad de cada una de estas cuestiones es, en realidad, una línea de investigación a desarrollar en el futuro.

6.1 Conclusiones

- En el desarrollo de esta tesis se ha puesto de manifiesto que a pesar de los defectos y limitaciones que poseen los métodos jerárquicos de clasificación, estos métodos siguen siendo los más utilizados en la realización de clasificaciones no paramétricas. Al realizar una revisión de los métodos de clasificación automática no paramétrica más usuales se ha podido constatar la existencia de tres elementos clave: las medidas de diferencia, las medidas de tendencia central y las medidas de dispersión. En esta tesis se ha expuesto que las medidas de diferencia, de tendencia central y de dispersión usuales no son coherentes con la naturaleza de los datos composicionales. En consecuencia su utilización puede conducirnos a resultados erróneos e incoherentes.

- En este trabajo de investigación se ha puesto de manifiesto que el símplex tiene estructura de espacio vectorial métrico sobre el cuerpo de los reales. Hemos estudiado en profundi-

dad las propiedades de las operaciones perturbación y producto por escalar. Entre estas propiedades destacamos las relacionadas con la transformación logratio centrada porque ponen de manifiesto que esta transformación establece una isometría entre el símplex $\mathcal{S}^D$ y el espacio real $\mathbb{R}^{D-1}$. En relación con las operaciones básicas definidas en el símplex, hemos revisado los requisitos que debe cumplir cualquier medida de diferencia entre datos composicionales. Se ha podido constatar que entre las medidas de diferencia más usuales, únicamente la *distancia de Aitchison* satisfacía los requerimientos.

- En esta tesis se ha realizado una revisión de las medidas de divergencia para distribuciones de probabilidad multinomiales. De este estudio ha surgido la propuesta de las medidas de diferencia que hemos denominado *medidas de divergencia composicionales*. Entre estas medidas se ha definido la *disimilitud de Kullback-Leibler composicional*. Se ha demostrado su coherencia con la transformación subcomposición y su compatibilidad con la operación perturbación. Un estudio en profundidad de la disimilitud de Kullback-Leibler composicional nos ha permitido demostrar que esta medida está íntimamente relacionada con la distancia de Aitchison a través de una aproximación por una transformación monótona. Esta característica nos permite asegurar que las dos medidas de diferencia darán lugar a clasificaciones aproximadamente iguales en la aplicación de métodos jerárquicos de clasificación invariantes por transformaciones monótonas. En el desarrollo de los casos prácticos presentados en esta tesis se ha podido constatar la total coincidencia de resultados al utilizar una u otra medida.

- En esta memoria se han expuesto medidas de tendencia central y de dispersión apropiadas para un conjunto de datos composicionales. Hemos demostrado que la media geométrica composicional es una medida de tendencia central adecuada para los conjuntos de datos composicionales. Al relacionar esta medida de tendencia central con la operación perturbación ha surgido de manera natural la operación *centrado de un conjunto de datos*. Esta operación ha resultado ser una transformación de enorme utilidad en el estudio práctico de conjuntos de datos con valores muy cercanos a cero. Hemos probado que la variabilidad total composicional es una medida de dispersión compatible con la estructura de espacio vectorial del símplex y hemos analizado su relación con la distancia de Aitchison.

- El estudio de medidas de diferencia, de tendencia central y de dispersión coherentes con la naturaleza composicional nos ha permitido adaptar los diferentes métodos de clasificación automática no paramétrica para su aplicación sobre datos composicionales. Es importante

recordar que en esta tesis se ha demostrado que si se elige la distancia de Aitchison, la clasificación de un conjunto de datos composicionales equivale a clasificar los datos clr-transformados utilizando la distancia euclídea. En consecuencia, todas las virtudes y todos los defectos de los diferentes métodos de clasificación aplicados al caso real se pueden trasladar a la clasificación de datos composicionales.

- En esta tesis se han expuesto los aspectos a tener en cuenta cuando se pretende realizar una clasificación de un conjunto de datos composicionales cuyas partes contienen valores iguales a cero. A este problema se le conoce como *el problema de los ceros*. Se han analizado los inconvenientes que presentan los métodos de reemplazamiento de ceros existentes cuando se aplican a datos composicionales y se ha puesto de manifiesto que estos métodos no son coherentes con las operaciones básicas definidas en el símplex. Se han revisado los métodos más usuales en el tratamiento de datos censurados en conjuntos de datos del espacio real. Hemos definido una nueva *fórmula de reemplazamiento de los ceros por redondeo* y hemos analizado en profundidad sus propiedades. A partir de dos casos prácticos hemos analizado la sensibilidad del proceso de substitución de ceros en la clasificación resultante.

6.2 Líneas de investigación futuras

A continuación presentamos una relación de cuestiones abiertas a estudio que nos han ido apareciendo durante el desarrollo de nuestro trabajo de investigación:

- ¿Qué relaciones, diferencias y/o semejanzas pueden establecerse entre el estudio de los datos composicionales y el estudio de datos categóricos multivariantes? ¿Pueden utilizarse en el estudio de los datos composicionales técnicas parecidas a las utilizadas en el estudio de las distancias, clasificaciones y agregaciones de variables de datos categóricos?

- ¿Qué peculiaridades podemos encontrar en las técnicas de clasificación paramétrica cuando se aplican a datos composicionales? ¿Qué distribuciones de probabilidad sobre el símplex podemos considerar a la hora de realizar una clasificación de este tipo?

- ¿Pueden relacionarse los elementos definidos por Aitchison para datos composicionales con el estudio de las distribuciones multivariantes discretas? ¿Hasta que punto es factible aplicar la distancia de Aitchison en el espacio de los parámetros de una distribución multinomial?

- ¿Qué resultados se obtendrían si se aplicasen a los datos composicionales los conceptos sobre análisis multivariante de datos basados en distancias expuestos en los trabajos de Cuadras y de Rao?

- ¿Puede definirse una transformación tipo Box-Cox que sea la generalización de la transformación log-ratio centrada? ¿Qué ventajas y qué desventajas tendría una transformación de este tipo?

Apéndice A

Conjuntos de datos

A.1 *Población ocupada por grupos profesionales (1991)*

Comarca	Prof. y téc.	Pers. Direct.	Serv. Adm.	Comerc. y vend.	Hostel. y otros	Agricul. y pesca	Indúst.	Fuer. Arm.
1. Alt Camp (ALC)	1231	243	1446	1420	875	1265	6286	25
2. Alt Empordà (ALE)	2948	793	5040	5510	4823	3509	12083	317
3. Alt Penedès (ALP)	2419	502	3667	3077	2000	1827	13118	36
4. Alt Urgell (ALU)	778	135	835	1020	798	1068	2777	79
5. Alta Ribagorça (ALR)	175	23	98	131	199	163	469	1
6. Anoia (ANO)	2764	614	3462	3556	2408	1124	17472	43
7. Bages (BAG)	6274	1022	6485	7095	4570	1755	28255	171
8. Baix Camp (BCM)	5699	989	6165	7029	5221	3270	18436	110
9. Baix Ebre (BEB)	2446	383	2311	2808	1994	3682	8846	65
10. Baix Empordà (BEM)	2810	737	3716	4900	4635	2747	14519	127
11. Baix Llobregat (BLL)	12371	4009	31296	26849	24955	2605	110826	274
12. Baix Penedès (BPE)	1116	320	1705	1997	1762	785	6305	49
13. Barcelonès (BCN)	146521	24845	182813	126740	95496	3462	274395	1258
14. Berguedà (BER)	1373	164	1207	1555	1131	1129	6910	78
15. Cerdanya (CER)	492	116	462	679	786	670	1695	38
16. Conca de Barberà (CBB)	563	124	636	631	488	1068	3018	7
17. Garraf (GRF)	3484	549	3419	3875	3559	836	11448	43
18. Garrigues (GAR)	539	79	524	619	424	2338	2286	13
19. Garrotxa (GRT)	1909	390	2064	2037	1420	1264	9712	32
20. Gironès (GIR)	7315	1187	8884	7173	5127	1727	19917	269
21. Maresme (MAR)	12837	3475	15056	15560	10867	4504	45818	189
22. Montsià (MON)	1329	282	1600	2046	1394	4588	7716	77
23. Noguera (NOG)	1131	185	931	1226	824	3215	7911	35
24. Osona (OSO)	4901	901	5277	5423	3238	3076	26436	50
25. Pallars Jussà (PLJ)	567	79	479	465	410	955	1530	101
26. Pallars Sobirà (PLS)	280	27	200	148	307	497	620	6
27. Pla d'Urgell (PUR)	863	169	1019	1020	597	2570	4200	24

Conjunto de datos de *Población ocupada* (continúa)

Comarca	Prof. y téc.	Pers. Direct.	Serv. Adm.	Comerc. y vend.	Hostel. y otros	Agricul. y pesca	Indúst.	Fuer. Arm.
28. Pla de l'Estany (PES)	923	187	1036	881	587	804	4004	8
29. Priorat (PRI)	287	34	245	255	232	1063	1179	10
30. Ribera d'Ebre (REB)	936	75	684	657	592	1318	3263	27
31. Ripollès (RIP)	1012	193	905	1106	1006	801	5908	27
32. Segarra (SRR)	654	125	653	560	415	1152	3023	6
33. Segrià (SEG)	7841	1279	8280	8294	6253	8678	18970	577
34. Selva (SEL)	2776	744	4106	4720	5758	2149	17562	66
35. Solsonès (SOL)	431	61	330	315	348	900	1854	6
36. Tarragonès (TAR)	8047	1201	9403	7294	7309	1640	21352	348
37. Terra Alta (TAL)	217	41	220	324	209	1757	1710	16
38. Urgell (URG)	1020	235	1099	1431	758	1991	4699	31
39. Vall d'Aran (VAR)	295	182	286	360	562	143	779	32
40. Vallès Occidental (VOC)	28614	5383	34772	31343	21310	1610	114191	231
41. Vallés Oriental (VOR)	9550	2250	13548	11619	8395	2499	54530	122

Grupos profesionales:

- Prof. y téc.: Profesionales y técnicos.
- Pers. Direct.: Personal Directivo.
- Serv. Adm.: Servicios Administrativos.
- Comerc. y vend.: Comerciantes y vendedores.
- Hostel. y otros: Hostelería y otros
- Agricul. y pesca: Agricultura y pesca.
- Indúst.: Indústria.
- Fuer. Arm.: Fuerzas Armadas.

A.2 *Glacial data set*

Observación	*red*	*gray*	*crystalline*	*miscellaneous*
1	0.9770	0.0230	0.0000	0.0000
2	0.5876	0.4124	0.0000	0.0000
3	0.9590	0.0410	0.0000	0.0000
4	0.1570	0.8430	0.0000	0.0000
5	0.0010	0.9990	0.0000	0.0000
6	0.0452	0.9548	0.0000	0.0000
7	0.9560	0.0420	0.0020	0.0000
8	0.7510	0.2460	0.0030	0.0000
9	0.2780	0.7190	0.0030	0.0000
10	0.1131	0.8829	0.0040	0.0000
11	0.2242	0.7698	0.0060	0.0000
12	0.1582	0.8358	0.0060	0.0000
13	0.7050	0.2880	0.0070	0.0000
14	0.6470	0.3450	0.0080	0.0000
15	0.8989	0.0931	0.0080	0.0000
16	0.1590	0.8330	0.0080	0.0000
17	0.8260	0.1650	0.0090	0.0000
18	0.9379	0.0521	0.0100	0.0000
19	0.8560	0.1340	0.0100	0.0000
20	0.9180	0.0710	0.0110	0.0000
21	0.8910	0.0980	0.0110	0.0000
22	0.9260	0.0630	0.0110	0.0000
23	0.9120	0.0770	0.0110	0.0000
24	0.9530	0.0360	0.0110	0.0000
25	0.2750	0.7140	0.0110	0.0000
26	0.1150	0.8740	0.0110	0.0000
27	0.9520	0.0360	0.0120	0.0000
28	0.9400	0.0480	0.0120	0.0000
29	0.8020	0.1840	0.0140	0.0000
30	0.8940	0.0920	0.0140	0.0000
31	0.8330	0.1510	0.0160	0.0000
32	0.8560	0.1200	0.0240	0.0000
33	0.3140	0.6590	0.0270	0.0000
34	0.2530	0.7110	0.0360	0.0000
35	0.5455	0.4094	0.0450	0.0000
36	0.3430	0.5980	0.0590	0.0000
37	0.8959	0.0981	0.0050	0.0010
38	0.9190	0.0750	0.0040	0.0020
39	0.8340	0.1570	0.0070	0.0020
40	0.1149	0.8601	0.0230	0.0020

Conjunto de datos *Glacial* (continúa)

Observación	*red*	*gray*	*crystalline*	*miscellaneous*
41	0.9101	0.0789	0.0080	0.0030
42	0.9009	0.0821	0.0130	0.0040
43	0.8042	0.1908	0.0000	0.0050
44	0.8890	0.1010	0.0050	0.0050
45	0.8859	0.0921	0.0170	0.0050
46	0.8320	0.1450	0.0160	0.0070
47	0.8162	0.1758	0.0000	0.0080
48	0.0780	0.9060	0.0080	0.0080
49	0.7790	0.2110	0.0000	0.0100
50	0.6720	0.3150	0.0030	0.0100
51	0.0340	0.9520	0.0040	0.0100
52	0.8050	0.1780	0.0070	0.0100
53	0.8730	0.1090	0.0080	0.0100
54	0.7900	0.1680	0.0300	0.0120
55	0.0704	0.8985	0.0181	0.0131
56	0.2012	0.7838	0.0010	0.0140
57	0.2700	0.7020	0.0130	0.0150
58	0.9150	0.0660	0.0030	0.0160
59	0.1870	0.7930	0.0040	0.0160
60	0.1950	0.7810	0.0080	0.0160
61	0.7093	0.2577	0.0150	0.0180
62	0.7253	0.2258	0.0310	0.0180
63	0.8460	0.1350	0.0000	0.0190
64	0.9613	0.0075	0.0108	0.0204
65	0.7310	0.2360	0.0140	0.0190
66	0.5110	0.4650	0.0030	0.0210
67	0.1520	0.8070	0.0190	0.0220
68	0.7968	0.1732	0.0070	0.0230
69	0.1950	0.7790	0.0000	0.0260
70	0.1502	0.8068	0.0170	0.0260
71	0.4980	0.4681	0.0070	0.0269
72	0.7562	0.2098	0.0060	0.0280
73	0.6249	0.3190	0.0281	0.0281
74	0.7417	0.2002	0.0290	0.0290
75	0.2150	0.7500	0.0040	0.0310

Conjunto de datos *Glacial* (continúa)

Observación	*red*	*gray*	*crystalline*	*miscellaneous*
76	0.7342	0.2156	0.0191	0.0311
77	0.8799	0.0851	0.0000	0.0350
78	0.1474	0.8134	0.0030	0.0361
79	0.5770	0.3510	0.0360	0.0360
80	0.7069	0.1506	0.1082	0.0343
81	0.8160	0.1000	0.0450	0.0390
82	0.6650	0.2870	0.0080	0.0400
83	0.1820	0.7660	0.0040	0.0480
84	0.6530	0.1990	0.0990	0.0490
85	0.2120	0.7280	0.0080	0.0520
86	0.7483	0.1469	0.0470	0.0579
87	0.1719	0.7558	0.0122	0.0600
88	0.4585	0.3333	0.0831	0.1251
89	0.3175	0.4709	0.0345	0.1772
90	0.5600	0.2630	0.0230	0.1540
91	0.4150	0.3640	0.0510	0.1700
92	0.3704	0.3614	0.0400	0.2282

Referencias

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London (GB): Chapman and Hall. 416 pp.

Aitchison, J. (1990). Comment on "Measures of variability for geological data" of Watson, D. F. and Philip, G. M. *Mathematical Geology 22*(2), 223–226.

Aitchison, J. (1991). Delusions of uniqueness and ineluctability. *Mathematical Geology 23*(2), 275–277.

Aitchison, J. (1992). On criteria for measurements of compositional difference. *Mathematical Geology 24*(4), 365–379.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97, The Third Annual Conference of the International Association for Mathematical Geology*, Volume 1, Barcelona (España), pp. 3–35. International Center for Numerical Methods in Engineering (CIMNE).

Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2000). Logratio analysis and compositional distance. *Mathematical Geology 32*(3), 271–275.

Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Reply to letter to the editor by S. Rehder and U. Zier on "Logratio analysis and compositional distance" by J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández and V. Pawlowsky-Glahn. *Mathematical Geology*. (En prensa).

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York (USA): Academic Press. 359 pp.

Art, D., R. Gnanadesikan, and R. Kettenring (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica 21A*, 75–99.

Barceló, C. (1996). *Mixtures of Compositional Data*. Ph. D. thesis, Universitat Politècnica de

Catalunya, Barcelona (E).

Barceló-Vidal, C. (2000, Septiembre). Fonamentació matemàtica de l'anàlisi de dades composicionals. Technical Report 00-02-RR, Departament d'Informàtica i Matemàtica Aplicada. Universitat de Girona, Girona (Spain). Report de Recerca.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (1999). Comment on "Singularity and Nonnormality in the Classification of Compositional Data" by Bohling, G. C. et al. *Mathematical Geology 31*(5), 581–586.

Bohling, G. C., J. C. Davis, R. A. Olea, and J. Harff (1996). Singularity and nonnormality in the classification of compositional data. *Mathematical Geology 30*(1), 5–20.

Bren, M. and V. Batagelj (1997, May). The metric index. Technical Report 35-561, Institute of Mathemtics, Physics and Mechanics. Department of Mathematics. University of Ljubljana, Ljubljana (Slovenia). Preprint Series.

Bren, M. and J. A. Martín-Fernández (1999). Measures of difference for compositional data. In *Abstracts of ICMS'99, International Conference on the Methodology and Statistics*, Preddvor (Slovenia), pp. 22. International Center for Numerical Methods in Engineering (CIMNE).

Burbea, J. (1983). J-divergences and related concepts. In *Encyclopedia of Statistical Sciences*, Volume 4, pp. 290–296. New York (USA): John Wiley and Sons.

Burbea, J. and C. R. Rao (1982). Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of multivariate analysis 12*, 575–596.

Chang, W. C. (1983). On using principal components before separating a mixture of two multivariate normal distribution. *Appl. Statist. 32*, 267–275. [Citado en McLachlan, G. J. 1992,*Discriminant Analysis and Statistical Pattern Recognition:* John Wiley & Sons, New York (USA), 526 pp.]

Chayes, F. (1983). Detecting nonrandom association between proportions by test of remaining-space variables. *Mathematical Geology 15*, 197–206. [Citado en Aitchison, J.,(1986)].

Cooper, M. C. and G. W. Milligan (1988). The effect of measurament error on determining the number of clusters in Cluster Analysis. In W. Gaul and M. Shader (Eds.), *Data Expert Knowledge and Decision*, pp. 319–328. Springer.

Cover, T. M. (1991). *Elements of information theory*. New York (USA): John Wiley & Sons. 542 pp.

Cuadras, C. M. (1989). Distancias estadísticas (con discusión). *Estadística Española 30*(119), 295–378.

Cuadras, C. M. (1991). *Métodos de análisis multivariante* (second ed.). Barcelona (E): PPU. 644 pp.

Cuadras, C. M. y C. Arenas (1997). Anàlisis multivariante basado en distancias. Technical report, Universitat de Barcelona, Barcelona (E). (2a edició).

Cuadras, C. M., J. Fortiana, and F. Oliva (1997). The proximity of an individual to a population with applications to discriminant analysis. *Journal of Classification 14*, 117–136.

Davis, J. C., J. Harff, R. Olea, and G. C. Bohling (1995). Regionalized classification of the Darss Sill sediments. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97, The Third Annual Conference of the International Association for Mathematical Geology*, Volume 1, Barcelona (España), pp. 145–150. International Center for Numerical Methods in Engineering (CIMNE).

Doveton, J. H. (1998). Beyond the Perfect Martini: Teaching the Mathematical Geology of petrophysical logs. In A. Buccianti, G. Nardi, and R. Potenza (Eds.), *Proceedings of IAMG'98*, Volume 1, Napoli (Italia), pp. 71–75. The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede Editore.

Everitt, B. S. (1993). *Cluster analysis* (third ed.). New York (USA): Edward Arnold. 170 pp.

Everitt, B. S. and G. Dunn (1991). *Applied multivariate data analysis*. London (GB): Edward Arnold. 304 pp.

Friedman, J. H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc. 82*(320), 249–266.

Gordon, A. D. (1998). Cluster validation. In C. Hayashi (Ed.), *Data Science, Classification, and Related Methods*, Tokyo (Japan), pp. 22–39. Springer.

Gordon, A. D. (1999). *Classification*. London (GB): Chapman & Hall. (second edition). 256 pp.

Gower, J. C. (1983). Measures of similarity, dissimilarity, and distance. In *Encyclopedia of Statistical Sciences*, Volume 5, pp. 397–405. New York (USA): John Wiley and Sons.

Gower, J. C. and D. J. Hand (1996). *Biplots*. London (GB): Chapman & Hall. 277 pp.

Harris, B. (1983). Entropy. In *Encyclopedia of Statistical Sciences*, Volume 2, pp. 512–516. New York (USA): John Wiley and Sons.

Jobson, J. D. (1992). *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Analysis*. New York (USA): Springer-Verlag. 731 pp.

Jones, M. C. and R. Sibson (1987). What is projection pursuit? (with discussion). *J. R. Statist. Soc. A 150*(1), 1–36.

Kaufman, L. and P. J. Rousseuw (1990). *Finding Groups in Data*. New York (USA): John Wiley & Sons, Inc. 342 pp.

Krzanowski, W. J. (1988a). Missing value imputation in multivariate data using the singular value descomposition of a matrix. *Biometrical Letters 25*(1,2), 31–38.

Krzanowski, W. J. (1988b). *Principles of Multivariate Analysis: A User's Perspective*. Oxford (GB): Clarendon Press. 563 pp (reprinted 1996).

Krzanowski, W. J. and F. H. C. Marriot (1994). *Multivariate Analysis. Part 1: Distributions, ordination and inference*. London (GB): Edward Arnold. 280 pp.

Krzanowski, W. J. and F. H. C. Marriot (1995). *Multivariate Analysis. Part 2: Classification, covariance structures and repeated measuraments*. London (GB): Edward Arnold. 280 pp.

Lance, G. N. and W. T. Williams (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal 9*, 373–380. [Citado en Cuadras(1991)].

Little, R. J. A. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. New York (USA): John Wiley and Sons. 278 pp.

MacQueens's, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings Symp. Mathematical Statist. and Probability, 5th.*, Volume 1, Berkeley, AD 669871, Univ. of California, pp. 281–297.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1992). *Multivariate Analysis*. London (GB): Academic Press. 518 pp.

Martín, M. C. (1996). Performance of eight dissimilarity coefficients to cluster a compositional data set. In *Abstracts of IFCS-96. Fifth Conference of International Federation of Classification Societies*, Volume 1, Kobe (Japan), pp. 215–217.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1997). Different classifications of the Darss Sill data set based on mixture models for compositional data. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97, The Third Annual Conference of the International Association for Mathematical Geology*, Volume 1, Barcelona (E), pp. 151–158. International Center for Numerical Methods in Engineering (CIMNE).

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998a). Measures of difference for compositional data and hierarchical clustering methods. In A. Buccianti, G. Nardi, and R. Potenza (Eds.), *Proceedings of IAMG'98*, Volume 2, Napoli (Italia), pp. 526–531. The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede Editore.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998b). A critical approach to non-parametric classification of compositional data. In A. Rizzi, M. Vichi, and H. H. Bock (Eds.), *Advances in Data Science and Classification. Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98)*, Berlin, Heidelberg, New York, pp. 49–56. Università La Sapienza, Roma: Springer-Verlag.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998c). Medida de diferencia Kullback-Leibler entre datos composicionales. In *Libro de actas del XXIV Congreso Nacional de Estadística e Investigación Operativa*, Almería (E), pp. 291–292.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Zero replacement in compositional data sets. In H. Kiers, J. Rasson, P. Groenen, and M. Shader (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization. Proceedings of IFCS'2000*, Namur (Belgium), pp. 155–160. The 7th Conference of the International federation of Classification Societies: Springer-Verlag.

Martín-Fernández, J. A., M. Bren, C. Barceló-Vidal, and V. Pawlowsky-Glahn (1999). A measure of difference for compositional data based on measures of divergence. In S. Lippard, A. Næss, and R. Sinding-Larsen (Eds.), *Proceedings of IAMG'99*, Volume 1, Trondheim (Norway), pp. 211–215. The Fifth Annual Conference of the International Association for Mathematical Geology: Tapir, Trondheim (N).

Martín-Fernández, J. A., R. Olea-Meneses, and V. Pawlowsky-Glahn (2001). Criteria to compare estimation methods of regionalized compositions. *Mathematical Geology*. (En prensa).

Mateu-Figueras, G., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998). Modeling compositional data with multivariate skew-normal distributions. In A. Buccianti, G. Nardi, and R. Potenza (Eds.), *Proceedings of IAMG'98*, Volume 1, Napoli (Italia), pp. 532–537. The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede Editore.

Medak, F. and N. Cressie (1991). Confidence regions in ternary diagrams based on the power-divergence statistics. *Mathematical Geology 23*(8), 1045–1057.

Murtagh, F. and A. Heck (1987). *Multivariate Data Analysis*. Dordrecht: D. Riedel. 210 pp.

Murtagh, F. and M. Hernández-Pajares (1995). The kohonen self-organizing map method: An assessment. *Journal of Classification 12*, 165–190.

Nason, G. (1995). Three-dimensional projection pursuit. *Applied Statistics 44*(4), 411–430.

Pawlowsky, V., G. Simarro, and J. A. Martín (1997). Spatial Cluster Analysis using a Generalised Mahalanobis Distance. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97, The Third Annual Conference of the International Association for Mathematical Geology*, Volume 1, Barcelona (E), pp. 175–180. International Center for Numerical Methods in Engineering (CIMNE).

Pearson, K. (1897). Mathematical contributions to the theory of evolution. on a form of spurious correlation which may arise when indices are used in the mesurements of organs. *Proc. R. Soc. 60*, 489–498.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology 21*, 24–43.

Rao, C. R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Qüestiió 19*(1,2 y 3), 23–63.

Rao, C. R. and S. K. Mitra (1971). *Generalized Inverse of Matrices and its Applications*. New York (USA): John Wiley and Sons. 240 pp.

Rayens, W. S. and C. Srinivasan (1991). Estimation in compositional data analysis. *J. of Chemometrics 5*, 361–374.

Sandford, R. F., C. T. Pierson, and R. A. Crovelli (1993). An objective replacement method for censored geochemical data. *Mathematical Geology 25*(1), 59–80.

Tauber, F. (1999). Spurious clusters in granulometric data caused by logratio transformation. *Mathematical Geology 31*(5), 491–504.

Vives, S. y A. Villarroya (1996). La combinació de técniques de geometria diferencial amb anàlisi multivariant clàssica: Una aplicació a la caracterització de les comarques catalanes. *Qüestiió 20*(3), 449–482.

Watson, D. F. (1991). Reply to "Delusions of uniqueness and ineluctability" of Aitchison, J. *Mathematical Geology 23*(2), 279.

Watson, D. F. and G. M. Philip (1989). Measures of variability for geological data. *Mathematical Geology 21*(2), 233–254.

Watson, D. F. and G. M. Philip (1990). Reply to comment on "Measures of variability for geological data" de Aitchison, J. *Mathematical Geology 22*(2), 227–231.

Zhou, D. (1997). Logratio statistical classification and estimation of hydrodynamic parameters from Darss Sill grain-size data. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97, The Third Annual Conference of the International Association for Mathematical Geology*, Volume 1, Barcelona (E), pp. 139–144. International Center for Numerical Methods in Engineering (CIMNE).

Zhou, D., H. Chen, and Y. Lou (1991). The logratio approach to the classification of modern sediments and sedimentary environments in northern south china sea. *Mathematical Geology 23*(2), 157–165.