Clasificación paramétrica de datos composicionales: aproximación metodológica

J. A. Martín-Fernández[1], *J. Daunis-i-Estadella*[1], *y G. Mateu-Figueras*[1]

[1]Dept. Informàtica i Matemàtica Aplicada, Universitat de Girona, josepantoni.martin@udg.es, josep.daunis@udg.es, gloria.mateu@udg.es

Resumen

Por su propia naturaleza los datos composicionales requieren la aplicación de técnicas estadísticas específicas. Presentamos una aproximación a la metodología a aplicar en las clasificaciones paramétricas, contemplando distribuciones de probabilidad y medidas de tendencia central, de dispersión y de distancia apropiadas a la naturaleza de los datos.

Palabras Clave: Datos de proporciones, diagrama ternario, distribución normal logística, perturbación, símplex, transformación logcociente

AMS: 62H30, 62P99

1. Introducción

En una clasificación paramétrica las fases de elección de la distribución de probabilidad y del método de clasificación desempeñan un papel crucial. Lógicamente, estas fases se apoyarán en los resultados obtenidos en una etapa descriptiva inicial del conjunto de datos en la que las técnicas de reducción de la dimensión desempeñarán un papel decisivo. A continuación, se deberá abordar ineludiblemente la cuestión del número de grupos a formar. Esta decisión se fundamentará en el cálculo de índices subjetivos y en una etapa descriptiva grupo a grupo. Queremos hacer especial hincapié en un hecho muy importante a tener en cuenta: las técnicas estadísticas que se utilizan al realizar una clasificación deben respetar la naturaleza específica de los datos a clasificar. En particular, las distribuciones de probabilidad, las técnicas de reducción de la dimensión, y las herramientas descriptivas deben ser adecuadas para la tipología de los datos.

Por dato composicional o dato de proporciones se entiende la realización de un vector aleatorio de componentes estrictamente positivas y de suma constante. El espacio soporte asociado a los datos composicionales es el símplex $\mathcal{S}^D$ donde se definen las operaciones básicas *perturbación* '$\oplus$', *potenciación* '$\otimes$', y *producto escalar* '$<,>$':

$$\mathcal{S}^D = \{[x_1, \ldots, x_D] : x_i > 0 \ (i = 1, \ldots, D);\ x_1 + \cdots + x_D = 1\}, \qquad (1)$$

$$\mathbf{x} \oplus \mathbf{y} = \left[\frac{x_1 y_1}{\sum x_k y_k}, \ldots, \frac{x_D y_D}{\sum x_k y_k}\right], \qquad a \otimes \mathbf{x} = \left[\frac{x_1^a}{\sum x_k^a}, \ldots, \frac{x_D^a}{\sum x_k^a}\right], \tag{2}$$

$$< \mathbf{x}, \mathbf{y} >= \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \tag{3}$$

donde $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, y a es un número real. Las operaciones $\oplus$, $\otimes$, y '$<,>$', dotan al símplex $\mathcal{S}^D$ de una estructura de espacio vectorial euclidiano de dimensión $D-1$. Esta estructura se debe tener en cuenta cuando se definen distribuciones de probabilidad y medidas de tendencia central, dispersión y distancia, si se desea que estos elementos sean compatibles con la tipología de los datos. Las características matemáticas del soporte de los datos composicionales, sus distribuciones de probabilidad y sus medidas adecuadas han sido analizadas en profundidad en muchos trabajos, entre los que destacamos [1], [2], [19], y [21].

Por su propia naturaleza los datos composicionales requieren la aplicación de técnicas estadísticas específicas. En la literatura no existe una metodología apropiada que paute la realización de una clasificación paramétrica para este tipo de datos. En este trabajo nos centramos en los métodos de clasificación paramétrica que en su formulación suponen que los datos se han generado por una mezcla de distribuciones de probabilidad. En consecuencia, la metodología propuesta debe contemplar distribuciones de probabilidad apropiadas a la naturaleza de los datos y únicamente deben considerarse distribuciones coherentes con las medidas de tendencia central y de dispersión.

En la siguiente sección presentamos una breve descripción de las principales características de las técnicas paramétricas de clasificación. La tercera sección está dedicada a desarrollar con más profundidad las propiedades específicas de los datos composicionales. La aproximación metodológica se expondrá en la cuarta sección, y, finalmente, en la quinta sección extraeremos las conclusiones de nuestro estudio y se propondrá la línea de investigación a seguir en futuros estudios.

2. Clasificación paramétrica: mezcla de distribuciones

En esta sección se presentan los conceptos básicos que subyacen a la aplicación de una técnica de clasificación automática. Exponemos una breve descripción del método paramétrico aglomerativo introducido en [3] y presentamos los fundamentos de la técnica de clasificación mediante mezcla de distribuciones basada en el algoritmo EM. Con el fin de no hacer muy repetitivas las referencias a la bibliografía, se ha optado por suprimirlas en su gran mayoría. Los aspectos generales relacionados con las técnicas de clasificación automática recogidos en esta sección aparecen en la mayoría de textos de Cluster Analysis. En nuestro trabajo se han consultado básicamente las obras [10], [15], [16], y [24]. Los aspectos más específicos relacionados con los métodos de clasificación

paramétricos se encuentran en los trabajos [3], [5], [11], [12], y [13].

2.1. Métodos de clasificación automática

El objetivo de las técnicas de *clasificación automática* (en inglés "cluster analysis") es realizar una agrupación. Es decir, a partir de una muestra representada por una matriz de datos (*individuos* $\times$ *variables*), asignar los N individuos a G grupos o *clusters*. Estos grupos, desconocidos a priori y que denominaremos $C_g(g = 1, ..., G)$, serán sugeridos por los datos, y se entenderá que hemos obtenido una *buena* clasificación si los grupos creados C_g 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 D variables observadas y, por el contrario, entre individuos pertenecientes a clases distintas pueden apreciarse características diferentes. 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:

CONJUNTO DE N INDIVIDUOS (D variables)

$\Downarrow$

ELECCIÓN de la MEDIDA DE DIFERENCIA, TENDENCIA CENTRAL, VARIABILIDAD, i/o DISTRIBUCIÓN DE PROBABILIDAD

$\Downarrow$

ELECCIÓN del MÉTODO DE CLASIFICACIÓN

$\Downarrow$

G GRUPOS DE INDIVIDUOS: $C_g(g = 1, ..., G)$

Siguiendo este planteamiento, cuando el método de clasificación elegido se base en técnicas no paramétricas será una necesidad fundamental establecer una o varias de las siguientes medidas:

- Una medida de diferencia entre dos datos.
- Una medida de tendencia central de un conjunto de datos.
- 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. La medida de tendencia central será útil para caracterizar los grupos, y la medida de dispersión nos permitirá medir la homogeneidad dentro del grupo y la heterogeneidad entre grupos. 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. El estudio y la adaptación de las técnicas no paramétricas de clasificación para conjuntos de datos composicionales se introdujo en los trabajos [17] y [18], para posteriormente desarrollarse en profundidad en [19].

En el presente trabajo abordamos una aproximación metodológica a las técnicas de clasificación paramétricas. Naturalmente, según el esquema anterior un elemento clave para este tipo de técnicas es la elección de la distribución de probabilidad que se presupone para los datos.

El supuesto principal de los métodos de clasificación paramétricos es que los datos proceden de una distribución multivariante. En consecuencia, es conveniente verificar que el modelo escogido ajusta razonablemente a los datos, pues la calidad de las inferencias que se hagan con los agrupamientos generados por esos modelos dependen de dicha distribución. En los métodos de clasificación paramétrica, como en la mayoría de métodos del análisis multivariante, el modelo más utilizado para datos de tipo continuo es la distribución normal. En la literatura ([14]) se encuentran diversas estrategias y diferentes contrastes que pueden ser de utilidad para analizar la bondad de ajuste. Por ejemplo, la prueba de normalidad basada en los coeficientes de asimetría y kurtosis ([24]) es un buen criterio para determinar normalidad multivariable. Más específicamente para datos composicionales, en [21] se encuentran las diferentes estrategias a seguir para realizar pruebas de bondad de ajuste de una distribución de probabilidad.

Sin embargo, necesariamente hay que ser conscientes que cuando en un conjunto de datos existen agrupaciones, el análisis del ajuste por una distribución se ve entorpecido y enmascarado por la propia existencia de los grupos. En consecuencia, aparece un problema difícil de resolver y, tradicionalmente, la solución pasa por relajar el requisito previo del ajuste. Naturalmente, el análisis del ajuste se podrá realizar en las etapas finales de la clasificación, cuando ya se dispone de una propuesta de agrupación que nos permite efectuar un análisis dentro de cada grupo.

2.2. Método paramétrico aglomerativo

Una vez escogida la distribución de probabilidad debe iniciarse la clasificación propiamente dicha. En esta sección no realizamos una presentación exhaustiva de todos los métodos existentes ni exponemos una relación de las técnicas más adecuadas. Hemos optado por realizar una breve descripción del

método paramétrico aglomerativo introducido en [3]. Esta decisión se ha tomado a la luz de los resultados de una exhaustiva búsqueda en la literatura. Se ha constatado que este método es el más referenciado por las obras más prestigiosas de clasificación y que es un método frecuentemente utilizado en los trabajos de investigación que incorporan una clasificación. En trabajos posteriores abordaremos el análisis y la adaptación de otros métodos paramétricos de aparición más reciente. Entre estos nuevos métodos destacamos la técnica de clasificación basada en el método MCMC vía Muestreo de Gibbs descrito en [24]; el método de las direcciones de proyección introducido en [23]; y el método SAR propuesto en [25].

El método de clasificación paramétrico aglomerativo introducido en [3] parte de la consideración que el conjunto de N individuos a clasificar consiste en G diferentes subpoblaciones cuyas respectivas densidades de probabilidad vienen dadas por la función $f_g(\mathbf{x}, \boldsymbol{\theta}_g)$ para $(g = 1, ..., G)$; donde $\boldsymbol{\theta}_g$ es el vector de parámetros. Se define la función de verosimilitud de la clasificación mediante la expresión

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}/\mathbf{X}) = \prod_{i=1}^{N} f_{\gamma_i}(\mathbf{x}_i, \boldsymbol{\theta}_{\gamma_i}), \tag{4}$$

donde $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_N)$ se utiliza para indexar la subpoblación a la que pertenece cada uno de los individuos. Es decir, se tiene que $\gamma_i = k$ cuando el individuo $\mathbf{x}_i$ pertenece a la subpoblación k-ésima.

Fraley and Raftery ([12]) nos hacen notar que la presencia de este elemento $\boldsymbol{\gamma}$ en la expresión (4) introduce un problema añadido al existir G^N combinaciones posibles que hacen impracticable la maximización exacta de la función de verosimilitud. La estrategia propuesta en [3] consiste en considerar inicialmente que cada individuo forma un grupo. A continuación, el algoritmo prosigue fusionando el par de grupos cuya unión suponga el mayor incremento en la función (4). Como es bien conocido, cuando las densidades de probabilidad $f_g(\mathbf{x}, \boldsymbol{\theta}_g)$ se asumen normales multivariantes de media $\boldsymbol{\mu}_g$ y matriz de covarianza $\boldsymbol{\Sigma}_g$ la expresión (4) se puede reformular en la función de log-verosimlitud

$$l(\boldsymbol{\theta}/\mathbf{X}) = -\frac{DN\log(2\pi)}{2} - \frac{1}{2}\sum_{g=1}^{G}\left[\mathrm{tr}\left(\mathbf{W}_g\boldsymbol{\Sigma}_g^{-1}\right) + N_g \log|\boldsymbol{\Sigma}_g|\right], \tag{5}$$

donde N_g representa el número de individuos del grupo C_g; se ha considerado que el estimador máximo verosímil de $\boldsymbol{\mu}_g$ es la media $\overline{\mathbf{x}}_g$; y la matriz $\mathbf{W}_g = \sum_{\mathbf{x}_i \in C_g}(\mathbf{x}_i - \overline{\mathbf{x}}_g)(\mathbf{x}_i - \overline{\mathbf{x}}_g)^t$ mide la variabilidad dentro del grupo C_g.

Por otra parte, cuando se realiza una agrupación, la forma y la disposición de los grupos existentes en el conjunto de datos afecta fuertemente al poder clasificador de cualquier método de clasificación. Con el objetivo de hacer más tratable la expresión (5) y teniendo en mente los aspectos geométricos de los grupos, se considera una parametrización de las matrices de covarianza $\mathbf{\Sigma}_g$ basada en su descomposición en valores y vectores propios

$$\mathbf{\Sigma}_g = \lambda_g \mathbf{V}_g \mathbf{A}_g \mathbf{V}_g^t, \tag{6}$$

donde $\mathbf{V}_g$ es la matriz ortogonal de vectores propios, $\mathbf{A}_g$ es una matriz diagonal cuyos elementos son proporcionales a los valores propios de $\mathbf{\Sigma}_g$, y $\lambda_g = |\mathbf{\Sigma}_g|^{1/D}$. La estrategia consiste en considerar λ_g, $\mathbf{V}_g$, y $\mathbf{A}_g$ como parámetros independientes y contemplar las diferentes posibilidades que se generan, desde que se considera que todos los parámetros toman valores iguales en todos los grupos hasta el caso de considerar que todos los parámetros toman valores diferentes en cada grupo. Este abanico de posibilidades que se abre tiene interpretaciones geométricas relacionadas con el papel que juega cada parámetro en la expresión (6). Obsérvese que los vectores propios que aparecen en $\mathbf{V}_g$ rigen la orientación del grupo (hiperelipsoide), la distribución de los valores de $\mathbf{A}_g$ nos informa de la forma del grupo, y el parámetro λ_g mide el tamaño o hipervolumen del grupo. Nótese que este valor λ_g, también conocido como varianza efectiva ([24]), no es más que la media geométrica de los valores propios de la matriz $\mathbf{\Sigma}_g$. Obviamente, las posibilidades que aparecen son muy numerosas e intentar abordarlas todas resulta nuevamente impracticable. El Cuadro 1 muestra algunas de las posibilidades más sencillas que se desarrollaron en [3], donde *Igual* o *Variable* indicará que los grupos coinciden o no, repectivamente, en tamaño, forma u orientación.

Cuadro 1: Parametrización en el método paramétrico aglomerativo

$\mathbf{\Sigma}_g$	Tamaño	Forma	Orientación
$\lambda\mathbf{I} = \sigma^2\mathbf{I}$	Igual	Igual, esférica	—
$\lambda_g\mathbf{I} = \sigma_g^2\mathbf{I}$	Variable	Igual, esférica	—
$\lambda\mathbf{V}\mathbf{A}\mathbf{V}^t$	Igual	Igual, elipsoidal	Igual
$\lambda_g\mathbf{V}_g\mathbf{A}_g\mathbf{V}_g^t$	Variable	Variable, elipsoidal	Variable

Es bien conocido ([10], [16], [24]) que la primera parametrización que se muestra en el Cuadro 1 equivale al método no paramétrico de la traza o método de Ward; la tercera parametrización equivale al método del determinante; y la cuarta al método Scott-Symons. Todos estos métodos tienen un fundamento no paramétrico basado en la optimización de un criterio numérico relacionado con la traza o el determinante de las matrices $\mathbf{W}_g$.

Además de las parametrizaciones que aparecen en el Cuadro 1 se han resuelto otras posibilidades más complejas. Sin embargo, resolver estas otras parametrizaciones requiere la incorporación del concepto de la mezcla de distribuciones de probabilidad y requiere la aplicación del algoritmo EM.

2.3. Clasificación mediante mezclas de distribuciones: algoritmo EM

Si los individuos a agrupar provienen de una mezcla de G distribuciones de probabilidad, entonces podemos expresar la densidad como

$$f_{mezcla}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\tau}) = \sum_{g=1}^{G} \tau_g f_g(\mathbf{x}, \boldsymbol{\theta}_g), \tag{7}$$

donde $\boldsymbol{\tau} = (\tau_1, ..., \tau_g);\ \tau_g \geq 0;\ \sum_{g=1}^{G} \tau_g = 1$, es el vector de probabilidades que un individuo pertenezca a la g-ésima componente de la mezcla . Si se tiene un conjunto $\mathbf{X}$ de N individuos de esta mezcla, entonces la función de verosimilitud es

$$L_{mezcla}(\boldsymbol{\theta}, \boldsymbol{\tau}/\mathbf{X}) = \prod_{i=1}^{N} \sum_{g=1}^{G} \tau_g f_g(\mathbf{x}, \boldsymbol{\theta}_g). \tag{8}$$

La resolución de las ecuaciones que nos proporcionan las estimaciones máximo verosímiles de los parámetros que aparecen en la función (8) lleva implícita la necesidad de conocer la probabilidad de que, una vez observado el individuo, el individuo pertenezca a la componente g-ésima de la mezcla. Estas probabilidades, también conocidas como probabilidades *a posteriori*, se podrían estimar si se conociesen los parámetros $\boldsymbol{\theta}, \boldsymbol{\tau}$ del modelo. La estrategia que aparece como más natural consiste en partir de una estimación inicial para, a continuación, realizar una fase iterativa hasta obtener convergencia. Esta estrategia es la que subyace en el funcionamiento del algoritmo EM.

A partir de los N individuos $\mathbf{x}_i$ a clasificar, en el algoritmo EM para mezclas de distribuciones, se consideran N nuevas observaciones multivariantes $(\mathbf{x}_i, \mathbf{z}_i)$, donde $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{iG})$ es un vector binario. Si el individuo $\mathbf{x}_i$ pertenece a la componente g-ésima de la mezcla, entonces el vector $\mathbf{z}_i$ toma el valor cero en todas las componentes excepto en la g-ésima que toma el valor uno $(z_{ig} = 1)$. Naturalmente, en el algoritmo EM las variables $\mathbf{z}$ representan el papel tradicional de variables *no observadas* y se asume que $\{\mathbf{z}_i\}$ son realizaciones *iid* según una distribución multinomial con probabilidades $\boldsymbol{\tau} = (\tau_1, ..., \tau_g)$. De estas consideraciones se deduce fácilmente que la función de densidad conjunta de las variables $(\mathbf{x}, \mathbf{z})$ es

$$f_{EM}(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\tau}) = \prod_{g=1}^{G} [\tau_g f_g(\mathbf{x}, \boldsymbol{\theta}_g)]^{z_g}. \tag{9}$$

Este planteamiento permite formular la función de log-verosimilitud

$$l_{EM}(\boldsymbol{\theta}, \boldsymbol{\tau}/\mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \left[\log(\tau_g) + \log\left(f_g(\mathbf{x}_i, \boldsymbol{\theta}_g)\right)\right], \tag{10}$$

que se utilizará como soporte para la estimación, mediante un proceso iterativo de estimación y maximización, de los valores no observados z_{ig} y de los parámetros de la mezcla. Para aplicar este proceso es necesario partir de una clasificación inicial que nos permita obtener una primera estimación $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\tau}})$ de los parámetros. A partir de esta estimación inicial, la etapa E del algoritmo EM nos proporciona una estimación de los valores no observados

$$\frac{\widehat{\tau}_g f_g(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_g)}{\sum_{k=1}^{G} \widehat{\tau}_k f_k(\mathbf{x}_i, \widehat{\boldsymbol{\theta}}_k)} \quad \longrightarrow \quad \widehat{z}_{ig}, \tag{11}$$

mientras que la etapa M del algoritmo, utilizando las estimaciones (11), nos proporciona el valor de los parámetros que maximiza la función (10). Para el caso de que en la mezcla se consideren distribuciones normales multivariantes se obtiene

$$\frac{N_g}{N} \rightarrow \widehat{\tau}_g; \quad \frac{\sum_{i=1}^{N} \widehat{z}_{ig} \mathbf{x}_i}{N_g} \rightarrow \widehat{\boldsymbol{\mu}}_g; \quad \frac{\sum_{i=1}^{N} \widehat{z}_{ig} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_g)^t}{N_g} \rightarrow \widehat{\boldsymbol{\Sigma}}_g; \tag{12}$$

donde $N_g \equiv \sum_{i=1}^{N} \widehat{z}_{ig}$. El cálculo de las matrices $\widehat{\boldsymbol{\Sigma}}_g$ que se propone en [13] depende de la parametrización de su descomposición en valores y vectores propios. Generalizando los casos que recoge el Cuadro 1, en el Cuadro 2 se muestran todas las posibilidades que se contemplan en la parametrización (6).

En [7] los autores desarrollan un estudio muy detallado sobre los diferentes cálculos a realizar para obtener la estimación (12) de la matriz $\widehat{\boldsymbol{\Sigma}}_g$ según la parametrización que se haya escogido del Cuadro 2 .

Como ya hemos mencionado, para empezar las iteraciones del algoritmo EM se necesita partir de una clasificación inicial. Esta primera clasificación se puede obtener mediante cualquier método jerárquico. El método escogido podrá ser no paramétrico (vecino más próximo, vecino más alejado, media, k-medias, ...) o podrá utilizarse uno de los métodos paramétricos cuya parametrización se recoge en el Cuadro 1. Sea cual sea nuestra elección deberemos decidir también el número G de grupos a construir. Este número de grupos será utilizado tanto en la agrupación inicial como en la del algoritmo EM. Como norma general ([24]), obsérvese que si se contemplan muchos grupos, con la parametrización más sencilla $\lambda\mathbf{I} = \sigma^2\mathbf{I}$ se podrá obtener una agrupación razonable. Por el contrario, si se contempla un número de grupos reducido parece lógico esperar

Cuadro 2: Parametrización para el algoritmo EM

$\mathbf{\Sigma}_g$	Distribución	Tamaño	Forma	Orientación
$\lambda\mathbf{I} = \sigma^2\mathbf{I}$	Esférica	Igual	Igual	—
$\lambda_g\mathbf{I} = \sigma_g^2\mathbf{I}$	Esférica	Variable	Igual	—
$\lambda\mathbf{A}$	Diagonal	Igual	Igual	Ejes coordenados
$\lambda_g\mathbf{A}$	Diagonal	Variable	Igual	Ejes coordenados
$\lambda\mathbf{A}_g$	Diagonal	Igual	Variable	Ejes coordenados
$\lambda_g\mathbf{A}_g$	Diagonal	Variable	Variable	Ejes coordenados
$\lambda\mathbf{V}\mathbf{A}\mathbf{V}^t$	Elipsoidal	Igual	Igual	Igual
$\lambda_g\mathbf{V}\mathbf{A}\mathbf{V}^t$	Elipsoidal	Variable	Igual	Igual
$\lambda\mathbf{V}\mathbf{A}_g\mathbf{V}^t$	Elipsoidal	Igual	Variable	Igual
$\lambda_g\mathbf{V}\mathbf{A}_g\mathbf{V}^t$	Elipsoidal	Variable	Variable	Igual
$\lambda\mathbf{V}_g\mathbf{A}\mathbf{V}_g^t$	Elipsoidal	Igual	Igual	Variable
$\lambda_g\mathbf{V}_g\mathbf{A}\mathbf{V}_g^t$	Elipsoidal	Variable	Igual	Variable
$\lambda\mathbf{V}_g\mathbf{A}_g\mathbf{V}_g^t$	Elipsoidal	Igual	Variable	Variable
$\lambda_g\mathbf{V}_g\mathbf{A}_g\mathbf{V}_g^t$	Elipsoidal	Variable	Variable	Variable

que la mejor agrupación será producida por una parametrización más general $\lambda_g\mathbf{V}_g\mathbf{A}_g\mathbf{V}_g^t$. En cualquier caso, constatamos que es conveniente realizar de manera conjunta la elección de la parametrización y la decisión del número de grupos. Uno de los índices numéricos que da mejores resultados ([6]) como ayuda para estas decisiones es el BIC (*Bayesian Information Criterion*)

$$\text{BIC} = 2\log(L_{mezcla}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\tau}}/\mathbf{X})) - \nu\log(N), \tag{13}$$

donde el valor que toma la función de verosimilitud L_{mezcla}, definida en (8), se calcula utilizando las estimaciones de los parámetros, y el valor de ν es el número de parámetros del modelo. Nuestro modelo, bajo hipótesis de normalidad, está caracterizado por el número G de grupos a construir y por el tipo de parametrización escogida de la matriz $\mathbf{\Sigma}_g$. En [7] se desarrolla un estudio detallado sobre las diferentes parametrizaciones y su número de parámetros correspondiente. En la práctica, una vez decidido un valor máximo para el número de grupos, y una vez escogidas las diferentes parametrizaciones a considerar, nos guiaremos por los valores resultantes del índice BIC en las correspondientes combinaciones para decidir las agrupaciones que pueden recoger razonablemente la estructura de nuestro conjunto de datos.

Sin embargo, todos los esfuerzos empleados en obtener una clasificación razonable de nuestros individuos pueden proporcionar un resultado baladí si la distribución de probabilidad elegida no es adecuada para la tipología de los datos. En la siguiente sección presentamos una breve descripción de los fundamentos de los datos composicionales.

3. Datos composicionales

3.1. Algunos principios fundamentales

En su formulación estadística, los datos composicionales son realizaciones de una composición, vector aleatorio cuyo recorrido está en el símplex (1). Tradicionalmente, a las D variables que forman parte de un vector composicional se las denomina componentes o partes. Los datos composicionales aparecen en áreas muy diversas como son, entre otras: geoquímica, volcanología, granulometría, economía, arqueometría, y medicina. La restricción de la suma constante ha sido considerada la fuente de todos los problemas pues impide la aplicación de los procedimientos estadísticos habituales que se utilizan para datos que no presentan esta restricción. Nótese, por ejemplo, que el cambio en una de las componentes de un dato composicional provoca necesariamente el cambio en como mínimo una de las otras partes.

Una de las dificultades más relevantes es la imposibilidad de interpretar correctamente las covarianzas y los coeficientes de correlación debido a la existencia de una falsa correlación entre las componentes de una composición. La matriz de correlaciones habitual no puede analizarse en el estudio de vectores de suma constante porque presenta necesariamente correlaciones negativas no nulas, determinadas precisamente por la mencionada restricción. Históricamente estas correlaciones sean calificado como "espurias", ya que falsean la imagen de las relaciones de dependencia y pueden conducir a interpretaciones erróneas. En particular, si analizamos la matriz de covarianzas usual entre las componentes de una composición, $\mathbf{K} = \{\text{cov}(x_i, x_j) : i, j = 1, 2, \ldots, D\}$, obtenemos que

$$\text{cov}(x_i, x_1) + \text{cov}(x_i, x_2) + \cdots + \text{cov}(x_i, x_D) = 0 \qquad i = 1, 2, \ldots, D \qquad (14)$$

a causa de la restricción $\sum_{i=1}^{D} x_i = 1$. Sabemos que $\text{cov}(x_i, x_i) = \text{var}(x_i) > 0$, excepto en la situación trivial que la componente x_i sea una constante. Este hecho provoca que necesariamente deba haber una covarianza $\text{cov}(x_i, x_j)$ $(i \neq j)$ de signo negativo. Vemos pues que estas covarianzas no son libres de tomar cualquier valor. Esto invalida la interpretación habitual de las covarianzas, y por ende de las correlaciones, pues a priori suponemos que deberían poder adquirir libremente valores nulos, positivos o negativos. Por el mismo motivo, el hecho que el coeficiente de correlación entre dos componentes cualesquiera de una composición sea igual a 0 no puede interpretarse, como es habitual, como indicio de independencia entre ambas partes.

En general tampoco es correcto aplicar las operaciones clásicas del espacio real vectorial a los datos composicionales. En [17] los autores muestran en detalle que la distancia euclidiana no es una medida de diferencia adecuada entre datos composicionales. Esto tiene consecuencias estadísticas importantes

porque existen multitud de conceptos y técnicas estadísticas que se fundamentan de forma más o menos explícita en la distancia euclidiana.

Otra de las dificultades importantes es la falta de familias paramétricas suficientemente flexibles para modelar los conjuntos de datos composicionales. Las distribuciones de Dirichlet y sus generalizaciones se obtienen mediante la clausura de vectores aleatorios con componentes independientes. Como consecuencia, sus partes son prácticamente independientes, puesto que su correlación está únicamente motivada por el hecho de haber dividido todas sus componentes por la suma de éstas. Esto impide su uso en la modelización de fenómenos con relaciones de dependencia no inducidas por la suma constante. A pesar de esta característica en [8] los autores defienden una clasificación paramétrica de datos composicionales usando la distribución de Dirichlet.

3.2. Metodología para el análisis de datos composicionales

La mayor aportación de la monografía de Aitchison ([1]) consistió en establecer que un estudio apropiado de la variación relativa en un conjunto de datos composicionales debe basarse en logcocientes. Según se reconoce en [14] esta aportación, conocida como "logratio analysis", ha sido el mayor avance reciente en el análisis de los datos composicionales. Aitchison argumenta que nuestra atención debe centrarse en la magnitud relativa de las componentes, es decir, en los cocientes x_i/x_j $(i, j = 1, 2, \ldots, D; i \neq j)$. Por lo tanto, diremos que un problema es composicional cuando reconozcamos que el valor en términos absolutos de las componentes es irrelevante. Trabajando con los cocientes desaparecen los problemas de las correlaciones espurias. La metodología de Aitchison se basa en la trasformación de los datos composicionales al espacio real multivariante. Si tomamos los logaritmos de los cocientes, el espacio final es todo el espacio real y por lo tanto podemos aplicar cualquier técnica estadística clásica. Existen diversas posibilidades para transformar los datos, todas ellas están basadas en los logaritmos de cocientes entre las componentes de un dato composicional.

La trasformación logcociente aditiva (alr)

$$\mathrm{alr}(\mathbf{x}) = \big(\ln(x_1/x_D), \ln(x_2/x_D), \ldots, \ln(x_{D-1}/x_D)\big)' \tag{15}$$

es una transformación biyectiva, pero no es simétrica en las partes de $\mathbf{x}$ ya que la componente del denominador adquiere un protagonismo especial respecto al resto.

La transformación logcociente centrada (clr)

$$\mathrm{clr}(\mathbf{x}) = \big(\ln(x_1/g(\mathbf{x})), \ln(x_2/g(\mathbf{x})), \ldots, \ln(x_D/g(\mathbf{x}))\big)', \tag{16}$$

donde $g(\mathbf{x})$ es la media geométrica de las D partes de $\mathbf{x}$. Esta trasformación es biyectiva y simétrica entre las partes. Su imagen es el hiperplano de $\Re^D$ que

pasa por el origen y es ortogonal al vector de unidades, es decir, la suma de las componentes del vector trasformado es igual a cero. Nos encontramos pues ante una nueva dificultad ya que la matriz de covarianzas del vector clr-trasformado será singular.

En [9] los autores introducen la transformación logcociente isométrica (ilr). Esta transformación tiene su fundamento en el hecho que las operaciones (2) y (3) dotan al símplex de estructura de espacio vectorial euclidiano con dimensión $D-1$. En consecuencia, si denotamos como $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ a una base ortonormal de $\mathcal{S}^D$, toda composición $\mathbf{x} \in \mathcal{S}^D$ está determinada de forma única por su vector de coordenadas

$$\text{ilr}(\mathbf{x}) = (< \mathbf{x}, \mathbf{e}_1 >_a, < \mathbf{x}, \mathbf{e}_2 >_a, \ldots, < \mathbf{x}, \mathbf{e}_{D-1} >_a)\,. \tag{17}$$

Esta transformación permite identificar cada elemento de $\mathcal{S}^D$ con su vector de coordenadas.

La existencia de más de una transformación nos lleva a la situación de deber elegir entre una de ellas como paso previo a la aplicación de un método estadístico multivariante. Ciertamente, las tres transformaciones estan relacionadas mediante expresiones matriciales que permiten obtener cada una de ellas a partir de cualquiera de las otras. En este trabajo no se reproducen estas relaciones matriciales por motivos de brevedad, para más detalles consúltese [1] y [9]. Naturalmente, será también nuestra misión investigar si los resultados de la clasificación paramétrica se ven o no afectados por la transformación elegida.

Tradicionalmente, en las aplicaciones que exigen simetría en el tratamiento de sus componentes, como por ejemplo una clasificación no paramétrica, se utiliza la trasformación clr. Para la modelización de conjuntos de datos composicionales con distribuciones multivariantes, se ha venido utilizando mayoritariamente la trasformación alr. De esta forma se evita trabajar con distribuciones degeneradas. Si se desea utilizar la transformación clr en trabajos que incluyan el modelo normal Barceló-Vidal et al. ([4]) demuestran que para salvar la dificultad de matrices de covarianzas degeneradas es suficiente con prescindir de una de las variables del conjunto de datos clr-transformados. Sin embargo, con cualquiera de las dos trasformaciones, se deberá analizar si los resultados del método aplicado son invariante por permutaciones de las componentes. Esta metodología ha permitido ampliar las familias de distribuciones sobre el símplex. Destacamos el modelo normal logístico aditivo ([1]) o el modelo normal asimétrico logístico aditivo ([20]). En la actualidad se están desarrollando ([22]) la definición de modelos paramétricos basados en la transformación ilr. Usando esta transformación únicamente queda la dificultad de constatar que los resultados no dependen de la base ortonormal escogida. Paralelamente se está trabajando en modelos definidos sin necesidad de recurrir a las transforma-

ciones. En el trabajo [22] los autores introducen el modelo normal en el símplex a partir de la función de densidad de su vector de coordenadas.

3.3. Distribuciones de probabilidad

Por su propia naturaleza las componentes de una composición toman sus valores en el intervalo $[0, 1]$. Esta naturaleza hace evidente que las distribuciones multivariantes tradicionales más usuales, como la distribución normal, pueden producir resultados erróneos si son aplicadas directamente a los datos. La estrategia propuesta en la metodología para el análisis de datos composicionales mediante transformaciones se basa en aplicar los métodos clásicos de estadística multivariante en el espacio logcociente transformado. Siguiendo esta estrategia, la definición de las distribuciones de probabilidad más usuales sobre el símplex aparecen de manera natural. Por este motivo, en este trabajo únicamente se reproduce la definición de la *distribución normal en* $\mathcal{S}^D$ mediante la transformación ilr. Si se desea profundizar en los aspectos relacionados con distribuciones de probabilidad para datos composicionales se puede consultar [1] y [21].

La composición aleatoria $\mathbf{x}$ tiene una distribución *normal en* $\mathcal{S}^D$si la función de densidad de su vector de coordenadas $\text{ilr}(\mathbf{x})$ es

$$f^*_{\mathbf{x}}(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2}}{|\,\boldsymbol{\Sigma}\,|^{1/2}} \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\text{ilr}(\mathbf{x}) - \boldsymbol{\mu}\right)\right],\ \mathbf{x} \in \mathcal{S}^D. \quad (18)$$

Obsérvese que la expresión (18) se corresponde con la densidad normal clásica en $\Re^{D-1}$. Por esta razón utilizamos la terminología *distribución normal en* $\mathcal{S}^D$. De la expresión (18) se deduce que los estimadores máximo verosímiles serán los usuales para la distribución normal multivariante siempre que se utilicen para su cálculo los datos ilr-transformados. Por otra parte, nótese que las diferentes parametrizaciones que aparecen en los Cuadros 1 y 2 serán consideradas sobre la matriz de covarianzas $\boldsymbol{\Sigma}$. Las características y propiedades de esta matriz serán determinantes en el momento de realizar la clasificación paramétrica basada en una distribución normal de los datos.

4. Aproximación metodológica

La metodología adecuada para la realización de una clasificación paramétrica no debe distar en exceso de la metodología usada en clasificaciones donde se aplican métodos no paramétricos ([17], [18], [19]). En una clasificación paramétrica la fase de elección de la distribución de probabilidad desempeña un papel crucial. Sin embargo, debemos ser plenamente conscientes de que no existe una distribución adecuada para todos los conjuntos de datos, y que, en general, puede suceder que para un conjunto de datos sea posible encontrar más de una distribución cuyo ajuste sea razonable. En el caso de conjuntos de datos composicionales deberemos escoger una distribución de probabilidad

entre aquellas distribuciones que son compatibles con la tipología especial de este tipo de datos.

Las fases a seguir en la realización de una clasificación paramétrica de datos composicionales se pueden resumir de manera esquemática en un diagrama como el que muestra la Figura 1. Si el método de clasificación escogido no se basa en el algoritmo EM para mezcla de distribuciones normales, entonces el esquema seguiría siendo válido adaptando la etapa intermedia de la clasificación automática. 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 se concentran básicamente en la etapa de elección de la distribución de probabilidad. Sin embargo, no hay que olvidar que las herramientas que se utilicen en las etapas descriptivas también deberán ser adecuadas a la tipología de los datos. La naturaleza inductiva-deductiva del proceso de clasificación es común a la gran mayoría de técnicas estadísticas y está en el fundamento del propio método estadístico. Nótese que 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 aquélla 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 distribución de probabilidad, la elección del método de clasificación o, en su caso, la elección del número de grupos a considerar.

5. Conclusiones y propuesta de trabajos futuros

La primera aproximación que se ha realizado en este trabajo nos ha permitido constatar que los métodos paramétricos de clasificación automática pueden constituir una familia de métodos muy útiles en problemas en los que el objetivo sea clasificar un conjunto de datos composicionales. De esta aproximación han surgido las pautas generales a seguir en el momento de realizar una agrupación y se ha hecho evidente que el aspecto crucial, para adaptar estos métodos a la tipología de los datos composicionales, radica en la elección de la distribución de probabilidad.

Por otro lado, pero íntimamente relacionado con los aspectos metodológicos, en la realización del presente estudio han surgido diversas e importantes cuestiones que deberán resolverse en trabajos de investigación futuros:

- abordar la adaptación de otros métodos de clasificación paramétricos de reciente aparición (entre los que destacamos la técnica de clasificación

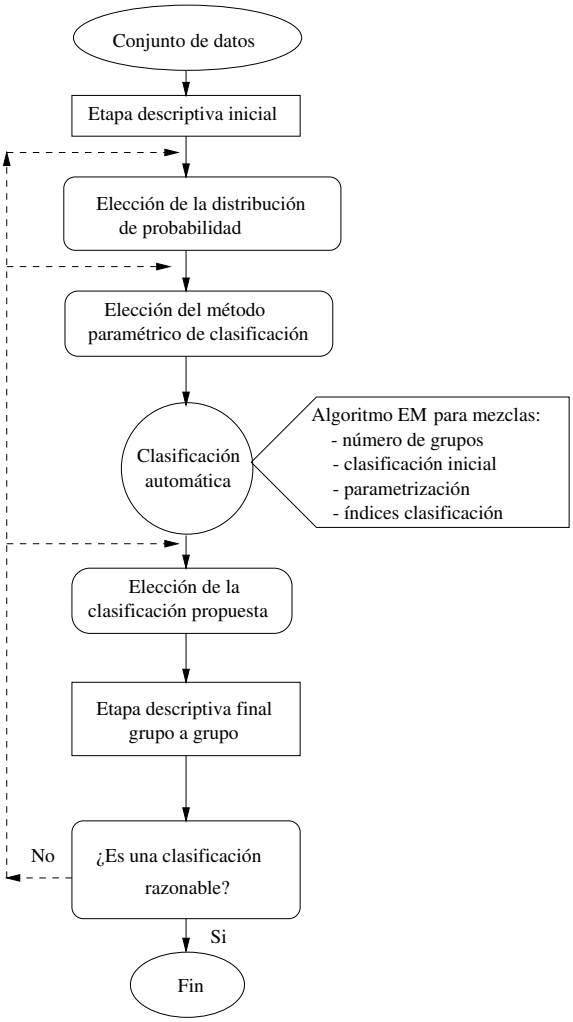

Figura 1: Esquema de las fases a seguir en la realización de una clasificación paramétrica de datos composicionales.

basada en el método MCMC vía Muestreo de Gibbs, el método de las direcciones de proyección y el método SAR, mencionados anteriormente).

- analizar si los resultados de una clasificación paramétrica es independiente de la transformación logcociente utilizada. En el caso que la transformación de datos elegida afecte los resultados de la clasificación, se deberá analizar si las permutaciones de componentes afecta cuando se escogen

las transformaciones alr y clr; y si la base ortonormal escogida en la transformación ilr incide en la agrupación.

- realizar el estudio de casos prácticos, basados en conjuntos de datos reales y en de datos simulados, con el objetivo de analizar el comportamiento de las diferentes distribuciones de probabilidad que tradicionalmente se utilizan en la modelización de datos composicionales.

Finalmente, queremos poner énfasis en la tenencia actual de modelización de datos composicionales basada en la propuesta de distribuciones de probabilidad que se definan directamente en el símplex sin la necesidad de realizar transformaciones. Deberemos recoger los resultados de estas aportaciones en nuestras propuestas metodológicas futuras.

6. Agradecimientos

Esta investigación ha sido parcialmente financiada por la Dirección General de Investigación (DGI) del Ministerio de Ciencia y Tecnología de España a través del proyecto BFM2003-05640/MATE; y parcialmente financiada por la Direcció General de Recerca de la Generalitat de Catalunya a través del proyecto 2003XT00079.

7. Bibliografía

[1] Aitchison, J. (1986). *The statistical analysis of compositional data.* Chapman and Hall, London, 416 p. Reprinted in 2003 by The Blackburn Press, Caldwell, NJ.

[2] Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V. (2000). *Logratio analysis and compositional distance. Mathematical Geology.* 32(3), 271–275.

[3] Banfield, J. D. and Raftery, A. E. (1993). *Model-Based Gaussian and Non-Gaussian Clustering.* Biometrics, vol. 49, 803–821.

[4] Barceló-Vidal, C. and Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (1999). *Comment on "Singularity and nonnormality in the classification of compositional data".* Mathematical Geology, vol. 31(5), 581–585.

[5] Bensmail, H. and Meulman, J. J., (2003). *Model-Based Clustering with Noise: Bayesian Inference and Estimation.* Journal of Classification, vol. 20, 49–76.

[6] Biernacki, C. and Govaert, G., (1999). *Choosing Models in Model-Based Clustering and Discriminant Analysis.* Journal of Statistical Computation and Simulation, vol. 64, 49–71.

[7] Celeux, G. and Govaert, G., (1995). *Gaussian Parsimonious Clustering Models.* Pattern Recognition, vol. 28, 781–793.

[8] DeSarbo, W.S. and Ramaswamy, V. and Lenk, P. (1993). *A Latent Class Procedure for the Structural Analysis of Two-Way Compositional Data. Journal of Classification,* vol. 10(3), 159–193.

[9] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barcelò-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, vol 35, no 3, pp. 279-300.

[10] Everitt, Brian S. (2001). *Cluster Analysis.* Edward Arnold, Cambridge (UK), 237 p., fourth edition.

[11] Fraley, C. and Raftery, A. E., (1999). *MCLUST: Software for Model-Based Cluster Analysis.* Journal of Classification, vol. 16, 297–306.

[12] Fraley, C. and Raftery, A. E., (2002) *Model-Based Clustering, Discriminant Analysis, and Density Estimation.* Journal of the American Statistical Association, vol. 97(458), 611–631.

[13] Fraley, C. and Raftery, A. E., (2003) *Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST.* Journal of Classification, vol. 20, 263–286.

[14] Krzanowski, W. J. and Marriott, F. H. C. (1994). *Multivariate Analysis, Part 1 - Distributions, ordination and inference.* Edward Arnold, London (UK), vol. 1, Kendall's Library of Statistics, 280 p.

[15] Krzanowski, W. J. and Marriott, F. H. C. (1995). *Multivariate Analysis, Part 2 - Classification, covariance structures and repeated measurements.* Edward Arnold, London (UK), vol. 2, Kendall's Library of Statistics, 280 p.

[16] Gordon, A. D. (1999). *Classification.* Chapman and Hall/CRC, Monographs on statisitcs and applied probability, 82, 256 p., 2d edition.

[17] Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998a). *Measures of Difference for Compositional Data and Hierarchical Clustering Methods.* In *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology*. Ed. A. Buccianti, G. Nard, and R. Potenza. Nápoles (I), Part 2, 526–531.

[18] Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998b). *A critical approach to non-parametric classification of compositional data.* In *Proc. of the 6th Conference of the International Federation of Classification Societies*. Ed. A. Rizzi, M. Vichi, and H.H. Bock. Springer-Verlag, Berlín (D), 49–56.

[19] Martín-Fernández, J. A. (2001) *Medidas de diferencia y técnicas de clasificación no paramétrica para datos composicionales.* Tesis doctoral (ISSBN: 84-699-5369-9) publicada en formato electrónico en www.tdcat.cesca.es/TDCat-0516101-135345/, 233 p.

[20] Mateu-Figueras, G., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998). *Modeling compositional data with multivariate skew-normal distributions.* In *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology*. Ed. A. Buccianti, G. Nard, and R. Potenza. Nápoles (I), Part 2, 532-537.

[21] Mateu Figueras, Glòria (2003). *Models de distribució sobre el símplex* Tesis doctoral (ISSBN: 84-688-6734-9) publicada en formato electrónico en *www.tdx.cesca.es/TDX-0427104-170301*, 202 pág.

[22] Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2004). *La distribución normal en $\mathcal{S}^D$ vs la distribución normal logística.* In *Proceedings of the 28th Congreso Nacional de Estadística e Investigación Operativa*, Cádiz (E), (*en esta publicación*).

[23] Peña, D. and Prieto, F. J. , (2001) *Cluster Identification using Projections.* Journal of the American Statistical Association, vol. 96(456), 1433–1445.

[24] Peña, D. (2002). *Análisis de datos multivariantes.* Mc-Graw Hill (E), 539 p.

[25] Peña, D. and Tiao, G. C. (2002). *Cluster Analysis by the SAR procedure.* Documento de trabajo, Universidad Carlos III, Madrid.