Otros espacios euclídeos

por

Juan José Egozcue, Vera Pawlowsky-Glahn y José Luis Díaz-Barrero

Resumen. Se presentan dos casos, $\mathbb{R}_+$ y el símplex $\mathcal{S}^n$, en que subconjuntos de un espacio euclídeo real se dotan de una estructura euclídea propia. Ambos casos corresponden a situaciones con aplicaciones prácticas en múltiples campos de las ciencias experimentales, ya que $\mathbb{R}_+$ corresponde a observaciones positivas y el símplex $\mathcal{S}^n$ a datos composicionales, como por ejemplo porcentajes o tantos por uno. El mérito de estas estructuras euclídeas reside en que las operaciones y la métrica propuestas son interpretables en la práctica.

1. Introducción

La acumulación de experiencias geométricas a lo largo de varios milenios ha permitido acuñar los modernos conceptos de espacio vectorial y espacio euclídeo. Los esfuerzos de formalización matemática realizados en el siglo XX han destilado lo esencial de la geometría del mundo que nos envuelve para dar lugar a las correspondientes definiciones y enunciar las propiedades principales de esos espacios.

La culminación de la formalización en el caso de los espacios euclídeos, o en un sentido más amplio, de los espacios de Hilbert, consiste en la identificación formal de todos los espacios con la misma dimensión. Todo espacio euclídeo de dimensión n es isomorfo a nuestro familiar $\mathbb{R}^n$. Desde el punto de vista formal todo se reduce al estudio del espacio real y sus propiedades pueden ser trasladadas automáticamente a cualquier otro espacio euclídeo.

Sin embargo, el esfuerzo formalizador de las matemáticas exige que la abstracción vaya acompañada de la interpretación de los resultados obtenidos en el proceso formal. El paso interpretativo también es objeto de las matemáticas. Las diferentes interpretaciones que pueden darse a un enunciado son precisamente su riqueza, porque devuelven a la experiencia el acerbo lógico asimilado y desarrollado en la formalización. De hecho, de poco sirve un enunciado formal del cual no se pueda dar más que un solo ejemplo de aplicación; el ejemplo bien pudiera sustituir al enunciado.

Estos comentarios vienen a incidir en el título de este artículo. Mientras la formalización nos llevaría a hablar de *El Espacio Euclídeo*, la interpretación en diferentes contextos nos permite hablar de diferentes espacios euclídeos, en especial aquellos *otros* que no son $\mathbb{R}^n$.

Una vez más nos referimos a la obviedad matemática que nos permite construir espacios euclídeos a partir de una biyección con $\mathbb{R}^n$. Consideremos un conjunto arbitrario, $\mathcal{S}$ y supongamos que φ es una aplicación biyectiva de $\mathcal{S}$ en $\mathbb{R}^n$. Resulta

inmediato definir operaciones y métrica en $\mathcal{S}$ de forma que este último quede constituido como espacio euclídeo. Basta definir la operación de grupo o suma, $\oplus$, y el producto por escalares reales, $\odot$, en $\mathcal{S}$, mediante

$$\mathbf{x} \oplus \mathbf{y} = \varphi^{-1}(\varphi(\mathbf{x}) + \varphi(\mathbf{y}))\,; \quad \alpha \odot \mathbf{x} = \varphi^{-1}(\alpha \cdot \varphi(\mathbf{x})),$$

donde $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ y $\alpha \in \mathbb{R}$. Esto estructura $\mathcal{S}$ como de espacio vectorial de dimensión n. De forma similar se puede exportar el producto escalar en $\mathbb{R}^n$ a $\mathcal{S}$ definiendo

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{S}} = \langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle,$$

donde $\langle \cdot, \cdot \rangle$ denota productos escalares, con subíndice $\mathcal{S}$ en ese espacio, y sin subíndice para el ordinario en $\mathbb{R}^n$. Estas definiciones se extienden a los demás conceptos usuales en espacios euclídeos: norma, distancia, medida. Con ello $\mathcal{S}$ tiene estructura de espacio euclídeo de dimensión n.

Una colección suficientemente extensa de aplicaciones biyectivas de conjuntos, elegidos con algunas precauciones, en $\mathbb{R}^n$, nos permitirían hablar de una familia pintoresca de espacios euclídeos. Sin embargo, el objetivo que nos ocupa consiste en destacar algunos casos (pocos), en que el espacio $\mathcal{S}$, las operaciones y la métrica inducidas mediante la biyección φ, tienen una interpretación concreta en la práctica. Este requerimiento es muy exigente, a pesar de su aparente laxitud. De hecho sólo nos referiremos a dos casos: $\mathbb{R}_+$, los reales positivos, y a $\mathcal{S}^n$, el símplex de n partes.

La mayor parte de los ejemplos de espacios euclídeos que admiten interpretaciones que difieran sustancialmente de las dadas en el propio $\mathbb{R}^n$ proceden de espacios funcionales de dimensión infinita. Son destacables los polinomios de grado n, con productos escalares definidos sobre un intervalo o sobre la circunferencia unidad, y el espacio generado por n variables aleatorias, tomando la covarianza como producto escalar. Menos habituales son los ejemplos en que un subconjunto de $\mathbb{R}$ o $\mathbb{R}^n$ se constituye en espacio euclídeo, sin que esa estructura sea un mero artificio. El ejemplo más simple se refiere a $\mathbb{R}_+$; véase [9]. Aunque el caso es perfectamente conocido, no es fácil hallarlo más que, como ejemplo, en unos apuntes de clase de álgebra o análisis. En la misma línea, aunque con una complejidad interpretativa mayor, el símplex también puede estructurarse como espacio euclídeo, con una interpretación perfectamente adaptada a situaciones reales y, además, frecuentes.

Antes de abordar el desarrollo en el símplex, y aún a modo de introducción, discutimos brevemente el caso de $\mathbb{R}_+$ en la próxima sección. La sección 3 se dedica a la descripción detallada de la estructura euclídea del símplex y se discute la interpretación que puede darse a la misma, con especial atención a la modelización de observaciones composicionales, como por ejemplo porcentajes o tantos por uno.

2. La semirrecta positiva como espacio euclídeo

Definimos $\mathbb{R}_+$ como el conjunto de números reales estrictamente positivos. Se considera la biyección $\varphi : \mathbb{R}_+ \to \mathbb{R}$ definida por $\varphi(x) = \ln(x)$. Utilizando esta

biyección se obtienen las siguientes definiciones, en las que x, $y \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$:

$$x \oplus y = x \cdot y\,; \quad \alpha \odot x = x^\alpha\,;$$
$$d_+(x,y) = |\ln x - \ln y|\,; \quad \|x\|_+ = |\ln x|\,; \quad \langle x, y\rangle_+ = \ln x \cdot \ln y\,,$$

donde la distancia, norma y producto escalar en $\mathbb{R}_+$ han sido subindicados con $+$.

Una propiedad que caracteriza los espacios euclídeos es la invariancia de la distancia frente a traslaciones y el escalado frente a la operación por escalares. En $\mathbb{R}_+$ la traslación toma la forma $y = x \oplus z$, donde y es el trasladado de x por la traslación $z \in \mathbb{R}_+$. Las propiedades mencionadas son

$$d_+(x \oplus z, y \oplus z) = d_+(x,y)\,; \quad d_+(\alpha \odot x, \alpha \odot y) = |\alpha|\, d_+(x,y)\,,$$

que se traducen a

$$|\ln(xz) - \ln(yz)| = |\ln x - \ln y|\,; \quad |\ln(x^\alpha) - \ln(y^\alpha)| = |\alpha| \cdot |\ln x - \ln y|$$

en la recta real.

Como $\mathbb{R}_+$ es de dimensión 1, una base la constituye un elemento de $\mathbb{R}_+$. Si se desea que sea de norma unidad, las posibilidades se reducen a e y e^{-1}. Si se adopta e, $\|e\|_+ = 1$, como base, la expresión $x = e^{\ln x}$ se debe interpretar como $x = \ln x \odot e$, es decir, $\ln x$ es la coordenada de x en la base e.

Aunque prescindible en la estructura euclídea, es lógico definir una medida de referencia en $\mathbb{R}_+$. La medida de referencia es necesaria para desarrollar una integración en el espacio. En particular, se requiere para un cálculo efectivo de probabilidades. En $\mathbb{R}$ se define como medida de referencia la de Lebesgue. La medida de Lebesgue en $\mathbb{R}$ asigna a cada intervalo (a,b) su longitud $d(a,b) = |b-a|$ y se escribe $\lambda\{(a,b)\} = |b-a|$. Esta asignación de medida a intervalos se puede extender a cualquier conjunto boreliano. En $\mathbb{R}_+$ se puede actuar de forma análoga. Así, definimos la medida en $\mathbb{R}_+$ como aquélla que asigna a un intervalo abierto en $\mathbb{R}_+$, (a,b), su medida $\lambda_+\{(a,b)\} = d_+(a,b) = |\ln b - \ln a|$; véase [8].

Como se ha mencionado, el valor de esta estructura euclídea de $\mathbb{R}_+$ depende de la interpretación que pueda darse de las operaciones y estructura métrica. La distancia en $\mathbb{R}_+$ es uno de los elementos más directamente interpretables. El siguiente ejemplo puede motivar la adopción de la estructura euclídea de $\mathbb{R}_+$ como base para la modelización de observaciones positivas.

En ingeniería marítima se describe el oleaje mediante la llamada *altura de ola significante* (h_s). Se trata del promedio de las alturas de olas individuales que superan el cuantil 2/3. La altura de ola individual se define como la distancia entre cresta y valle consecutivos. Obviamente h_s es un valor positivo. En primer lugar se observa que $h_s = 0$ correspondería a la situación físicamente imposible de un mar sin oscilación alguna. A continuación intentamos valorar la distancia entre dos mares que tienen, respectivamente, h_s igual a 0.1 m y 0.2 m (quizá en el interior de un puerto). La respuesta que esperamos es que 0.2 m es el doble que 0.1 m y que la agitación dentro de los dos supuestos puertos debe diferir bastante. Sin embargo, si se pregunta sobre la distancia entre un gran temporal con $h_s = 10.0$ m y otro con

$h_s = 10.1\,\mathrm{m}$, la respuesta inmediata es que la diferencia es casi inapreciable y, desde luego, intuitivamente mucho menor que en la situación citada antes. Se concluye que la distancia ordinaria en $\mathbb{R}$, que es la misma en ambas situaciones, representa mal la apreciación subjetiva de h_s. En concreto, esa distancia asigna una distancia finita y pequeña desde 0 hasta 0.1, mientras partimos del hecho que $h_s = 0$ corresponde a un suceso imposible. Contrariamente, la distancia logarítmica $d_+(\cdot,\cdot)$ refleja mucho mejor la situación física a la que nos referimos. Por ello, los errores de medida en esos contextos suelen expresarse como *error relativo*: ciertamente la diferencia en $\mathbb{R}_+$ es el logaritmo del cociente.

Así se observa que tanto la operación suma como la distancia en $\mathbb{R}_+$ se adaptan a las apreciaciones físicas y subjetivas mejor que las correspondientes en $\mathbb{R}$. La situación comentada se repite con bastante frecuencia cuando se trata con medidas positivas como la temperatura absoluta, la precipitación, el caudal de un río, etc.

3. El símplex

Suele definirse el símplex de n partes como el conjunto de vectores reales positivos cuya suma es constante. Aparece en contextos muy diversos, desde teoría de la probabilidad hasta problemas de optimización o ecuaciones diferenciales y sistemas dinámicos. El desarrollo y notación que se emplea aquí proviene del análisis estadístico de datos composicionales [1, 9].

Formalmente, el símplex de n partes se define como

$$\mathcal{S}^n = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^n \;\middle|\; x_i > 0,\ i = 1, \ldots, n;\ \sum_{i=1}^{n} x_i = 1 \right\}.$$

En otros contextos, el superíndice n (n componentes o partes) hubiera sido sustituido por $n-1$ (la dimensión de la variedad o grados de libertad). El corchete en el que se indican las componentes de $\mathbf{x}$ indica que se trata de un vector fila; este hecho no tiene ninguna importancia, pero corresponde a la práctica usual en estadística de representar individuos por filas y muestras por columnas. Las condiciones $x_i > 0$ se sustituyen a veces por $x_i \geq 0$. En el presente desarrollo, los elementos en que alguna componente es nula corresponden a los puntos del infinito, por lo que excluirlos de la definición facilita el desarrollo. Finalmente, se ha supuesto que la suma de componentes es 1. No hay inconveniente alguno en suponer cualquier otro valor positivo.

Existen multitud de ejemplos en que aparecen vectores positivos de suma constante. En casi todas las ciencias experimentales (química, geología, biología, física) intervienen datos expresados en porcentajes o en tantos por uno; en algunos casos esos datos, llamados composicionales, están en formas disimuladas, como en química cuando se expresan las concentraciones como molaridades o partes por millón. Pero también en matemáticas aparecen frecuentemente restricciones de suma constante. Un ejemplo obvio es el análisis convexo.

La definición de una operación clausura facilita las expresiones en componentes. Consiste en dividir las componentes de un vector positivo por la suma de todas ellas,

para reducirlas a suma unitaria:

$$\mathcal{C}[x_1, x_2, \ldots, x_n] = \left[\frac{x_1}{\sum x_i}, \frac{x_2}{\sum x_i}, \ldots, \frac{x_n}{\sum x_i}\right],$$

donde se supone que $x_i > 0$, $i = 1, \ldots, n$.

En el caso $\mathcal{S}^2$ es fácil definir biyecciones con $\mathbb{R}$. En particular, la transformación logit permite inducir la estructura de espacio euclídeo. Pero la extensión a dimensiones mayores que 1 tiene dificultades. El desarrollo de la teoría en el símplex a lo largo de tres décadas no ha procedido por la vía de definir toda la estructura euclídea por medio de biyecciones con $\mathbb{R}^n$. De hecho, el proceso se inicia definiendo las operaciones de espacio vectorial mediante una biyección, pero no así la distancia, que se justifica mediante razonamientos heurísticos [1]. Por ello, Aitchison se ve obligado a trabajar con dos transformaciones, denominadas respectivamente alr (additive log-ratio) y clr (centred log-ratio), del símplex $\mathcal{S}^n$ en $\mathbb{R}^{n-1}$, respectivamente $\mathbb{R}^n$. El motivo subyacente es que la primera, siendo biyectiva, no es isometría, mientras que la segunda, conservando la distancia, no es biyección. Sólo a finales de la década de los 90 [9, 4, 3] se completa la estructura euclídea del símplex compatible con la distancia introducida por J. Aitchison [1]. Las biyecciones que inducen la estructura euclídea se publicaron poco más tarde [7]. El hecho de que las definiciones de operación de grupo (llamada perturbación) y la distancia en el símplex (o de Aitchison) precedieran a la estructura de espacio euclídeo es reflejo de la motivación experimental de ambas definiciones independientemente de la estructura que inducen. Finalmente, la estructura euclídea del símplex ha sido extendida al espacio de densidades de medidas positivas cuyo logaritmo es de cuadrado integrable [6], dando lugar a un espacio de Hilbert.

Una operación frecuente es la extracción de una subcomposición. El término proviene del análisis de datos composicionales. A los elementos del símplex se les llama composiciones. Si se descartan componentes del vector composición $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, reteniendo solamente algunas de ellas, por ejemplo, las r primeras, se puede construir el vector $\text{sub}\{\mathbf{x}; R\} = \mathcal{C}[x_1, x_2, \ldots, x_r]$, que es un elemento de $\mathcal{S}^r$ y que se llama subcomposición con índices en el conjunto $R = \{1, 2, \ldots, r\}$.

Espacio vectorial

Como se ha comentado, la operación de grupo conmutativo que definimos en el símplex se llama perturbación. Si $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ son elementos de $\mathcal{S}^n$ su perturbación se define por

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \ldots, x_n y_n].$$

Es fácil comprobar las propiedades de grupo conmutativo. El elemento neutro es

$$\mathbf{n} = \mathcal{C}[1, 1, \ldots, 1] = [n^{-1}, n^{-1}, \ldots, n^{-1}].$$

La operación externa con escalares reales es la potenciación, que se define por

$$\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, x_2^\alpha, \ldots, x_n^\alpha],$$

donde $\alpha \in \mathbb{R}$. El elemento unitario es 1. Para denotar a los elementos opuestos por $\oplus$ se utiliza el signo $\ominus$, es decir,

$$\ominus \mathbf{x} = -1 \odot \mathbf{x} = \mathcal{C}[x_1^{-1}, x_2^{-1}, \ldots, x_n^{-1}].$$

Comprobar todas las propiedades de espacio vectorial para estas operaciones se reduce a un mero ejercicio. En resumen, tenemos que $(\mathcal{S}^n, \oplus, \odot)$ es un espacio vectorial sobre $\mathbb{R}$. La dimensión es $n-1$, lo que quedará justificado cuando se hallen bases del espacio.

La estructura de espacio vectorial en el símplex permite definir variedades lineales. Llamaremos rectas a las variedades lineales de dimensión 1. La ecuación de la recta que pasa por el punto $\mathbf{x}_0 \in \mathcal{S}^n$ y con la dirección de $\mathbf{c} \in \mathcal{S}^n$ es

$$\mathbf{x}(t) = \mathbf{x}_0 \oplus (t \odot \mathbf{c}), \quad t \in \mathbb{R}. \tag{1}$$

Las operaciones de perturbación y potenciación definidas anteriormente quedan justificadas si podemos dar una interpretación de las rectas en el símplex. El crecimiento de masa bacteriana sin interacción proporciona un buen ejemplo. También pueden hallarse ejemplos en física o geología (desintegración radiactiva, deposición de materiales en suspensión, etc.). Supongamos pues que n especies de bacterias crecen sin interacción en un medio con nutrientes ilimitados. Supongamos que $\mathbf{z}(t)$ es el vector cuyas componentes son las masas de cada especie en el instante t. El crecimiento de masa se supone proporcional a la masa presente en cada instante. Las ecuaciones diferenciales correspondientes se pueden agrupar en un sistema de ecuaciones que podemos representar por $d\mathbf{z}(t)/dt = \boldsymbol{\lambda}\,\mathbf{z}(t)$, donde el producto de vectores $\boldsymbol{\lambda}\,\mathbf{z}(t)$ se interpreta como producto componente a componente y el vector $\boldsymbol{\lambda}$ contiene las constantes de crecimiento de cada una de las especies. Estas constantes son reales (en este caso positivas) y dependen de las características de cada especie. Las soluciones representan un crecimiento exponencial de masa

$$\mathbf{z}(t) = \mathbf{z}(0)\, \exp(\boldsymbol{\lambda} t). \tag{2}$$

Para centrar nuestra atención sobre los tantos por uno de masa de cada especie, bastará dividir las masas por el total en cada momento. El modelo de crecimiento exponencial de masas (2) se reduce entonces a la expresión composicional

$$\mathcal{C}[\mathbf{z}(t)] = \mathcal{C}[\mathbf{z}(0)\, \exp(\boldsymbol{\lambda} t)] = \mathcal{C}[\mathbf{z}(0)] \oplus (t \odot \mathcal{C}[\exp(\boldsymbol{\lambda})])\,,$$

en donde se identifica claramente una recta composicional. Es decir, los procesos de crecimiento o decrecimiento exponencial de masa, examinados desde el punto de vista composicional, evolucionan sobre rectas en el símplex.

Aparte de la perturbación, hay otras operaciones en el símplex que admiten interpretaciones inmediatas. Quizá la más importante es la combinación lineal convexa de elementos del símplex, que corresponde a la mezcla de dos o más composiciones. Sean $\mathbf{x}_i \in \mathcal{S}^n$, $i = 1, 2, \ldots, k$, y sea $\mathbf{y} \in \mathcal{S}^k$. Se define la mezcla, $\mathbf{z} \in \mathcal{S}^n$, de las x_i's según $\mathbf{y}$, como

$$\mathbf{z} = \sum_{i=1}^{k} y_i \cdot \mathbf{x}_i\,.$$

Esta operación, aun fijado el valor de $\mathbf{y}$, no tiene carácter lineal en el espacio vectorial de $\mathcal{S}^n$ definido previamente. Este hecho indica que deben tomarse precauciones cuando se utilizan operaciones de $\mathbb{R}^n$ entre vectores de $\mathcal{S}^n$, a pesar del carácter intuitivo de la operación mezcla.

Métrica

En el desarrollo de la estructura euclídea de $\mathcal{S}^n$, se introdujo la distancia, llamada de Aitchison, antes que el producto escalar [1]. Desde el punto de vista formal resulta más directo definir el producto escalar [9, 4, 3] y, a partir de él, obtener la norma y la distancia asociadas.

Si $\mathbf{x}$, $\mathbf{y} \in \mathcal{S}^n$, se define el producto escalar de Aitchison por

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \ln \frac{x_i}{x_j} \cdot \ln \frac{y_i}{y_j} = \sum_{i=1}^{n} \ln \frac{x_i}{g(\mathbf{x})} \cdot \ln \frac{y_i}{g(\mathbf{y})} \,, \tag{3}$$

donde $g(\cdot)$ denota la media geométrica de las componentes del vector en el argumento. También se puede expresar (3) como una forma cuadrática en $\mathbb{R}^n$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{n} (\ln \mathbf{x})\; M\; (\ln \mathbf{y})' \,, \quad M = \operatorname{diag}(n, n, \ldots, n) - \mathbf{1} \,, \tag{4}$$

donde los logaritmos se aplican componente a componente, $\mathbf{1}$ es la matriz $(n \times n)$ con todos sus elementos unitarios, y $(\cdot)'$ denota trasposición de un vector fila. La expresión (4) del producto escalar de Aitchison como forma cuadrática degenerada permite comprender que se cumplen todos los requisitos.

Con la definición del producto escalar, una vez comprobadas sus propiedades, se completa la estructura de espacio euclídeo de $\mathcal{S}^n$ y estamos en condiciones de desarrollar toda la geometría que le corresponde. La geometría euclídea basada en la perturbación como grupo y el producto escalar (3) la llamaremos geometría de Aitchison, para distinguirla de la euclídea ordinaria que le corresponde a $\mathcal{S}^n$ como subconjunto de $\mathbb{R}^n$.

El producto escalar permite de forma inmediata definir tanto la norma como la distancia de Aitchison asociadas,

$$\|\mathbf{x}\|_a = \big(\langle \mathbf{x}, \mathbf{x} \rangle_a\big)^{1/2}, \quad d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a \,.$$

A pesar de la sencillez de estas últimas definiciones, es conveniente dar expresiones más explícitas de la distancia de Aitchison, más utilizadas en la práctica:

$$d_a^2(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2 = \sum_{i=1}^{n} \left(\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{y_i}{g(\mathbf{y})} \right)^2 .$$

Las propiedades de la distancia y de la norma se satisfacen automáticamente porque se han construido a partir del producto escalar. Sin embargo, cabe mencionar las propiedades de la distancia de Aitchison que, antes de que se estableciese el producto escalar, avalaron su adopción [2]. Estas son:

1. Sólo depende de los cocientes entre las distintas componentes de cada uno de los vectores del símplex. Corresponde a la idea de que la información aportada por una composición es, precisamente, esos cocientes.
2. Es invariante por perturbación. Es decir,

$$d_a(\mathbf{x} \oplus \mathbf{z}, \mathbf{y} \oplus \mathbf{z}) = d_a(\mathbf{x}, \mathbf{y}).$$

3. Garantiza la dominancia subcomposicional. Corresponde a la propiedad de que la distancia entre $\mathbf{x}$, $\mathbf{y} \in \mathcal{S}^n$ es mayor que entre sus respectivas subcomposiciones respecto a un conjunto de índices R, es decir,

$$d_a(\mathbf{x}, \mathbf{y}) \geq d_a(\text{sub}\{\mathbf{x}; R\}, \text{sub}\{\mathbf{y}; R\}).$$

4. Es invariante frente a permutaciones de las componentes de los vectores composicionales.

La definición del producto escalar (3) permite establecer bases ortonormales del símplex. Una diferencia sustancial, desde el punto de vista interpretativo, de la geometría del símplex respecto a la de $\mathbb{R}^n$, es la dificultad para elegir una base canónica. Una de las bases ortonormales más sencillas de $\mathcal{S}^n$ es

$$\mathbf{e}_i = \mathcal{C}\left[\exp\left[\underbrace{\sqrt{\frac{1}{i(i+1)}}, \ldots, \sqrt{\frac{1}{i(i+1)}}}_{i \text{ elementos}}, -\sqrt{\frac{i}{i+1}}, 0, \ldots, 0\right]\right], \; i = 1, 2, \ldots, n-1, \tag{5}$$

y las coordenadas de $\mathbf{x} \in \mathcal{S}^n$ respecto a ella son

$$x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a = \sqrt{\frac{i}{i+1}} \ln\left(\frac{g(x_1, \ldots, x_i)}{x_{i+1}}\right), \; i = 1, 2, \ldots, n-1, \tag{6}$$

donde nuevamente $g(\cdot)$ denota la media geométrica de los argumentos.

Una vez definida una base ortonormal como (5), la función que asigna a $\mathbf{x}$ sus coordenadas $\mathbf{x}^* = [x_1^*, x_2^*, \ldots, x_{n-1}^*]$ según (6) es una biyección $\boldsymbol{\varphi} : \mathcal{S}^n \to \mathbb{R}^{n-1}$. Además se puede demostrar que es una isometría, es decir, el producto escalar, norma y distancia ordinarias en $\mathbb{R}^{n-1}$ inducen, a través de $\boldsymbol{\varphi}^{-1}$, los productos escalares, normas y distancias de Aitchison en $\mathcal{S}^n$. Por esta razón, $\boldsymbol{\varphi}$ fue llamada *isometric log-ratio transformation* [7]. Pero debe destacarse que cada base ortonormal define una nueva isometría con las mismas propiedades.

Las coordenadas, $\mathbf{x}^*$, en una base ortonormal como (5), permiten obtener todos los elementos de la geometría de Aitchison en el símplex $\mathcal{S}^n$ a partir de los correspondientes de la geometría euclídea ordinaria en $\mathbb{R}^{n-1}$. Las siguientes figuras nos permiten ilustrar algunos de ellos. A efectos de representación nos restringimos al símplex de 3 partes, $\mathcal{S}^3$, que habitualmente se representa en un diagrama ternario; el diagrama ternario consiste en representar $[x_1, x_2, x_3] \in \mathcal{S}^3$ por un punto en el interior de un triángulo equilátero, de forma que las componentes x_i son proporcionales

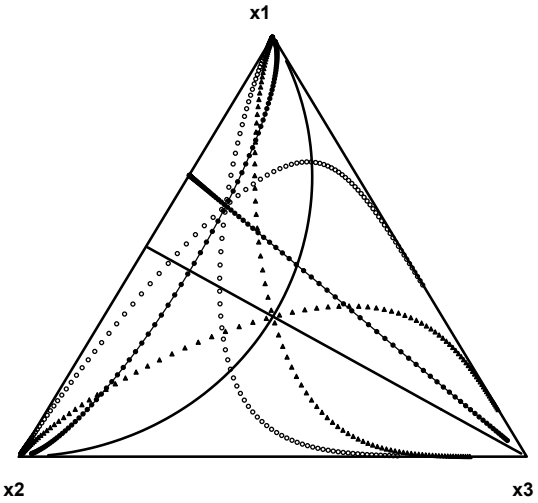

Figura 1: Representación de dos haces de rectas en el diagrama ternario que pasan por los puntos $[1/3, 1/3, 1/3]$ y $[0.60, 0.29, 0.11]$.

a las distancias a cada uno de los lados. Como se sabe, la suma de esas distancias es constante. En la figura 1 se han representado dos haces de cuatro rectas cada uno. El primero de estos haces lo constituyen los dos ejes ortogonales de la base (5), en línea continua, y sus dos bisectrices, con símbolos triangulares. El segundo haz es un desplazamiento paralelo del primero, de forma que el origen (0,0) se ha desplazado al punto de coordenadas $(0.5, 1.1)$. Las rectas paralelas a los ejes se han indicado con círculos rellenos sobre línea continua, mientras que sus correspondientes bisectrices se han indicado mediante marcadores circulares vacíos. El origen de coordenadas $(0, 0)$ corresponde, en el símplex de tres partes, a $[1/3, 1/3, 1/3]$ y el punto de coordenadas $(0.5, 1.1)$ a $[0.60, 0.29, 0.11]$ aproximadamente. La figura 2 representa diversos elementos geométricos en el diagrama ternario (izquierda) y en el plano de coordenadas (derecha). Cada elemento se ha caracterizado con los mismos símbolos en ambas representaciones.

Medida, integración y derivación

La definición de una medida de referencia en un espacio euclídeo no es un elemento esencial de su estructura. Sin embargo, es importante para establecer una integración acorde con la métrica del espacio. En un caso como el símplex $\mathcal{S}^n$, cuyos elementos constituyen un subconjunto de $\mathbb{R}^n$, se podría mantener, como referencia, la medida de Lebesgue en $\mathbb{R}^n$. Pero la forma natural de definir una medida de referencia es representar los elementos de $\mathcal{S}^n$ por sus coordenadas en alguna base ortonormal e inducir la medida a partir de la medida de Lebesgue en el espacio de las coordenadas que es $\mathbb{R}^{n-1}$. Denotamos nuevamente la función que asigna a cada elemento de $\mathcal{S}^n$ sus coordenadas en $\mathbb{R}^{n-1}$ por $\boldsymbol{\varphi}$. Supongamos que $B^* \subset \mathbb{R}^{n-1}$ es un conjunto

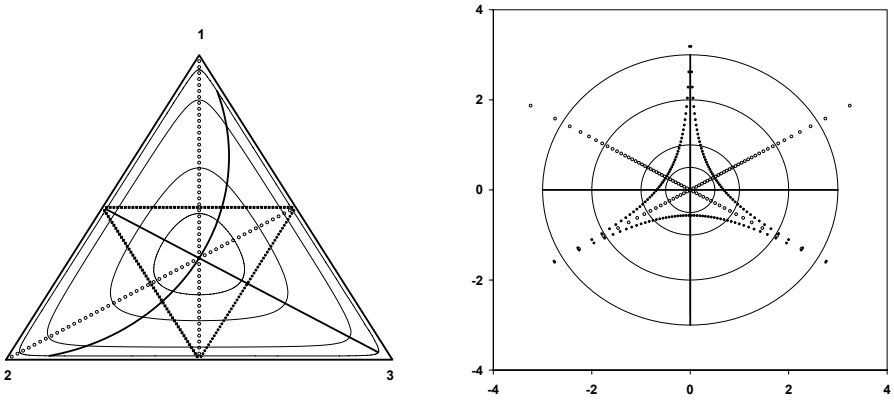

Figura 2: Izquierda: $\mathcal{S}^3$, diagrama ternario. Derecha: espacio de coordenadas $\mathbb{R}^2$.

boreliano con medida de Lebesgue $\lambda(B^*)$. En primer lugar, es inmediato comprobar que los conjuntos anti-imagen $B = \boldsymbol{\varphi}^{-1}(B^*)$ constituyen una sigma-álgebra de $\mathcal{S}^n$. A continuación, basta definir la medida de referencia en $\mathcal{S}^n$ como $\mu(B) = \lambda(B^*)$. Llamaremos a μ medida de Aitchison en el símplex [8].

La integración de funciones reales definidas en el símplex respecto de μ es una integral ordinaria. Si $h : \mathcal{S}^n \to \mathbb{R}$ y $\mathbf{x} \in \mathcal{S}^n$, se tiene

$$\int h(\mathbf{x})\, d\mu = \int h(\boldsymbol{\varphi}^{-1}(\mathbf{x}^*))\, \frac{d\mu}{d\lambda}\, d\lambda,$$

donde $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{x}^*$ son las coordenadas de $\mathbf{x}$, λ denota la medida de Lebesgue en $\mathbb{R}^{n-1}$ y $d\mu/d\lambda$ es la densidad de μ respecto a λ en $\mathbb{R}^{n-1}$. Esta densidad es el inverso del Jacobiano de $\boldsymbol{\varphi}$, cuyo valor es

$$J_{\boldsymbol{\varphi}}(\mathbf{x}) = \frac{1}{\sqrt{n}} \cdot \frac{1}{x_1 \cdot x_2 \cdots x_n}\,,$$

y por tanto

$$\frac{d\mu}{d\lambda} = \sqrt{n} \prod_{i=1}^{n} \varphi_i^{-1}(\mathbf{x}^*)\,,$$

donde las φ_i^{-1} denotan las componentes de la función inversa de $\boldsymbol{\varphi}$.

Un ejemplo en que aparecen integraciones de este tipo se presenta en el cálculo de probabilidades sobre el símplex. Supongamos que $\mathbf{X}$ es una variable aleatoria sobre el símplex; por ejemplo, una composición química aleatoria de un mineral. La densidad de probabilidad de $\mathbf{X}$ se define respecto a la medida de referencia μ, es decir, $f_{\mathbf{X}} = d\mathrm{P}/d\mu$, donde P denota la medida de probabilidad sobre el símplex. La probabilidad de un suceso $B \subset \mathcal{S}^n$ es, entonces,

$$\mathrm{P}[\mathbf{X} \in B] = \int_B f_{\mathbf{X}}(\mathbf{x})\, d\mu = \int_{B^*} f_{\mathbf{X}}(\boldsymbol{\varphi}^{-1}(\mathbf{x}^*))\, \sqrt{n} \prod_{i=1}^{n} \varphi_i^{-1}(\mathbf{x}^*)\, d\mathbf{x}^*\,,$$

donde nuevamente $\mathbf{x}^* = \boldsymbol{\varphi}(\mathbf{x})$, $B^* = \boldsymbol{\varphi}(B)$ y $d\mathbf{x}^*$ denota el elemento de medida (Lebesgue) en $\mathbb{R}^{n-1}$, que es el espacio de coordenadas. Una revisión de la expresión anterior nos lleva a identificar el último integrando con la densidad de probabilidad definida en el espacio de coordenadas respecto a λ,

$$\frac{d\mathrm{P}}{d\lambda} = f^*_{\mathbf{X}}(\mathbf{x}^*) = f_{\mathbf{X}}(\boldsymbol{\varphi}^{-1}(\mathbf{x}^*))\,\sqrt{n}\prod_{i=1}^{n}\varphi_i^{-1}(\mathbf{x}^*)\,,$$

y, por consiguiente,

$$\mathrm{P}[\mathbf{X}\in B] = \int_{B^*} f^*_{\mathbf{X}}(\mathbf{x}^*)\,d\mathbf{x}^*,$$

que es la expresión de la probabilidad si hubiéramos optado por caracterizar la variable aleatoria $\mathbf{X}$ mediante la densidad de probabilidad de sus coordenadas.

La integración de funciones con imágenes en el símplex implica nuevos elementos. Supongamos que $\mathbf{h} : \mathbb{R} \to \mathcal{S}^n$. Es lógico definir la integral de $\mathbf{h}$ como un elemento de $\mathcal{S}^n$ y, la operación subyacente a la misma, la perturbación, que juega el papel de la suma en $\mathcal{S}^n$. Si mantenemos la notación $\boldsymbol{\varphi}$ para la aplicación que da las coordenadas de los elementos del símplex, definimos la integral

$$\int^{\oplus} \mathbf{h}(t)\,dt = \boldsymbol{\varphi}^{-1}\left(\int \boldsymbol{\varphi}(\mathbf{h}(t))\,dt\right), \tag{7}$$

donde el superíndice de la integral indica el carácter de esa integral. Por ejemplo, las sumas de Riemann de $\int^{\oplus}$ toman la forma

$$\int^{\oplus} \mathbf{h}(t)\,dt \approx \bigoplus_i (t_{i+1} - t_i) \odot \mathbf{h}(t_i'),$$

donde los puntos t_i generan una partición del intervalo de integración y t_i' indica un punto del intervalo $(t_i, t_{i+1}]$.

Un ejemplo de este tipo de integración se presenta en el cálculo de valores medios. Por ejemplo, el valor medio de una composición química, $\mathbf{h}(t)$, que evoluciona con el tiempo t en un intervalo T, de longitud T, sería

$$\bar{\mathbf{h}} = \frac{1}{T} \odot \int_{\mathrm{T}}^{\oplus} \mathbf{h}(t)\,dt = \boldsymbol{\varphi}^{-1}\left(\frac{1}{T}\cdot\int_{\mathrm{T}} \mathbf{h}^*(t)\,dt\right), \tag{8}$$

donde $\mathbf{h}^*(t) = \boldsymbol{\varphi}(\mathbf{h}(t))$ son las coordenadas de $\mathbf{h}(t)$. Este valor medio contrasta con el valor medio que se hubiera tomado en el caso de ignorar la estructura euclídea del símplex, cuyo valor sería

$$\frac{1}{T}\cdot\int_{\mathrm{T}} \mathbf{h}(t)\,dt. \tag{9}$$

La elección de uno u otro valor medio parece una cuestión de preferencias. Sin embargo, se puede comprobar que el primer valor medio (8) es una operación lineal en el símplex, es decir, el valor medio de $(\alpha \odot \mathbf{h}(t)) \oplus \mathbf{z}$ es $(\alpha \odot \bar{\mathbf{h}}) \oplus \mathbf{z}$. Esto no es cierto para el valor medio (9).

La expresión de la integral (7) puede simplificarse, sustituyendo las funciones coordenadas $\boldsymbol{\varphi}$, de la forma

$$\int^{\oplus} \mathbf{h}(t)\, dt = \mathcal{C}\left[\exp\left(\int \ln(\mathbf{h}(t))\, dt\right)\right],$$

donde se supone que las funciones ln y exp se aplican a cada componente del vector del argumento.

La derivación de funciones $\mathbf{h} : \mathbb{R} \to \mathcal{S}^n$ puede definirse mediante el límite,

$$D^{\oplus}\mathbf{h}(t) = \lim_{\tau \to 0} \frac{1}{\tau} \odot (\mathbf{h}(t+\tau) \ominus \mathbf{h}(t)),$$

donde la notación para la derivada $D^{\oplus}$ se ha adoptado por coherencia con la de la integral. Como en el caso de la integración, puede comprobarse que la definición corresponde a representar $\mathbf{h}(t)$ por sus coordenadas $\mathbf{h}^*(t) = \boldsymbol{\varphi}(\mathbf{h}(t))$, realizar la derivación ordinaria de las mismas, y recuperar la función derivada en el símplex mediante $\boldsymbol{\varphi}^{-1}$. Análogamente a la integración, la función de coordenadas puede sustituirse por una trasformación logarítmica más simple. Estos resultados se resumen en la expresión

$$D^{\oplus}\mathbf{h}(t) = \boldsymbol{\varphi}^{-1}\left(\frac{d}{dt}\, \boldsymbol{\varphi}(\mathbf{h}(t))\right) = \mathcal{C}\left[\exp\left(\frac{d}{dt}\, \ln(\mathbf{h}(t))\right)\right].$$

Las derivadas elementales de funciones reales tienen sus homólogas en el símplex. Por ejemplo, es fácil comprobar que $D^{\oplus}\mathbf{c} = \mathbf{n}$ y que $D^{\oplus}(t \odot \mathbf{c}) = \mathbf{c}$, donde $\mathbf{c} \in \mathcal{S}^n$ y $\mathbf{n}$ es el elemento neutro. Esto permite resolver las ecuaciones diferenciales más simples en $\mathcal{S}^n$. Por ejemplo, la ecuación diferencial $D^{\oplus}\mathbf{x}(t) = \mathbf{c}$, donde $\mathbf{c} \in \mathcal{S}^n$ tiene por solución la expresión (1); es decir, las soluciones son rectas en el símplex.

3.1. Elementos interpretativos adicionales

Se ha indicado que el proceso de crecimiento (o decrecimiento) exponencial de masa (2), una vez clausurado como elemento del símplex, evoluciona siguiendo una recta del símplex. El proceso de crecimiento de la masa puede descomponerse en varias componentes de forma natural que, desde la perspectiva del símplex, corresponden a direcciones ortogonales. Esto indicará una forma natural de definir bases ortonormales en el símplex ligadas a la interpretación del problema que se plantea. Se sigue el desarrollo presentado en [5].

Retomamos el ejemplo del crecimiento de masa de n especies diferentes de bacterias que crecen sin interacción en un medio rico en nutrientes. La masa de la especie i-ésima evoluciona según $z_i = \exp(\alpha_i + \lambda_i t)$. Suponemos que las bacterias presentes pueden dividirse en dos grupos de especies afines. Denotamos por G_0 el conjunto de índices que corresponden al primer grupo y Q_0 los subíndices del segundo, con $G_0 \cap Q_0 = \emptyset$. A su vez, suponemos que este segundo grupo se subdivide en otros dos subgrupos, siendo G_1, G_2 los conjuntos de subíndices de estos subgrupos, con $Q_0 = G_1 \cup G_2$ y $G_1 \cap G_2 = \emptyset$. Denotamos por γ_0, γ_1 y γ_2, $\gamma_0 + \gamma_1 + \gamma_2 = n$, el

número de subíndices contenido en cada uno de los grupos. Supondremos, por simplicidad, que las especies del grupo G_0 corresponden a los subíndices iniciales, y a continuación siguen los del grupo G_1 y los del grupo G_2.

Basándonos en la supuesta afinidad de las especies dentro de cada uno de los grupos G_0 y Q_0, aproximamos la evolución de las masas de las especies por $z_i \simeq \exp(\beta_j + \nu_j t)$, donde $\beta_j = \sum_{i\in G_j} \alpha_i/\gamma_j$ y $\nu_j = \sum_{i\in G_j} \lambda_i/\gamma_j$. Es decir, las masas de especies dentro del mismo grupo las suponemos aproximadamente iguales, con la constante de crecimiento igual al promedio dentro del grupo. Esta aproximación elimina toda la información que permite distinguir entre masas de especies del mismo grupo. En cambio, las diferencias de evolución entre los diferentes grupos queda capturada por

$$
\begin{aligned}
z_i &\simeq \exp\left(\beta_0 + \nu_0 t\right) && \cdot \exp\left(0 \cdot t\right), && \text{si } i \in G_0\,,\\
z_i &\simeq \exp\left(\tfrac{\gamma_1(\beta_1+\nu_1 t)+\gamma_2(\beta_2+\nu_2 t)}{\gamma_1+\gamma_2}\right) && \cdot \exp\left(\beta_1+\nu_1 t - \tfrac{\gamma_1(\beta_1+\nu_1 t)+\gamma_2(\beta_2+\nu_2 t)}{\gamma_1+\gamma_2}\right), && \text{si } i \in G_1\,,\\
z_i &\simeq \exp\left(\tfrac{\gamma_1(\beta_1+\nu_1 t)+\gamma_2(\beta_2+\nu_2 t)}{\gamma_1+\gamma_2}\right) && \cdot \exp\left(\beta_2+\nu_2 t - \tfrac{\gamma_1(\beta_1+\nu_1 t)+\gamma_2(\beta_2+\nu_2 t)}{\gamma_1+\gamma_2}\right), && \text{si } i \in G_2\,.
\end{aligned}
$$

Esta expresión descompone, salvo aproximaciones intra-grupos, el proceso de crecimiento exponencial de masas en el producto de dos de tales procesos. En el primero de ellos, representado por la primera columna de exponenciales, se observa que las constantes del proceso son iguales en cada especie perteneciente a cada uno de los grupos G_0 y Q_0 de la primera partición. En la segunda columna de exponenciales, las constantes del proceso se mantienen constantes en el grupo G_0, pero quedan diferenciadas las de los subgrupos G_1 y G_2. Es fácil imaginarse que sucesivas particiones de un grupo en otros dos daría lugar a una nueva columna de exponenciales de forma similar a la exhibida.

Si se toma el punto de vista composicional, estamos interesados en los tantos por uno de masa correspondientes a cada especie presentes en cada instante. Para ello basta clausurar el vector de masas $\mathbf{z}(t)$. La descomposición anterior puede verse ahora como la perturbación (suma composicional) de dos procesos que siguen rectas en el símplex. Por tanto, podemos interpretar la diferenciación de dos grupos de especies con una dirección en el símplex. Pero es interesante identificar las direcciones de estas rectas. Puede demostrarse que éstas son las de los vectores unitarios en el símplex [7, 5]

$$
\mathbf{e}_1 = \mathcal{C}\left[\exp\left[\underbrace{\sqrt{\frac{q_0}{\gamma_0(\gamma_0+q_0)}},\ldots}_{\gamma_0 \text{ términos}}\ \underbrace{-\sqrt{\frac{\gamma_0}{q_0(\gamma_0+q_0)}},\ldots}_{q_0 \text{ términos}}\right]\right], \qquad q_0 = \gamma_1+\gamma_2\,,
$$

$$
\mathbf{e}_2 = \mathcal{C}\left[\exp\left[\underbrace{0,\ldots,0}_{\gamma_0 \text{ términos}}\ \underbrace{\sqrt{\frac{\gamma_2}{\gamma_1(\gamma_1+\gamma_2)}},\ldots}_{\gamma_1 \text{ términos}}\ \underbrace{-\sqrt{\frac{\gamma_1}{\gamma_2(\gamma_1+\gamma_2)}},\ldots}_{\gamma_2 \text{ términos}}\right]\right].
$$

La propiedad más importante de estos vectores es que son ortogonales. Lo que indica que las dos aproximaciones, que se han efectuado separando sucesivamente en

dos grupos las especies de un grupo previo, conducen a una perturbación (suma composicional) sucesiva de términos lineales ortogonales.

Esta idea permite la construcción de bases ortonormales que quedan asociadas a las particiones binarias sucesivas de las componentes de un vector composicional. En particular, esta aproximación al proceso de crecimiento de masas corresponde a una proyección ortogonal en el subespacio de dimensión 2 cuya base son los vectores $\mathbf{e}_1$ y $\mathbf{e}_2$.

La descripción de los comportamientos diferenciados de las masas de las diferentes especies dentro de un grupo se obtiene también mediante una proyección ortogonal. Esto indica que el análisis de una subcomposición independiente del resto de las componentes es una operación totalmente compatible con la geometría de Aitchison en el símplex.

4. Conclusión

Se han presentado dos casos, $\mathbb{R}_+$ y el símplex $\mathcal{S}^n$, en que subconjuntos de un espacio euclídeo real han sido dotados de una estructura euclídea propia. El mérito de estas estructuras euclídeas reside en que las operaciones y la métrica propuestas son interpretables en la práctica y por ello pueden considerarse destacadas frente a otras posibles opciones.

La principal dificultad, inherente al uso de estos planteamientos, procede de que se exigen interpretaciones de los resultados en dos representaciones alternativas de los vectores: la representación en el subconjunto del espacio real, normalmente asociados a datos tal cual son obtenidos; y la representación en coordenadas del espacio euclídeo correspondiente, en donde tienen sentido operaciones, distancias y proyecciones ortogonales que pueden estar alejadas de la intuición inicial sobre las observaciones.

Referencias

[1] J. Aitchison, The statistical analysis of compositional data (with discussion), *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **44**(2): 139–177, 1982.

[2] J. Aitchison, *The Statistical Analysis of Compositional Data*, Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London, 1986. (Reimpreso en 2003, con material adicional, por The Blackburn Press.)

[3] J. Aitchison, C. Barceló-Vidal, J. J. Egozcue y V. Pawlowsky-Glahn, A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. En U. Bayer, H. Burger y W. Skala (editores), *Proceedings of IAMG'02 – The eigth annual conference of the International Association for Mathematical Geology*, volume I and II, pp. 387–392. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 2002.

[4] D. Billheimer, P. Guttorp y W. F. Fagan, Statistical interpretation of species composition, *Journal of the American Statistical Association* **96**(456): 1205–1214, 2001.

[5] J. J. Egozcue y V. Pawlowsky-Glahn, Groups of parts and their balances in compositional data analysis, *Mathematical Geology* **37**(7): 795–828, 2005.

[6] J. J. Egozcue, J. L. Díaz-Barrero y V. Pawlowsky-Glahn, Hilbert space of probability density functions based on Aitchison geometry, *Acta Mathematica Sinica (English Series)* **22**(4): 1175–1182, 2006.

[7] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras y C. Barceló-Vidal, Isometric logratio transformations for compositional data analysis, *Mathematical Geology* **35**(3): 279–300, 2003.

[8] V. Pawlowsky-Glahn, Statistical modelling on coordinates. En S. Thió-Henestrosa y J. A. Martín-Fernández (editores), *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*, Universitat de Girona, ISBN 84-8458-111-X, `http://ima.udg.es/Activitats/CoDaWork03/`, 2003.

[9] V. Pawlowsky-Glahn y J. J. Egozcue, Geometric approach to statistical analysis on the simplex, *Stochastic Environmental Research and Risk Assessment (SERRA)* **15**(5): 384–398, 2001.

J. J. Egozcue y J. L. Díaz-Barrero, Dpto. de Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Barcelona, Spain
Correo electrónico: `juan.jose.egozcue@upc.edu` y `jose.luis.diaz@upc.edu`

V. Pawlowsky-Glahn, Dpto. de Infomàtica i Matemàtica Aplicada, Universitat de Girona, Girona, Spain
Correo electrónico: `vera.pawlowsky@udg.edu`