*27 Congreso Nacional de Estadística e Investigación Operativa*
*Lleida, 8–11 de abril de 2003*

# FUNDAMENTACIÓN MATEMÁTICA DEL ANÁLISIS ESTADÍSTICO DE DATOS COMPOSICIONALES

Carles Barceló-Vidal[1], Josep A. Martín-Fernández[1], Vera Pawlowsky-Glahn[1],

[1]Departament d'Informàtica i Matemàtica Aplicada
Universitat de Girona, Campus de Montilivi, E-17071 Girona, España
E-mail: carles.barcelo@udg.es, josepantoni.martin@udg.es, vera.pawlowsky@udg.es

**RESUMEN**

Por vector composicional se entiende un vector cuyas componentes son todas estrictamente positivas y de suma constante. Revisamos esta definición partiendo de una relación de equivalencia en el octante positivo del espacio real que permite identificar las composiciones con clases de equivalencia. La transformación logarítmica entre el conjunto de las composiciones y un espacio vectorial real euclidiano permite demostrar que el símplex es una representación posible del conjunto de clases y que tiene estructura de espacio vectorial real euclidiano. Además, la metodología de Aitchison es compatible con esta estructura y no depende de la representación.

**Palabras y frases clave**: clases de equivalencia, espacio soporte, estructura algebraica y geométrica del espacio soporte, símplex.
**Clasificación AMS**: 46N30.

# 1 Introducción

Históricamente, los *datos composicionales* han sido identificados con *datos clausurados*, y el símplex ha sido considerado el espacio muestral natural o soporte de este tipo de datos. A nuestro entender, ha sido precisamente el énfasis sobre la restricción debida a la suma constante el que ha contribuido a enmascarar su naturaleza real. Pensamos además que es la razón subyacente a la mayoría de las controversias suscitadas por la metodología introducida por Aitchison (1986). Véase a título de ejemplo las polémicas recogidas en los siguientes artículos: Aitchison (1989a, 1989b, 1990a, 1991, 1992), Aitchison et al. (2000, 2001), Barceló-Vidal et al. (1999), Baxter (1993), Bohling et al. (1996), Rehder and Zier (2001), Tangri and Wright (1993), Tauber (1999), Watson (1990, 1991), Watson and Philip (1989), Whitten (1995), Woronow (1997a, 1997b), Zier and Rehder (1998).

Sin embargo, más crucial que la propiedad de la restricción es la *propiedad de la invariancia por cambios de escala*. En efecto, cuando consideramos únicamente algunas partes de una composición, nuestras observaciones siguen siendo composicionales, a pesar de no satisfacer la restricción de suma constante. Este hecho fue reconocido por Aitchison (1992) al argumentar que todo análisis estadístico consistente de datos composicionales debiera basarse en logcocientes, y está en la base de los desarrollos ulteriores de esta metodología, recogidos esencialmente en las siguientes publicaciones: Aitchison (1990b, 1997, 1999, 2002), Aitchison and Bacon-Shone (1999), Aitchison and Greenacre (2002), Aitchison and Thomas (1998), Barceló-Vidal (1996), Barceló and Pawlowsky (1994), Barceló et al. (1995, 1996), Martín-Fernández (2001), Martín-Fernández et al. (1997, 1998a, 1998b, 1998c, 1999, 2000), Mateu-Figueras et al. (1998), Pawlowsky-Glahn and Barceló-Vidal (1999), Pawlowsky-Glahn and Egozcue (2001, 2002).
Para hacer justicia a este hecho, pensamos que es necesario dar una definición más amplia del concepto de composición, y así lo hacemos en la sección 2 del presente trabajo, donde se introduce la *relación de equivalencia composicional* en el ortante positivo del espacio real $D$-dimensional $\mathbb{R}^D$. Obtenemos así que el espacio de todas las composiciones —el *espacio composicional* $\mathcal{C}$— es un espacio cociente, y que el *símplex* $\mathcal{S}^D$ es una forma entre muchas de representar $\mathcal{C}$. A resultas de ello es todavía más evidente que todo análisis de datos composicionales puede y debe ser independiente de su representación, y por ende de su suma constante.
Partiendo de esta definición más amplia y recurriendo a las transformaciones exponencial y logarítmica, desarrollamos en las secciones 3 a 6 los sucesivos pasos para definir una estructura euclidiana en el espacio composicional $\mathcal{C}$. Primero, definimos una estructura euclidiana sobre un espacio vectorial cociente real $\mathcal{L}$ de dimensión $D-1$, y luego transferimos dicha estructura a $\mathcal{C}$ mediante la transformación exponencial. Todos estos resultados han sido ampliamente desarrollados en Barceló-Vidal (2000).

# 2 El espacio de las composiciones

## 2.1 Primeras definiciones y propiedades

Denominaremos *vector $D$-observacional* a todo vector $D \times 1$ real $\mathbf{w} = (w_1, \ldots, w_D)'$ cuyas componentes o partes son estrictamente positivas. El conjunto de todos estos vectores es el ortante positivo $\mathbb{R}^D_+$, subconjunto del espacio real $\mathbb{R}^D$. Diremos que dos vectores $D$-observacionales $\mathbf{w}$ y $\mathbf{w}^*$ son *composicionalmente equivalentes*, y escribiremos $\mathbf{w} \sim \mathbf{w}^*$, si existe una constante positiva $k$ tal que $\mathbf{w} = k\mathbf{w}^*$. Esta relación de equivalencia sobre $\mathbb{R}^D_+$ divide el espacio en clases de equivalencia denominadas *composiciones con $D$-partes* o, brevemente, *composiciones*. Representaremos la clase de equivalencia generada por un vector $\mathbf{w} \in \mathbb{R}^D_+$ por

$$\underline{\mathbf{w}} = \{k\mathbf{w} : k \in \mathbb{R}^+\}.$$

El conjunto de todas las composiciones, es decir, el espacio cociente $\mathbb{R}_+^D/\sim$, se denomina *espacio composicional*, y se representa por $\mathcal{C}$. La aplicación cociente de $\mathbb{R}_+^D$ en $\mathcal{C}$ que asigna a cada vector $\mathbf{w}$ su clase $\underline{\mathbf{w}}$, se representa por ccl (de *compositional class*):

$$\operatorname{ccl} \mathbf{w} = \underline{\mathbf{w}} \qquad (\mathbf{w} \in \mathbb{R}_+^D).$$

Una composición con $D$-partes puede interpretarse geométricamente como una semirrecta que parte del origen de coordenadas en el ortante positivo de $\mathbb{R}^D$ (ver Fig. 1a).

Nótese que la relación de equivalencia puede reformularse a partir de cocientes de las componentes de vectores observacionales, pues es inmediato ver que dos vectores $\mathbf{w} = (w_1, \ldots, w_D)'$ y $\mathbf{w}^* = (w_1^*, \ldots, w_D^*)'$ son composicionalmente equivalentes sí, y sólo sí,

$$\frac{w_i}{w_j} = \frac{w_i^*}{w_j^*} \quad \forall\ i, j = 1, \ldots, D.$$

## 2.2 Criterios de selección

Toda composición $\underline{\mathbf{w}}$ queda completamente determinada por un vector observacional arbitrario de la clase de equivalencia. En consecuencia, pueden utilizarse diversos criterios para elegir un vector observacional específico como representante de una composición, dando lugar a resultados interesantes.

### Criterio lineal

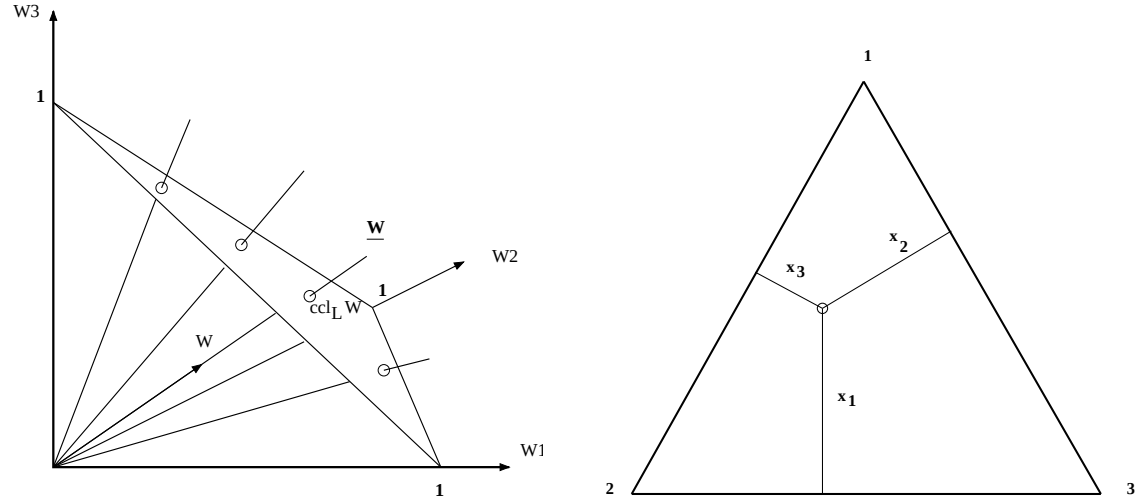
Representaremos por $\operatorname{ccl}_L$ el operador que transforma cada vector $\mathbf{w} \in \mathbb{R}_+^D$ en el vector de suma unidad $\mathbf{w}/\sum_{i=1}^D w_i$. Este operador corresponde al *operador clausura* introducido en Aitchison (1986, p. 31). Es inmediato que $\mathbf{w} \sim \operatorname{ccl}_L \mathbf{w}$, y que el operador $\operatorname{ccl}_L$ es constante sobre los vectores de una misma clase de equivalencia composicional, de modo que si $\mathbf{w} \sim \mathbf{w}^*$, entonces $\operatorname{ccl}_L \mathbf{w} = \operatorname{ccl}_L \mathbf{w}^*$.

**Definición 1.** *La operación que selecciona de cada composición* $\underline{\mathbf{w}}$ *como representante el vector observacional de suma unidad* $\operatorname{ccl}_L \mathbf{w}$ *se denomina* criterio lineal.

Geométricamente, $\operatorname{ccl}_L \mathbf{w}$ es la intersección de la semirrecta $\underline{\mathbf{w}}$, que parte del origen, con el hiperplano de $\mathbb{R}^D$ definido por la ecuación $\sum_{i=1}^D w_i = 1$ (ver Fig. 1a). El conjunto de todos estos puntos es el *símplex*:

$$\mathcal{S}^D = \{(w_1, \ldots, w_D)' : w_i > 0, \forall i = 1, \ldots, D; \sum_{i=1}^D w_i = 1\}.$$

El símplex $\mathcal{S}^3$ se conoce también como *diagrama ternario*, un triángulo equilátero de altura unidad. Para cada punto $P$ del triángulo 123 (ver Fig. 1b) las perpendiculares $w_1$, $w_2$ y $w_3$ de $P$ a los lados opuestos a 23, 13 y 12, respectivamente, satisfacen $w_1 + w_2 + w_3 = 1$. Análogamente, el símplex $\mathcal{S}^4$ corresponde a un tetraedro regular 1234 de altura unidad.




(a) (b)

Figura 1: (a) Las composiciones con 3-partes se interpretan como semirrectas que parten del origen de $\mathbb{R}^3_+$. Criterio de selección lineal. (b) El símplex $\mathcal{S}^3$.

### Criterio esférico

Representaremos por $\mathrm{ccl}_E$ el operador que transforma cada vector $\mathbf{w} \in \mathbb{R}^D_+$ en el vector de norma unidad $\mathbf{w}/\|\mathbf{w}\|$. Es inmediato ver que $\mathbf{w} \sim \mathrm{ccl}_E\, \mathbf{w}$, y que el operador $\mathrm{ccl}_E$ es constante sobre los vectores de una misma clase de equivalencia composicional, de modo que si $\mathbf{w} \sim \mathbf{w}^*$, entonces $\mathrm{ccl}_E\, \mathbf{w} = \mathrm{ccl}_E\, \mathbf{w}^*$.

**Definición 2.** *La operación que selecciona de cada composición* $\underline{\mathbf{w}}$ *como representante el vector observacional de norma unidad* $\mathrm{ccl}_E\, \mathbf{w}$ *se denomina* criterio esférico.

Geométricamente, $\mathrm{ccl}_E\, \mathbf{w}$ es la intersección de la semirrecta $\underline{\mathbf{w}}$, que parte del origen, con la esfera unidad de $\mathbb{R}^D$ centrada en el origen (ver Fig. 2a).

### Criterio hiperbólico

Representaremos por $\mathrm{ccl}_H$ el operador que transforma cada vector $\mathbf{w} \in \mathbb{R}^D_+$ en el vector de producto unidad $\mathbf{w}/g(\mathbf{w})$, donde $g(\mathbf{w}) = (\prod_{i=1}^D w_i)^{1/D}$ es la media geométrica de las componentes del vector $\mathbf{w}$. Es inmediato que $\mathbf{w} \sim \mathrm{ccl}_H\, \mathbf{w}$, y que el operador $\mathrm{ccl}_H$ es constante sobre los vectores de una misma clase de equivalencia composicional, de modo que si $\mathbf{w} \sim \mathbf{w}^*$, entonces $\mathrm{ccl}_H\, \mathbf{w} = \mathrm{ccl}_H\, \mathbf{w}^*$.

**Definición 3.** *La operación que selecciona de cada composición* $\underline{\mathbf{w}}$ *como representante el vector observacional de producto unidad* $\mathrm{ccl}_H\, \mathbf{w}$ *se denomina* criterio hiperbólico.

Geométricamente, $\mathrm{ccl}_H\, \mathbf{w}$ es la intersección de la semirrecta $\mathbf{w}$, que parte del origen, con la superficie hiperbólica $Hip_D$ de $\mathbb{R}^D_+$ definida por la ecuación $\prod_{i=1}^D w_i = 1$ (ver

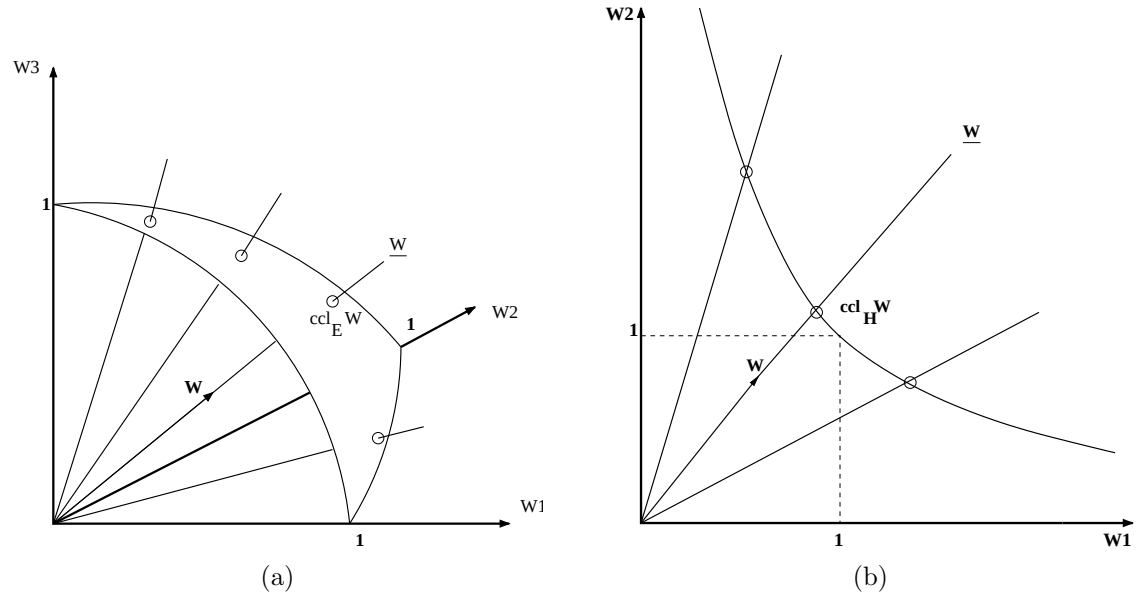



Figura 2: (a) Criterio de selección esférico (caso D=3); (b) Criterio de selección hiperbólico (caso D=2).

Fig. 2b). La Figura 3 muestra los tres criterios de selección para $D = 2$ en una única gráfica.

### Otras representaciones

En las definiciones precedentes, hemos hecho referencia a vectores de suma unidad, de norma unidad, y de producto unidad. Es evidente que la unidad puede sustituirse por cualquier constante arbitraria, tal y como suele hacerse con frecuencia en la práctica, por ejemplo al trabajar con tantos por cien, partes por millón, u otras unidades similares, en lugar de con tantos por uno. Pero no es ésta la única forma de obtener otros representantes de un dato composicional. De hecho, basta hallar una representación única, que en su versión más sencilla será una superficie que tenga intersección única con cada clase de equivalencia, y ello se da por ejemplo con planos inclinados. Resulta entonces que la característica de algo constante (suma, producto, norma) no es esencial a la representación. Luego, cualquier conjunto de datos cuyos vectores observacionales están en el ortante positivo de $\mathbb{R}_+^D$ es candidato a ser considerado un conjunto de datos composicionales. Surge entonces la pregunta referente a cómo determinar cuando son composicionales y cuando no, pregunta a la que intentamos responder en el siguiente apartado.

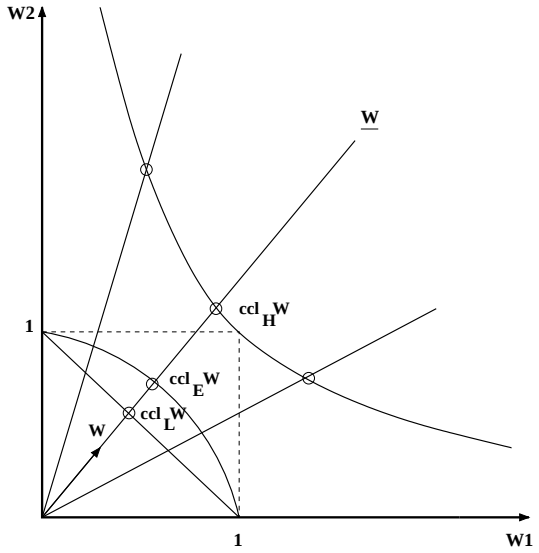



Figura 3: Tres criterios de selección de representantes de una composición (caso $D = 2$).

## 2.3 Naturaleza composicional de un conjunto de datos

Cuando las componentes de los vectores observacionales de un conjunto de datos representan cocientes de un total dado, es decir, cuando representan magnitudes *relativas*, los datos serán con toda seguridad *composicionales*, pues las componentes tan sólo nos aportan información relativa y no absoluta. En estos casos, sólo los *cocientes* entre componentes tienen sentido, y esos cocientes son independientes del total arbitrario.

Ocasionalmente, las componentes de los vectores observacionales representan magnitudes absolutas y tienen por tanto un significado en sí mismas. Sin embargo, a pesar de ello, podemos optar por tener en cuenta únicamente la información aportada por los cocientes entre componentes. En este caso, asumimos que los vectores $\mathbf{w}$ y $k\mathbf{w}$, para todo $k > 0$, nos aportan la misma información. Por ello, en este caso, estamos interpretando nuestros datos como datos *composicionales*.

En todo caso, tanto si nuestro conjunto de datos es de naturaleza composicional de origen, como si elegimos analizarlos como tales, nuestro análisis debe ser independiente de los representantes elegidos para representar las composiciones. Esto es equivalente a decir que todo análisis que pretenda tener un sentido debe poder expresarse en términos de cocientes de componentes de los vectores composicionales. En términos matemáticos, es equivalente a trabajar con clases de equivalencia del espacio composicional $\mathcal{C}$. Por ello, aunque la representación del espacio composicional en el símplex se considere óptima por ser la más sencilla, debemos ser capaces de realizar nuestro análisis independientemente de la representación.

## 2.4 Subcomposiciones

Ocasionalmente, necesitamos centrar nuestra atención sobre las magnitudes relativas de un subconjunto de las partes.

**Definición 4.** *Dada una composición* $\underline{\mathbf{w}} \in \mathcal{C}$*, toda composición obtenida de la selección de dos o más partes de* $\mathbf{w}$ *se denomina* subcomposición *de* $\underline{\mathbf{w}}$.

Sea $C$ el número de partes seleccionadas, con $2 \leq C < D$. Representamos por $S$ el subconjunto ordenado de índices de las partes seleccionadas de $\underline{\mathbf{w}}$ a incluir en la subcomposición, y por $\underline{\mathbf{w}}_S$ la subcomposición resultante, que pertenece al espacio composicional $\mathcal{C}_S$.

**Definición 5.** *Dado un conjunto ordenado* $S$ *compuesto por* $C$ *índices distintos de* $\{1, \ldots, D\}$*, la acción de extraer una subcomposición es una transformación* $\mathrm{sub}_S$ *de* $\mathcal{C}$ *en* $\mathcal{C}_S$ *dada por*

$$\begin{array}{rccc} \mathrm{sub}_S: & \mathcal{C} & \to & \mathcal{C}_S \\ & \underline{\mathbf{w}} & \to & \underline{\mathbf{w}}_S \end{array} \quad . \tag{1}$$

Es evidente que $\mathrm{sub}_S\underline{\mathbf{w}}$ no depende del vector observacional elegido para representar la composición $\underline{\mathbf{w}}$. Geométricamente, formar una subcomposición $\underline{\mathbf{w}}_S$ a partir de una composición $\underline{\mathbf{w}}$ con $D$-partes, corresponde a la proyección ortogonal de la semirrecta asociada a $\underline{\mathbf{w}}$ en $\mathbb{R}^D_+$ sobre un subespacio de dimensión $C$. Este subespacio está generado por los ejes de coordenadas asociados a las partes seleccionadas para formar la subcomposición (ver Fig. 4).
La aplicación $\mathrm{sub}_S$ de $\mathcal{C}$ en $\mathcal{C}_S$ no es inyectiva. A pesar de ello, sería interesante definir una aplicación de $\mathcal{C}_S$ en $\mathcal{C}$ que de forma unívoca asocie toda composición de $\mathcal{C}_S$ a una composición de $\mathcal{C}$. Con el fin de no complicar la notación de forma innecesaria, asumiremos sin pérdida de generalidad que $S = \{D - C + 1, \ldots, D\}$.

**Definición 6.** *Dado un vector* $\underline{\mathbf{w}}_S \in \mathcal{C}_S$*, definimos la aplicación* $\mathrm{inc}_S$ *de* $\mathcal{C}_S$ *en* $\mathcal{C}$ *por*

$$\mathrm{inc}_S\, \underline{\mathbf{w}}_S = \mathrm{ccl}\left(1, \overset{D-C}{\ldots}, 1, \frac{w_1}{g(\mathbf{w}_S)}, \ldots, \frac{w_C}{g(\mathbf{w}_S)}\right)', \tag{2}$$

*donde* $\mathbf{w}_S = (w_1, \ldots, w_C)'$ *es un representante arbitrario de* $\underline{\mathbf{w}}_S$.

Es inmediato que $\mathrm{inc}_S\, \underline{\mathbf{w}}_S$ no depende del vector observacional elegido para representar la composición $\underline{\mathbf{w}}_S$.

**Proposición 1.** *La aplicación* $\mathrm{inc}_S$ *de* $\mathcal{C}_S$ *en* $\mathcal{C}$ *es inyectiva. Además, la aplicación compuesta* $\mathrm{sub}_S \circ \mathrm{inc}_S$ *es la aplicación identidad en* $\mathcal{C}_S$.

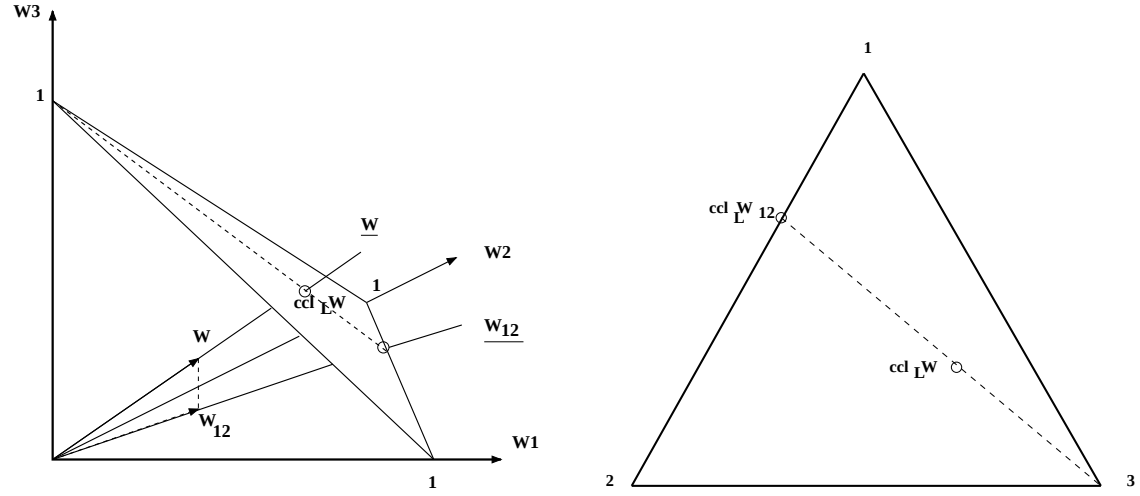



Figura 4: Interpretación geométrica de la formación de una subcomposición $\underline{\mathbf{w}_{12}}$ a partir de una composición $\underline{\mathbf{w}}$: (a) En $\mathbb{R}^3_+$; (b) En $\mathcal{S}^3$.

# 3 Transformaciones del espacio composicional

## 3.1 Un espacio cociente en $\mathbb{R}^D$

La transformación logaritmo de $\mathbb{R}^D_+$ en $\mathbb{R}^D$, siendo biyectiva, sugiere definir en $\mathbb{R}^D$ una relación de equivalencia relacionada con la relación de equivalencia composicional definida en $\mathbb{R}^D_+$, pues tenemos que si $\mathbf{w} \sim \mathbf{w}^*$, entonces $\log \mathbf{w} - \log \mathbf{w}^*$ es un múltiplo del vector de unidades $\mathbf{1}_D = (1, \ldots, 1)' \in \mathbb{R}^D$.

**Definición 7.** *Diremos que dos vectores* $\mathbf{z}$ *y* $\mathbf{z}^*$ *de* $\mathbb{R}^D$ *son* equivalentes*, y escribiremos* $\mathbf{z} \equiv \mathbf{z}^*$*, sí y sólo sí existe una constante* $\lambda$ *tal que* $\mathbf{z}^* = \mathbf{z} + \lambda \mathbf{1}_D$*. Teniendo en cuenta que* $U = \{\lambda \mathbf{1}_D : \lambda \in \mathbb{R}\}$ *es un subespacio unidimensional de* $\mathbb{R}^D$*, la relación de equivalencia anterior puede escribirse asimismo*

$$\mathbf{z} \equiv \mathbf{z}^* \quad \Longleftrightarrow \quad \mathbf{z} - \mathbf{z}^* \in U.$$

En consecuencia, es natural representar por $\mathbf{z} + U$ la clase de equivalencia generada por el vector $\mathbf{z}$ en $\mathbb{R}^D$. El conjunto de todas estas clases es el espacio cociente $\mathbb{R}^D/U$ y lo representaremos por $\mathcal{L}^d$. La aplicación cociente de $\mathbb{R}^D$ en $\mathcal{L}^d$ que asigna la clase $\mathbf{z} + U$ a cada vector $\mathbf{z} \in \mathbb{R}^D$ se indicará mediante ucl:

$$\operatorname{ucl} \mathbf{z} = \mathbf{z} + U \qquad (\mathbf{z} \in \mathbb{R}^D).$$

De la Figura 5 se desprende que los clases $\mathbf{z}+U$ pueden interpretarse geométricamente como rectas paralelas a $\mathbf{1}_D$. Así pues, parece natural representar una clase de equivalencia $\mathbf{z} + U$ por el punto de intersección de la recta asociada a dicha clase con el

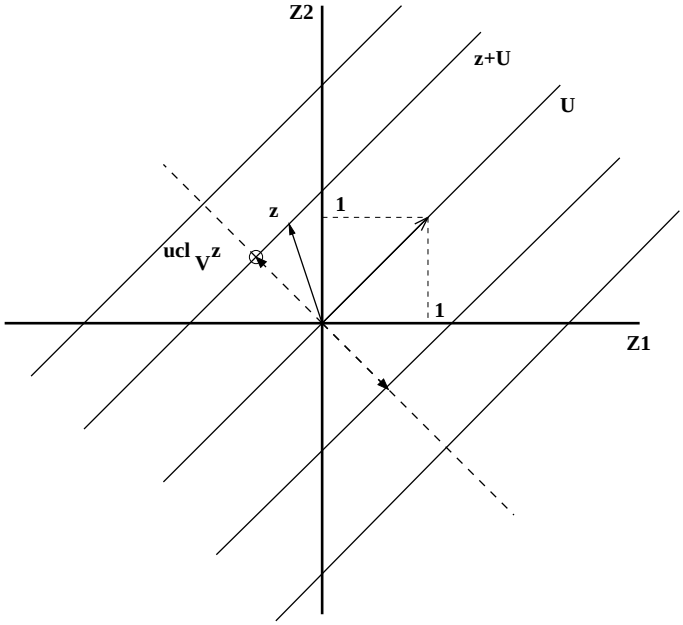



Figura 5: Interpretación geométrica de clases de equivalencia en $\mathcal{L}^1 = \mathbb{R}^2/U$

hiperplano $V = \{\mathbf{z} \in \mathbb{R}^D : \mathbf{z}'\mathbf{1}_D = 0\} \subset \mathbb{R}^D$, que pasa por el origen y es ortogonal a $\mathbf{1}_D$. Este punto de intersección puede interpretarse asimismo como la proyección ortogonal sobre $V$ de todos los vectores pertenecientes a la clase $\mathbf{z} + U$. Representaremos por $\mathrm{ucl}_V$ el operador que transforma cada vector $\mathbf{z}$ de $\mathbb{R}^D$ en su proyección ortogonal sobre el hiperplano $V$. Es evidente que $\mathbf{z} \equiv \mathrm{ucl}_V\, \mathbf{z}$, y que el operador $\mathrm{ucl}_V$ es constante sobre los vectores pertenecientes a una misma clase. Es fácil demostrar que

$$\mathrm{ucl}_V\, \mathbf{z} = \mathbf{z} - \frac{\sum_{j=1}^{D} z_j}{D}\mathbf{1}_D = \mathbf{H}_D\mathbf{z},$$

donde $\mathbf{H}_D$ es la *matriz de centrado* de orden $D \times D$ habitual (Mardia et al. 1979). Recordemos que esta matriz es igual a $\mathbf{I}_D - D^{-1}\mathbf{J}_D$, donde $\mathbf{I}_D$ es la matriz identidad de orden $D \times D$, y $\mathbf{J}_D = \mathbf{1}_D\mathbf{1}_D'$.

## 3.2 Transformaciones entre espacios cocientes

La transformación logaritmo de $\mathbb{R}_+^D$ en $\mathbb{R}^D$ y su inversa, la transformación exponencial, son compatibles con las relaciones de equivalencia $\sim$ y $\equiv$ definidas en $\mathbb{R}_+^D$ y $\mathbb{R}^D$, respectivamente, es decir,

$$\mathbf{w} \sim \mathbf{w}^* \text{ in } \mathbb{R}_+^D \quad \Longrightarrow \quad \log \mathbf{w} \equiv \log \mathbf{w}^* \text{ in } \mathbb{R}^D,$$

$$\mathbf{z} \equiv \mathbf{z}^* \text{ in } \mathbb{R}^D \quad \Longrightarrow \quad \exp \mathbf{z} \sim \exp \mathbf{z}^* \text{ in } \mathbb{R}_+^D.$$

Por ello, es posible extender estas transformaciones a los espacios cocientes $\mathcal{C}$ y $\mathcal{L}^d$. Representaremos por logc la transformación de $\mathcal{C}$ en $\mathcal{L}^d$,

$$\operatorname{logc} \underline{\mathbf{w}} = \log \mathbf{w} + U \qquad (\underline{\mathbf{w}} \in \mathcal{C}),$$

y por expc la transformación inversa de $\mathcal{L}^d$ en $\mathcal{C}$,

$$\operatorname{expc} (\mathbf{z} + U) = \operatorname{ccl} (\exp \mathbf{z}) \qquad (\mathbf{z} + U \in \mathcal{L}^d).$$

Entonces, el vector proyección de la clase $\operatorname{logc} \underline{\mathbf{w}}$, vendrá dado por

$$\operatorname{ucl}_V (\log \mathbf{w}) = \mathbf{H}_D \log \mathbf{w} = \log \frac{\mathbf{w}}{g(\mathbf{w})}.$$

**Definición 8.** *Denominaremos* transformación logcociente centrada *a la función biunívoca del espacio composicional $\mathcal{C}$ en el subespacio $V$ de $\mathbb{R}^D$, definida por*

$$\operatorname{clr} \underline{\mathbf{w}} = \log \frac{\mathbf{w}}{g(\mathbf{w})} \qquad (\underline{\mathbf{w}} \in \mathcal{C}).$$

*La transformación inversa, de $V$ en $\mathcal{C}$, viene dada por*

$$\operatorname{clr}^{-1} \mathbf{z} = \operatorname{ccl} (\exp \mathbf{z}) \qquad (\mathbf{z} \in V).$$

Nótese que las transformaciones logaritmo y exponencial establecen una correspondencia biunívoca entre la superficie hiperbólica $Hip_D$ en $\mathbb{R}^D_+$ y el hiperplano $V$ en $\mathbb{R}^D$.

# 4 Estructura vectorial del espacio composicional

Puesto que $U$ es un subespacio vectorial unidimensional de $\mathbb{R}^D$, es posible dotar al espacio cociente $\mathcal{L}^d = \mathbb{R}^D/U$ de estructura de espacio vectorial real de dimensión $d = D - 1$. Para ello, se define la suma de dos clases de equivalencia $\mathbf{z} + U$ y $\mathbf{z}^* + U$ por

$$(\mathbf{z} + U) + (\mathbf{z}^* + U) = (\mathbf{z} + \mathbf{z}^*) + U,$$

y el producto de una clase $\mathbf{z} + U$ por un escalar $\lambda \in \mathbb{R}$ por

$$\lambda(\mathbf{z} + U) = \lambda \mathbf{z} + U.$$

La clase $\mathbf{1}_D + U$ es entonces el elemento neutro y el elemento inverso de $\mathbf{z} + U$ es la clase $(-\mathbf{z}) + U$. La correspondencia biunívoca entre $\mathcal{C}$ y $\mathcal{L}^d$ permite entonces dotar a $\mathcal{C}$ de estructura de espacio vectorial real isomorfo a $\mathcal{L}^d$. En efecto:

**Definición 9.** *En correspondencia con la suma en $\mathcal{L}^d$, se define una operación interna $\oplus$ en $\mathcal{C}$ por*

$$\underline{\mathbf{w}} \oplus \underline{\mathbf{w}}^* = \operatorname{expc} (\operatorname{logc} \underline{\mathbf{w}} + \operatorname{logc} \underline{\mathbf{w}}^*) = \operatorname{ccl} (w_1 w_1^*, \ldots, w_D w_D^*)' \qquad (\underline{\mathbf{w}}, \underline{\mathbf{w}}^* \in \mathcal{C}).$$

*Asimismo, en correspondencia con el producto por un escalar en $\mathcal{L}^d$, se define una operación externa $\otimes$ en $\mathcal{C}$ por*

$$\lambda \otimes \underline{\mathbf{w}} = \operatorname{expc} (\lambda \operatorname{logc} \underline{\mathbf{w}}) = \operatorname{ccl} (w_1^\lambda, \ldots, w_D^\lambda)' \qquad (\underline{\mathbf{w}} \in \mathcal{C}) \ (\lambda \in \mathbb{R}).$$

De esta forma $(\mathcal{C}, \oplus, \otimes)$ deviene un espacio vectorial real, isomorfo al espacio cociente $\mathcal{L}^d$. En el grupo conmutativo $(\mathcal{C}, \oplus)$, la composición $\underline{\mathbf{1}}_D = \text{ccl}\,(1, \ldots, 1)'$ es el elemento neutro, y la composición inversa de $\underline{\mathbf{w}} = \text{ccl}\,(w_1, \ldots, w_D)'$ es la composición $\underline{\mathbf{w}}^{-1} = \text{ccl}\,(1/w_1, \ldots, 1/w_D)'$.
Puesto que $(\mathcal{C}, \oplus, \otimes)$ es un espacio vectorial real, puede entenderse como un espacio afín cuando el grupo $(\mathcal{C}, \oplus)$ opera sobre $\mathcal{C}$ como grupo de transformaciones.

**Definición 10.** *Dada una composición* $\mathbf{p} \in \mathcal{C}$*, la* perturbación *asociada a* $\mathbf{p}$ *es la transformación de* $\mathcal{C}$ *en* $\mathcal{C}$ *definida por*

$$\mathbf{c} \to \mathbf{p} \oplus \mathbf{c} \qquad (\mathbf{c} \in \mathcal{C}).$$

*Diremos que* $\mathbf{p} \oplus \mathbf{c}$ *es la composición resultante de aplicar la* perturbación $\mathbf{p}$ *a la composición* $\mathbf{c}$*.*

La perturbación juega en el espacio composicional el mismo rol que la traslación en el espacio real. Del mismo modo, el conjunto de todas las perturbaciones en $\mathcal{C}$ es un grupo conmutativo isomorfo a $(\mathcal{C}, \oplus)$. En consecuencia, la composición de dos perturbaciones $\mathbf{p}_1$ y $\mathbf{p}_2$ es la perturbación asociada a $\mathbf{p}_1 \oplus \mathbf{p}_2$. Además, la perturbación asociada a $\underline{\mathbf{1}}_D$ es la perturbación identidad y para toda perturbación $\mathbf{p}$ existe la perturbación inversa $\mathbf{p}^{-1}$. Finalmente, dadas dos composiciones

$$\underline{\mathbf{w}} = \text{ccl}\,(w_1, \ldots, w_D)' \text{ y } \underline{\mathbf{w}}^* = \text{ccl}\,(w_1^*, \ldots, w_D^*)' \in \mathcal{C},$$

existe una única perturbación $\mathbf{p}$ que transforma $\underline{\mathbf{w}}$ en $\underline{\mathbf{w}}^*$ y es

$$\mathbf{p} = \underline{\mathbf{w}}^* \oplus \underline{\mathbf{w}}^{-1} = \text{ccl}\,(\frac{w_1^*}{w_1}, \ldots, \frac{w_D^*}{w_D})'.$$

Nótese que estos resultados son válidos por igual para cualquier representación posible del espacio composicional $\mathcal{C}$. Resulta así que tanto el símplex, como la intersección de la esfera con $\mathbb{R}_+^D$, como la superficie hiperbólica de la sección 2.2, tienen estructura de espacio vectorial con las operaciones indicadas compuestas con el criterio de selección correspondiente.
La hipótesis de que el grupo de perturbaciones es el grupo de operaciones en el espacio composicional es el elemento clave de la metodología introducida por Aitchison (1986). En efecto, implica aceptar que la diferencia entre dos composiciones $\underline{\mathbf{w}} = \text{ccl}\,(w_1, \ldots, w_D)'$ y $\underline{\mathbf{w}}^* = \text{ccl}\,(w_1^*, \ldots, w_D^*)'$ se basa en las razones $w_j^*/w_j$ entre partes y no en las diferencias $w_j^* - w_j$.

# 5 Estructura euclidiana del espacio composicional

## 5.1 $\mathcal{L}^d$ como espacio euclidiano

Puesto que los elementos de $\mathcal{L}^d$ pueden interpretarse como rectas paralelas al vector $\mathbf{1}_D$, parece lógico definir la distancia entre dos clases $\mathbf{z} + U$ y $\mathbf{z}^* + U$ de $\mathcal{L}^d$ como la

distancia euclidiana entre dichas rectas en $\mathbb{R}^D$. Esa distancia será igual a la norma del vector diferencia $\mathrm{ucl}_V\,\mathbf{z}^* - \mathrm{ucl}_V\,\mathbf{z}$, donde $\mathrm{ucl}_V\,\mathbf{z}$ y $\mathrm{ucl}_V\,\mathbf{z}^*$ son los puntos de intersección de las rectas con el hiperplano ortogonal $V$ (ver Fig. 6). En consecuencia,

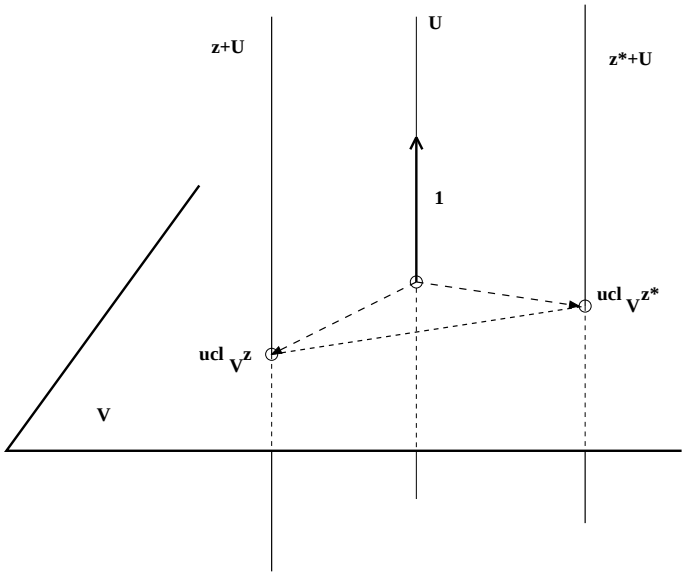



Figura 6: Distancia entre dos clases de equivalencia $\mathbf{z}+U$ y $\mathbf{z}^*+U$ de $\mathcal{L}^d$.

es inmediato transferir la estructura euclidiana de $V \subset \mathbb{R}^D$ a $\mathcal{L}^d$.

**Definición 11.** *Dados $\mathbf{z}+U \in \mathcal{L}^d$ y $\mathbf{z}^*+U \in \mathcal{L}^d$, definimos el $\mathcal{L}$-producto interno $<\mathbf{z}+U, \mathbf{z}^*+U>_{\mathcal{L}}$ como el producto usual $<\mathrm{ucl}_V\,\mathbf{z}, \mathrm{ucl}_V\,\mathbf{z}^*>$ en $\mathbb{R}^D$.*

Es fácil demostrar que

$$<\mathbf{z}+U, \mathbf{z}^*+U>_{\mathcal{L}} = \sum_{j=1}^{D} z_j z_j^* - \frac{1}{D}\left(\sum_{j=1}^{D} z_j\right)\left(\sum_{j=1}^{D} z_j^*\right) = \mathbf{z}'\mathbf{H}_D\mathbf{z}^*,$$

para todo $\mathbf{z}+U, \mathbf{z}^*+U \in \mathcal{L}^d$. A partir de aquí, se definen de forma habitual una norma y una distancia en $\mathcal{L}^d$. La $\mathcal{L}$-norma de un clase $\mathbf{z}+U \in \mathcal{L}^d$ tiene por expresión

$$\|\mathbf{z}+U\|_{\mathcal{L}} = (<\mathbf{z}+U, \mathbf{z}+U>_{\mathcal{L}})^{1/2} = \left[\sum_{j=1}^{D} z_j^2 - \frac{1}{D}\left(\sum_{j=1}^{D} z_j\right)^2\right]^{1/2} = (\mathbf{z}'\mathbf{H}_D\mathbf{z})^{1/2},$$

y se cumple que $\|\mathbf{z}+U\|_{\mathcal{L}} = \|\mathrm{ucl}_V\,\mathbf{z}\|$.
Análogamente, la $\mathcal{L}$-distancia entre dos clases $\mathbf{z}+U$ y $\mathbf{z}^*+U$ viene dada por

$$d_{\mathcal{L}}(\mathbf{z}+U, \mathbf{z}^*+U) = \|(\mathbf{z}^*+U)-(\mathbf{z}+U)\|_{\mathcal{L}} = \left[\sum_{j=1}^{D}(z_j^*-z_j)^2 - \frac{1}{D}\left(\sum_{j=1}^{D}(z_j^*-z_j)\right)^2\right]^{1/2}.$$

Esta expresión puede escribirse en forma matricial como

$$d_{\mathcal{L}}(\mathbf{z}+U,\mathbf{z}^*+U)=\left[(\mathbf{z}-\mathbf{z}^*)'\mathbf{H}_D(\mathbf{z}-\mathbf{z}^*)\right]^{1/2},$$

y se cumple que $d_{\mathcal{L}}(\mathbf{z}+U,\mathbf{z}^*+U)=d(\mathrm{ucl}_V\mathbf{z},\mathrm{ucl}_V\mathbf{z}^*)$. De este modo, el espacio cociente $\mathcal{L}^d$ deviene un espacio euclidiano.

## 5.2 $\mathcal{C}$ como espacio euclidiano

Las transformaciones biunívocas logc y expc entre $\mathcal{C}$ y $\mathcal{L}^d$ permiten transferir a $\mathcal{C}$ la estructura de espacio real euclidiano definida en $\mathcal{L}^d$.

**Definición 12.** *Dadas dos composiciones* $\underline{\mathbf{w}}$ *y* $\underline{\mathbf{w}}^*$*, definimos el* producto interno composicional *por*

$$<\underline{\mathbf{w}},\underline{\mathbf{w}}^*>_{\mathcal{C}}=<\log\mathbf{w}+U,\log\mathbf{w}^*+U>_{\mathcal{L}}.$$

Es fácil demostrar que

$$<\underline{\mathbf{w}},\underline{\mathbf{w}}^*>_{\mathcal{C}}=\sum_{j=1}^{D}\log w_j\log w_j^*-\frac{1}{D}\left(\sum_{j=1}^{D}\log w_j\right)\left(\sum_{j=1}^{D}\log w_j^*\right)=(\log\mathbf{w})'\mathbf{H}_D\log\mathbf{w}^*,$$

y que

$$<\underline{\mathbf{w}},\underline{\mathbf{w}}^*>_{\mathcal{C}}=\sum_{j=1}^{D}\log\frac{w_j}{g(\mathbf{w})}\log\frac{w_j^*}{g(\mathbf{w}^*)}=<\mathrm{clr}\,\underline{\mathbf{w}},\mathrm{clr}\,\underline{\mathbf{w}}^*>,$$

para cada $\underline{\mathbf{w}},\underline{\mathbf{w}}^*\in\mathcal{C}$, resultando que el $\mathcal{C}$-producto interno de $\underline{\mathbf{w}}$ y $\underline{\mathbf{w}}^*$ en $\mathcal{C}$ coincide con el producto interno ordinario de $\mathrm{clr}\,\underline{\mathbf{w}}$ y $\mathrm{clr}\,\underline{\mathbf{w}}^*$ en $\mathbb{R}^D$. Como es habitual, diremos que dos composiciones $\underline{\mathbf{w}}$ y $\underline{\mathbf{w}}^*$ son $\mathcal{C}$*-ortogonales*, o simplemente *ortogonales*, sí y sólo sí $<\underline{\mathbf{w}},\underline{\mathbf{w}}^*>_{\mathcal{C}}=0$.
A partir de este producto interno en $\mathcal{C}$ podemos definir una norma y una distancia en el espacio composicional. La *norma composicional* de una composición $\underline{\mathbf{w}}\in\mathcal{C}$ vendrá dada por

$$\begin{aligned}\|\underline{\mathbf{w}}\|_{\mathcal{C}} &= (<\underline{\mathbf{w}},\underline{\mathbf{w}}>_{\mathcal{C}})^{1/2}=\left[\sum_{j=1}^{D}(\log w_j)^2-\frac{1}{D}\left(\sum_{j=1}^{D}\log w_j\right)^2\right]^{1/2}\\ &= \left[(\log\mathbf{w})'\mathbf{H}_D\log\mathbf{w}\right]^{1/2}.\end{aligned}$$

En consecuencia, la $\mathcal{C}$-norma de una composición coincide con la norma euclidiana en $\mathbb{R}^D$ del vector clr-transformado:

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}}=\|\mathrm{clr}\,\underline{\mathbf{w}}\|\qquad(\underline{\mathbf{w}}\in\mathcal{C}).$$

Otra expresión posible para la $\mathcal{C}$-norma de una composición viene dada por

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}}^2 = \frac{1}{D} \sum_{1 \le i < j \le D} \left( \log \frac{w_i}{w_j} \right)^2 \quad (\underline{\mathbf{w}} \in \mathcal{C}).$$

Una composición $\underline{\mathbf{w}}$ diremos que es $\mathcal{C}$-*unitaria* sí y sólo sí $\|\underline{\mathbf{w}}\|_{\mathcal{C}} = 1$.
La *distancia composicional* entre dos composiciones $\underline{\mathbf{w}}$ y $\underline{\mathbf{w}}^*$ viene dada por la $\mathcal{C}$-norma de la composición $\underline{\mathbf{w}}^* \oplus \underline{\mathbf{w}}^{-1} = \text{ccl}\ (w_1^*/w_1, \ldots, w_D^*/w_D)'$, es decir,

$$d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}}^*) = \left[ \sum_{j=1}^{D} \left( \log \frac{w_j^*}{w_j} \right)^2 - \frac{1}{D} \left( \sum_{j=1}^{D} \log \frac{w_j^*}{w_j} \right)^2 \right]^{1/2}.$$

Esta distancia puede expresarse en forma matricial por

$$d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}}^*) = \left[ (\log \mathbf{w}^* - \log \mathbf{w})' \mathbf{H}_D (\log \mathbf{w}^* - \log \mathbf{w}) \right]^{1/2}.$$

Por ello, la $\mathcal{C}$-distancia entre dos composiciones coincide con la distancia euclidiana ordinaria en $\mathbb{R}^D$ entre los correspondientes vectores clr-transformados:

$$d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}}^*) = d(\text{clr}\, \underline{\mathbf{w}}, \text{clr}\, \underline{\mathbf{w}}^*) \qquad (\underline{\mathbf{w}}, \underline{\mathbf{w}}^* \in \mathcal{C}).$$

Así pues, la $\mathcal{C}$-distancia definida convierte el espacio composicional $\mathcal{C}$ en un espacio euclidiano isométrico al espacio euclidiano $\mathcal{L}^d$. Además, la transformación logcociente centrada clr es una isometría natural entre $\mathcal{C}$ y el subespacio $V$ de $\mathbb{R}^D$.
Esta distancia en $\mathcal{C}$ cumple todas las propiedades habituales de las distancias euclidianas. En particular, está relacionada con las operaciones del espacio composicional por las siguientes identidades:

$$d_{\mathcal{C}}(\mathbf{a}, \mathbf{b}) = d_{\mathcal{C}}(\mathbf{a} \oplus \mathbf{c}, \mathbf{b} \oplus \mathbf{c}) \quad (\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathcal{C}),$$

y

$$d_{\mathcal{C}}(\lambda \otimes \mathbf{a}, \lambda \otimes \mathbf{b}) = |\lambda| d_{\mathcal{C}}(\mathbf{a}, \mathbf{b}) \quad (\mathbf{a}, \mathbf{b} \in \mathcal{C})\ (\lambda \in \mathbb{R}).$$

Otra propiedad importante de esta distancia composicional está relacionada con las subcomposiciones. Se basa en el hecho de que la aplicación $\text{sub}_S$ introducida en la definición 5, que transforma una composición $\underline{\mathbf{w}} \in \mathcal{C}$ en una subcomposición $\underline{\mathbf{w}}_S \in \mathcal{C}_S$, es una aplicación lineal entre los espacios vectoriales reales $(\mathcal{C}, \oplus, \otimes)$ y $(\mathcal{C}_S, \oplus, \otimes)$.

**Proposición 2.** *La distancia composicional es* subcomposicionalmente dominante, *pues*

$$d_{\mathcal{C}}(\underline{\mathbf{w}}, \underline{\mathbf{w}^*}) \ge d_{\mathcal{C}}(\underline{\mathbf{w}}_S, \underline{\mathbf{w}_S^*}) \quad (\underline{\mathbf{w}}, \underline{\mathbf{w}^*} \in \mathcal{C}),$$

*o, equivalentemente,*

$$\|\underline{\mathbf{w}}\|_{\mathcal{C}} \ \ge \ \|\underline{\mathbf{w}}_S\|_{\mathcal{C}} \quad (\underline{\mathbf{w}} \in \mathcal{C}).$$

Tenemos además que la aplicación $\mathrm{inc}_S$ introducida en la definición 6 es una aplicación lineal entre los espacios vectoriales reales $(\mathcal{C}_S, \oplus, \otimes)$ y $(\mathcal{C}, \oplus, \otimes)$, y que preserva la $\mathcal{C}$-distancia, puesto que cumple

$$\|\mathrm{ccl}\,(w_1, \ldots, w_C)'\|_{\mathcal{C}} = \left\|\mathrm{ccl}\,\left(1, \overset{D-C}{\ldots}, 1, \frac{w_1}{g(\mathbf{w})}, \ldots, \frac{w_C}{g(\mathbf{w})}\right)'\right\|_{\mathcal{C}} \quad (\underline{\mathbf{w}} \in \mathcal{C}_S).$$

La proposición 2 implica que la $\mathcal{C}$-distancia entre dos subcomposiciones nunca puede ser mayor que la $\mathcal{C}$-distancia entre las correspondientes composiciones. Es razonable exigir que esta propiedad sea satisfecha por toda distancia definida sobre el espacio composicional.

# 6 Bases en el espacio composicional

## 6.1 Bases naturales en $\mathcal{L}^d$

Sea $\mathbf{e}_1 = (1, 0, \ldots, 0)', \mathbf{e}_2 = (0, 1, 0, \ldots, 0)', \ldots, \mathbf{e}_D = (0, \ldots, 0, 1)'$ la base canónica de $\mathbb{R}^D$. Sea $\mathcal{B}^{\mathcal{L}}$ el conjunto ordenado $\{\mathbf{e}_1 + U, \ldots, \mathbf{e}_D + U\}$ de clases de equivalencia de $\mathcal{L}^d$. Para cada $j = 1, \ldots, D$, representaremos por $\mathcal{B}^{\mathcal{L}}_{-j}$ el conjunto ordenado $\mathcal{B}^{\mathcal{L}} - \{\mathbf{e}_j + U\}$. Es evidente que el conjunto $\mathcal{B}^{\mathcal{L}}_{-j}$ es una base de $\mathcal{L}^d$ para cada $j = 1, \ldots, D$. En particular, si $\mathbf{z} = (z_1, \ldots, z_D)' \in \mathbb{R}^D$, el vector $\mathbf{y}$ cuyas coordenadas son las de $\mathbf{z} + U$ en la base $\mathcal{B}^{\mathcal{L}}_{-D}$ es

$$\mathbf{y} = (z_1 - z_D, \ldots, z_d - z_D)' = \mathbf{F}\mathbf{z},$$

donde $\mathbf{F}$ es la $d \times D$ matriz $[\mathbf{I}_d : -\mathbf{1}_d]$. Para esta base se cumple que

$$\|\mathbf{e}_i + U\|^2_{\mathcal{L}} = \frac{D-1}{D} \quad (i = 1, \ldots, D),$$

y

$$<\mathbf{e}_i + U, \mathbf{e}_j + U>_{\mathcal{L}} = -\frac{1}{D} \quad (i, j = 1, \ldots, D;\ i \neq j).$$

Por consiguiente, ninguna de las bases $\mathcal{B}^{\mathcal{L}}_{-j}$ de $\mathcal{L}^d$ es $\mathcal{L}$-ortonormal. La matriz $d \times d$ que expresa la $\mathcal{L}$-métrica en la base $\mathcal{B}^{\mathcal{L}}_{-j}$ de $\mathcal{L}^d$ es igual a $\mathbf{M} = \mathbf{I}_d - D^{-1}\mathbf{J}_d$, independientemente del índice $j = 1, \ldots, D$. Esta matriz puede expresarse como una función de $\mathbf{F}$, pues se cumple que:

$$\mathbf{M} = (\mathbf{F}\mathbf{F}')^{-1}. \tag{3}$$

## 6.2 Bases ortonormales de $\mathcal{L}^d$

El subespacio $V = \{\mathbf{z} \in \mathbb{R}^D : \mathbf{z}'\mathbf{1}_D = 0\}$ de $\mathbb{R}^D$ tiene dimensión $d = D - 1$. Sea $\mathbf{v}_1 = (v_{11}, \ldots, v_{1D})', \ldots, \mathbf{v}_d = (v_{d1}, \ldots, v_{dD})'$ una base ortonormal de $V$, y sea $\mathbf{V}$ la

$D \times d$ matriz $[\mathbf{v}_1 : \ldots : \mathbf{v}_d]$. Es inmediato comprobar que esta matriz cumple las igualdades

$$\text{(i)}\ \ \mathbf{V}'\mathbf{V} = \mathbf{I}_d; \qquad \text{y} \qquad \text{(ii)}\ \ \mathbf{V}\mathbf{V}' = \mathbf{H}_D. \tag{4}$$

Recíprocamente, si $\mathbf{V}$ es una matriz $D \times d$ que satisface las igualdades (4), sus vectores columna son una base ortonormal del subespacio $V$. Entonces, el conjunto ordenado $\mathcal{V}^{\mathcal{L}} = \{\mathbf{v}_1 + U, \ldots, \mathbf{v}_d + U\}$ es una base $\mathcal{L}$-ortonormal de $\mathcal{L}^d$. Si $\mathbf{z} \in \mathbb{R}^D$, el vector $\mathbf{u}$ de $\mathbb{R}^d$ cuyas componentes son las de la clase $\mathbf{z} + U$ en la base $\mathcal{V}^{\mathcal{L}}$ es

$$\mathbf{u} = (\mathbf{F}\mathbf{V})^{-1}\mathbf{F}\mathbf{z}.$$

Recordemos que dadas dos bases, siempre existe una transformación lineal que permite pasar de una a otra. En consecuencia, los elementos de $\mathcal{V}^{\mathcal{L}}$ serán combinación lineal de los elementos de $\mathcal{B}^{\mathcal{L}}_{-j}$.

## 6.3 Bases naturales en $\mathcal{C}$

Partiendo de las clases $\mathbf{e}_1 + U, \ldots, \mathbf{e}_D + U$ de $\mathcal{L}^d$, definimos las correspondientes composiciones en $\mathcal{C}$:

$$\tilde{\mathbf{e}}_1 = \operatorname{expc}(\mathbf{e}_1 + U) = \operatorname{ccl}(e, 1, \ldots, 1, 1)', \ldots, \tilde{\mathbf{e}}_D = \operatorname{expc}(\mathbf{e}_D + U) = \operatorname{ccl}(1, 1, \ldots, 1, e)'.$$

**Definición 13.** *Si $\mathcal{B}$ representa el conjunto ordenado $\{\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_D\}$, el conjunto $\mathcal{B}_{-j} = \mathcal{B} - \{\tilde{\mathbf{e}}_j\}$ es una base del espacio vectorial real $(\mathcal{C}, \oplus, \otimes)$ para todo índice $j = 1, \ldots, D$. Estas bases se denominan* bases naturales *de $\mathcal{C}$.*

Entonces, si $\mathbf{w} = (w_1, \ldots, w_D)' \in \mathbb{R}^D_+$, el vector $\mathbf{y}$ de $\mathbb{R}^d$ cuyas componentes son las de la composición $\underline{\mathbf{w}}$ en la base $\mathcal{B}_{-D}$ es igual a

$$\mathbf{y} = (\log \frac{w_1}{w_D}, \ldots, \log \frac{w_d}{w_D})'.$$

En general, si $\mathbf{w}_{-j}$ representa el vector $\mathbf{w}$ sin la componente $w_j$, las components de $\underline{\mathbf{w}}$ en la base $\mathcal{B}_{-j}$ son las del vector $\log(\mathbf{w}_{-j}/w_j)$.

**Definición 14.** *La* transformación logcociente aditiva *de índice $j$ ($j = 1, \ldots, D$) —denotada por* $\operatorname{alr}_j$ *— es una transformación biunívoca de $\mathcal{C}$ en $\mathbb{R}^d$ que asigna a cada composición $\underline{\mathbf{w}}$ sus componentes en la base $\mathcal{B}_{-j}$:*

$$\underline{\mathbf{w}} \longrightarrow \operatorname{alr}_j \underline{\mathbf{w}} = \log \frac{\mathbf{w}_{-j}}{w_j}.$$

La transformación inversa de $\operatorname{alr}_j$, de $\mathbb{R}^d$ en $\mathcal{C}$, viene dada por

$$\operatorname{alr}_j^{-1} \mathbf{y} = \operatorname{ccl}(\exp y_1, \ldots, \exp y_{j-1}, 1, \exp y_j, \ldots, \exp y_d)' \quad (\mathbf{y} \in \mathbb{R}^d).$$

En particular, cuando $j = D$, estas transformaciones pueden expresarse fácilmente en notación matricial por

$$\operatorname{alr}_D \underline{\mathbf{w}} = \mathbf{F} \log \mathbf{w}, \qquad \text{y} \qquad \operatorname{alr}_D^{-1} \mathbf{y} = \operatorname{ccl}\left\{\exp\left[\mathbf{F}'(\mathbf{F}\mathbf{F}')^{-1}\mathbf{y}\right]\right\}.$$

## 6.4 Bases ortonormales en $\mathcal{C}$

Lo mismo que ocurre con las bases $\mathcal{B}_{-j}^{\mathcal{L}}$ de $\mathcal{L}^d$, ninguna de las bases $\mathcal{B}_{-j}$ de $\mathcal{C}$ es $\mathcal{C}$-ortonormal. La matriz $\mathbf{M}$ introducida en (3) es la matriz que determina la $\mathcal{C}$-métrica en estas bases. La matriz $\mathbf{M}$ corresponde a la matriz $\mathbf{H}^{-1}$ definida en Aitchison (1986, p. 343).

**Definición 15.** *De toda $D \times d$ matriz $\mathbf{V} = [\mathbf{v}_1 : \ldots : \mathbf{v}_d]$ que verifica las dos igualdades (4), podemos definir las composiciones*

$$\tilde{\mathbf{v}}_1 = \operatorname{expc}(\mathbf{v}_1 + U), \ldots, \tilde{\mathbf{v}}_d = \operatorname{expc}(\mathbf{v}_d + U).$$

*Entonces, el conjunto ordenado $\mathcal{V} = \{\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_d\}$ es una base $\mathcal{C}$-ortonormal de $\mathcal{C}$.*

Por consiguiente, si $\mathbf{w}$ es un vector observacional de $\mathbb{R}_+^D$, el vector $\mathbf{u}$ de $\mathbb{R}^d$ cuyas componentes son las de la composición $\underline{\mathbf{w}}$ en la base $\mathcal{V}$ es igual a

$$\mathbf{u} = (\mathbf{FV})^{-1}\mathbf{F} \log \mathbf{w}.$$

**Definición 16.** *Dada una $D \times d$ matriz $\mathbf{V} = [\mathbf{v}_1 : \ldots : \mathbf{v}_d]$ que satisface las condiciones (4), la* transformación logcociente isométrica *—denotada por* $\operatorname{ilr}_{\mathrm{V}}$ *— asociada a esta matriz $\mathbf{V}$, es la transformación biunívoca de $\mathcal{C}$ en $\mathbb{R}^d$ que asigna a cada composición $\underline{\mathbf{w}}$ sus componentes en la base $\mathcal{V}$:*

$$\underline{\mathbf{w}} \longrightarrow \operatorname{ilr}_{\mathrm{V}} \underline{\mathbf{w}} = (\mathbf{FV})^{-1}\mathbf{F} \log \mathbf{w}.$$

La transformación inversa de $\operatorname{ilr}_{\mathrm{V}}$ viene dada por

$$\operatorname{ilr}_{\mathrm{V}}^{-1} \mathbf{x} = \operatorname{ccl}\left(\exp\left\{\left[(\mathbf{FV})^{-1}\mathbf{F}\right]' \mathbf{x}\right\}\right) \quad (\mathbf{x} \in \mathbb{R}^d).$$

Nótese que, por construcción, la transformación $\operatorname{ilr}_{\mathrm{V}}$ es una isometría entre los espacios métricos $\mathcal{C}$ y $\mathbb{R}^d$, justificando así el término *transformación logcociente isométrica*. Este término fue acuñado por J.J. Egozcue (comunicación personal).

**Proposición 3.** *Si $\mathbf{V} = [\mathbf{v}_1 : \ldots : \mathbf{v}_d]$ y $\mathbf{V}^* = [\mathbf{v}_1^* : \ldots : \mathbf{v}_d^*]$ son dos $D \times d$ matrices que satisfacen las condiciones (4), entonces las transformaciones logcociente isométricas* $\operatorname{ilr}_{\mathrm{V}}$ *y* $\operatorname{ilr}_{V^*}$ *asociadas a $\mathbf{V}$ y $\mathbf{V}^*$, respectivamente, se relacionan por la siguiente igualdad:*

$$\operatorname{ilr}_{\mathrm{V}} \underline{\mathbf{w}} = \mathbf{V}'\mathbf{V}^* \operatorname{ilr}_{V^*} \underline{\mathbf{w}} \quad (\underline{\mathbf{w}} \in \mathcal{C}).$$

## 6.5 Representación de una composición

Como consecuencia de los resultados previos, una composición $\underline{\mathbf{w}} \in \mathcal{C}$ puede expresarse de diversas formas:

(i) Dando un vector $D$-observacional de $\underline{\mathbf{w}}$.

(ii) Dando las coordenadas $(y_1, \ldots, y_d) = \mathbf{y}'$ de $\underline{\mathbf{w}}$ en la base $\mathcal{B}_{-D}$ de $\mathcal{C}$. Si fuera necesario, podemos elegir las componentes de cualquier otro logcociente $\mathrm{alr}_j \underline{\mathbf{w}}$ $(j \neq D)$.

(iii) Dando las coordenadas $(z_1, \ldots, z_D)' = \mathbf{z}$ del vector transformado $\mathrm{clr}\, \underline{\mathbf{w}}$. Puesto que $\mathbf{z}$ pertenece al subespacio $V$ de $\mathbb{R}^D$, sus componentes están sujetas a la condición $z_1 + \ldots + z_D = 0$.

(iv) Dando las coordenadas $(u_1, \ldots, u_d)' = \mathbf{u}$ de $\underline{\mathbf{w}}$ en una base ortonormal $\mathcal{V}$ de $\mathcal{C}$. En este caso, es necesario conocer la matriz $\mathbf{V}$ que individualiza la base $\mathcal{V}$.

La relación entre las diversas expresiones es sencilla.

**Proposición 4.** *Los vectores* $\mathbf{u}$, $\mathbf{y}$ *y* $\mathbf{z}$ *asociados a una misma composición* $\underline{\mathbf{w}}$ *están relacionadas por las siguientes ecuaciones:*

1. $\mathbf{u} = (\mathbf{FV})^{-1}\mathbf{y}$, y $\mathbf{u} = (\mathbf{FV})^{-1}\mathbf{Fz}$.
2. $\mathbf{y} = \mathbf{FVu}$, y $\mathbf{y} = \mathbf{Fz}$.
3. $\mathbf{z} = \left[(\mathbf{FV})^{-1}\mathbf{F}\right]' \mathbf{u}$, y $\mathbf{z} = \mathbf{F}'(\mathbf{FF}')^{-1}\mathbf{y}$.

La opción (i) ofrece la ventaja de que las componentes $w_1, \ldots, w_D$ del vector $\mathbf{w}$ son directamente interpretables, pues son las componentes *observadas*, aunque los resultados se expresen en términos de clases de equivalencia. Por ello, habitualmente, suele elegirse el vector observacional $\mathrm{ccl}_L\, \mathbf{w}$ perteneciente al símplex $\mathcal{S}^D$, pues en este caso sus componentes expresan partes de un total.
En la opción (ii), las componentes $y_j$ son asimismo interpretables, pues representan simples logcocientes, $y_j = \log(w_j/w_D)$ $(j = 1, \ldots, d)$, y resulta muy sencillo calcular a partir de ellos cualquier otro logcociente:

$$\log \frac{w_i}{w_j} = y_i - y_j \ (i, j = 1, \ldots, d), \qquad \text{y} \qquad \log \frac{w_D}{w_j} = -y_j \ (j = 1, \ldots, d)$$

Es más difícil dar una interpretación directa las componentes clr, $z_j = \log\left[w_j/g(\mathbf{w})\right]$ $(j = 1, \ldots, D)$, de la opción (iii). Ello se debe a la presencia de la media geométrica $g(\mathbf{w})$ en el denominador de estos logcocientes. La componente $z_j$ da, en escala logarítmica, información sobre el valor de la parte $j$ con respecto al valor global de las otras partes. Sin embargo, es muy sencillo calcular a partir de las componentes clr cualquier otro logcociente, pues $\log(w_i/w_j) = z_i - z_j$ $(i, j = 1 \ldots D)$.
Finalmente, las componentes del vector $\mathbf{u}$ en la opción (iv) no son, en general, directamente interpretables, pues dependen de una matriz $\mathbf{V}$ elegida de forma arbitraria que satisfaga las condiciones (4). Sin embargo, hemos visto que siempre serán combinación lineal de los elementos de la base $\mathcal{B}_{-D}$, y con ello expresable en términos de logcocientes. Ello permite obtener fácilmente cualquier logcociente a partir de los elementos de la base $\mathcal{V}$. Esta representación es además muy útil cuando necesitamos analizar las relaciones métricas en un conjunto de composiciones, pues la relación

entre componentes de una composición en la base $\mathcal{V}$ son euclidianas si consideramos en $\mathcal{C}$ la estructura métrica asociada a la $\mathcal{C}$-distancia definida previamente. Pero el interés principal de una base ortonormal radica en que permite aplicar todos los resultados de la estadística sobre $\mathbb{R}^d$ a los coeficientes, dando una fundamentación matemática rigurosa al análisis estadístico de datos composicionales.

# 7 Conclusiones

La metodología desarrollada por Aitchison (1986) para el análisis estadístico de datos composicionales se basa esencialmente en el concepto de perturbación, y en las transformaciones logcociente centrada y logcociente aditiva. Hemos demostrado que estos conceptos y transformaciones no son arbitrarios. En efecto, desde una perspectiva matemática, vienen inducidos por la naturaleza de los datos composicionales si presuponemos que estos datos se caracterizan por su invariancia por cambios de escala. En consecuencia, la metodología propuesta por Aitchison (1986) no puede ser rechazada con argumentos matemáticos, pues es plenamente compatible con la naturaleza composicional de los datos y es, además, independiente de la representación utilizada para manejar los datos. Asimismo, el análisis de las subcomposiciones es totalmente coherente con el análisis de la composición completa.

Los detractores de esta metodología, que abogan por un análisis estándar de este tipo de datos, rechazan implícitamente el *cociente* como la forma natural de comparar dos composiciones o dos partes de la misma composición. Asimismo, aceptan implícitamente la *diferencia* usual como la forma lógica de realizar estas comparaciones. Este tipo de análisis depende completamente de la representación utilizada para manejar los datos y muchos de los resultados pueden inducir a error, tal y como señaló Pearson (1897) hace ya más de un siglo.

# 8 Agradecimientos

Este trabajo de investigación ha sido parcialmente financiado por la DGESIC (Ref.: BFM2000-0540) y por el Dept. de Informàtica i Matemàtica Aplicada de la Universitat de Girona (UdG).

# Referencias


Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). 416 p.

Aitchison, J. (1989a). Measures of location of compositional data sets. *Mathematical Geology 21*(7), 787–790.

Aitchison, J. (1989b). Reply to "Interpreting and testing compositional data" by A. Woronow, K. M. Love and J. C. Butler. *Mathematical Geology 21*(1), 65–71.

Aitchison, J. (1990a). Comment on "Measures of variability for geological data" by D. F. Watson and G. M. Philip. *Mathematical Geology 22*(2), 223–226.

Aitchison, J. (1990b). Relative variation diagrams for describing patterns of compositional variability. *Mathematical Geology 22*(4), 487–511.

Aitchison, J. (1991). Delusions of uniqueness and ineluctability. *Mathematical Geology 23*(2), 275–277.

Aitchison, J. (1992). On criteria for measures 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. See Pawlowsky-Glahn (1997), pp. 3–35.

Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology 131*(5), 563–580.

Aitchison, J. (2002). Simplicial inference. See Viana and Richards (2002), pp. 1–22.

Aitchison, J. and J. Bacon-Shone (1999). Convex linear combination of compositions. *Biometrika 86*(2), 351–364.

Aitchison, J., C. Barceló-Vidal, J. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Reply to Letter to the Editor by S. Rehder and U. Zier. *Mathematical Geology 33*(7), 849–860.

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. and M. Greenacre (2002). Biplots for compositional data. *Applied Statistics 51*(4), 375–392.

Aitchison, J. and C. W. Thomas (1998). Differential perturbation processes: a tool for the study of compositional processes. See Buccianti, Nardi, and Potenza (1998), pp. 499–504.

Barceló, C. and V. Pawlowsky (1994). Finite mixtures of compositional data. *Science de la Terre, Ser. inf. 32*, 29–48.

Barceló, C., V. Pawlowsky, and E. Grunsky (1995). Classification problems of samples of finite mixtures of compositions. *Mathematical Geology 27*(1), 129–148.

Barceló, C., V. Pawlowsky, and E. Grunsky (1996). Some aspects of transformations of compositional data and the identification of outliers. *Mathematical Geology 28*(4), 501–518.

Barceló-Vidal, C. (1996). *Mixturas de Datos Composicionales*. Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona (E). 261 p.

Barceló-Vidal, C. (2000). Fundamentación matemática del análisis de datos composicionales. Technical Report IMA 00-02-RR, Departament d'Informática i Matemática Aplicada, Universitat de Girona, Spain. 77 p.

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

Baxter, M. (1993). Comment on D. Tangri and R. V. S. Wright "Multivariate analysis of compositional data ...". *Archaeometry 35*(1), 112–115.

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.

Buccianti, A., G. Nardi, and R. Potenza (Eds.) (1998). *Proceedings of IAMG'98 — The fourth annual conference of the International Association for Mathematical Geology*, Volume I and II. De Frede Editore, Napoli (I). 969 p.

Kiers, H., J. Rasson, P. Groenen, and M. Shader (Eds.) (2000). *Studies in Classification, Data Analysis, and Knowledge Organization (Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000), University of Namur, Namur, 11-14 July*. Springer-Verlag, Berlin (D). 428 p.

Lippard, S. J., A. Næss, and R. Sinding-Larsen (Eds.) (1999). *Proceedings of IAMG'99 — The fifth annual conference of the International Association for Mathematical Geology*, Volume I and II. Tapir, Trondheim (N). 784 p.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, London (GB). 518 p.

Martín-Fernández, J. A. (2001). *Medidas de diferencia y clasificación no paramétrica de datos composicionales*. Ph. D. thesis.

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. See Pawlowsky-Glahn (1997), pp. 151–156.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998c). A critical approach to non-parametric classification of compositional data. See Rizzi, Vichi, and Bock (1998), pp. 49–56.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998a). Measures of difference for compositional data and hierarchical clustering methods. See Buccianti, Nardi, and Potenza (1998), pp. 526–531.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998b). Medida de diferencia de Kullback-Leibler entre datos composicionales. See SEIO (1998), pp. 291–292.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Zero replacement in compositional data sets. See Kiers, Rasson, Groenen, and Shader (2000), pp. 155–160.

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. See Lippard, Næss, and Sinding-Larsen (1999), pp. 211–216.

Mateu-Figueras, G., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998). Modeling compositional data with multivariate skew-normal distributions. See Buccianti, Nardi, and Potenza (1998), pp. 532–537.

Pawlowsky-Glahn, V. (Ed.) (1997). *Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology*, Volume I, II and addendum. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E). 1100 p.

Pawlowsky-Glahn, V. and C. Barceló-Vidal (1999). Confidence regions in ternary diagrams. *Terra Nostra (Schriften der Alfred-Wegener-Stiftung) (99)*(1), 37–47.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *SERRA 15*(5), 384–398.

Pawlowsky-Glahn, V. and J. J. Egozcue (2002). BLU estimators and compositional data. *Mathematical Geology 34*(3), 259–274.

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 measurement of organs. *Proceedings of the Royal Society of London LX*, 489–502.

Rehder, S. and U. Zier (2001). Letter to the Editor: Comment on "Logratio analysis and compositional distance" by J. Aitchison, C. Barceló-Vidal, J.A. Martín-Fernández and V. Pawlowsky-Glahn. *Mathematical Geology 33*(7), 845–848.

Rizzi, A., M. Vichi, and H.-H. Bock (Eds.) (1998). *Advances in Data Science and Classification (Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS'98), Università "La Sapienza", Rome, 21–24 July*. Springer-Verlag, Berlin (D). 677 p.

SEIO (1998). *Actas del XXIV Congreso Nacional de la Sociedad de Estadística e Investigación Operativa (SEIO)*. Sociedad Española de Estadística e Investigación Operativa, Almería (E).

Tangri, D. and V. Wright (1993). Multivariate analysis of compositional data: Applied comparisons favour standard principal components analysis over Aitchison's loglinear constrast method. *Archaeometry 35*(1), 103–115.

Tauber, F. (1999). Spurious clusters in granulometric data caused by logratio transformation. *Mathematical Geology 31*(5), 491–504.

Viana, M. A. G. and D. S. P. Richards (Eds.) (2002). *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics Series*. American Mathematical Society, Providence, Rhode Island (USA).

Watson, D. F. (1990). Reply to *Comment on "Measures of variability for geological data" by D.F. Watson and G.M. Philip. Mathematical Geology 22*(2), 227–231.

Watson, D. F. (1991). Reply to *"Delusions of uniqueness and ineluctability" by J. Aitchison. 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.

Whitten, E. H. T. (1995). Open and closed compositional data in petrology. *Mathematical Geology 27*(6), 789–806.

Woronow, A. (1997a). The elusive benefits of logratios. See Pawlowsky-Glahn (1997), pp. 97–101.

Woronow, A. (1997b). Regression and discrimination analysis using raw compositional data: Is it really a problem? See Pawlowsky-Glahn (1997), pp. 157–162.

Zier, U. and S. Rehder (1998). Grain-size analysis—a closed data problem. See Buccianti, Nardi, and Potenza (1998), pp. 555–558.