XXIV Congreso Nacional de Estadística e Investigación Operativa
Almería 20–23 de octubre de 1998

Medida de diferencia Kullback-Leibler entre datos composicionales.

J. A. Martín-Fernández[1], C. Barceló-Vidal[1], V. Pawlowsky-Glahn[2].

[1] Dept. d'Informàtica i Matemàtica Aplicada. Universitat de Girona

[2] Dept. de Matemàtica Aplicada III. Universitat Politècnica de Catalunya

RESUMEN

Cualquier medida de diferencia entre individuos u observaciones debe tener en cuenta la propia naturaleza de los datos. En este trabajo se exponen los requerimientos que debe cumplir una medida de diferencia entre datos composicionales y se propone una medida basada en la disimilitud de Kullback-Leibler.

Palabras y frases clave: Datos composicionales, distancia, disimilitud, divergencia.
Clasificación AMS: 62H25, 62H30.

1 Introducción

Cualquier vector $\mathbf{x} = (x_1, x_2, ..., x_D)$ cuyos componentes no negativos representen proporciones de un total, está sujeto a la restricción $x_1+x_2+...+x_D = 1$. En Aitchison (1986) se encuentra un estudio detallado de este tipo de vectores de proporciones o composiciones, conocidos también como datos composicionales. En el trabajo de Aitchison queda plenamente justificado que el hecho de ignorar la restricción antes mencionada en el momento de aplicar las técnicas estadísticas de análisis multivariante más usuales da origen a un análisis de los datos erróneo o irrelevante . En este trabajo exponemos los requerimientos específicos que debe cumplir una medida de diferencia entre datos composicionales (Aitchison, 1992) y presentamos una medida basada en la disimilitud de Kulback-Leibler utilizada en Teoria de la Información (Cover, 1991).

2 Medida de diferencia entre datos composicionales

Analogamente al papel jugado por el grupo de las traslaciones cuando el espacio muestral es $\mathbf{R}^D$, Aitchison (1986) propone el grupo de las perturbaciones para caracterizar la '*diferencia*' entre dos composiciones en el espacio muestral de las composiciones o símplex $\mathbf{S}^D = \{(x_1, x_2, \ldots, x_D) : x_j > 0 \ (j = 1, 2, \ldots, D),\ x_1 + x_2 + \ldots + x_D = 1\}$. Si convenimos en usar el símbolo '$\circ$' para indicar la operación '*perturbación*', entonces la perturbación $\mathbf{p} \in \mathbf{S}^D$ aplicada a la composición $\mathbf{x}$ es la composición

$\mathbf{p} \circ \mathbf{x} = (p_1 x_1, p_2 x_2, ..., p_D x_D) / \sum_{j=1}^{D} p_j x_j$. Los requerimientos que debe verificar toda medida de diferencia entre dos composiciones son (Aitchison, 1992): invarianza por cambio de escala, invarianza por permutaciones, invarianza por perturbaciones y dominancia respecto a subcomposiciones. Una distancia factible entre dos composiciones $\mathbf{x}, \mathbf{y} \in \mathbf{S}^D$ es $\Delta(\mathbf{x}, \mathbf{y}) = d_{eu}(clr(\mathbf{x}), clr(\mathbf{y}))$, donde d_{eu} representa la distancia euclídea y $clr(\mathbf{x}) = (log\,(x_1/g(\mathbf{x}))\,, log\,(x_2/g(\mathbf{x}))\,, ..., log\,(x_D/g(\mathbf{x})))$, con $g(\mathbf{x})$ la media geométrica de la composición $\mathbf{x}$. Esta distancia (Martín et al., 1998) es equivalente a la distancia propuesta por Aitchison (1992).

3 Una medida de Kullback-Leibler en $\mathbf{S}^D$

Proposición. Sea $d_K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{D} x_j log\left(\frac{x_j}{y_j}\right)$ la medida de diferencia Kullback-Leibler entre dos composiciones, y $\mathbf{e} = (1/D, 1/D, ..., 1/D)$ el centro del símplex. La medida de diferencia entre dos composiciones definida por

$$d(\mathbf{x}, \mathbf{y}) = \frac{D}{2}\left(d_K(\mathbf{e}, \mathbf{x} \circ \mathbf{y}^{-1}) + d_K(\mathbf{e}, \mathbf{y} \circ \mathbf{x}^{-1})\right), \tag{1}$$

cumple las siguientes propiedades:

P1. $d(\mathbf{x}, \mathbf{y}) = \frac{D}{2} log\left(\frac{A_D(\mathbf{x}/\mathbf{y})}{H_D(\mathbf{x}/\mathbf{y})}\right)$, donde $A_D(\mathbf{x}/\mathbf{y})$ y $H_D(\mathbf{x}/\mathbf{y})$ simbolizan la media aritmética y la media armónica, respectivamente, del vector de ratios $\mathbf{x}/\mathbf{y}$.

P2. $d(\mathbf{x}, \mathbf{y}) \geq 0, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{S}^D$, y , $d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$.

P3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{S}^D$.

P4. $d(\mathbf{p} \circ \mathbf{x}, \mathbf{p} \circ \mathbf{y}) = d(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{p} \in \mathbf{S}^D$.

P5. $d(\mathbf{x}_s, \mathbf{y}_s) \leq d(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{S}^D$, para toda subcomposición s que se considere.

Esta medida significa una alternativa a las distancias mencionadas en el apartado 2 y abren una nueva vía de investigación para las técnicas de clasificación sobre el símplex basadas en los conceptos de distancia o disimilitud.

Referencias

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall. New York (USA), 416 pp.

Aitchison, J. (1992). 'On Criteria for Measures of Compositional Difference'. *Math. Geology*, vol. 24, No. 4, pp. 365-379.

Cover, T.M. (1991). *Elements of information theory*. John Wiley & Sons. New York (USA), 542 pp.

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998). 'A Critical Approach to Non-parametric Classification of Compositional Data', en: Proceedings of IFCS'98. The Sixth Conference of the International Federation of Classification Societies. (in press).