Contribuciones a la Estadística
y a la Investigación Operativa
Sicilia et. al. (editores)

ISBN 84-689-8552-X
páginas 1253–1269

Técnicas paramétricas de clasificación automática para datos composicionales: resultados preliminares

J. A. Martín-Fernández[1], *J. Daunis-i-Estadella*[2], *G. Mateu-Figueras*[3]

[1]josepantoni.martin@udg.es, Dept. d'Informàtica i Matemàtica Aplicada, Universitat de Girona,
[2]josep.daunis@udg.es, Dept. IMA, UdG,
[3]gloria.mateu@udg.es,Dept. IMA, UdG.

Abstract

Los últimos avances teóricos nos aseguran que el símplex tiene estructura de espacio Euclidiano y que la transformación log-cociente isométrica nos permite trabajar en coordenadas ortonormales. En este trabajo mostramos los resultados preliminares cuando se aplican estas nuevas propiedades en una clasificación paramétrica de datos composicionales.

Palabras Clave: log-ratio, datos de proporciones, cluster.

AMS: 62H30, 62P99.

1. Introducción

Un dato composicional expresa las proporciones de las partes respecto a un todo. Este tipo de datos está presente en multitud de problemas prácticos de diferentes áreas como son, entre otros: las composiciones químicas de cerámicas (arqueometría); las composiciones de contaminantes (ciencias medioambientales); la composición de óxidos mayores en rocas y la composición de sedimentos (geología); las composiciones de elementos en sangre, orina, y cálculos renales (medicina); y la distribución del presupuesto familiar (economía). El espacio muestral de los datos composicionales ([1]) es el símplex $\mathcal{S}^D$ definido por $\mathcal{S}^D = \{\mathbf{x} = [x_1, ..., x_D] : x_i > 0, x_1 + ... + x_D = c\}$, donde c puede valer 1, 100, 10^6 o cualquier otro valor constante que refleja las unidades de medida. Se ha constatado ([11]) que el símplex $\mathcal{S}^D$ tiene estructura de espacio Euclídeo. La operación interna es la *perturbación* definida como $\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}[x_1 x_1^*, \ldots, x_D x_D^*]$.

La operación externa es la *potenciación* definida por $\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$. El *producto escalar* se define como $\langle \mathbf{x}, \mathbf{x}^* \rangle_a = \frac{1}{D} \sum_{i<j} \ln(x_i/x_j) \ln(x_i^*/x_j^*)$. Aquí $\mathcal{C}$ es el operador clausura definido por $\mathcal{C}(\mathbf{w}) = [w_1/\sum w_j, \ldots, w_D/\sum w_j]$, para $\mathbf{w} \in \mathcal{R}_+^D$.

Desde [1] la comunidad científica acepta mayoritariamente que los datos composicionales reflejan magnitudes relativas. En consecuencia, las distribuciones de probabilidad definidas sobre el símplex deben ser compatibles con la estructura de $\mathcal{S}^D$ y deben modelar estos datos teniendo en cuenta que se está interesado en los cambios relativos y no en los cambios absolutos. La metodología log-cociente nos permite definir este tipo de distribuciones mediante funciones de densidad expresadas en términos de los logaritmos de los cocientes de las componentes ([1]; [9]). Las distribuciones de probabilidad definidas en el símplex que han proporcionado los resultados más razonables se definen a través de las transformaciones *additive log-ratio* (alr) y *centred log-ratio* (clr), introducidas en [1] e *isometric log-ratio* (ilr) introducida en [6]. Estas transformaciones se definen por

$$\mathrm{alr}(\mathbf{x}) = \left[\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}\right], \qquad \mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right], \quad (549)$$

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{y}, \quad \text{con} \quad y_i = \frac{1}{\sqrt{i(i+1)}} \ln \left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i} \right). \quad (550)$$

donde $g(\mathbf{x}) = (x_1 x_2 \ldots x_D)^{1/D}$ es la media geométrica de la composición $\mathbf{x}$. Todas estas transformaciones son biyectivas entre el símplex y el espacio real y permiten la aplicación de los métodos estadísticos multivariantes usuales. Sin embargo, se constatan diferencias importantes entre las tres transformaciones. Es muy sencillo constatar que la transformación clr es una transformación isométrica pero tiene la dificultad de proporcionar conjuntos de datos transformados cuya matriz de covarianzas es singular. Por otro lado, se debe proceder con mucha cautela cuando se aplica la transformación alr porque es una transformación no isométrica y que otorga un papel relevante a la componente que aparece en el denominador de su formulación (549). La transformación ilr evita estas dificultades. Esta transformación es isométrica y la matriz de covarianzas de los datos ilr-transformados no es singular. En esencia la transformación ilr es equivalente a la transformación clr expresada en términos de una base ortonormal. En [6] se construye una base ortonormal respecto de la cual la transformación ilr se expresa como (550). El hecho de no que no exista una única base ortonormal implica que cuando se utilice la transformación ilr se deberá analizar si los resultados del estudio estadístico son invariantes por un cambio de base ortonormal.

En [8] se realiza una primera propuesta de la metodología para las técnicas paramétricas de clasificación de conjuntos de datos composicionales. La me-

todología presentada, si bien de carácter general, se basa en las técnicas de clasificación que combinan el algoritmo EM y la mezcla de distribuciones. En la siguiente sección del presente trabajo presentamos brevemente los fundamentos del método de clasificación paramétrico basado en mezcla de distribuciones. Pasaremos, a continuación, a mostrar los resultados preliminares que se obtienen cuando se aplican estas técnicas paramétricas de clasificación usuales en combinación con las transformaciones alr, clr e ilr. Aquí ponemos especial énfasis en describir el comportamiento de las tres transformaciones en relación a las diferentes parametrizaciones posibles de la matriz de covarianzas en el método EM. Para finalizar, y a modo de conclusión, indicaremos los desarrollos futuros a realizar.

2. Clasificación paramétrica mediante mezcla de distribuciones

2.1. Métodos de clasificación automática

El objetivo de las técnicas de *clasificación automática* 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. Estos grupos, desconocidos a priori, serán sugeridos por los datos, y se entenderá que hemos obtenido una clasificación *razonable* si los grupos creados son homogéneos en su interior y heterogéneos entre si. Es decir, 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.

Cuando el método de clasificación considerado no presupone la existencia de un modelo de distribución de probabilidad para las observaciones objeto de la agrupación, diremos que es un método no paramétrico. El estudio y la adaptación de las técnicas no paramétricas de clasificación para conjuntos de datos composicionales se encuentran desarrollados en [7]. En los métodos de clasificación paramétrica el modelo más utilizado para datos de tipo continuo es la distribución normal. Naturalmente, 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. Sin embargo, 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, tradicionalmente se relaja 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.

No pretendemos realizar una presentación exhaustiva de los métodos de clasi-

ficación paramétrica existentes. Hemos optado por analizar el comportamiento de la clasificación paramétrica mediante mezclas de distribuciones vía el algoritmo EM. Esta decisión se ha tomado después de una exhaustiva búsqueda en la literatura donde se ha constatado que es el método más utilizado y referenciado en los trabajos de investigación que incorporan una clasificación. En trabajos posteriores se abordará 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 [13]; el método de las direcciones de proyección introducido en [12]; y el método SAR propuesto en [14].

2.2. Clasificación mediante el algoritmo EM

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

$$f_{mezcla}(\mathbf{x}, \theta, \tau) = \sum_{g=1}^{G} \tau_g f_g(\mathbf{x}, \theta_g), \tag{551}$$

donde $\tau = (\tau_1, ..., \tau_g)$; $\tau_g \geq 0$; $\sum_{g=1}^{G} \tau_g = 1$, 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

$$L_{mezcla}(\theta, \tau/\mathbf{X}) = \prod_{i=1}^{N} \sum_{g=1}^{G} \tau_g f_g(\mathbf{x}_i, \theta_g). \tag{552}$$

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. 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 $\tau = (\tau_1, ..., \tau_g)$. De estas consideraciones se deduce fácilmente que la función de densidad conjunta de las variables $(\mathbf{x}, \mathbf{z})$ es

$$f_{EM}(\mathbf{x}, \mathbf{z}, \theta, \tau) = \prod_{g=1}^{G} \left[\tau_g f_g(\mathbf{x}, \theta_g)\right]^{z_g}. \tag{553}$$

Este planteamiento permite formular la función de log-verosimilitud

$$l_{EM}(\theta, \tau/\mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \left[\log(\tau_g) + \log\left(f_g(\mathbf{x}_i, \theta_g)\right)\right], \tag{554}$$

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 $(\widehat{\theta}, \widehat{\tau})$ de los parámetros. A partir de esta estimación inicial, la etapa E del algoritmo EM nos proporciona una estimación de los valores no observados $\widehat{z}_{ig}$. La etapa M del algoritmo, utilizando las estimaciones $\widehat{z}_{ig}$, nos proporciona el valor de los parámetros que maximiza la función (554). Para el caso de distribuciones normales multivariantes los parámetros del modelo son el vector de valores esperados y la matriz de covarianzas: $\theta_g = (\widehat{\mu}_g, \widehat{\mathbf{\Sigma}}_g)$.

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 (554) y teniendo en mente los aspectos geométricos de los grupos, en [2] se considera una parametrización de las matrices de covarianza $\mathbf{\Sigma}_g$ basada en su descomposición en valores y vectores propios

$$\mathbf{\Sigma}_g = \lambda_g \mathbf{V}_g \mathbf{A}_g \mathbf{V}_g^t, \tag{555}$$

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, 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 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. Estas posibilidades tienen interpretaciones geométricas relacionadas con el papel que juega cada parámetro en la expresión (555). 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 hipervolumen del grupo. Nótese que este valor λ_g, también conocido como varianza efectiva ([13]), no es más que la media geométrica de los valores propios de la matriz $\mathbf{\Sigma}_g$. En el Cuadro 1 se muestran todas las posibilidades que se contemplan en la parametrización (555), donde *Igual* o *Variable* indicará que los grupos coinciden o no, respectivamente, en tamaño, forma u orientación.

En [5] se presenta un estudio muy detallado sobre los diferentes cálculos a realizar en (554) para obtener la estimación de la matriz $\widehat{\mathbf{\Sigma}}_g$ según la parametrización que se haya escogido del Cuadro 1.

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 aglomerativos introducidos

Cuadro 1: Parametrización de la matriz $\mathbf{\Sigma}_g$ para el algoritmo EM

$\mathbf{\Sigma}_g$	Distribución	Tamaño	Forma	Orientación	Modelo
$\lambda\mathbf{I} = \sigma^2\mathbf{I}$	Esférica	Igual	Igual	—	A
$\lambda_g\mathbf{I} = \sigma_g^2\mathbf{I}$	Esférica	Variable	Igual	—	B
$\lambda\mathbf{A}$	Diagonal	Igual	Igual	Ejes coord.	C
$\lambda_g\mathbf{A}$	Diagonal	Variable	Igual	Ejes coord.	D
$\lambda\mathbf{A}_g$	Diagonal	Igual	Variable	Ejes coord.	E
$\lambda_g\mathbf{A}_g$	Diagonal	Variable	Variable	Ejes coord.	F
$\lambda\mathbf{V}\mathbf{A}\mathbf{V}^t$	Elipsoidal	Igual	Igual	Igual	G
$\lambda_g\mathbf{V}\mathbf{A}\mathbf{V}^t$	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	H
$\lambda_g\mathbf{V}_g\mathbf{A}\mathbf{V}_g^t$	Elipsoidal	Variable	Igual	Variable	I
$\lambda\mathbf{V}_g\mathbf{A}_g\mathbf{V}_g^t$	Elipsoidal	Igual	Variable	Variable	—
$\lambda_g\mathbf{V}_g\mathbf{A}_g\mathbf{V}_g^t$	Elipsoidal	Variable	Variable	Variable	J

en [2]. Sea cual sea nuestra elección deberemos decidir también el número G de grupos a construir. Como norma general ([13]), obsérvese que si se contemplan muchos grupos, con la parametrización más sencilla $\lambda\mathbf{I} = \sigma^2\mathbf{I}$ o modelo A se podrá obtener una agrupación razonable. Por el contrario, si se contempla un número de grupos reducido parece lógico esperar que la mejor agrupación será producida por el modelo J o 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 ([4]) como ayuda para estas decisiones es el BIC (*Bayesian Information Criterion*)

$$\text{BIC} = 2\log(L_{mezcla}(\widehat{\theta}, \widehat{\tau}/\mathbf{X})) - \nu\log(N), \tag{556}$$

donde el valor que toma la función de verosimilitud L_{mezcla} (552) se calcula utilizando las estimaciones de los parámetros, y el valor de ν es el número de parámetros del modelo. Bajo hipótesis de normalidad, el número de parámetros está caracterizado ([5]) por el número G de grupos a construir y por el tipo de parametrización escogida de la matriz $\mathbf{\Sigma}_g$. 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.

3. Clasificación paramétrica de datos composicionales

La existencia de tres transformaciones log-cociente alr, clr, e ilr, nos lleva a la situación de deber elegir entre una de ellas como paso previo a la aplicación de un método estadístico multivariante. Esta elección se debe realizar teniendo en cuenta si el método que se quiere aplicar es invariante por permutaciones de las componentes para el caso de la transformación alr, o si el método se ve afectado por la singularidad de la matriz de covarianzas para el caso de la transformación clr, o si el método es invariante por cambio de base en el caso de la transformación ilr.

Para la modelización de conjuntos de datos composicionales con distribuciones multivariantes, se ha venido utilizando mayoritariamente la transformación alr. Como máximo exponente podemos resaltar la definición de la distribución aln o normal logístico aditiva ([1], p. 113) que ha sido habitualmente utilizada para modelar la normalidad en conjuntos de datos alr-transformados. En este trabajo únicamente analizaremos los resultados obtenidos cuando se aplica las transformaciones alr e ilr. En [3] los autores demuestran que es posible utilizar la transformación clr en trabajos que incluyan el modelo normal salvando la dificultad de matrices de covarianzas degeneradas. Para ello es suficiente con prescindir de una de las variables del conjunto de datos clr-transformados. Los autores demuestran que esta estrategia produce exactamente los mismos resultados que utilizando la transformación alr.

La metodología log-cociente ha permitido ampliar las familias de distribuciones sobre el símplex. En la actualidad se están desarrollando ([10]) la definición de modelos paramétricos basados en la transformación ilr. Únicamente queda la dificultad de constatar que los resultados no dependen de la base ortonormal escogida.

3.1. 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 log-cociente 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 el vector $\text{ilr}(\mathbf{x})$. Si se desea profundizar en los aspectos relacionados con distribuciones de probabilidad para datos composicionales se puede consultar [1] y

[9].

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

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2}}{|\,\boldsymbol{\Sigma}\,|^{1/2}} \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \mu\right)' \boldsymbol{\Sigma}^{-1} \left(\text{ilr}(\mathbf{x}) - \mu\right)\right],\ \mathbf{x} \in \mathcal{S}^D. \tag{557}$$

Obsérvese que la expresión (557) se corresponde con la densidad normal clásica en $\mathcal{R}^{D-1}$. Por esta razón utilizamos la terminología *distribución normal en* $\mathcal{S}^D$. De la expresión (557) 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 el Cuadro 1 serán consideradas sobre la matriz de covarianzas $\boldsymbol{\Sigma}$.

3.2. Resultados para un caso real

Los datos del conjunto $\mathbf{X}$ ([1], p. 363) corresponden a las proporciones de tres metabolitos en las secreciones urinarias diarias de 67 personas. Las proporciones de estos metabolitos son de utilidad en el diagnóstico de un síndrome raro (*Cushing's syndrome*) y se ha constatado que el patrón de las proporciones difiere de adultos a niños. En $\mathbf{X}$ se encuentran 30 muestras de niños y 37 de adultos.

En la Figura 1A se muestra el conjunto de datos original y en la Figura 1B el conjunto de datos centrado. El centro del conjunto de datos que se obtiene mediante esta operación de centrado composicional ([7]) es el baricentro del símplex. Esta operación de centrado se basa en la operación perturbación y es coherente con la estructura de espacio vectorial. En la Figura 1B se aprecia claramente el patrón diferenciado en las proporciones de metabolitos entre el grupo de niños y el de adultos. La Figura 1C muestra la representación dos a dos de las componentes del conjunto de datos $\mathbf{X}$. Nuevamente se aprecia claramente la diferencia de patrones entre el grupo niños y el grupo adultos. Si se calculan los centros composicionales de cada grupo (media geométrica en porcentajes) y se representan en un diagrama de barras (Fig. 1D) se observa que los adultos tienen, en término medio, proporciones inferiores en los dos primeros metabolitos y superiores en el tercer metabolito. Por otra parte, en lo que se refiere la hipótesis de normalidad de los datos, cuando se aplican los contrastes de normalidad multivariante coherentes con la metodología log-cociente ([9]) a los dos grupos por separado se obtienen p-valores siempre superiores a 0.1.

Con el propósito de comparar el comportamiento de las diferentes metodologías y estrategias, vamos a utilizar la misma preclasificación para el algoritmo EM en todos los casos. Procediendo de esta manera, podremos apreciar estrictamente las diferencias debidas a la clasificación paramétrica realizada por el algoritmo

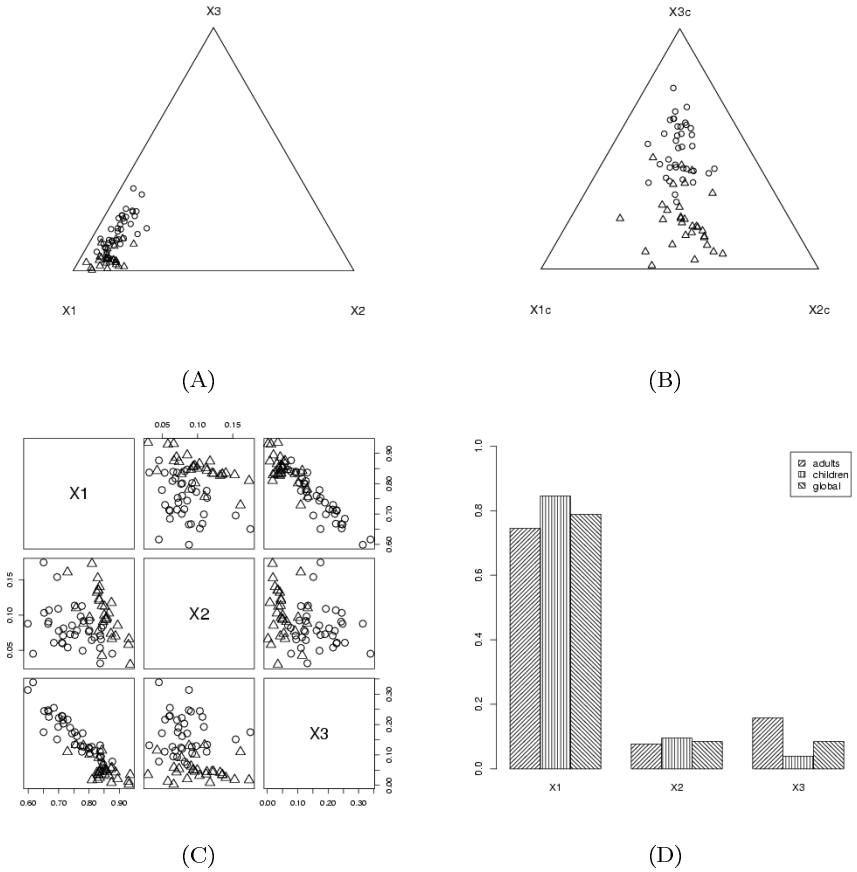

Figura 1: Diagramas ternarios (adultos=círculos; niños=triángulos): (A) Conjunto **X**; (B) Conjunto **X** centrado; (C) Variables *dos a dos*; (D) Centros composicionales de cada grupo y global.

EM y eliminamos el efecto que pueda producir la preclasificación.

Aún a riesgo de confundir al lector, presentamos en la Figura 2 los resultados de la clasificación obtenida cuando se aplica el método EM a los datos de **X** sin transformar, es decir sin utilizar la metodología log-cociente. Observamos en la Figura 2 que los valores BIC para esta clasificación muestran que no ha sido posible realizar la clasificación paramétrica para los modelos G, H, I, J (Cuadro 1) que trabajan con la matriz de covarianza muestral. Corresponden a formar grupos elípticos con orientación no paralela a los ejes coordenados. Este hecho está causado por la singularidad de la matriz de covarianzas de los

datos en porcentajes. Las otras parametrizaciones sugieren que la clasificación óptima es la que considera nueve grupos formados con la parametrización A que construye grupos esféricos de igual tamaño. Sin embargo, debido a que el conjunto de datos sólo está formado por 67 muestras y a la vista de los gráficos de los datos (Fig. 1) parece excesivo considerar nueve grupos. Si consideramos un número más reducido de grupos, la Figura 2 nos sugiere que la clasificación más razonable consta de cuatro grupos mediante la parametrización B. Se concluye que esta clasificación no ha detectado la estructura en dos grupos (adultos y niños) existente en el conjunto $\mathbf{X}$. Con este ejemplo se constata que la metodología usual aplicada a los datos sin transformar puede proporcionar clasificaciones no razonables.

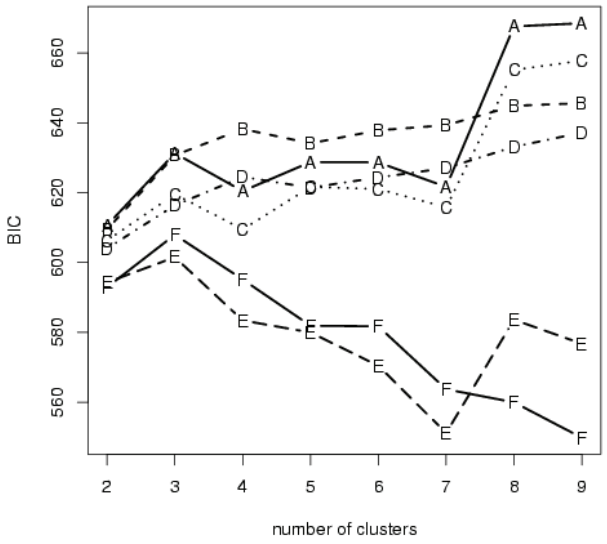

Figura 2: Diagrama BIC para la clasificación sin metodología log-cociente (Cuadro 1).

El conjunto de datos de metabolitos pertenece a $\mathcal{S}^3$, en consecuencia, en el momento de aplicar la metodología log-cociente mediante la transformación alr podemos escoger entre tres componentes para que jueguen el papel del denominador en (549). En la Figura 3A, 3B, y 3C se muestran los diagramas BIC cuando, respectivamente, se escoge como denominador la variable $\mathbf{X}_1$, $\mathbf{X}_2$, y $\mathbf{X}_3$. Un estudio en detalle de los valores numéricos de BIC pone de relieve diferencias entre las tres gráficas en todas las posibles parametrizaciones. Por ejemplo, si se consideran las clasificaciones formadas por dos o más grupos, según si se utiliza cómo denominador una componente u otra, el valor óptimo

de BIC sugiere una clasificación con dos grupos (denominador $\mathbf{X}_2$ o $\mathbf{X}_3$) o con tres grupos (denominador $\mathbf{X}_1$). Únicamente coinciden las tres gráficas de BIC para el caso de considerar sólo un grupo y las parametrizaciones elípticas. Este hecho se debe a qué en estas parametrizaciones se trabaja con la matriz de covarianza completa y única. Sólo en este caso el *efecto denominador* no altera la clasificación. En consecuencia, se constata que la clasificación resultante depende del denominador escogido en (549). Este hecho nos hace desaconsejar imperativamente el uso de la aplicación alr en las clasificaciones paramétricas. Obsérvese que partiendo de una misma preclasificación, dos investigadores que usen la transformación alr (549) con diferente denominador pueden obtener clasificaciones diferentes aún usando la misma parametrización de la matriz de covarianzas. En la Figura 3D se muestra el diagrama ternario del conjunto de datos centrado cuando se considera la clasificación que ha proporcionado la parametrización B (Cuadro 1) con denominador $\mathbf{X}_1$ en (549). Esta clasificación propone agrupar los datos en tres grupos. Únicamente una muestra de adulto (círculos) se clasifica en el grupo 3 (cuadrados); seis muestras de niños se clasifican cómo adultos (círculos); y ocho muestras de niños se clasifican en el grupo 3 (cuadrados).

Cuando se pretende utilizar la transformación ilr se debe constatar que los resultados son independientes de la base ortonormal utilizada para formar la expresión (550). Por este motivo realizaremos dos clasificaciones diferentes: una, a los datos ilr-transformados; y otra, a los datos que se obtienen aplicando la transformación ilr seguida de una rotación de sesenta grados. Llamamos $\mathbf{X}^*$ al conjunto correspondiente a los datos rotados. En la Figuras 4A y 4B se muestran los diagramas BIC para los conjuntos $\mathbf{X}$ y $\mathbf{X}^*$ cuando aplicamos la transformación ilr. Un estudio detallado de los valores numéricos BIC muestra que se aprecian diferencias únicamente para aquellas parametrizaciones (Cuadro 1) que consideran elipses cuyos ejes son paralelos a los ejes coordenados: modelos C, D, E, y F. En las parametrizaciones que corresponden a grupos esféricos (A, B) y en las parametrizaciones que buscan grupos elípticos con ejes inclinados (G, H, I, y J) no existen diferencias. En las Figuras 4C y 4D se muestran los diagramas ternarios de los conjuntos $\mathbf{X}$ y $\mathbf{X}^*$ centrados. Los dos grupos definidos en cada caso corresponden a la parametrización coincidente G. Obviamente las agrupaciones resultantes coinciden plenamente, y se observa que se clasifican mal 15 muestras de adultos porque se asignan al grupo de muestras de niños.

Todos los resultados que se muestran en esta sección se han obtenido del estudio del conjunto de datos real de proporciones de metabolitos. Sin embargo, es importante mencionar que estos mismos cálculos se han repetido con otros conjuntos de datos reales y con multitud de conjuntos de datos simulados. Los resultados en todos los casos nos han sugerido las mismas conclusiones.

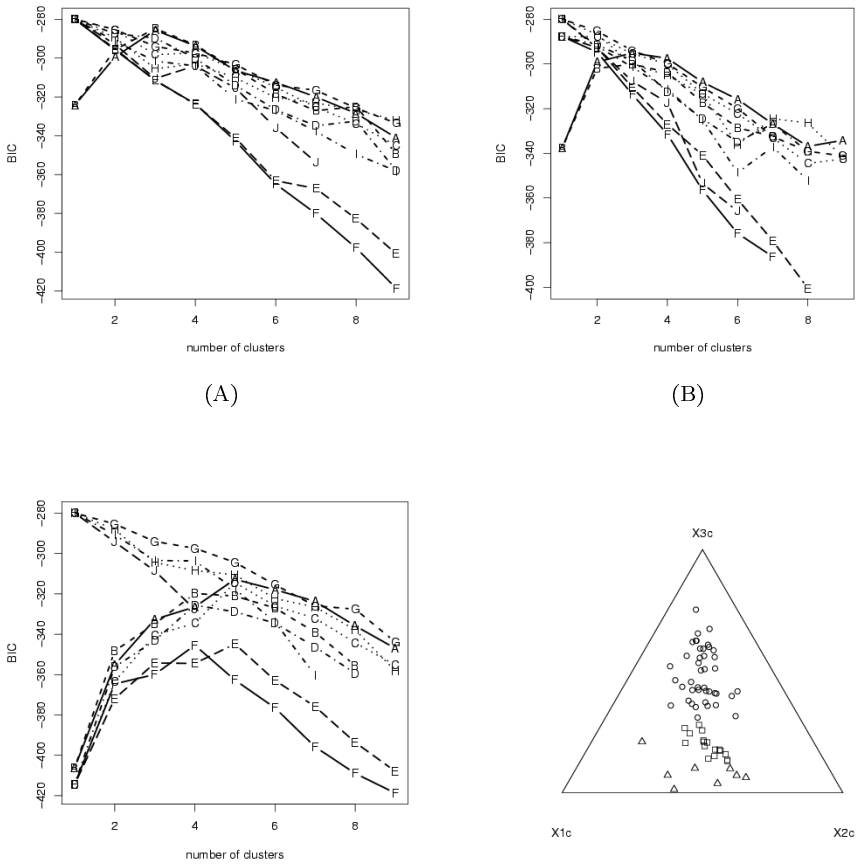

Figura 3: Diagramas BIC y diagrama ternario centrado de clasificación (Cuadro 1) vía transformación alr (A) Denominador $\mathbf{X_1}$; (B) Denominador $\mathbf{X_2}$; (C) Denominador $\mathbf{X_3}$; (D) Clasificación B vía alr de denominador $\mathbf{X_1}$ (grupo 1=círculos ; grupo 2= triángulos; grupo 3= cuadrados)

4. Conclusiones y desarrollos futuros

En este trabajo se ha constatado que la metodología log-cociente para datos composicionales ofrece resultados razonables cuando se aplica en una clasificación paramétrica mediante el algoritmo EM. Estos resultados son manifiestamente mejores que los resultados que se obtienen cuando se aplica la metodología usual a los datos sin transformar. Sin embargo, no se recomienda la metodología log-cociente vía la transformación alr debido a la influencia en los resultados de la elección del denominador de la transformación. Tampoco se

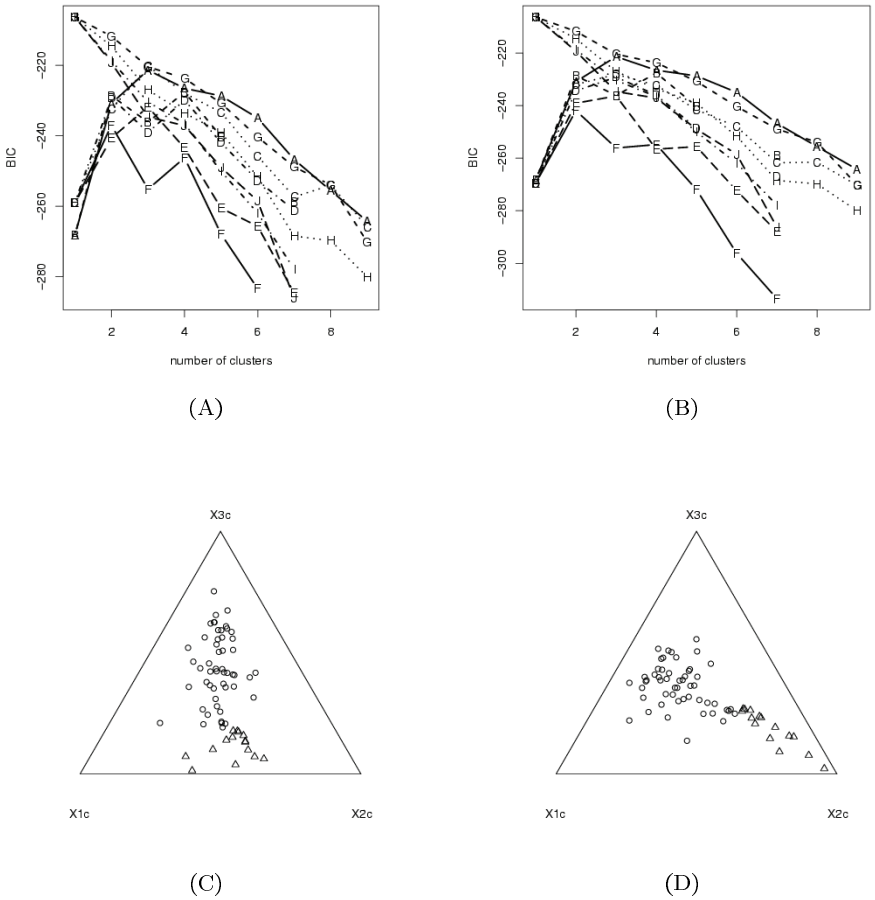

Figura 4: Diagramas BIC (Cuadro 1) y diagramas ternarios (grupo 1=círculos ; grupo 2=triángulos) vía transformación ilr:(A) BIC datos $\mathbf{X}$; (B) BIC datos $\mathbf{X}^*$; (C) Clasificación G datos $\mathbf{X}$ centrados; (D) Clasificación G datos $\mathbf{X}^*$ centrados

recomienda la utilización de la transformación clr por ser los resultados totalmente equivalentes a los obtenidos mediante la transformación alr. Se concluye que la mejor estrategia consiste en la utilización de la transformación ilr en combinación con aquellas parametrizaciones de la matriz de covarianzas que respetan la invarianza por cambio de base ortonormal.

La justificación de estas conclusiones se desarrollará en un futuro trabajo mediante el análisis en profundidad del algoritmo EM para cada parametrización diferente ([5]). De esta manera se justificará la influencia del denominador en (549) o de la base en (550) en cada una de las diferentes parametrizaciones de

la matriz de covarianza.

5. Agradecimientos

Este trabajo ha recibido financiación mediante el proyecto BFM2003-05640/MATE (Dir. General de Investigación; Minis. de Ciencia y Tecnología).

6. Bibliografía

[1] Aitchison J, 1986, The Statistical Analysis of Compositional Data. London: Chapman and Hall. Reprinted in 2003 by Blackburn Press.

[2] Banfield, J. D. and Raftery, A. E., 1993, Model-Based Gaussian and Non-Gaussian Clustering: Biometrics, v. 49, p. 803–821.

[3] Barceló-Vidal, C. and Martín-Fernández, J. A. and Pawlowsky-Glahn, V., 1999, Comment on "Singularity and nonnormality in the classification of compositional data": Math. Geol., v. 31, no. 5, p. 581–585.

[4] Biernacki, C. and Govaert, G., 1999, Choosing Models in Model-Based Clustering and Discriminant Analysis: Journal of Statistical Computation and Simulation, v. 64, p. 49–71.

[5] Celeux, G. and Govaert, G., 1995, Gaussian Parsimonious Clustering Models: Pattern Recognition, v. 28, p. 781–793.

[6] Egozcue, J.J., Pawlowsky-Glahn V., Mateu-Figueras G., Barceló-Vidal C., 2003, Isometric logratio transformations for compositional data analysis: Math. Geol., v. 35, no.3, p. 279-300.

[7] Martín-Fernández, J.A., 2001, Measures of difference and non-parametric cluster analysis for compositional data, Ph.D. thesis, Universitat Politècnica Catalunya, ISBN: 84-699-5369-9, available at http://www.tdcat.cesca.es/TDCat-0516101-135345/, 233p.

[8] Martín-Fernández, J. A., Daunis-i-Estadella, J., and Mateu-Figueras, G., 2004, Clasificación paramétrica de datos composicionales: aproximación metodológica. In: Dept. EIO, UCA (Eds.), XXVIII Congreso Nacional de Estad. e Invest. Oper., Cádiz (E), CD-ROM, ISBN: 84-689-0438-4, 18 p. (electronic publication).

[9] Mateu-Figueras G., 2003, Distribution Models on the Simplex;, Ph.D. thesis, Universitat Politècnica Catalunya, ISBN: 84-688-6734-9, available at http://www.tdcat.cesca.es/TDX-0427104-170301/index.html, 202p.

[10] Mateu-Figueras, G. and Pawlowsky-Glahn, V., 2004, La distribución normal en $\mathcal{S}^D$ vs la distribución normal logística, *in*: Dept. EIO, UCA (eds.), Proceedings of XXVIII Congreso Nacional de Estadística e Investigación Operativa, Cádiz (E), CD-ROM, ISBN: 84-689-0438-4, 21p.

[11] Pawlowsky-Glahn, V. and Egozcue, J.J., 2001, Geometrical Approach to Statistical Analysis on the Simplex: Stochastic Environmental Research and Risk Assessment, v. 15, no. 5, p. 384-398.

[12] Peña, D. and Prieto, F. J. , 2001, Cluster Identification using Projections: Journal of the American Statistical Association, v. 96, no. 456, p. 1433–1445.

[13] Peña, D., 2002, Análisis de datos multivariantes. Madrid: Mc-Graw Hill, 539p.

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