VNiVERSiDAD D SALAMANCA

TRABAJO DE FIN DE GRADO EN ESTADÍSTICA

Curso 2021/2022

DATOS COMPOSICIONALES: GEOMETRÍA DE AITCHISON

Autor: Sofía Martínez García

Tutor: María Jesús Rivas López

TRABAJO DE FIN DE GRADO EN ESTADÍSTICA

Curso 2021/2022

DATOS COMPOSICIONALES: GEOMETRÍA DE AITCHISON

Autor: Sofía Martínez García

Tutor: María Jesús Rivas López

VNiVERSiDAD D SALAMANCA

Firma del autor/a

Firma del tutor/a

Agradecimientos

Quisiera agradecer a mi tutora por apoyarme y guiarme en la realización de este trabajo. Por motivarme y enseñarme tanto durante los años de la carrera así como inspirarme a continuar sin miedo en los estudios y en la vida.

También quiero agradecer al resto del profesorado que enseñan con ganas, con ello motivan al alumnado a seguir trabajando aun cuando se encuentran perdidos.

Por último, quiero agradecer a mi familia por haber estado siempre conmigo y ser un gran apoyo para todo.

ÍNDICE

INTRODUCCIÓN

Los datos composicionales son un caso especial en el campo de la Estadística. Consisten en vectores en los que cada componente es no negativa con la propiedad de que sus valores suman una constante, normalmente estandarizada a 1 o 100, en el que cada componente muestra la importancia relativa de una parte en un total (Aitchison, 1986). Debido a esta característica especial, es necesario manejar e interpretar los datos composicionales de manera diferente a cómo se tratan vectores que no tienen dicha restricción de suma constante.

A lo largo del siglo XX se han ido aumentando los conocimientos sobre el correcto tratamiento y análisis de este tipo de datos. En (Pearson, 1897), Karl Pearson publicó la siguiente advertencia: *"Cuidado con los intentos de interpretar correlaciones entre cocientes cuyos numeradores y denominadores contienen partes comunes"*. Pero hasta 1960 no se tuvo realmente en cuenta, lo que llevó a muchos investigadores a implementar métodos de análisis multivariante estándar diseñados para datos multivariantes sin restricciones, que llevaban a conclusiones poco apropiadas sobre los datos.

Además, ese mismo año (Chayes, 1960) realizó una crítica hacia la interpretación que se estaba realizando de las correlaciones entre partes de una misma composición, ya que se generaban correlaciones negativas entre partes sin dicha correlación. Aun siendo conscientes de dicho problema, los investigadores se centraron en la distorsión de las técnicas multivariantes estándar y no en encontrar una correcta metodología para el estudio de los datos composicionales.

En la década de 1980, Aitchison se percató de que los datos composicionales proporcionan información únicamente de las magnitudes relativas de las partes, no de sus valores absolutos, ya que el total no es de interés. Esto permite expresar un dato composicional en términos de proporciones de los componentes, lo que se conoce como cocientes o ratios.

En (Aitchison & Shen, 1980), presentaron un análisis de la relación entre estos cocientes para utilizarlos como solución a los problemas en el tratamiento de datos composicionales, culminando en la monografía de (Aitchison, 1986) donde se demostró la importancia de trabajar con estos cocientes, ya que mantienen una correspondencia uno a uno con las componentes y conservan las relaciones de las partes.

Mostremos un ejemplo para presentar el problema de posibles correlaciones negativas falsas si no se utilizan los cocientes entre las partes. Supongamos 2 investigadores que miden las mismas partes (a, b, c, d) pero el primero decide expresar todas las partes en porcentaje para que sumen al 100%, mientras que el segundo decide tomar únicamente tres de los cuatro datos que cree relevantes (a, b, c) y los ajusta de igual modo al 100%.

Datos Originales	Primer Investigador	Segundo Investigador
[5, 5, 10, 30]	[10, 10, 20, 60]	[25, 25, 50]
[3, 10, 7, 1]	[14.29, 47.62, 33.33, 4.76]	[15, 50, 35]
[1, 2, 2, 5]	[10, 20, 20, 50]	[20, 40, 40]

Con estos registros, podemos comprobar que los investigadores se encuentran estudiando las mismas partes con el mismo origen de datos pero estarán llegando a diferentes resultados, ya que el primer investigador obtiene una correlación positiva de 0.96 entre las variables b y c, mientras que el segundo investigador obtiene una correlación negativa de -0.99. Por ello, el método fundamental para el correcto análisis de datos composicionales es la utilización de los cocientes que es invariable para las partes elegidas.

Al deber utilizarse los cocientes entre las partes, (Aitchison, 1981, 1982, 1983, 1984) desarrolló una serie de transformaciones (alr y clr) basadas en logaritmos de las mismas. La ventaja de estas transformaciones no solo es que resultan más fáciles de manejar que los cocientes, sino que además eliminan el problema de un espacio vectorial restringido, el símplex, y pasan a un espacio no restringido,

un espacio real $\mathbb{R}^D$, permitiendo el uso de la estadística multivariante estándar sin restricciones aplicada a los datos composicionales transformados.

Posteriormente, con numerosos estudios como (Pawlowsky-Glahn & Egozcue, 2001), se comprobó que las diferentes operaciones y medidas definidas por Aitchison para el símplex: la operación interna de perturbación, la operación externa de potenciación y la métrica simplicial, permiten estructurar el símplex como un espacio vectorial métrico. Esto ha permitido que muchos problemas de los datos composicionales pudiesen estudiarse dentro de su espacio y con su propia estructura, con ello se generó una perspectiva de permanencia en el símplex que propone representar las composiciones por coordenadas e interpretarlas a partir de sus representaciones en el símplex.

En el grado de Estadística solo tenemos un conocimiento básico mínimo sobre la estructuración de un espacio, por tanto, en este trabajo comenzamos presentando la estructura en un espacio vectorial real. Y se pretende llegar a dotar de estructura de espacio vectorial al espacio del símplex.

Primero se definirán las propiedades que posee el espacio real $\mathbb{R}^D$, así como también las operaciones que pueden definirse sobre él y que le dotan de estructura de espacio vectorial. Posteriormente se trasladarán dichas operaciones al símplex para demostrar que éste verifica las condiciones para ser también un espacio vectorial.

A continuación se definirán varias transformaciones, alr y clr, siendo no isométrica e isométrica respectivamente, aunque existen numerosas transformaciones, cada una con diferentes características y ventajas. Son transformaciones entre el espacio vectorial del símplex y $\mathbb{R}^D$ que conservan las operaciones. De este modo es posible elegir si tratar los datos composicionales en un espacio u otro.

Por último, se definirán conceptos estadísticos específicos para el tratamiento de datos como los conceptos de media muestral o varianza para ver cómo se definen en muestras de datos composicionales. Además, se realizarán ejemplos numéricos del cálculo de estos y visualizaciones con sus representaciones en el símplex.

CONCEPTOS BÁSICOS

Los datos composicionales poseen un carácter especial ya que el total no es de interés y los valores que forman el dato composicional dependen de la elección particular de las partes que importen. Supongamos un dato composicional de dimensión 3, al añadirle una parte más (de modo que ahora tenga dimensión 4) al menos una de las partes anteriores debe disminuir puesto que se ven influenciadas por la restricción de suma constante.

La clave de los datos composicionales reside en que la geometría del espacio muestral sobre el que se define un vector de proporciones es diferente de la clásica geometría Euclídea de $\mathbb{R}^D$. Por ello, las técnicas multivariantes habitualmente utilizadas y fundamentadas en esta geometría, no son directamente aplicables.

Definición. Un dato composicional con *D* partes de suma constante *k*, es un vector con *D* componentes no negativas que aportan información relativa y cuyo espacio muestral es el símplex S^D, definido por

$$S^D(k) = \left\{ \boldsymbol{x} \in \mathbb{R}^D / \, \boldsymbol{x} = (x_1, x_2, \dots, x_D) \, : \, x_i \geq 0; \sum_{i=1}^{D} x_i = k \right\}$$

Definición. Se denomina cierre o clausura a la aplicación sobre un dato composicional que le asigna el valor de la suma constante que posee. Suponiendo un vector $\boldsymbol{x} = (x_1, \dots, x_D) \in \mathbb{R}^{D+}$ el cierre de $\boldsymbol{x}$ a $k > 0$ consiste en

$$C(\boldsymbol{x}) = \frac{k\boldsymbol{x}}{(x_1 + x_2 + \cdots + x_D)} = \frac{k\boldsymbol{x}}{\sum_{i=1}^{D} x_i}$$

Por tanto, dos datos composicionales serán equivalentes si sus cierres para suma *k* son iguales:

$$\boldsymbol{x} \sim \boldsymbol{y} \Leftrightarrow C(\boldsymbol{x}) = C(\boldsymbol{y}), \;\; \forall k > 0$$

La constante *k* es de libre elección, pero para simplificar cuentas, se utilizarán datos composicionales con cierre a 1 y como el espacio muestral asociado es un símplex, los datos composicionales de *D* partes con suma 1 se encuentran en el Simplex Unitario:

$$S^D = \left\{ \boldsymbol{x} \in \mathbb{R}^D / \, \boldsymbol{x} = (x_1, x_2, \dots, x_D) \, : \, x_i \geq 0; \sum_{i=1}^{D} x_i = 1 \right\}$$

En ocasiones, el interés está enfocado en un subconjunto de los componentes. Este es el concepto de subcomposición que consiste en proyecciones del símplex S^D sobre un sub-símplex de dimensión menor.

Definición. Se define como subcomposición $\boldsymbol{x}_T \in S^T$ de un dato composicional $\boldsymbol{x} \in S^D$ $\boldsymbol{x} = \{x_1, \dots, x_D\}$, a un subvector formado por *T* partes seleccionadas de entre las *D* partes de $\boldsymbol{x}$, tal que $T \leq D$, $\boldsymbol{x}_T \subset \boldsymbol{x}$.

PRINCIPIOS BÁSICOS

Aitchison (Aitchison, 1994) desarrolla una serie de principios que deben cumplir los datos composicionales. De ellos, tres se consideran como principios básicos que siempre se deben respetar:

- Invarianza de escala:
  Los datos de composición sólo aportan información relativa, por lo que cualquier cambio de la escala de los datos originales no tiene ningún efecto.

  $f: S^D \to \mathbb{R}^n$ presenta invarianza de escala si $\forall \lambda \in \mathbb{R}^+,\ x \in S^D \ \Rightarrow f(\lambda \boldsymbol{x}) = f(\boldsymbol{x})$;

  Es decir, produce el mismo resultado para todos los vectores composicionales equivalentes.

  Por ejemplo, supongamos un dato composicional $\boldsymbol{w} = (1, 2, 5)$ al que multiplicamos por $\lambda = 2$. Si observamos la aplicación de cierre sobre ambos datos composicionales obtenemos:

$$\boldsymbol{w} = (1, 2, 1) \overset{C}{\to} \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$

$$\lambda \cdot \boldsymbol{w} = (2, 4, 2) \overset{C}{\to} \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$

  Así que, $\boldsymbol{w} = \lambda \boldsymbol{w}$ puesto que $C(\boldsymbol{w}) = C\ (\lambda \boldsymbol{w})$ y por ello la función de cierre presenta invarianza de escala.

  En otras palabras, cualquier función *f* aplicada sobre datos composicionales debe poder expresarse en términos de cocientes entre sus partes o componentes (Mateu-Figueras et al., 2003).

  Una función *f* de denomina log contraste cuando:

$$f: S^D \to \mathbb{R}$$

$$f(\boldsymbol{x}) = \sum_{i=1}^{D} \alpha_i \ln(x_i) = \sum_{i=i}^{D} \ln(x_i)^{\alpha_i} = \ln\left(\prod_{i=1}^{D} x_i^{\alpha_i}\right)$$

  donde $\alpha_i \in \mathbb{R}, x \in S^D, x = (x_1, \dots, x_D)$ y con $\sum_{i=1}^{D} \alpha_i = 0$ (para que se mantenga la invarianza de escala)

  De este modo,

$$f(\lambda \boldsymbol{x}) = \sum_{i=1}^{D} \alpha_i \ln(\lambda\, x_i) = \sum_{i=1}^{D} \alpha_i (\ln(\lambda) + \ln(x_i)) = \ln(\lambda) \sum_{i=1}^{D} \alpha_i + \sum_{i=1}^{D} \alpha_i \ln(x_i) =$$

$$= f(\boldsymbol{x})$$

- Invarianza de permutación:
  El orden en que aparecen las partes en una composición no afecta a los resultados.

  En un conjunto de datos composicionales, las partes deben están ordenadas de la misma manera para cada muestra, pero deberían poder reordenarse en todo el conjunto de datos sin afectar a los resultados.

  $f: S^D \to \mathbb{R}$ es invariante en la permutación si produce resultados equivalentes cuando el orden de las partes del dato composicional cambia.

- Coherencia subcomposicional:
  Los resultados obtenidos para una subcomposición de una composición deben seguir siendo los mismos que en la composición.

  Supongamos dos datos composicionales de las mismas partes, pero uno con dimensión 4 (a, b, c, d) y otro con dimensión 3 (a, b, c) y que están midiendo lo mismo, $[\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4]$ y $[s_1, s_2, s_3]$. Entonces, cualquier comparación entre las partes de los datos composicionales debe coincidir.

  Para ello las subcomposiciones deben comportarse como proyecciones ortogonales en un espacio real, es decir:

  - La distancia entre dos datos composicionales debe ser mayor o igual que la distancia entre dos subcomposiciones de ellas.
  - Se debe mantener la invarianza escalar, es decir, los ratios entre partes cualesquiera de una subcomposicion deben ser iguales a los ratios correspondientes en el dato composicional original.

GEOMETRÍA DE AITCHISON

ESPACIO VECTORIAL

En álgebra lineal, un espacio vectorial es una estructura algebraica creada a partir de:

- un conjunto no vacío.
- una operación interna (denominada como suma y definida para los elementos del propio conjunto).
- una operación externa (denominada como producto por un escalar y definida entre dicho conjunto y otro conjunto, con estructura de cuerpo).

A los elementos de este espacio se les llama vectores y a los elementos del cuerpo se les conoce como escalares.

Un espacio vectorial sobre un cuerpo Q es un conjunto no vacío dotado de dos operaciones para las cuales será cerrado, por ejemplo, el cuerpo de los números reales está dotado con la suma y el producto por escalares.

Veamos esto usando el espacio $\mathbb{R}^D$, como ejemplo utilizaremos $\mathbb{R}^3$.

SUMA

La suma es una operación interna $\mathbb{R}^D \times \mathbb{R}^D \xrightarrow{+} \mathbb{R}^D$ tal que $(\mathbb{R}^D, +)$ es grupo abeliano, es decir, se cumple:

- Propiedad conmutativa
$$\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u} \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D$$
- Propiedad asociativa
$$\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w} \quad \forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^D$$
- Exista el elemento neutro
$$\exists n \in \mathbb{R}^D : \boldsymbol{u} + e = \boldsymbol{u} \quad \forall \boldsymbol{u} \in \mathbb{R}^D$$
- Exista el elemento simétrico
$$\forall \boldsymbol{u} \in \mathbb{R}^D \quad \exists -\boldsymbol{u} \in \mathbb{R}^D : \boldsymbol{u} + (-\boldsymbol{u}) = n$$

PRODUCTO POR ESCALAR

El producto por escalar, donde el cuerpo de escalares es $\mathbb{R}$, es una operación externa $\mathbb{R} \times \mathbb{R}^D \xrightarrow{\cdot} \mathbb{R}^D$ tal que cumple:

- Propiedad asociativa
$$a \cdot (b \cdot \boldsymbol{u}) = (a \cdot b) \cdot \boldsymbol{u} \quad \forall a, b \in \mathbb{R}, \quad \forall \boldsymbol{u} \in \mathbb{R}^D$$
- Propiedad distributiva respecto de la suma vectorial
$$a \cdot (\boldsymbol{u} + \boldsymbol{v}) = a \cdot \boldsymbol{u} + a \cdot \boldsymbol{v} \quad \forall a \in \mathbb{R}, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D$$
- Propiedad distributiva respecto de la suma escalar
$$(a + b) \cdot \boldsymbol{u} = a \cdot \boldsymbol{u} + b \cdot \boldsymbol{u} \quad \forall a, b \in \mathbb{R}, \quad \forall \boldsymbol{u} \in \mathbb{R}^D$$
- Exista el elemento neutro, $n = 1$
$$\exists n \in \mathbb{R} : n \cdot \boldsymbol{u} = \boldsymbol{u} \quad \forall \boldsymbol{u} \in \mathbb{R}^D$$

Una vez se tenga un espacio vectorial es posible obtener una estructura de espacio de medida si se define un producto interior. Una medida sobre un conjunto es una la aplicación que, de manera sistemática y rigurosa, asigna un valor numérico a cada subconjunto de dicho conjunto.

PRODUCTO INTERIOR

El producto interior es una operación algebraica que toma dos secuencias de números con la misma dimensión y devuelve un único número, es decir, $\mathbb{R}^D \times \mathbb{R}^D \xrightarrow{< \ , \ >} \mathbb{R}$

$$< (x_1, \dots, x_D), (y_1, \dots, y_D) > = \sum_{i=1}^{D} x_i \, y_i$$

Veamos un ejemplo en $\mathbb{R}^3$,

$$(2,1,3) \cdot (5,7,9) = 10 + 7 + 27 = 44$$

Una vez definido el concepto de producto escalar, los siguientes operadores se obtienen de éste por definición.

NORMA

La norma es un operador que determina la longitud o magnitud de un vector bajo consideración. Debe cumplir con unas condiciones básicas que son:

- Debe ser no negativa e independiente de la orientación.
- La longitud debe ser directamente proporcional al tamaño.
- La longitud entre dos puntos será siempre menor o igual que la suma de longitudes desde esos mismos dos puntos a un tercero diferente de ellos (desigualdad triangular: la suma de dos lados de un triángulo nunca es menor que el tercer lado)

En un espacio de dimensión *D* con un vector $(x_1, \dots, x_D)$, la norma al cuadrado consiste en el producto escalar del vector consigo mismo, es decir, viene dada por:

$$\|\boldsymbol{x}\|^2 = < (x_1, \dots, x_D), (x_1, \dots, x_D) > = \sum_{i=1}^{D} x_i^2$$

Usando el anterior ejemplo en $\mathbb{R}^3$,

$$\|(2,1,3)\|^2 = < (2,1,3), (2,1,3) > = 14$$

DISTANCIA

Por último, la distancia se entiende como "el camino más corto" entre dos puntos (o vectores en espacios de dimensión superior a 1) expresado numéricamente. En un $\mathbb{R}^D$, para dos vectores $(x_1, \dots, x_D)$ y $(y_1, \dots, y_D)$ se define como:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\|\boldsymbol{x} - \boldsymbol{y}\|}$$

Pongamos un ejemplo en $\mathbb{R}^3$,

$$d\big((2,1,3),(5,7,9)\big) = \sqrt{\|(2,1,3) - (5,7,9)\|} = \sqrt{\|(-3,-6,-6)\|} = \sqrt{3^2 + 6^2 + 6^2} = \sqrt{81} = 9$$

BASES

Además, todo espacio vectorial posee bases *B,* que consisten en un subconjunto del espacio que cumple las siguientes condiciones:

- Todos los elementos de *B* pertenecen al espacio.
- *B* es un sistema generador; es decir, todo elemento del espacio se puede escribir como combinación lineal de los elementos de la base *B.* Con esta propiedad, una base nos permite generar todo el espacio. Por ejemplo, en $\mathbb{R}^3$ tenemos un sistema generador dado por el conjunto de vectores $\{(0,1,3), (1,0,-1), (3,1,0), (1,2,4)\}$
- Todos los elementos de *B* forman un sistema linealmente independiente. Por ejemplo, en $\mathbb{R}^3$ el conjunto de vectores $\{(0,0,3), (0,0,-1), (3,0,0)\}$ es linealmente independiente ya que no puedes obtener ninguno de ellos con combinaciones de los otros dos.

El número de vectores que forman la base determina la dimensión del espacio vectorial.

Definición. Se denomina dimensión de un espacio vectorial al número de elementos necesarios para poder expresar todos los elementos de ese espacio en términos de combinaciones lineales.

En el espacio vectorial $\mathbb{R}^3$ la dimensión es 3 porque se necesitan 3 vectores linealmente independientes para poder obtener todos los elementos de $\mathbb{R}^3$. Sin embargo, en el espacio del símplex S^3, un dato composicional poseerá 3 partes, pero solo 2 serán linealmente independientes debido a la restricción de suma constante. Por este motivo el espacio del símplex presenta siempre una dimensión menos.

Además, una base se denomina ortonormal $(\overrightarrow{b_1}, \overrightarrow{b_2}, \overrightarrow{b_3})$ cuando está formada por un conjunto de elementos que son mutuamente ortogonales y normales, es decir, el producto escalar entre cualquier pareja de ellos es 0 y tienen norma unitaria. Siguiendo con $\mathbb{R}^3$, una base ortonormal es $\{(1,0,0), (0,1,0), (0,0,1)\}$ ya que cualquier otro vector del espacio puede expresarse como combinación lineal de esos y cada vector tiene norma uno:

$$\boldsymbol{x} = (x_1, x_2, x_3) = (2,1,3) = 2 \cdot (1,0,0) + 1 \cdot (0,1,0) + 3 \cdot (0,0,1) = \sum_{i=1}^{3} x_i \overrightarrow{b_i}$$

$$|(1,0,0)| = \sqrt{1^2 + 0^2 + 0^2} = 1$$
$$|(0,1,0)| = \sqrt{0^2 + 1^2 + 0^2} = 1$$
$$|(0,0,1)| = \sqrt{0^2 + 0^2 + 1^2} = 1$$

GEOMETRÍA DEL SÍMPLEX

La restricción de la suma fija en los datos composicionales conduce a una representación geométrica especial en el espacio del símplex, S^D. La estructura más simple de un símplex es un triángulo, conteniendo un dato composicional con tres partes $\boldsymbol{x} = (x_1, x_2, x_3)$, donde cada una corresponde con el punto que dista x_1, x_2 y x_3 respectivamente de los lados opuestos a los vértices 1, 2 y 3. Los datos composicionales con cuatro partes se representan en un tetraedro y aquellos con más partes no pueden ser directamente visualizados.

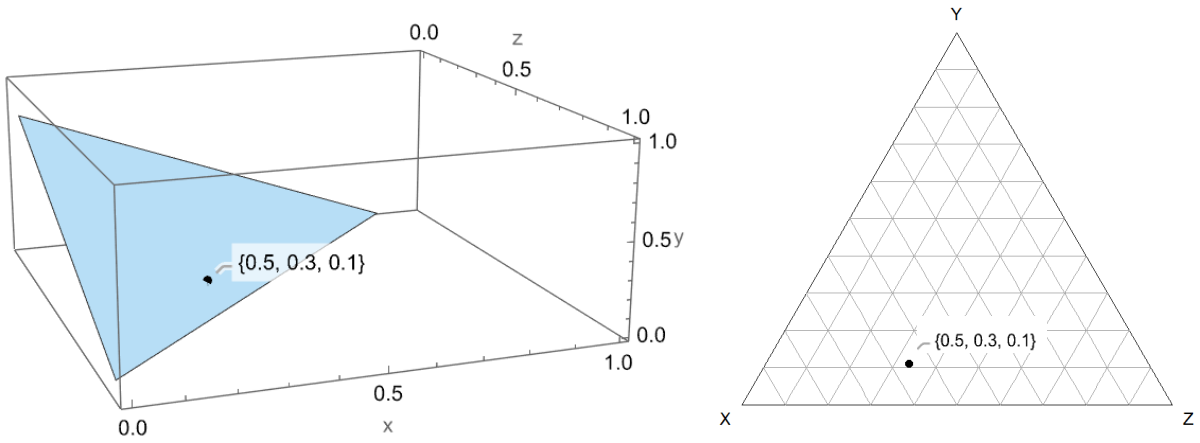

Gráfico 1.

Como se ha explicado en la introducción, un dato composicional con *D* partes pertenece al espacio $\mathbb{R}^D$. Al realizar el cierre de un dato composicional, este ahora se encuentra en el espacio del símplex.

$$\boldsymbol{x} = (x_1, x_2, \dots, x_D) \in \mathbb{R}^D \quad \rightarrow \quad C(\boldsymbol{x}) \in S^D$$

Además, hay que tener en cuenta que la suma de todas las partes es una constante, es decir, una de las partes está en función del resto y, por lo tanto, S^D posee una dimensión menos $(D-1)$.

En el anterior apartado hemos definido los conceptos y operadores necesarios para formar un espacio vectorial y un espacio de medida. Teniendo en cuenta sus definiciones, Aitchison consiguió crear unos operadores dentro del espacio del símplex que funcionen de la misma forma. De este modo, se genera una geometría propia del símplex que se asemeja a la geometría estándar euclídea y que, con unas transformaciones, se puede trasladar a $\mathbb{R}^D$.

Esta geometría del símplex la desarrollaremos ahora de manera similar a cómo se ha desarrollado para un espacio vectorial. Así pues, primero debemos definir una operación interior en el símplex equivalente a la suma en el espacio vectorial y una operación exterior equivalente al producto escalar en el espacio vectorial.

PERTURBACIÓN

La perturbación, $(S^D, \oplus)$ consiste en una operación interna $(S^D \oplus S^D \overset{\oplus}{\rightarrow} S^D)$ tal que el grupo es abeliano y la operación es equivalente a la suma en el espacio vectorial.

Definición. Dadas dos composiciones de *D* partes $\mathbf{x}, \mathbf{y} \in \mathrm{S^D}$, la perturbación se define como:

$$\boldsymbol{x} \oplus \boldsymbol{y} = C[x_1 y_1, \dots, x_D y_D] = \frac{[x_1 y_1, \dots, x_D y_D]}{(x_1 y_1 + \cdots + x_D y_D)}$$

Para que sea grupo abeliano debe cumplir:

- Propiedad conmutativa

$$\boldsymbol{x} \oplus \boldsymbol{y} = \boldsymbol{y} \oplus \boldsymbol{x} \quad \forall \boldsymbol{x}, \boldsymbol{y} \in S^D$$

Ejemplo para un espacio S^3, $\boldsymbol{x} = (1,2,1) \quad \boldsymbol{y} = (3,1,4)$

$$\boldsymbol{x} \oplus \boldsymbol{y} = \frac{(3,2,4)}{9} = \left(\frac{1}{3}, \frac{2}{9}, \frac{4}{9}\right)$$
$$\boldsymbol{y} \oplus \boldsymbol{x} = \frac{(3,2,4)}{9} = \left(\frac{1}{3}, \frac{2}{9}, \frac{4}{9}\right)$$

- Propiedad asociativa

$$\boldsymbol{x} \oplus (\boldsymbol{y} \oplus \boldsymbol{w}) = (\boldsymbol{x} \oplus \boldsymbol{y}) \oplus \boldsymbol{w} \quad \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w} \in S^D$$

Ejemplo para un espacio S^3, $\boldsymbol{w} = (1,2,3)$

$$\boldsymbol{x} \oplus (\boldsymbol{y} \oplus \boldsymbol{w}) = \boldsymbol{x} \oplus \frac{(3,2,12)}{17} = \boldsymbol{x} \oplus \left(\frac{3}{17}, \frac{2}{17}, \frac{12}{17}\right) = \frac{\left(\frac{3}{17}, \frac{4}{17}, \frac{12}{17}\right)}{\frac{19}{17}} = \left(\frac{3}{19}, \frac{4}{19}, \frac{12}{19}\right)$$

$$(\boldsymbol{x} \oplus \boldsymbol{y}) \oplus \boldsymbol{w} = \left(\frac{1}{3}, \frac{2}{9}, \frac{4}{9}\right) \oplus \boldsymbol{w} = \frac{\left(\frac{1}{3}, \frac{4}{9}, \frac{4}{3}\right)}{\frac{19}{9}} = \left(\frac{3}{19}, \frac{4}{19}, \frac{12}{19}\right)$$

- Exista el elemento neutro

$$\exists \boldsymbol{n} \in S^D: \ \boldsymbol{x} \oplus \boldsymbol{n} = \boldsymbol{x} \quad \forall \boldsymbol{x} \in S^D$$

De modo que $\boldsymbol{n} = (1, \dots, 1) \rightarrow C(\boldsymbol{n}) = \left(\frac{1}{D}, \dots, \frac{1}{D}\right)$ conteniendo *D* partes

Ejemplo para un espacio S^3, $\boldsymbol{n} = (1,1,1)$

$$\boldsymbol{x} \oplus \boldsymbol{n} = C(\boldsymbol{xn}) = \frac{(1,2,1)}{4} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$

Pero hay que tener en cuenta que $\boldsymbol{x} \notin S^D$ sino que $\boldsymbol{x} \in \mathbb{R}^D$, en este caso a $\mathbb{R}^3$. Para que $\boldsymbol{x} \in S^D$ debemos hacer el cierre de $\boldsymbol{x}$:

$$C(\boldsymbol{x}) = \frac{(1,2,1)}{4} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$

- Exista el elemento simétrico

$$\forall \boldsymbol{x} \in S^D \quad \exists -\boldsymbol{x} \in S^D: \ \boldsymbol{x} \oplus (-\boldsymbol{x}) = \boldsymbol{n}$$

Siguiendo con el ejemplo anterior,

$$\boldsymbol{x} = (1,2,1), -\boldsymbol{x} = \left(\frac{1}{1}, \frac{1}{2}, \frac{1}{1}\right)$$

$$\boldsymbol{x} \oplus -\boldsymbol{x} = \frac{(1,1,1)}{3} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) = C(\boldsymbol{n})$$

El operador $\ominus$ se define como la inversa de la perturbación y se define como:

$$\boldsymbol{x} \ominus \boldsymbol{y} = C\left(\frac{x_1}{y_1}, \dots, \frac{x_D}{y_D}\right) = \frac{\left(\frac{x_1}{y_1}, \dots, \frac{x_D}{y_D}\right)}{\left(\frac{x_1}{y_1} + \dots + \frac{x_D}{y_D}\right)}$$

Este operador es importante en el análisis de datos composicionales, por ejemplo, para la construcción de residuos composicionales.

POTENCIACIÓN

La potenciación, $(S^D, \odot)$ consiste en una operación externa $(\mathbb{R} \odot S^D \xrightarrow{\odot} S^D)$ tal que a sea un escalar, $a \in \mathbb{R}$, y la operación es equivalente al producto por un escalar en el espacio vectorial.

Definición. Dado un dato composicional de D partes $\boldsymbol{x} \in S^D$, y un escalar $a \in \mathbb{R}$, la potenciación se define como:

$$a \odot \boldsymbol{x} = C(x_1^a, \dots, x_D^a) = \frac{(x_1^a, \dots, x_D^a)}{(x_1^a + \dots + x_D^a)}$$

Debe cumplir las siguientes propiedades:

- Propiedad asociativa

  $$a \odot (b \odot \boldsymbol{x}) = (a \cdot b) \odot \boldsymbol{x} \quad \forall a, b \in \mathbb{R}, \quad \forall \boldsymbol{x} \in S^D$$

  Ejemplo para un espacio S^3, $\boldsymbol{x} = (1,2,1), a = 3, b = 2$

  $$a \odot (b \odot \boldsymbol{x}) = 3 \odot (2 \odot \boldsymbol{x}) = 3 \odot \left(\frac{(1^2, 2^2, 1^2)}{1^2 + 2^2 + 1^2}\right) = 3 \odot \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right) = \frac{\left(\frac{1}{216}, \frac{8}{27}, \frac{1}{216}\right)}{\frac{11}{36}} =$$

  $$= \left(\frac{1}{66}, \frac{32}{33}, \frac{1}{66}\right)$$

  $$(b \cdot a) \odot \boldsymbol{x} = 6 \odot \boldsymbol{x} = \left(\frac{(1^6, 2^6, 1^6)}{1^6 + 2^6 + 1^6}\right) = \left(\frac{1}{66}, \frac{32}{33}, \frac{1}{66}\right)$$

- Propiedad distributiva respecto de la perturbación

  $$a \odot (\boldsymbol{x} \oplus \boldsymbol{y}) = (a \odot \boldsymbol{x}) \oplus (a \odot \boldsymbol{y}) \quad \forall a \in \mathbb{R}, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in S^D$$

  Ejemplo para un espacio S^3, $\boldsymbol{x} = (1,2,1), \boldsymbol{y} = (3,1,4), a = 3$

  $$a \odot (\boldsymbol{x} \oplus \boldsymbol{y}) = 3 \odot (\boldsymbol{x} \oplus \boldsymbol{y}) = 3 \odot \left(\frac{1}{3}, \frac{2}{9}, \frac{4}{9}\right) = \frac{\left(\frac{1}{27}, \frac{8}{729}, \frac{64}{729}\right)}{\frac{11}{81}} = \left(\frac{3}{11}, \frac{8}{99}, \frac{64}{99}\right)$$

  $$(a \odot \boldsymbol{x}) \oplus (a \odot \boldsymbol{y}) = 3 \odot (1,2,1) \oplus 3 \odot (3,1,4) = \left(\frac{1}{10}, \frac{4}{5}, \frac{1}{10}\right) \oplus \left(\frac{27}{92}, \frac{1}{92}, \frac{16}{23}\right) =$$

  $$= C\left(\frac{27}{920}, \frac{1}{115}, \frac{16}{230}\right) = \left(\frac{3}{11}, \frac{8}{99}, \frac{64}{99}\right)$$

- Propiedad distributiva respecto de la suma escalar

$$(a+b)\odot \boldsymbol{x} = (a\odot \boldsymbol{x}) \oplus (b\odot \boldsymbol{x}) \quad \forall a, b \in \mathbb{R}, \quad \forall x \in S^D$$

Ejemplo para un espacio S^3,

$$(a+b)\odot \boldsymbol{x} = 5\odot x = \frac{(1{,}32{,}1)}{34} = \left(\frac{1}{34}, \frac{16}{17}, \frac{1}{34}\right)$$

$$(a\odot \boldsymbol{x}) \oplus (b\odot \boldsymbol{x}) = (2\odot \boldsymbol{x}) \oplus (3\odot \boldsymbol{x}) = \left(\frac{1}{10}, \frac{4}{5}, \frac{1}{10}\right) \oplus \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right) = \left(\frac{1}{34}, \frac{16}{17}, \frac{1}{34}\right)$$

- Exista el elemento neutro

$$\exists n \in \mathbb{R}:\ n\odot \boldsymbol{x} = \boldsymbol{x} \quad \forall \boldsymbol{x} \in S^D$$

n al tener que ser un escalar para que sea elemento neutro y no cambie al vector $\boldsymbol{x}$, *n=1*. Así que:

$$n = 1 \rightarrow 1\odot \boldsymbol{x} = \boldsymbol{x}$$

De igual modo que pudimos definir un espacio vectorial como un espacio de medida, el espacio del símplex se puede definir también como espacio de medida si conseguimos definir unas operaciones equivalentes al producto interno, la norma y la distancia en el propio espacio.

PRODUCTO INTERIOR

Definición. Se define como producto interno a una aplicación que a todo par de composiciones $\boldsymbol{x}, \boldsymbol{y} \in S^D$ se asocia un número, es decir, $S^D \times S^D \xrightarrow{<\ ,\ >} \mathbb{R}$

$$< \boldsymbol{x}, \boldsymbol{y} >_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j\neq i}^{D} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right)$$

Satisface las siguientes propiedades, dados un $h \in \mathbb{R}$ y $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in S^D$:

- Conmutativa
  $< \boldsymbol{u}, \boldsymbol{v} >_a = < \boldsymbol{v}, \boldsymbol{u} >_a$

- Asociativa
  $< \boldsymbol{u} + \boldsymbol{v}, \boldsymbol{w} >_a = < \boldsymbol{u}, \boldsymbol{w} >_a + < \boldsymbol{v}, \boldsymbol{w} >_a$

- $< h\boldsymbol{u}, \boldsymbol{v} >_a = h < \boldsymbol{u}, \boldsymbol{v} >_a$

Continuando con los ejemplos anteriores, siendo $\boldsymbol{x} = (1{,}2{,}1), \boldsymbol{y} = (3{,}1{,}4)$

$$< \boldsymbol{x}, \boldsymbol{y} >_a = \frac{1}{6}\left[\sum_{i=1}^{3} \sum_{j=1}^{3} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right)\right] = -0.574$$

Los productos internos desempeñan el papel de logaritmos de contraste, conocidos como las "combinaciones lineales" compositivas requeridas en muchos tipos de análisis de datos composicionales, como el análisis de componentes principales.

NORMA

En un espacio no euclídeo, el camino más corto entre dos puntos no es necesariamente una línea recta, por ello, se utilizan las propiedades de la norma vectorial comentadas anteriormente para extraer las condiciones que se deben cumplir en la norma en un espacio vectorial cualquiera.

Definición. Se denomina norma de un dato composicional $\boldsymbol{x} \in S^D$ como $||\boldsymbol{x}||_a$ y se define:

$$||\boldsymbol{x}||_a = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\frac{x_i}{x_j}\right)^2}$$

Observando la ecuación, vemos que la norma de $\boldsymbol{x}$ al cuadrado, $||\boldsymbol{x}||_a{}^2$ y es igual del producto interno del vector consigo mismo: $<\boldsymbol{x},\boldsymbol{x}>_a$. Además, la distancia entre dos vectores, $\boldsymbol{x}$ e $\boldsymbol{y}$, se define por $||\boldsymbol{x}-\boldsymbol{y}||_a$ y cuando ambos son distintos del vector cero, se dice que son ortogonales, $<\boldsymbol{x},\boldsymbol{y}>_a = 0$

Propiedades de la norma:

1. Es no negativa. Toma valor cero únicamente cuando el vector es cero.
$$||\boldsymbol{x}||_a = 0 \Leftrightarrow x = 0$$

2. La norma del múltiplo escalar de un vector, $\boldsymbol{kx}$, es k veces la norma de $\boldsymbol{x}$.
$$||\boldsymbol{kx}||_a = k||\boldsymbol{x}||_a$$

3. El valor absoluto del producto interno de $\boldsymbol{x}$ e $\boldsymbol{y}$ no excede al producto de las normas de $\boldsymbol{x}$ e $\boldsymbol{y}$.
$$||<\boldsymbol{x},\boldsymbol{y}>||_a \leq ||\boldsymbol{x}||_a \cdot ||\boldsymbol{y}||_a$$

4. La norma de la suma de $\boldsymbol{x}$ e $\boldsymbol{y}$ no excede la de suma de sus respectivas normas.
$$||\boldsymbol{x}+\boldsymbol{y}||_a \leq ||\boldsymbol{x}||_a + ||\boldsymbol{y}||_a$$

Ejemplo con $\boldsymbol{x} = (1,2,1)$

$$||\boldsymbol{x}||_a = \sqrt{\frac{1}{6}\left[\left(\ln\frac{1}{1}\right)^2+\left(\ln\frac{1}{2}\right)^2+\left(\ln\frac{1}{1}\right)^2+\left(\ln\frac{2}{1}\right)^2+\left(\ln\frac{2}{2}\right)^2+\left(\ln\frac{2}{1}\right)^2+\left(\ln\frac{1}{1}\right)^2+\left(\ln\frac{1}{2}\right)^2+\left(\ln\frac{1}{1}\right)^2\right]} =$$
$$= 0.566$$

DISTANCIA

Mide la distancia entre dos composiciones si $\boldsymbol{x}$ e $\boldsymbol{y}$ son vectores.

Definición. Sean $\boldsymbol{x},\boldsymbol{y} \in S^D$, se define como distancia a

$$d_a(\boldsymbol{x},\boldsymbol{y}) = ||\boldsymbol{x} \ominus \boldsymbol{y}||_a$$

siendo $\boldsymbol{x} \ominus \boldsymbol{y} = \boldsymbol{x} \oplus \boldsymbol{y}^{-1} = C\left(\frac{x_1}{y_1},\ldots,\frac{x_D}{y_D}\right)$

De modo que,

$$d_a(\boldsymbol{x},\boldsymbol{y}) = ||\boldsymbol{x} \ominus \boldsymbol{y}||_a = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\left(\frac{x_i}{x_j}\right)-\ln\left(\frac{y_i}{y_j}\right)\right)^2} = \sqrt{\sum_{i=1}^{D}\left(\ln\left(\frac{x_i}{g(x)}\right)-\ln\left(\frac{y_i}{g(y)}\right)\right)^2}$$

Con *g(**x**), g(**y**)* siendo las medias geométrica de las *D* partes de $\boldsymbol{x}$ e $\boldsymbol{y}$ respectivamente:

$$g(\boldsymbol{x}) = \left[\prod_{j=1}^{D} x_j\right]^{1/D} \tag{1}$$

Ejemplo con $\boldsymbol{x} = (1,2,1), \boldsymbol{y} = (3,1,4)$

$$g(\boldsymbol{x}) = \left[\prod_{j} x_j\right]^{1/D} = (1 \cdot 2 \cdot 1)^{\frac{1}{3}} = 1.26$$

$$g(\boldsymbol{y}) = \left[\prod_{j} y_j\right]^{1/D} = (3 \cdot 1 \cdot 4)^{\frac{1}{3}} = 2.29$$

$$d_a(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\left(\ln\left(\frac{1}{1.26}\right) - \ln\left(\frac{3}{2.29}\right)\right)^2 + \left(\ln\left(\frac{2}{1.26}\right) - \ln\left(\frac{1}{2.29}\right)\right)^2 + \left(\ln\left(\frac{1}{1.26}\right) - \ln\left(\frac{4}{2.29}\right)\right)^2} =$$

$$= 1.594 \tag{2}$$

Observando la ecuación de la distancia, podemos ver que la norma de $\left|\left|\boldsymbol{x}\right|\right|_a = d_a(\boldsymbol{x},\boldsymbol{n})$, siendo n el elemento neutro del símplex. Recordemos $\left|\left|\boldsymbol{x}\right|\right|_a = 0.566, g(\boldsymbol{x}) = 1.26, \boldsymbol{n} = (1,1,1), g(\boldsymbol{n}) = 1$

$$d_a(\boldsymbol{x},\boldsymbol{n}) = \sqrt{\left(\ln\left(\frac{1}{1.26}\right) - \ln(1)\right)^2 + \left(\ln\left(\frac{2}{1.26}\right) - \ln(1)\right)^2 + \left(\ln\left(\frac{1}{1.26}\right) - \ln(1)\right)^2} = 0.566$$

Una vez definidos todos estos conceptos, hemos dotado al espacio del símplex de una estructura de espacio de medida, $(S^D, \mathbb{R}, \oplus, \odot)$. Como para cualquier espacio vectorial, los vectores generadores, las bases, la dependencia lineal son esenciales y esto ocurre de igual modo para el espacio vectorial del símplex.

En el símplex, una base ortonormal $\vec{\boldsymbol{b}}$ es un conjunto de elementos que son mutuamente ortogonales y normales, es decir, el producto interior entre ellos es 0 y tienen norma unitaria. La idea es que cada dato composicional $\boldsymbol{x} \in S^D$ pudiese expresarse como "combinación lineal" de la base del símplex tal que:

$$\boldsymbol{x} = (x_1, \dots, x_D) = \oplus_{i=1}^{D} u_i \vec{\boldsymbol{b}}_i$$

En estos conceptos, el equivalente de una "combinación lineal" es una "combinación de potencia-perturbación" como:

$$\boldsymbol{x} = \left(u_1 \odot \vec{\boldsymbol{b}}_1\right) \oplus \dots \oplus \left(u_C \odot \vec{\boldsymbol{b}}_C\right)$$

Las $\vec{\boldsymbol{b}}$ son composiciones consideradas como generadores, y la combinación genera algún subespacio del símplex unitario al variar el coeficiente del número real u.

Cuando este subespacio es la totalidad del simplex unitario, las $\vec{\boldsymbol{b}}$ forman una base. Por lo general, la base debe ser elegida de forma que los generadores sean "linealmente independientes", $\vec{\boldsymbol{b}}_1, \dots, \vec{\boldsymbol{b}}_C$ son linealmente independientes si y sólo si

$$(u_1 \odot \beta_1) \oplus \dots \oplus (u_C \odot \beta_C) = e \Leftrightarrow u_1 = \dots = u_c = 0$$

Para el espacio S^D, que es un espacio de $D-1$ dimensiones, una base linealmente independiente tiene $D-1$ generadores, y entre ellas las más importantes son aquellas que forman una base ortonormal. Supongamos los generadores $\beta_1, \dots, \beta_{D-1}$ que tienen medida 1 tal que $\left|\left|\beta_i\right|\right| = 1\ (i = 1, \dots, D-1)$ y sean ortogonales tal que $< \beta_i, \beta_j >= 0\ \ (i \neq j)$.

TRANSFORMACIONES LOG-COCIENTE

Como tenemos dos espacios vectoriales de dimensión $D-1$: por una parte, el símplex $(S^D, \mathbb{R}, \oplus, \odot)$ y por otra el espacio $(\mathbb{R}^{D-1}, \mathbb{R}, +, \cdot)$, vamos a intentar dar una aplicación biyectiva entre ellos que conserve las operaciones de ambos, $S^D \xrightarrow{f} \mathbb{R}^{D-1}$, y se pueda invertir.

Esto nos permitirá trasladar los datos composicionales, puntos del símplex, a un espacio real, poder aplicar análisis multivariantes y trasladar los resultados de nuevo al símplex.

Una función *f* entre espacios vectoriales, por ejemplo, A $\xrightarrow{f}$ B, se define como aplicación si y sólo si todo elemento de A tiene imagen en B. Una aplicación biyectiva es toda aquella que es inyectiva y epiyectiva a su vez.

- Inyectiva: si y sólo si todos los elementos del origen, A, tienen distintas imágenes en B, es decir implica que si ocurre que $f(a_1) = f(a_2)$ entonces $a_1 = a_2$
- Epiyectiva: si y sólo si todo elemento de B tiene anti imagen.

Una transformación log-cociente es aquella que manda un punto de un símplex al logaritmo de un cociente que involucra a las partes de la composición.

Teniendo en cuenta las propiedades de los logaritmos, los cocientes que se comparan de manera multiplicativa pasan a compararse de forma aditiva (Tolosana-Delgado, 2011). Por ejemplo, supongamos la relación

$$\frac{A}{B} = 0.8 \qquad \frac{C}{B} = 0.4$$

Entonces,

$$\frac{A}{C} = 2$$

Podemos expresarlo tal que:

$$\frac{A}{B} = \frac{A}{C} \cdot \frac{C}{B} = 2 \cdot \frac{C}{B}$$

Añadiendo logaritmos vemos como esa relación multiplicativa se vuelve aditiva

$$\log\left(\frac{A}{B}\right) = \log\left(2 \cdot \frac{C}{B}\right) = \log(2) + [\log(C) - \log(B)]$$

Hay diferentes variantes de la transformación log-cociente que hay que considerar por las diferentes propiedades prácticas y teóricas de cada una. (Aitchison, 1984) propone dos tipos de transformaciones basadas en los logaritmos de cocientes entre las partes de un dato composicional. Por un lado, una bastante utilizada es la transformación log-cociente aditiva (alr), que es útil sobre todo para facilitar varios cálculos, y es una transformación entre espacios vectoriales (de forma que conserva las operaciones), aunque otra opción es la transformación log-cociente centrada (clr) que presenta la ventaja de ser además isometría entre ambos espacios, de forma que no solo conserva las operaciones sino también las distancias.

Transformación log-cociente aditiva

Definición. La transformación log-cociente aditiva (alr) consiste en una función biyectiva de $S^D \overset{alr}{\longrightarrow} \mathbb{R}^{D-1}$, se define como:

$$\boldsymbol{x}' = alr(\boldsymbol{x}) = \left(ln\,\frac{x_1}{x_s}, ln\,\frac{x_2}{x_s}, \ldots, ln\,\frac{x_{D-1}}{x_s}\right) = (ln(x_1) - ln(x_s), \ldots, ln(x_{D-1}) - ln(x_s))$$

donde $\boldsymbol{x} \in S^D$, $\boldsymbol{x}' \in \mathbb{R}^{D-1}$ y x_s puede ser cualquiera de las partes del dato composicional que se quiera utilizar, ya sea por la naturaleza del dato composicional o por un criterio estadístico. Por simplicidad nosotros tomaremos la última de las componentes, x_D, de modo que la transformación que se utilizará sea

$$\boldsymbol{x}' = alr(\boldsymbol{x}) = (ln(x_1) - ln(x_D), \ldots, ln(x_{D-1}) - ln(x_D))$$

La transformación inversa $\mathbb{R}^{D-1} \overset{alr^{-1}}{\longrightarrow} S^D$ está definida por:

$$\boldsymbol{x} = alr^{-1}(\boldsymbol{x}') = C(exp(x'_1), exp(x'_2), \ldots, \exp(x'_{D-1}), 1)$$

Demostración:

$$alr^{-1}(\boldsymbol{x}') = C\left(\exp\left(ln\,\frac{x_1}{x_D}\right), \ldots, \exp\left(ln\,\frac{x_{D-1}}{x_D}\right), 1\right) = C\left(\frac{x_1}{x_D}, \ldots, \frac{x_{D-1}}{x_D}, 1\right) = \frac{\left(\frac{x_1}{x_D}, \ldots, \frac{x_{D-1}}{x_D}, 1\right)}{\frac{x_1}{x_D} + \cdots + \frac{x_{D-1}}{x_D} + 1} =$$

$$= \frac{x_D\left(\frac{x_1}{x_D}, \ldots, \frac{x_{D-1}}{x_D}, 1\right)}{x_1 + \cdots + x_{D-1} + x_D} = \frac{(x_1, \ldots, x_{D-1}, x_D)}{x_1 + \cdots + x_D} = C(\boldsymbol{x})$$

Al pasar del símplex al espacio real ha de añadirse un 1 como última componente. Esto ocurre debido a que estamos pasando de un espacio vectorial con dimensión $D-1$ a uno de dimensión D que, en realidad visto como hiperplano del espacio $\mathbb{R}^D$ es el conjunto de los puntos de $\mathbb{R}^D$cuya última coordenada es 0. Al realizar la inversa queda $e^0 = 1$ y por eso se añade el término 1.

Esta transformación verifica:

$$S^D \overset{alr}{\longrightarrow} \mathbb{R}^{D-1}$$
$$\boldsymbol{x} \overset{alr}{\longrightarrow} alr(\boldsymbol{x})$$
$$\boldsymbol{y} \overset{alr}{\longrightarrow} alr(\boldsymbol{y})$$
$$\boldsymbol{x} \oplus \boldsymbol{y} \overset{alr}{\longrightarrow} alr(\boldsymbol{x}) + alr(\boldsymbol{y})$$
$$\alpha \odot \boldsymbol{x} \overset{alr}{\longrightarrow} \alpha \cdot alr(\boldsymbol{x})$$

Demostración:

- $\boldsymbol{x} \oplus \boldsymbol{y} \overset{alr}{\longrightarrow} alr(\boldsymbol{x}) + alr(\boldsymbol{y})$

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathrm{C}(\mathbf{xy}) = \frac{(x_1y_1, \ldots x_Dy_D)}{x_1y_1 + \cdots + x_Dy_D} = \left(\frac{x_1y_1}{x_1y_1 + \cdots + x_Dy_D}, \ldots, \frac{x_Dy_D}{x_1y_1 + \cdots + x_Dy_D}\right)$$

$$alr\big(C(\boldsymbol{xy})\big) = \left(\ln\left(\frac{x_1y_1}{x_Dy_D}\right), \ldots, \ln\left(\frac{x_{D-1}y_{D-1}}{x_Dy_D}\right)\right)$$

$$alr(\boldsymbol{x}) + alr(\boldsymbol{y}) = alr(x_1, \dots x_D) + alr(y_1, \dots, y_D) =$$
$$= \left(\ln\left(\frac{x_1}{x_D}\right), \dots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right) + \left(\ln\left(\frac{y_1}{y_D}\right), \dots, \ln\left(\frac{y_{D-1}}{y_D}\right)\right) =$$
$$= \left(\ln\left(\frac{x_1}{x_D}\right) + \ln\left(\frac{y_1}{y_D}\right), \dots, \ln\left(\frac{x_{D-1}}{x_D}\right) + \ln\left(\frac{y_{D-1}}{y_D}\right)\right) =$$
$$= \left(\ln\left(\frac{x_1 y_1}{x_D y_D}\right), \dots, \ln\left(\frac{x_{D-1} y_{D-1}}{x_D y_D}\right)\right)$$

Veamos un ejemplo, $S^3 \overset{alr}{\longrightarrow} \mathbb{R}^2, \boldsymbol{x} = (1,2,1),\ \boldsymbol{y} = (3,1,4)$

$$C(\boldsymbol{x}) = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) \rightarrow\ alr(\boldsymbol{x}) = \left(\ln\left(\frac{\frac{1}{4}}{\frac{1}{4}}\right), \ln\left(\frac{\frac{1}{2}}{\frac{1}{4}}\right)\right) = (0, 0.693) \qquad (3)$$

$$C(\boldsymbol{y}) = \left(\frac{3}{8}, \frac{1}{8}, \frac{1}{2}\right) \rightarrow\ alr(\boldsymbol{y}) = \left(\ln\left(\frac{\frac{3}{8}}{\frac{1}{2}}\right), \ln\left(\frac{\frac{1}{8}}{\frac{1}{2}}\right)\right) = (-0.288, -1.386)$$

Queremos comprobar que se cumple: $\boldsymbol{x} \oplus \boldsymbol{y} \overset{alr}{\longrightarrow} alr(\boldsymbol{x}) + alr(\boldsymbol{y})$

$$\boldsymbol{x} \oplus \boldsymbol{y} = \boldsymbol{z} = \left(\frac{1}{3}, \frac{2}{9}, \frac{4}{9}\right)$$

$$alr(C(\boldsymbol{z})) = \left(\ln\left(\frac{\frac{1}{3}}{\frac{4}{9}}\right), \ln\left(\frac{\frac{2}{9}}{\frac{4}{9}}\right)\right) = (-0.288, -0.693)$$

$$alr(\boldsymbol{x}) + alr(\boldsymbol{y}) = (0, 0.693) + (-0.288, -1.386) = (-0.288, -0.693)$$

❖ $\alpha \odot \boldsymbol{x} \overset{\boldsymbol{alr}}{\longrightarrow} \alpha \cdot \boldsymbol{alr}(\boldsymbol{x})$

$$\alpha \odot \boldsymbol{x} = (x_1^\alpha, \dots, x_D^\alpha)$$

$$alr(\alpha \odot \boldsymbol{x}) = \left(\ln\left(\frac{x_1^\alpha}{x_D^\alpha}\right), \dots, ln\left(\frac{x_{D-1}^\alpha}{x_D^\alpha}\right)\right)$$
$$\alpha \cdot alr(\boldsymbol{x}) = \alpha \cdot \left(\ln\left(\frac{x_1}{x_D}\right), \dots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right) = \left(\ln\left(\frac{x_1}{x_D}\right)^\alpha, \dots, \ln\left(\frac{x_{D-1}}{x_D}\right)^\alpha\right) =$$
$$= \left(\ln\left(\frac{x_1^\alpha}{x_D^\alpha}\right), \dots, ln\left(\frac{x_{D-1}^\alpha}{x_D^\alpha}\right)\right)$$

Continuando el ejemplo anterior ahora con $\boldsymbol{x} = (1,2,1), alr(\boldsymbol{x}) = (0, 0.693)$ calculado en (3) y $\alpha = 2$.

$$\alpha \odot \boldsymbol{x} = 2 \odot \boldsymbol{x} = \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right)$$

$$alr\left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right) = \left(\ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right), \ln\left(\frac{\frac{2}{3}}{\frac{1}{6}}\right)\right) = (0, 1.386)$$

$$\alpha \cdot alr(\boldsymbol{x}) = 2 \cdot (0, 0.693) = (0, 1.386)$$

Sin embargo, la transformación alr aun facilitando los cálculos y la interpretación en la práctica, presenta un inconveniente en su asimetría respecto a las partes de la composición, ya que la parte del denominador cobra especial protagonismo. Además, esta no es una transformación isométrica, es decir, los ángulos y distancias en el símplex no pueden asociarse con ángulos y distancias en el espacio real.

Demostremos esto último; es decir, si la alr es una isometría entre espacios vectoriales.
Para que lo sea se tendría que cumplir que $d_a^2(\boldsymbol{x},\boldsymbol{y}) \overset{alr}{\longrightarrow} d^2(alr(\boldsymbol{x}), alr(\boldsymbol{y}))$

$$d_a^2(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a = \sum_{i=1}^{D}\left(\ln\left(\frac{x_i}{g(x)}\right) - \ln\left(\frac{y_i}{g(y)}\right)\right)^2 =$$

$$= \sum_{i=1}^{D} \ln^2\left(\frac{\frac{x_i}{g(\boldsymbol{x})}}{\frac{y_i}{g(\boldsymbol{y})}}\right) = \sum_{i=1}^{D} \ln^2\left(\frac{x_i g(\boldsymbol{y})}{y_i g(\boldsymbol{x})}\right) = \sum_{i=1}^{D}\left(\ln\left(\frac{x_i}{y_i}\right) + \ln\left(\frac{g(\boldsymbol{y})}{g(\boldsymbol{x})}\right)\right)^2$$

$$d^2\big(alr(\boldsymbol{x}), alr(\boldsymbol{y})\big) = \sum_{i=1}^{D-1}\left(\ln\left(\frac{x_i}{x_D}\right) - \ln\left(\frac{y_i}{y_D}\right)\right)^2 = \sum_{i=1}^{D-1} \ln^2\left(\frac{x_i y_D}{y_i x_D}\right) = \sum_{i=1}^{D-1}\left(\ln\left(\frac{x_i}{y_i}\right) + \ln\left(\frac{y_D}{x_D}\right)\right)^2$$

Así que $d_a^2(\boldsymbol{x},\boldsymbol{y}) = d^2\big(alr(\boldsymbol{x}), alr(\boldsymbol{y})\big)$ sí y sólo si $g(\boldsymbol{x}) = x_D$ y $g(\boldsymbol{y}) = y_D$, pero esto es imposible ya que las medias geométricas son:

$$g(\boldsymbol{x}) = \left[\prod_{j=1}^{D} x_j\right]^{1/D}$$

Por ello, podemos afirmar que la transformación alr no mantiene las distancias.

Utilizando un ejemplo numérico vemos que:

$$d_a(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a = \sqrt{\Sigma_{i=1}^{3}\left(\ln\left(\frac{x_i}{g(x)}\right) - \ln\left(\frac{y_i}{g(y)}\right)\right)^2} = 1.594 \qquad \text{cuenta realizada en (2)}$$

$$d\big(alr(\boldsymbol{x}), alr(\boldsymbol{y})\big) = \sqrt{\|alr(x) - alr(y)\|} = \sqrt{\|(0{,}0.693) - (-0.288, -1.386)\|} =$$

$$= \sqrt{\|(0.288{,}2.079)\|} = \sqrt{0.288^2 + 2.079^2} = 2.098$$

Como no es isometría, busquemos otra transformación que lo sea.

Transformación log-cociente centrada

Definición. La transformación log-cociente centrada (clr) $S^D \overset{clr}{\longrightarrow} U^{D-1}$, donde U^{D-1} es un hiperplano de $\mathbb{R}^D$, $\boldsymbol{x} \in S^D$, $\boldsymbol{z} \in \mathbb{R}^D$ se define como:

$$\boldsymbol{z} = clr(\boldsymbol{x}) = \left(\ln\frac{x_1}{g(\boldsymbol{x})}, \ln\frac{x_2}{g(\boldsymbol{x})}, \dots, \ln\frac{x_D}{g(\boldsymbol{x})}\right) = (\ln x_1 - \ln g(\boldsymbol{x}), \dots, \ln x_D - \ln g(\boldsymbol{x})) \qquad (4)$$

donde $g(\boldsymbol{x})$ es la media geométrica de las D partes de $\boldsymbol{x}$

$$g(\boldsymbol{x}) = \left[\prod_{j=1}^{D} x_j\right]^{1/D}$$

La inversa de esta transformación consiste en:

$$\boldsymbol{x} = clr^{-1}(\boldsymbol{z}) = C(\exp(z_1), \dots, \exp(z_D))$$

Demostración:

$$clr^{-1}(\boldsymbol{z}) = C(\exp(z_1), \dots, \exp(z_D)) = C\left(\exp\left(\ln\frac{x_1}{g(\boldsymbol{x})}\right), \dots, \exp\left(\ln\frac{x_D}{g(\boldsymbol{x})}\right)\right) =$$

$$= C\left(\frac{x_1}{g(\boldsymbol{x})}, \dots, \frac{x_D}{g(\boldsymbol{x})}\right) = \frac{\left(\frac{x_1}{g(\boldsymbol{x})}, \dots, \frac{x_D}{g(\boldsymbol{x})}\right)}{\frac{x_1}{g(\boldsymbol{x})} + \dots + \frac{x_D}{g(\boldsymbol{x})}} = \frac{g(\boldsymbol{x})\left(\frac{x_1}{g(\boldsymbol{x})}, \dots, \frac{x_D}{g(\boldsymbol{x})}\right)}{x_1 + \dots + x_D} = \frac{(x_1, \dots, x_D)}{x_1 + \dots + x_D} =$$

$$= C(\boldsymbol{x})$$

Esta transformación presenta propiedades teóricas interesantes que son útiles a la hora de computar ya que es simétrica como la transformación alr. Pero, al contrario, es isométrica, es decir, mantiene los ángulos y por ello las distancias al pasar del símplex al espacio real. Por este motivo con esta transformación se mantiene que:

$$S^D \overset{clr}{\rightarrow} \mathbb{R}^D$$
$$\boldsymbol{x} \overset{clr}{\rightarrow} clr(\boldsymbol{x})$$
$$\boldsymbol{y} \overset{clr}{\rightarrow} clr(\boldsymbol{y})$$
$$\boldsymbol{x} \oplus \boldsymbol{y} \overset{clr}{\rightarrow} clr(\boldsymbol{x}) + clr(\boldsymbol{y})$$
$$\alpha \odot \boldsymbol{x} \overset{clr}{\rightarrow} \alpha \cdot clr(\boldsymbol{x})$$
$$d_a^2(\boldsymbol{x}, \boldsymbol{y}) \overset{clr}{\rightarrow} d^2(clr(\boldsymbol{x}), clr(\boldsymbol{y}))$$

Comprobemos que se cumplen estas relaciones.

- $\boldsymbol{x} \oplus \boldsymbol{y} \overset{\boldsymbol{clr}}{\rightarrow} \boldsymbol{clr}(\boldsymbol{x}) + \boldsymbol{clr}(\boldsymbol{y})$

$$\boldsymbol{x} \oplus \boldsymbol{y} = (x_1 y_1, \dots, x_D y_D)$$

$$clr(\boldsymbol{x} \oplus \boldsymbol{y}) = \left(\ln\frac{x_1 y_1}{g(\boldsymbol{x} \oplus \boldsymbol{y})}, \dots, \ln\frac{x_D y_D}{g(\boldsymbol{x} \oplus \boldsymbol{y})}\right) = \left(\ln\frac{x_1 y_1}{g(\boldsymbol{x})g(\boldsymbol{y})}, \dots, \ln\frac{x_D y_D}{g(\boldsymbol{x})g(\boldsymbol{y})}\right)$$

$$g(\boldsymbol{x} \oplus \boldsymbol{y}) = \left[\prod_{j=1}^{D} x_j y_j\right]^{1/D} = \left[\prod_{j=1}^{D} x_j\right]^{1/D} \left[\prod_{j=1}^{D} y_j\right]^{1/D} = g(\boldsymbol{x})g(\boldsymbol{y})$$

$$clr(\boldsymbol{x}) + clr(\boldsymbol{y}) = \left(\ln\frac{x_1}{g(\boldsymbol{x})}, \dots, \ln\frac{x_D}{g(\boldsymbol{x})}\right) + \left(\ln\frac{y_1}{g(\boldsymbol{y})}, \dots, \ln\frac{y_D}{g(\boldsymbol{y})}\right) =$$
$$= \left(\ln\frac{x_1 y_1}{g(\boldsymbol{x})g(\boldsymbol{y})}, \dots, \ln\frac{x_D y_D}{g(\boldsymbol{x})g(\boldsymbol{y})}\right)$$

Veamos un ejemplo, $S^3 \overset{clr}{\rightarrow} \mathbb{R}^3$, $\boldsymbol{x} = (1,2,1)$, $\boldsymbol{y} = (3,1,4)$

$$C(\boldsymbol{x}) = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) \qquad g(\boldsymbol{x}) = \left(\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{4}\right)^{\frac{1}{3}} = 0.31498$$

$$C(\boldsymbol{y}) = \left(\frac{3}{8}, \frac{1}{8}, \frac{1}{2}\right) \qquad g(\boldsymbol{y}) = \left(\frac{3}{8} \cdot \frac{1}{8} \cdot \frac{1}{2}\right)^{\frac{1}{3}} = 0.286179$$

$$clr(\boldsymbol{x}) = clr(C(\boldsymbol{x})) = \left(\ln\frac{\frac{1}{4}}{0.31498}, \ln\frac{\frac{1}{2}}{0.31498}, \ln\frac{\frac{1}{4}}{0.31498}\right) = (-0.23105, 0.4621, -0.23105)$$

$$clr(\boldsymbol{y}) = clr\big(C(\boldsymbol{y})\big) = \left(\ln\frac{\frac{3}{8}}{0.286179}, \ln\frac{\frac{1}{8}}{0.286179}, \ln\frac{\frac{1}{2}}{0.286179}\right) =$$
$$= (0.270309, -0.8283, 0.55799)$$

$$\boldsymbol{x} \oplus \boldsymbol{y} = C(\boldsymbol{xy}) = \left(\frac{1}{3}, \frac{2}{9}, \frac{4}{9}\right) \qquad g(\boldsymbol{xy}) = \left(\frac{1}{3}\cdot\frac{2}{9}\cdot\frac{4}{9}\right)^{\frac{1}{3}} = 0.32049$$

$$clr(C(\boldsymbol{xy})) = \left(\ln\frac{\frac{1}{3}}{0.32049}, \ln\frac{\frac{2}{9}}{0.32049}, \ln\frac{\frac{4}{9}}{0.32049}\right) = (0.0393, -0.3662, 0.3269)$$

$$clr(\boldsymbol{x}) + clr(\boldsymbol{y}) = (-0.23105, 0.4621, -0.23105) + (0.270309, -0.8283, 0.55799) =$$
$$= (0.039259, -0.3662, 0.32694)$$

❖ $\boldsymbol{\alpha \odot x} \overset{\boldsymbol{clr}}{\rightarrow} \boldsymbol{\alpha \cdot clr(x)}$

$$\alpha \cdot clr(\boldsymbol{x}) = \alpha \cdot \left(ln\frac{x_1}{g(\boldsymbol{x})}, \dots, ln\ \frac{x_D}{g(\boldsymbol{x})}\right)$$

$$\alpha \odot \boldsymbol{x} = (x_1^{\alpha}, \dots, x_D^{\alpha})$$

$$clr(\alpha \odot \boldsymbol{x}) = \left(\ln\frac{x_1^{\alpha}}{g(\alpha \odot \boldsymbol{x})}, \dots, \ln\frac{x_D^{\alpha}}{g(\alpha \odot \boldsymbol{x})}\right) = \left(\ln\frac{x_1^{\alpha}}{g(\boldsymbol{x})^{\alpha}}, \dots, \ln\frac{x_D^{\alpha}}{g(\boldsymbol{x})^{\alpha}}\right) =$$
$$= \left(\ln\left(\frac{x_1}{g(\boldsymbol{x})}\right)^{\alpha}, \dots, \ln\left(\frac{x_D}{g(\boldsymbol{x})}\right)^{\alpha}\right) = \alpha \cdot \left(ln\frac{x_1}{g(\boldsymbol{x})}, \dots, ln\ \frac{x_D}{g(\boldsymbol{x})}\right) = \alpha \cdot clr(\boldsymbol{x})$$

$$g(\alpha \odot \boldsymbol{x}) = \left[\prod_{j=1}^{D} x_j^{\alpha}\right]^{1/D} = \left(\left[\prod_{j=1}^{D} x_j\right]^{\frac{1}{D}}\right)^{\alpha} = g(\boldsymbol{x})^{\alpha}$$

Mostremos esto con un ejemplo, $\alpha = 2$

$$\alpha \odot \boldsymbol{x} = 2 \odot \boldsymbol{x} = \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right) = \boldsymbol{z} \qquad g(\boldsymbol{z}) = \left(\frac{1}{6}\cdot\frac{2}{3}\cdot\frac{1}{6}\right)^{\frac{1}{3}} = 0.264567$$

$$clr(\boldsymbol{z}) = \left(\ln\frac{\frac{1}{6}}{0.264567}, \ln\frac{\frac{2}{3}}{0.264567}, \ln\frac{\frac{1}{6}}{0.264567}\right) = (-0.4621, 0.9242, -0.4621)$$

$$\alpha \cdot clr(\boldsymbol{x}) = 2 \cdot (-0.23105, 0.4621, -0.23105) = (-0.4621, 0.9242, -0.4621)$$

❖ Veamos que en comparación con la transformación alr, con la clr si se mantienen las distancias.

$$d_a^2(\boldsymbol{x},\boldsymbol{y}) \overset{clr}{\rightarrow} \; d^2(clr(\boldsymbol{x}), clr(\boldsymbol{y}))$$

$$d_a^2(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a = \sum_{i=1}^{D}\left(\ln\left(\frac{x_i}{g(\boldsymbol{x})}\right) - \ln\left(\frac{y_i}{g(\boldsymbol{y})}\right)\right)^2 =$$

$$= \sum_{i=1}^{D}\ln^2\left(\frac{\frac{x_i}{g(\boldsymbol{x})}}{\frac{y_i}{g(\boldsymbol{y})}}\right) = \sum_{i=1}^{D}\ln^2\left(\frac{x_i g(\boldsymbol{y})}{y_i g(\boldsymbol{x})}\right) = \sum_{i=1}^{D}\left(\ln\left(\frac{x_i}{y_i}\right) + \ln\left(\frac{g(\boldsymbol{y})}{g(\boldsymbol{x})}\right)\right)^2$$

$$d^2\big(clr(\boldsymbol{x}), clr(\boldsymbol{y})\big) = \sum_{i=1}^{D}\left(\ln\left(\frac{x_i}{g(\boldsymbol{x})}\right) - \ln\left(\frac{y_i}{g(\boldsymbol{y})}\right)\right)^2 = \sum_{i=1}^{D}\ln^2\left(\frac{x_i g(\boldsymbol{y})}{y_i g(\boldsymbol{x})}\right) =$$

$$= \sum_{i=1}^{D}\left(\ln\left(\frac{x_i}{y_i}\right) + \ln\left(\frac{g(\boldsymbol{y})}{g(\boldsymbol{x})}\right)\right)^2$$

Utilizando un ejemplo numérico,

$$d_a(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a = \sqrt{\sum_{i=1}^{3}\left(\ln\left(\frac{x_i}{g(x)}\right) - \ln\left(\frac{y_i}{g(y)}\right)\right)^2} = 1.594 \qquad \text{cuenta realizada en (2)}$$

$$d\big(clr(\boldsymbol{x}), clr(\boldsymbol{y})\big) = \sqrt{\|clr(x) - clr(y)\|} = \sqrt{\|(-0.50136, 1.2904, -0.78904)\|} =$$

$$= \sqrt{(-0.50136)^2 + (1.2904)^2 + (-0.78904)^2} = 1.594$$

Inconvenientes de la transformación clr

- Esta transformación pasa de $S^D \overset{clr}{\rightarrow} \mathbb{R}^D$, existiendo en S^D la restricción de suma constante, $\sum_{i=1}^{D} x_i = 1$. Al pasar un dato composicional a $\mathbb{R}^D$, esta restricción sigue existiendo pero al aplicar clr ahora las partes del vector en $\mathbb{R}^D$, u_i deben cumplir: $\sum_{i=1}^{D} u_i = 0$

  Demostración con un ejemplo,

  $\boldsymbol{x} \in S^D \quad \overset{clr}{\rightarrow} \quad \boldsymbol{u} \in U^{D-1} \subset \mathbb{R}^D$

  $\sum_{i=1}^{D} x_i = 1$

  $\boldsymbol{x} = (0.1, 0.2, 0.3, 0.4) \quad \overset{clr}{\rightarrow} \quad \boldsymbol{u} = \left(\ln\left(\frac{0.1}{0.221}\right), \ln\left(\frac{0.2}{0.221}\right), \ln\left(\frac{0.3}{0.221}\right), \ln\left(\frac{0.4}{0.221}\right)\right) =$

  $g(\boldsymbol{x}) = 0.221 \qquad = (-0.79299, -0.09985, 0.30562, 0.5933)$

  $\sum_{i=1}^{D} u_i = 0$

- Los coeficientes resultantes al aplicar clr no son subcomposicionalmente coherentes, es decir, incumplen el tercer principio básico de los datos composicionales. Esto se debe a que la media geométrica de un dato composicional, $g(\boldsymbol{x})$, no coincide con la media geométrica de una subcomposición de ese dato composicional, $\boldsymbol{x}_T \subset \boldsymbol{x}$.

  Demostración con un ejemplo, $\boldsymbol{c} = (1,2,3,4), \boldsymbol{x} \in S^D,\ \boldsymbol{x}_T \in S^T \subset \boldsymbol{x}$

  $\boldsymbol{x} = (0.1, 0.2, 0.3, 0.4) \qquad \boldsymbol{x}_T = C(1,2,3) = \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2}\right)$

  $g(\boldsymbol{x}) = 0.221 \qquad g(\boldsymbol{x}_T) = 0.303$

BÚSQUEDA DE BASES EN EL SÍMPLEX

Con todo esto, vemos que hemos conseguido obtener una isometría entre dos espacios de medida y por tanto se puede generar un sistema generador en el símplex a partir de uno en el espacio real $D-$ dimensional de modo que los datos composicionales se puedan expresar mediante coordenadas en el símplex de manera similar a las coordenadas de un vector en el espacio real: $\sum_{i=1}^{D} x_i \vec{\boldsymbol{b}}_i$ siendo $\boldsymbol{B}$ un conjunto de vectores generador del espacio, preferentemente base y ortonormal.

Por tanto, el objetivo principal es encontrar una base ortonormal, para ello intentemos encontrar antes un sistema generador con la transformación clr^{-1}.

Un sistema generador se define como un conjunto de vectores de $\mathbb{R}^D$, $\boldsymbol{B} = (\vec{\boldsymbol{b}}_1, \dots, \vec{\boldsymbol{b}}_N)$, a partir del cual puedo conseguir cualquier otro vector del espacio como combinación lineal de ellos

$$\vec{\boldsymbol{x}} = \sum_{i=1}^{N} x_i \, \vec{\boldsymbol{b}}_{\boldsymbol{i}} \;\; , \forall \vec{\boldsymbol{x}} \in \mathbb{R}^D$$

Se define como base al conjunto de vectores que son sistema generador y linealmente independientes, es decir,

$$\mathbf{x} = \sum_{i=1}^{D} x_i \, \vec{\boldsymbol{b}}_{\boldsymbol{i}} \;\; \forall \vec{\boldsymbol{x}} \in \mathbb{R}^D$$

Recordemos que por la definición de dimensión el número de vectores necesarios para formar una base marca la dimensión del espacio vectorial.

Observando en $\mathbb{R}^3$, se puede reescribir como:

$$\boldsymbol{x} = (x_1, x_2, x_3) \qquad base = \begin{cases} \vec{\boldsymbol{b}}_{\boldsymbol{1}} = (b_{11}, b_{12}, b_{13}) \\ \vec{\boldsymbol{b}}_{\boldsymbol{2}} = (b_{21}, b_{22}, b_{23}) \\ \vec{\boldsymbol{b}}_{\boldsymbol{3}} = (b_{31}, b_{32}, b_{33}) \end{cases}$$

$\boldsymbol{x} = x_1\vec{\boldsymbol{b}}_1 + x_2\vec{\boldsymbol{b}}_2 + x_3\vec{\boldsymbol{b}}_3$
donde cada x_i se puede expresar como el producto interior $x_i =< \vec{\boldsymbol{x}}, \vec{\boldsymbol{b}}_{\boldsymbol{i}} >$

Usemos un ejemplo, la base canónica de $\mathbb{R}^3$: $\vec{\boldsymbol{b}} = \{(1,0,0), (0,1,0), (0,0,1)\}$ y $\boldsymbol{x} = (3,4,5)$

$(3,4,5) = 3(1,0,0) + 4(0,1,0) + 5(0,0,1)$

$< (3,4,5), (1,0,0) >= 3$
$< (3,4,5), (0,1,0) >= 4$
$< (3,4,5), (0,0,1) >= 5$

Así pues, $\vec{\boldsymbol{x}} = x_1\vec{\boldsymbol{b}}_1 + x_2\vec{\boldsymbol{b}}_2 + x_3\vec{\boldsymbol{b}}_3 = \sum_{i=1}^{D} x_i \, \vec{\boldsymbol{b}}_i = \sum_{i=1}^{D} < \boldsymbol{x}, \vec{\boldsymbol{b}}_i > \cdot \, \vec{\boldsymbol{b}}_i$ en un espacio vectorial.

Utilizando la inversa de transformación isométrica clr para pasar una base de $\mathbb{R}^D$ al espacio del símplex con las similitudes entre operaciones que hemos ido mencionando tenemos que una componente se debería poder expresar tal que:

$$\vec{\boldsymbol{c}}_i = \oplus_{\boldsymbol{i=1}}^{\boldsymbol{D}} c_i \odot \vec{\boldsymbol{w}}_i$$

donde $\vec{\boldsymbol{w}}_i$ sea base de S^D y cumpla $\boldsymbol{w}_i = clr^{-1}(\vec{\boldsymbol{b}}_i)$ y $c_i =< \boldsymbol{c}, \vec{\boldsymbol{w}}_i >_a$

Esta expresión crea un sistema generador en el espacio vectorial del símplex. Como hemos mencionado, nuestro objetivo es encontrar una base ortonormal. Veamos si tomando una base ortonormal en el espacio vectorial y aplicando la transformación clr inversa (que mantiene las distancias) llegamos a una base ortonormal adecuada en el símplex.

$$\mathbb{R}^D \xrightarrow{clr^{-1}} S^D$$

Base ortonormal tal que:

$$e_i = (0 \dots \underset{\text{i)}}{1} \dots 0) \xrightarrow{clr^{-1}} w_i = C[\exp(e_i)] = C[(1 \dots e \dots 1)]$$

w_i es sistema generador en S^D puesto que una base de $\mathbb{R}^D$ es sistema generador de un hiperplano contenido, U^D, y por antiimagen es sistema generador del símplex. No podemos asegurar que sea base porque la isometría es con un hiperplano de $\mathbb{R}^D$. Al pasar de $\mathbb{R}^D$ a S^D tenemos un sistema generador con D vectores y las bases de S^D tienen $D-1$ vectores por su dimensión, por tanto no son linealmente independientes y por ello no es base.

CONCEPTOS ESTADÍSTICOS

Una vez creado un espacio de medida que permite trabajar con datos composicionales, se pueden trasportar conceptos estadísticos de modo que se puedan aplicar a los propios datos composicionales de manera adecuada. No es posible aplicar los estadísticos descriptivos estándar ya que no se ajustan a la geometría de Aitchison como medidas de tendencia central y dispersión (Aitchison, 1986). La finalidad de realizar esto es poder estudiar el conjunto de datos composicionales y ser capaces de realizar análisis exploratorios sobre ellos.

Como se ha demostrado en el apartado de Conceptos Básicos, debido a la restricción de suma constante de los datos composicionales, es necesario el uso de cocientes para que no aparezca un sesgo negativo (negative bias), es decir, correlaciones negativas entre partes que no poseen dicha correlación. Además, la correlación entre partes puede cambiar dependiendo de si se utiliza un dato composicional o una subcomposición, lo que se conoce como correlación espúrea.

Así pues, antes de poder analizar el conjunto de datos composicionales hace falta realizar unos pasos (desarrollados en (Pawlowsky-Glahn et al., 2011)), de los cuales nos centraremos en los 3 primeros:

1. Calcular los estadísticos descriptivos adecuados para el espacio del símplex. Estos son la media geométrica, la matriz de variación y la varianza total
2. La centralización y estandarización del conjunto de datos composicionales.
3. Visualizar el Biplot correspondiente a los datos para analizar patrones.

Los conceptos estadísticos han de aplicarse sobre un conjunto de datos, así que en el símplex estarán definidos para una matriz de datos composicionales. Matemáticamente, podemos representar un conjunto de datos composicionales en forma de matriz $\boldsymbol{X}(n \times D)$, con n filas donde cada una es un dato composicional y D columnas que son las partes. La restricción de cierre mantendrá que la suma de cada fila sea igual a la constante k (genéricamente 1):

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \dots \\ \boldsymbol{x}_n \end{pmatrix} \in \mathbb{R}^{n\,x\,D}, \quad \boldsymbol{x}_i \in S^D \quad i = 1, \dots, n$$

$$\sum_{j=1}^{D} x_{ij} = k$$

Antes de comenzar el análisis hay que comprobar si el conjunto de datos composicionales presenta errores, outliers (valores atípicos respecto a una distribución determinada), o la presencia de ceros. Esta última puede solucionarse con diferentes métodos desarrollados en multitud de estudios como (Martín-Fernández et al., 2000).

1. ESTADÍSTICOS DESCRIPTIVOS

Aitchison definió en (Aitchison, 1986, p. 198) y (Aitchison & Pawlowsky-Glahn, 1997) unos conceptos apropiados que se ajustasen a la geometría de datos composicionales y que los definiesen de manera estadística así como la media y la varianza lo hacen a un conjunto de datos en el espacio vectorial común.

MEDIA GEOMÉTRICA

Una medida de tendencia central es la media geométrica, definida ya en (1) para calcular la distancia de Aitchison. Aplicada para una matriz de datos composicionales se expresa como:

$$\hat{g}(\boldsymbol{X}) = C(\hat{g}_1, \dots, \hat{g}_D)$$

$$\hat{g}_i = \left[\prod_{j=1}^{n} x_{ji}\right]^{1/n} \qquad i = 1, \dots, D$$

Esta medida representa el centro de gravedad de la nube de puntos composicionales y se ajusta mejor que aplicando la media aritmética. Esto se puede observar en el siguiente gráfico donde se visualiza un conjunto de datos composicionales obtenidos por (*CoDaWeb*, s. f.) donde se representa su media geométrica composicional, representada en granate, y la media aritmética del conjunto, representada en azul oscuro.

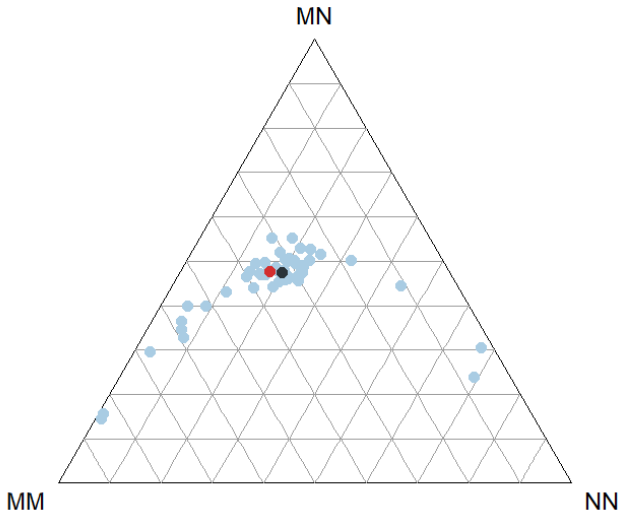

Gráfico 2.

Veamos un ejemplo con una matriz $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 2 & 3 \end{pmatrix}$,

$$\hat{g}_1 = \left[\prod_{j=1}^{3} X_{j1}\right]^{\frac{1}{3}} = (1 \cdot 3 \cdot 1)^{\frac{1}{3}} = 1.442 \qquad \hat{g}_2 = \left[\prod_{j=1}^{3} X_{j2}\right]^{\frac{1}{3}} = (2 \cdot 1 \cdot 2)^{\frac{1}{3}} = 1.587$$

$$\hat{g}_3 = \left[\prod_{j=1}^{3} X_{j3}\right]^{\frac{1}{3}} = (1 \cdot 4 \cdot 3)^{\frac{1}{3}} = 2.289$$

$$\hat{g}(\boldsymbol{X}) = C(\hat{g}_1, \hat{g}_2, \hat{g}_3) = \frac{(1.442, 1.587, 2.289)}{5.318} = (0.2712, 0.2984, 0.43)$$

MATRIZ DE VARIACIÓN

Una medida de dispersión en un espacio vectorial normal para representa la variabilidad de un conjunto de datos con respecto a su media es la varianza, aunque también se puede utilizar la desviación típica puesto que es la raíz cuadrada de la varianza.

Este concepto se puede trasladar al espacio del símplex conociéndose como la matriz de variación:

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & \dots & t_{1D} \\ \dots & \dots & \dots \\ t_{D1} & \dots & t_{DD} \end{pmatrix} \in \mathbb{R}^{D\,x\,D}, \qquad t_{ij} = Var\left(\ln\left(\frac{x_{ki}}{x_{kj}}\right)\right) = Var\big(\ln(x_{ki}) - \ln(x_{kj})\big), \quad k = 1 \dots n$$

Esta matriz es simétrica y tiene ceros en la diagonal principal. Apliquémosla con un ejemplo,

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 2 & 3 \end{pmatrix} \qquad \boldsymbol{T} = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix} \in \mathbb{R}^{n\,x\,D}$$

$$t_{11} = Var\left(\ln\left(\frac{x_{k1}}{x_{k1}}\right)\right) = Var(\ln x_{k1} - \ln x_{k1}) = 0$$

$$t_{12} = Var\left(\ln\left(\frac{x_{k1}}{x_{k2}}\right)\right) = Var(\ln x_{k1} - \ln x_{k2}) =$$

$$= Var[(\ln x_{11}, \ln x_{21}, \ln x_{33}) - (\ln x_{12}, \ln x_{22}, \ln x_{32}] =$$

$$= Var\left[\ln\left(\frac{x_{11}}{x_{12}}\right), \ln\left(\frac{x_{21}}{x_{22}}\right), \ln\left(\frac{x_{31}}{x_{32}}\right)\right] = Var\left[\ln\left(\frac{1}{2}\right), \ln\left(\frac{3}{1}\right), \ln\left(\frac{1}{2}\right)\right] =$$

$$= Var(-0.693, 1.099, -0.693) = 1.07$$

$$t_{13} = Var\left[\ln\left(\frac{x_{11}}{x_{13}}\right), \ln\left(\frac{x_{21}}{x_{23}}\right), \ln\left(\frac{x_{31}}{x_{33}}\right)\right] = Var\left[\ln\left(\frac{1}{1}\right), \ln\left(\frac{3}{4}\right), \ln\left(\frac{1}{3}\right)\right] = Var(0, -0.288, -1.099) =$$
$$= 0.325$$

$$t_{21} = Var\left[\ln\left(\frac{x_{12}}{x_{11}}\right), \ln\left(\frac{x_{22}}{x_{21}}\right), \ln\left(\frac{x_{32}}{x_{31}}\right)\right] = Var\left[\ln\left(\frac{2}{1}\right), \ln\left(\frac{1}{3}\right), \ln\left(\frac{2}{1}\right)\right] =$$
$$= Var(0.693, -1.099, 0.693) = 1.07$$

$$t_{22} = Var(\ln x_{k2} - \ln x_{k2}) = 0$$

$$t_{23} = Var\left[\ln\left(\frac{x_{12}}{x_{13}}\right), \ln\left(\frac{x_{22}}{x_{23}}\right), \ln\left(\frac{x_{32}}{x_{33}}\right)\right] = Var\left[\ln\left(\frac{2}{1}\right), \ln\left(\frac{1}{4}\right), \ln\left(\frac{2}{3}\right)\right] =$$
$$= Var(0.693, -1.386, -0.405) = 1.082$$

$$t_{31} = Var\left[\ln\left(\frac{x_{13}}{x_{11}}\right), \ln\left(\frac{x_{23}}{x_{21}}\right), \ln\left(\frac{x_{33}}{x_{31}}\right)\right] = Var\left[\ln\left(\frac{1}{1}\right), \ln\left(\frac{4}{3}\right), \ln\left(\frac{3}{1}\right)\right] = Var(0, 0.288, 1.099) =$$
$$= 0.325$$

$$t_{32} = Var\left[\ln\left(\frac{x_{13}}{x_{12}}\right), \ln\left(\frac{x_{23}}{x_{22}}\right), \ln\left(\frac{x_{33}}{x_{32}}\right)\right] = Var\left[\ln\left(\frac{1}{2}\right), \ln\left(\frac{4}{1}\right), \ln\left(\frac{3}{2}\right)\right] =$$
$$= Var(-0.693, 1.386, 0.405) = 1.082$$

$$t_{33} = Var(\ln x_{k3} - \ln x_{k3}) = 0$$

Así pues la matriz de variación de $\boldsymbol{X}$ es $\boldsymbol{T} = \begin{pmatrix} 0 & 1.07 & 0.325 \\ 1.07 & 0 & 1.082 \\ 0.325 & 1.082 & 0 \end{pmatrix}$

Podemos normalizar la matriz de variación de modo que tengamos la matriz de variación normalizada:

$$\boldsymbol{T}^* = \begin{pmatrix} {t_{11}}^* & \dots & {t_{1D}}^* \\ \dots & \dots & \dots \\ {t_{D1}}^* & \dots & {t_{DD}}^* \end{pmatrix} \in \mathbb{R}^{D\,x\,D}, \qquad {t_{ij}}^* = Var\left(\frac{1}{\sqrt{2}}\ln\left(\frac{x_i}{x_j}\right)\right)$$

Esto también se puede expresar como $\boldsymbol{T}^* = \frac{1}{2}\boldsymbol{T}$.

Veamos un ejemplo de normalizar una matriz de variación utilizando la calculada anteriormente,

$$\boldsymbol{T} = \begin{pmatrix} 0 & 1.07 & 0.325 \\ 1.07 & 0 & 1.082 \\ 0.325 & 1.082 & 0 \end{pmatrix}$$

$$\boldsymbol{T}^* = \frac{1}{2}\begin{pmatrix} 0 & 1.07 & 0.325 \\ 1.07 & 0 & 1.082 \\ 0.325 & 1.082 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0.535 & 0.162 \\ 0.535 & 0 & 0.541 \\ 0.162 & 0.541 & 0 \end{pmatrix}$$

VARIANZA TOTAL

Una medida de dispersión global es la varianza total.

$$\text{totvar}[\mathbf{X}] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} Var\left(\ln\left(\frac{x_i}{x_j}\right)\right) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} t_{ij}$$

Esta medida resume tanto la matriz de variación como la normalizada en una única cantidad y es posible definirla porque todas las partes de una composición comparten una escala común (Pawlowsky-Glahn et al., 2011).

Veamos un ejemplo numérico de su cálculo,

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 2 & 3 \end{pmatrix}, \boldsymbol{T} = \begin{pmatrix} 0 & 1.07 & 0.325 \\ 1.07 & 0 & 1.082 \\ 0.325 & 1.082 & 0 \end{pmatrix}, D = 3$$

$$\text{totvar}[\mathbf{X}] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} t_{ij} = \frac{1}{6}(0 + 1.07 + 0.325 + 1.07 + 0 + 1.082 + 0.325 + 1.082 + 0) = \\ = 0.8256 \qquad (5)$$

Tanto la varianza total como la matriz de variación no se ven restringidos por la constante *k* asociada a cada dato composicional. Debido a esto, el cambio de escala en dichos elementos no tiene ningún efecto.

2. CENTRALIZACIÓN Y ESTANDARIZACIÓN

El proceso de centralizar un conjunto de datos en el espacio real consiste en una transformación lineal que convierte las puntuaciones de unos vectores o variables en forma de desviación típica de modo que la media de cada vector sea igual a cero. Con ello se consigue trasportar el centro de un conjunto de datos de manera que los valores medios del conjunto están representados por un cero y los que estaban por encima o por debajo tendrán valores positivos o negativos.

Utilicemos un ejemplo en $\mathbb{R}^3$ con los vectores: $\boldsymbol{x} = (1,2,1)$ $\boldsymbol{y} = (3,5,1)$ $\boldsymbol{w} = (2,2,4)$
Primero debemos calcular la media del conjunto de datos,

$$m = \left(\frac{1+3+2}{3}, \frac{2+5+2}{3}, \frac{1+1+4}{3}\right) = (2,3,2) \qquad (6)$$

Los nuevos datos centralizados se obtienen de la resta de cada dato con la media de modo que,

$$\boldsymbol{x}^c = (1,2,1) - m = (-1,-1,-1)$$
$$\boldsymbol{y}^c = (3,5,1) - m = (1,2,-1)$$
$$\boldsymbol{w}^c = (2,2,1) - m = (0,-1,-1)$$

En el espacio del símplex, (Martin Fernández et al., s. f.) introdujo la propiedad de que la perturbación mantiene la estructura de las líneas rectas. Esto permite que al aplicar la operación de perturbación se consiga mover cualquier composición al baricentro del conjunto y gravitar los datos alrededor de él de la misma forma que la traslación mueve datos reales en el espacio real.

Así pues, para centralizar un conjunto de datos composicionales se perturba cada fila de la matriz de datos, $x_i = (x_{i1}, \dots, x_{iD})$ por la inversa de la media geométrica, es decir:

$$\boldsymbol{X}^C = \boldsymbol{X} \oplus \hat{g}^{-1}(\boldsymbol{X}) = \boldsymbol{X} \ominus \hat{g}(\boldsymbol{X}) = \boldsymbol{X} \ominus C(\hat{g}_1, \dots, \hat{g}_D)$$

Veamos un ejemplo numérico, $\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 2 & 3 \end{pmatrix}$, $\hat{g}(\boldsymbol{X}) = (0.2712, 0.2984, 0.43)$

$$\boldsymbol{X}^C = \boldsymbol{X} \ominus \hat{g}(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{X}_1 \ominus \hat{g}(\boldsymbol{X}) \\ \boldsymbol{X}_2 \ominus \hat{g}(\boldsymbol{X}) \\ \boldsymbol{X}_3 \ominus \hat{g}(\boldsymbol{X}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^C \\ \boldsymbol{X}_2^C \\ \boldsymbol{X}_3^C \end{pmatrix}$$

$$\boldsymbol{X}_1^C = \boldsymbol{X}_1 \ominus \hat{g}(\boldsymbol{X}) = C\left(\frac{X_{11}}{\hat{g}_1}, \frac{X_{12}}{\hat{g}_2}, \frac{X_{31}}{\hat{g}_3}\right) = C\left(\frac{1}{0.2712}, \frac{2}{0.2984}, \frac{1}{0.43}\right) = C(3.687, 6.702, 2.326) =$$

$$= \frac{(3.687, 6.702, 2.326)}{12.715} = (0.29, 0.527, 0.183)$$

$$\boldsymbol{X}_2^C = \boldsymbol{X}_2 \ominus \hat{g}(\boldsymbol{X}) = C\left(\frac{X_{21}}{\hat{g}_1}, \frac{X_{22}}{\hat{g}_2}, \frac{X_{23}}{\hat{g}_3}\right) = C\left(\frac{3}{0.2712}, \frac{1}{0.2984}, \frac{4}{0.43}\right) = C(11.062, 3.351, 9.302) =$$

$$= \frac{(11.062, 3.351, 9.302)}{23.715} = (0.467, 0.141, 0.392)$$

$$\boldsymbol{X}_3^C = \boldsymbol{X}_3 \ominus \hat{g}(\boldsymbol{X}) = C\left(\frac{X_{31}}{\hat{g}_1}, \frac{X_{32}}{\hat{g}_2}, \frac{X_{33}}{\hat{g}_3}\right) = C\left(\frac{1}{0.2712}, \frac{2}{0.2984}, \frac{3}{0.43}\right) = C(3.687, 6.702, 6.977) =$$

$$= \frac{(3.687, 6.702, 6.977)}{17.366} = (0.212, 0.386, 0.402)$$

$$\boldsymbol{X}^C = \begin{pmatrix} \boldsymbol{X}_1^C \\ \boldsymbol{X}_2^C \\ \boldsymbol{X}_3^C \end{pmatrix} = \begin{pmatrix} 0.29 & 0.527 & 0.183 \\ 0.467 & 0.141 & 0.392 \\ 0.212 & 0.386 & 0.402 \end{pmatrix} \qquad (7)$$

Podemos comprobar que la matriz de datos centralizados es correcta si su media geométrica es el baricentro de un triángulo equilátero, es decir, se encuentra a la misma distancia de todos los lados. Dicho punto en el espacio del símplex tiene como coordenadas:

$$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) = (0.333, 0.333, 0.333)$$

En este caso:

$$\hat{g}_1(\boldsymbol{X}^C) = (0.29 \cdot 0.527 \cdot 0.183)^{\frac{1}{3}} = 0.3$$

$$\hat{g}_2(\boldsymbol{X}^C) = (0.467 \cdot 0.141 \cdot 0.392)^{\frac{1}{3}} = 0.295$$

$$\hat{g}_3(\boldsymbol{X}^C) = (0.212 \cdot 0.386 \cdot 0.402)^{\frac{1}{3}} = 0.32$$

$$\hat{g}(\boldsymbol{X}^C) = C(\hat{g}_1, \dots, \hat{g}_D) = \frac{(0.3, 0.295, 0.32)}{0.915} = (0.33, 0.33, 0.33)$$

En el siguiente conjunto de gráficos se puede observar el proceso de centralización aplicado a un conjunto de datos composicionales aportado por (*CoDaWeb*, s. f.).

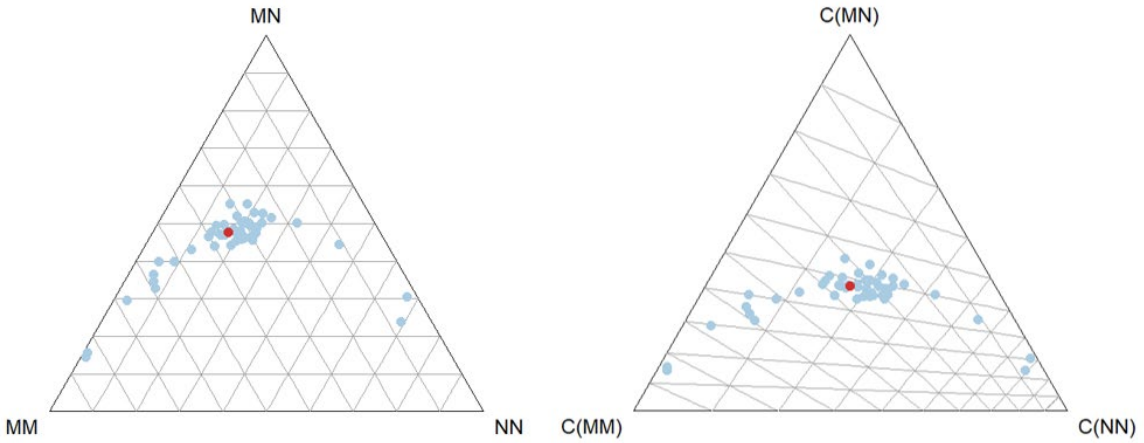

Gráfico 3.Izq. conjunto de datos no centralizados. Der. conjunto de datos centralizados. Media geométrica de cada conjunto marcada en rojo.

Por otro lado, la estandarización consiste en el proceso de ajuste o adaptación de valores o datos para que se asemejen a un modelo común con el objetivo de ofrecer una mayor facilidad en su acceso y tratamiento.

En el espacio real, este procedimiento transforma los valores originales restándoles a cada uno el valor de la media y dividiéndolos entre la desviación típica del grupo. De modo que si tenemos un conjunto de valores $Y = [Y_1, \dots, Y_D]$, los datos estandarizados se calculan tal que:

$$Y_i^E = \frac{y_i - \bar{Y}}{\sigma_Y}$$

Utilicemos un ejemplo en $\mathbb{R}^3$ con los vectores: $\boldsymbol{x} = (1,2,1)$ $\boldsymbol{y} = (3,5,1)$ $\boldsymbol{w} = (2,2,4)$. Debemos calcular la media del conjunto de datos, ya calculada en (6), $m = (2,3,2)$, y la desviación típica de cada conjunto de partes, es decir, de cada grupo de valores que ocupan la misma posición en los vectores. Calcularemos la desviación típica con su fórmula:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{D}(y_i - \bar{Y})^2}$$

Calculándola para cada grupo siendo: partes en la primera posición [1,3,2], segunda posición [2,5,2] y tercera posición [1,1,4].

$$\sigma_1 = \sqrt{\frac{1}{3-1}[(1-2)^2 + (3-2)^2 + (2-2)^2]} = 1$$

$$\sigma_2 = \sqrt{\frac{1}{3-1}[(2-3)^2 + (5-3)^2 + (2-3)^2]} = 3$$

$$\sigma_3 = \sqrt{\frac{1}{3-1}[(1-2)^2 + (1-2)^2 + (4-2)^2]} = 3$$

Los datos estandarizados quedarán

$$\boldsymbol{x}^E = \left(\frac{1-2}{1}, \frac{2-3}{3}, \frac{1-2}{3}\right) = (-1, -0.333, -0.333)$$

$$\boldsymbol{y}^E = \left(\frac{3-2}{1}, \frac{5-3}{3}, \frac{1-2}{3}\right) = (1, 0.666, -0.333)$$

$$\boldsymbol{w}^E = \left(\frac{2-2}{1}, \frac{2-3}{3}, \frac{4-2}{3}\right) = (0, -0.333, 0.666)$$

De manera similar, un conjunto de datos de composición centralizados $\boldsymbol{X}$ se puede estandarizar o escalar realizando una potenciación con su varianza total. Es decir:

$$\boldsymbol{X}^E = \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \odot \boldsymbol{X}^C = C \begin{pmatrix} X_{11}^C \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} & \cdots & X_{1D}^C \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \\ \vdots & \ddots & \vdots \\ X_{n1}^C \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} & \cdots & X_{nD}^C \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \end{pmatrix}$$

A continuación se muestra el proceso de estandarización aplicado al mismo conjunto de datos composicionales que hemos centralizado anteriormente.

Recordemos la matriz $\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 2 & 3 \end{pmatrix}$, con su varianza total ya calculada en (5), $\text{totvar}[\mathbf{X}] = 0.8256$

y la matriz centralizada en (7) $\boldsymbol{X}^C = \begin{pmatrix} \boldsymbol{X}_1^C \\ \boldsymbol{X}_2^C \\ \boldsymbol{X}_3^C \end{pmatrix} = \begin{pmatrix} 0.29 & 0.527 & 0.183 \\ 0.467 & 0.141 & 0.392 \\ 0.212 & 0.386 & 0.402 \end{pmatrix}$

$$\boldsymbol{X}^E = \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \odot \boldsymbol{X}^C = \begin{pmatrix} \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \odot \boldsymbol{X}_1^C \\ \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \odot \boldsymbol{X}_2^C \\ \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \odot \boldsymbol{X}_3^C \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{0.8256}} \odot \boldsymbol{X}_1^C \\ \frac{1}{\sqrt{0.8256}} \odot \boldsymbol{X}_2^C \\ \frac{1}{\sqrt{0.8256}} \odot \boldsymbol{X}_3^C \end{pmatrix} = \begin{pmatrix} 1.1 \odot \boldsymbol{X}_1^C \\ 1.1 \odot \boldsymbol{X}_2^C \\ 1.1 \odot \boldsymbol{X}_3^C \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^E \\ \boldsymbol{X}_2^E \\ \boldsymbol{X}_3^E \end{pmatrix}$$

$$\boldsymbol{X}_1^E = 1.1 \odot \boldsymbol{X}_1^C = C(0.29^{1.1}, 0.527^{1.1}, 0.183^{1.1}) = \frac{(0.256, 0.494, 0.154)}{0.904} = (0.283, 0.546, 0.17)$$

$$\boldsymbol{X}_2^E = 1.1 \odot \boldsymbol{X}_2^C = C(0.467^{1.1}, 0.141^{1.1}, 0.392^{1.1}) = \frac{(0.433, 0.116, 0.357)}{0.906} = (0.478, 0.128, 0.394)$$

$$\boldsymbol{X}_3^E = 1.1 \odot \boldsymbol{X}_3^C = C(0.29^{1.1}, 0.527^{1.1}, 0.183^{1.1}) = \frac{(0.182, 0.351, 0.366)}{0.899} = (0.202, 0.39, 0.407)$$

$$\boldsymbol{X}^E = \begin{pmatrix} \boldsymbol{X}_1^E \\ \boldsymbol{X}_2^E \\ \boldsymbol{X}_3^E \end{pmatrix} = \begin{pmatrix} 0.283 & 0.546 & 0.17 \\ 0.478 & 0.128 & 0.394 \\ 0.202 & 0.39 & 0.407 \end{pmatrix}$$

Podemos comprobar que dicho conjunto de datos está correctamente estandarizado si su varianza total es 1:

$$\boldsymbol{X}^E = \begin{pmatrix} 0.283 & 0.546 & 0.17 \\ 0.478 & 0.128 & 0.394 \\ 0.202 & 0.39 & 0.407 \end{pmatrix}, \boldsymbol{T}^E = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ t_{31} & t_{32} & t_{33} \end{pmatrix} \in \mathbb{R}^{n\,x\,D}$$

$$t_{11} = 0$$

$$t_{12} = Var\left[\ln\left(\frac{x_{11}}{x_{12}}\right), \ln\left(\frac{x_{21}}{x_{22}}\right), \ln\left(\frac{x_{31}}{x_{32}}\right)\right] = Var\left[\ln\left(\frac{0.283}{0.546}\right), \ln\left(\frac{0.478}{0.128}\right), \ln\left(\frac{0.202}{0.39}\right)\right] =$$
$$= Var(-0.657, 1.318, -0.657) = 1.296$$

$$t_{13} = Var\left[\ln\left(\frac{x_{11}}{x_{13}}\right), \ln\left(\frac{x_{21}}{x_{23}}\right), \ln\left(\frac{x_{31}}{x_{33}}\right)\right] = Var\left[\ln\left(\frac{0.283}{0.17}\right), \ln\left(\frac{0.478}{0.394}\right), \ln\left(\frac{0.202}{0.407}\right)\right] =$$
$$= Var(0.51, 0.193, -0.7) = 0.393$$

$$t_{21} = Var\left[\ln\left(\frac{x_{12}}{x_{11}}\right), \ln\left(\frac{x_{22}}{x_{21}}\right), \ln\left(\frac{x_{32}}{x_{31}}\right)\right] = Var\left[\ln\left(\frac{0.546}{0.283}\right), \ln\left(\frac{0.128}{0.478}\right), \ln\left(\frac{0.39}{0.202}\right)\right] =$$
$$= Var(0.657, -1.318, 0.657) = 1.296$$

$$t_{22} = 0$$

$$t_{23} = Var\left[\ln\left(\frac{x_{12}}{x_{13}}\right), \ln\left(\frac{x_{22}}{x_{23}}\right), \ln\left(\frac{x_{32}}{x_{33}}\right)\right] = Var\left[\ln\left(\frac{0.546}{0.17}\right), \ln\left(\frac{0.128}{0.394}\right), \ln\left(\frac{0.39}{0.407}\right)\right] =$$
$$= Var(1.167, -1.124, -0.04) = 1.311$$

$$t_{31} = Var\left[\ln\left(\frac{x_{13}}{x_{11}}\right), \ln\left(\frac{x_{23}}{x_{21}}\right), \ln\left(\frac{x_{33}}{x_{31}}\right)\right] = Var\left[\ln\left(\frac{0.17}{0.283}\right), \ln\left(\frac{0.394}{0.478}\right), \ln\left(\frac{0.407}{0.202}\right)\right] =$$
$$= Var(-0.51, -0.193, 0.7) = 0.393$$

$$t_{32} = Var\left[\ln\left(\frac{x_{13}}{x_{12}}\right), \ln\left(\frac{x_{23}}{x_{22}}\right), \ln\left(\frac{x_{33}}{x_{32}}\right)\right] = Var\left[\ln\left(\frac{0.17}{0.546}\right), \ln\left(\frac{0.394}{0.478}\right), \ln\left(\frac{0.407}{0.39}\right)\right] =$$
$$= Var(-1.167, 1.124, 0.04) = 1.311$$

$$t_{33} = 0$$

Así que la matriz de variación es $\boldsymbol{T}^E = \begin{pmatrix} 0 & 1.296 & 0.393 \\ 1.296 & 0 & 1.311 \\ 0.393 & 1.311 & 0 \end{pmatrix}$. Calculemos la varianza total:

$$\text{totvar}[\boldsymbol{X}^E] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} t_{ij} = \frac{1}{6}(0 + 1.296 + 0.393 + 1.296 + 0 + 1.311 + 0.396 + 1.311 + 0)$$
$$\approx 1$$

En el siguiente conjunto de gráficos se puede observar el proceso de estandarización aplicado al mismo conjunto de datos composicionales que visualizamos en la centralización.

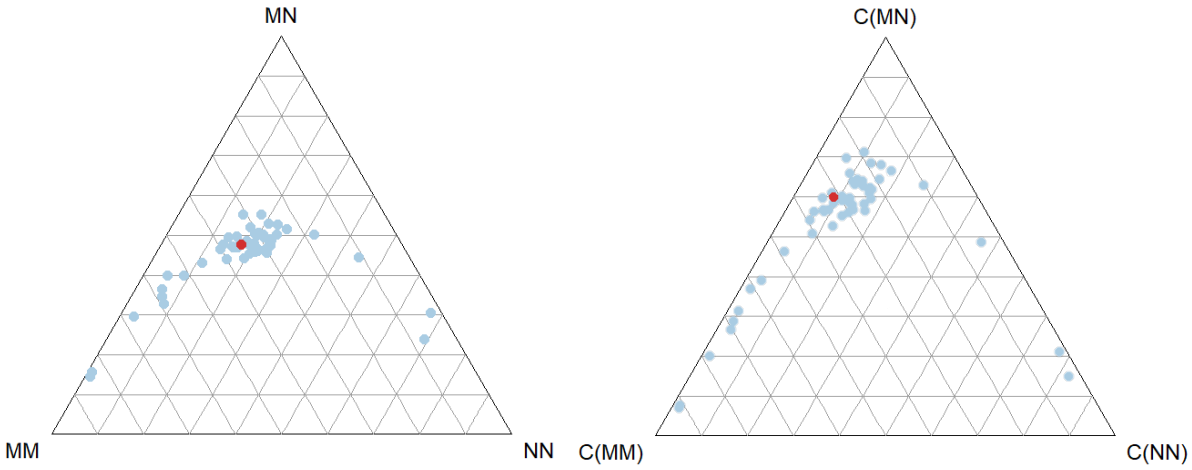

Gráfico 4. Izq. conjunto de datos no estandarizados. Der. conjunto de datos estandarizados. Media geométrica de cada conjunto marcada en rojo.

3. BIPLOTS

En 1971 se introdujo por primera vez el Biplot, un gráfico exploratorio que permite representar simultáneamente las filas y columnas de cualquier matriz utilizando una aproximación de rango 2. Las filas se representan como puntos y las variables mediante vectores o regiones de predicción. (Aitchison, 1992) (Aitchison, 1994) lo adaptó para poder utilizarlo en el símplex aplicándolo así aun conjunto de datos composicionales y usándolo como herramienta exploratoria.

Primero desarrollaremos teóricamente la construcción de un Biplot en el símplex, después su interpretación y por último realizaremos un ejemplo de las cuentas numéricas y otro visual.

CREACIÓN

Considerando una matriz de datos composicionales con *n* filas y *D* columnas, $\boldsymbol{X}(n \times D)$, se han tomado *D* mediciones en cada una de las *n* muestras.

Primeramente, se debe centralizar dicha matriz según se ha explicado en el punto 2 de esta sección y obtener los coeficientes de sus coordenadas clr (4) creando la matriz $\boldsymbol{Z}$. Siguiendo con nuestra nomenclatura y en forma de matriz se calcula tal que:

$$\boldsymbol{Z} = \begin{pmatrix} \ln \dfrac{X_{11}^C}{\hat{g}_1(\boldsymbol{X})} & \cdots & \ln \dfrac{X_{1D}^C}{\hat{g}_1(\boldsymbol{X})} \\ \vdots & \ddots & \vdots \\ \ln \dfrac{X_{n1}^C}{\hat{g}_n(\boldsymbol{X})} & \cdots & \ln \dfrac{X_{nD}^C}{\hat{g}_n(\boldsymbol{X})} \end{pmatrix}$$

Ahora se debe descomponer esta matriz utilizando la descomposición de valores singulares de $\boldsymbol{Z}$, obteniendo así la siguiente forma.

$$\boldsymbol{Z} = \boldsymbol{U} \begin{pmatrix} {\lambda_1}^{\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & {\lambda_2}^{\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & {\lambda_s}^{\frac{1}{2}} \end{pmatrix} \boldsymbol{V}'$$

Siendo:

- Las raíces cuadradas de los s valores propios positivos $\lambda_1, \dots, \lambda_s$ de $\boldsymbol{Z}\boldsymbol{Z}'$ o $\boldsymbol{Z}'\boldsymbol{Z}$, con $s \leq \min\{(D-1), n\}$ siendo el rango de $\boldsymbol{Z}$ y estando $\lambda_1, \dots, \lambda_s$ en orden descendente de magnitud.
- La matriz de valores propios de $\boldsymbol{Z}\boldsymbol{Z}'$, $\boldsymbol{U}$, con dimensión (*n, s*).
- La matriz de valores propios de $\boldsymbol{Z}'\boldsymbol{Z}, \boldsymbol{V}$, con dimensión (*D, s*).

Las matrices $\boldsymbol{U}$ y $\boldsymbol{V}$ son ortonormales, es decir, $\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}_S$ y $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{I}_S$. Además, cada fila de la matriz $\boldsymbol{V}'$ es el clr de un elemento de una base ortonormal del símplex. El producto entre la matriz $\boldsymbol{U}$ y la diagonal $({\lambda_1}^{\frac{1}{2}}, \dots, {\lambda_s}^{\frac{1}{2}})$ es una matriz que contiene las coordenadas de cada dato composicional con respecto a la base ortonormal descrita por $\boldsymbol{V}'$.

Como hemos demostrado anteriormente, la transformación clr mantiene las distancias. Debido a esto podemos aproximar la matriz $\boldsymbol{Z}$ con una matriz de rango 2, siendo la mejor aproximación la dada por la descomposición de valores singulares. Con el objetivo de reducir la dimensionalidad podemos suprimir algunas coordenadas ortogonales, normalmente las que tiene una baja varianza. Supongamos que conservamos *t* valores singulares, ${\lambda_1}^{\frac{1}{2}}, \dots, {\lambda_t}^{\frac{1}{2}}$ entonces la proporción de la varianza conservada es:

$$\frac{\lambda_1 + \cdots + \lambda_t}{\lambda_1 + \cdots + \lambda_s}$$

Generalmente, para el Biplot se conservan dos dimensiones siempre que la proporción de varianza explicada se mantenga alta. La aproximación de rango 2, $\boldsymbol{Y}$, se obtiene simplemente sustituyendo los valores singulares de las posiciones mayores que las que se deseen (en este caso 2) por ceros, recordemos que en la matriz diagonal de $\boldsymbol{X}$ están ordenados descendentemente. De modo que queda que:

$$\boldsymbol{Y} = \begin{pmatrix} u_{11} & u_{12} \\ \dots & \dots \\ u_{n1} & u_{n2} \end{pmatrix} \begin{pmatrix} {\lambda_1}^{1/2} & 0 \\ 0 & {\lambda_2}^{1/2} \end{pmatrix} \begin{pmatrix} v_{11} & \dots & v_{D1} \\ v_{12} & \dots & u_{D2} \end{pmatrix}$$

Conservando la proporción de varianza siguiente:

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{s} \lambda_i}$$

Pero para obtener el Biplot debemos expresar esta matriz $\boldsymbol{Y}$ como producto de otras, a las cuales llamaremos $\boldsymbol{G}$ y $\boldsymbol{H}$, de modo que $\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{H}'$. La matriz $\boldsymbol{G}$ va a tener una dimensión de (n, 2) y la matriz $\boldsymbol{H}$ de (D, 2). Para obtener estas matrices hay multitud de formas como por ejemplo:

$$\mathbf{Y} = \begin{pmatrix} \sqrt{n-1}(\sqrt{\lambda_1})^{\alpha}\, u_{11} & \sqrt{n-1}\,(\sqrt{\lambda_2})^{\alpha} u_{12} \\ \dots & \dots \\ \sqrt{n-1}(\sqrt{\lambda_1})^{\alpha}\, u_{n1} & \sqrt{n-1}\,(\sqrt{\lambda_2})^{\alpha} u_{n2} \end{pmatrix} \begin{pmatrix} \frac{(\sqrt{\lambda_1})^{1-\alpha}\, v_{11}}{\sqrt{n-1}} & \dots & \frac{(\sqrt{\lambda_1})^{1-\alpha}\, v_{1D}}{\sqrt{n-1}} \\ \frac{(\sqrt{\lambda_2})^{1-\alpha}\, v_{21}}{\sqrt{n-1}} & \dots & \frac{(\sqrt{\lambda_2})^{1-\alpha}\, v_{2D}}{\sqrt{n-1}} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \\ \dots \\ \boldsymbol{g}_n \end{pmatrix} (\boldsymbol{h}_1 \quad \boldsymbol{h}_2 \quad \dots \quad \boldsymbol{h}_D)$$

Los vectores $(\boldsymbol{g}_1, \dots, \boldsymbol{g}_n)$ se denominan marcadores de fila y corresponden a las proyecciones de las n muestras sobre el plano definido por los dos primeros vectores propios de $\boldsymbol{Z}\boldsymbol{Z}'$. Por otro lado, los vectores $(\boldsymbol{h}_1, \dots, \boldsymbol{h}_D)$ se denominan marcadores de columna y corresponden a las proyecciones de las partes D clr en el plano definido por los dos primeros vectores propios de $\boldsymbol{Z}'\boldsymbol{Z}$.

El Biplot consigue representar todos los vectores en un plano, pudiendo superponerlos para visualizar la relación entre las muestras y las piezas.

INTERPRETACIÓN

Un Biplot para datos composicionales consta de las siguientes partes (Pawlowsky-Glahn et al., 2011):

- Un origen O que representa el centro del conjunto de datos composicionales.
- Un vértice con posición en $\boldsymbol{h}_i$ para cada una de las D partes.
- Un punto con posición en $\boldsymbol{g}_i$ para cada una de las n muestras.

Se conoce a la unión de O con un vértice $\boldsymbol{h}_i$ como el rayo $\overline{O\boldsymbol{h}_s}$ y a la unión de dos vértices $\boldsymbol{h}_i$ y $\boldsymbol{h}_k$ como el enlace o link $\boldsymbol{h}_i\boldsymbol{h}_k$.

Existen unas propiedades principales para la interpretación de la variabilidad composicional:

1. Los links y los rayos proporcionan información sobre la variabilidad relativa de un conjunto de datos composicionales de modo que:

$$\overline{\boldsymbol{h}_s\boldsymbol{h}_k}^{\,2} \approx Var\left(\ln\frac{x_s}{x_k}\right) \qquad \text{y} \qquad \overline{\boldsymbol{O}\boldsymbol{h}_s}^{\,2} \approx Var\left(\ln\frac{x_s}{g(x)}\right)$$

Hay que tener cuidado al interpretar los rayos ya que dependen de la composición completa debido a la presencia de *g(x)* y varía cuando se considera una subcomposición.

2. Los links proporcionan información sobre la correlación de las subcomposiciones. Si los links $\overline{\boldsymbol{h}_s\boldsymbol{h}_k}$ y $\overline{\boldsymbol{h}_r\boldsymbol{h}_l}$ se cruzan en *A*, entonces

$$\cos(\boldsymbol{h}_s A \boldsymbol{h}_r) \approx corr\left(\ln\frac{x_s}{x_k}, \ln\frac{x_r}{x_l}\right).$$

   En el caso de que los links generen un ángulo recto entonces $\cos(\boldsymbol{h}_s A \boldsymbol{h}_r) \approx 0$ y por tanto se espera una correlación de 0.

3. Análisis subcomposicional:

   El origen *O* es el centroide de los *D* vértices $\boldsymbol{h}_1, \dots, \boldsymbol{h}_D$ y por ello los cocientes se mantienen para subcomposiciones del propio conjunto.
   Por esto, los Biplot de subcomposiciones se forman seleccionando los vértices correspondientes a las partes de dicha subcomposición y el centro *O* se desplaza al centroide de dichos vértices seleccionados.

4. Vértices coincidentes:

   Si dos vértices coinciden entonces su $Var\left(\ln\frac{x_s}{x_k}\right) = 0$ y por tanto $\frac{x_s}{x_k}$ tiene que ser una constante.

   Cuando esto sucede se puede suponer que x_s y x_k son redundantes debido a que sí coinciden y la proporción de la varianza capturada por el Biplot no es muy alta, esto sugiere que $\ln\frac{x_s}{x_k}$ es ortogonal al plano del Biplot, lo que puede indicar una posible independencia de ese logaritmo y las dos primeras direcciones principales de la descomposición del valor singular (que marcan la dirección del Biplot).

5. Vértices colineales:

   Si un subconjunto de vértices es colineal, se puede decir que la subcomposición asociada tiene un Biplot unidimensional, lo que puede indicar que la subcomposición tiene variabilidad unidimensional, es decir, las composiciones se sitúan a lo largo de una línea de composición.

Con las propiedades anteriores se puede llegar a la conclusión de que los enlaces son los componentes centrales de un Biplot. La longitud de estos corresponde aproximadamente a la varianza de las relaciones logarítmicas simples entre elementos individuales.

En la interpretación de un Biplot se describe su geometría interna por lo que no resulta afectada por cualquier rotación o imagen de espejo de los datos.

EJEMPLO

Cálculos

Teniendo la siguiente matriz de datos composicionales primero deberemos centralizarla, ya calculado en (7) de modo que:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 2 & 3 \end{pmatrix}, \hat{g}(\boldsymbol{X}) = (0.23, 0.43, 0.34), \boldsymbol{X}^C = \begin{pmatrix} 0.29 & 0.527 & 0.183 \\ 0.467 & 0.141 & 0.392 \\ 0.212 & 0.386 & 0.402 \end{pmatrix}$$

A partir de $\boldsymbol{X}^C$ debemos calcular la matriz $\boldsymbol{Z}$,

$$\boldsymbol{Z} = \begin{pmatrix} \ln \frac{X_{11}^C}{g(\boldsymbol{X}_1^C)} & \cdots & \ln \frac{X_{1D}^C}{g(\boldsymbol{X}_1^C)} \\ \vdots & \ddots & \vdots \\ \ln \frac{X_{n1}^C}{g(\boldsymbol{X}_n^C)} & \cdots & \ln \frac{X_{nD}^C}{g(\boldsymbol{X}_n^C)} \end{pmatrix} = \begin{pmatrix} clr(\boldsymbol{X}_1^C) \\ clr(\boldsymbol{X}_2^C) \\ clr(\boldsymbol{X}_3^C) \end{pmatrix}$$

Recordando las fórmulas, podremos obtener los coeficientes de sus coordenadas clr de forma que:

$$clr(\boldsymbol{X}_i^C) = \left(\ln \frac{\boldsymbol{X}_{11}^C}{g(\boldsymbol{x})}, \ln \frac{\boldsymbol{X}_i^C}{g(\boldsymbol{x})}, \ldots, \ln \frac{\boldsymbol{X}_i^C}{g(\boldsymbol{x})} \right)$$

$$g(\boldsymbol{X}_i^C) = \left[\prod_{j=1}^{D} x_i \right]^{1/D}$$

$$\hat{g}(\boldsymbol{X}_1^C) = (0.29 \cdot 0.527 \cdot 0.183)^{\frac{1}{3}} = 0.306$$

$$clr(\boldsymbol{X}_1^C) = clr(0.29, 0.527, 0.183) = \left(\ln \frac{0.29}{0.306}, \ln \frac{0.527}{0.306}, \ln \frac{0.183}{0.306} \right) = (-0.045, 0.552, -0.507)$$

$$\hat{g}(\boldsymbol{X}_2^C) = (0.467 \cdot 0.141 \cdot 0.391)^{\frac{1}{3}} = 0.295$$

$$clr(\boldsymbol{X}_2^C) = clr(0.467, 0.141, 0.391) = \left(\ln \frac{0.467}{0.295}, \ln \frac{0.141}{0.295}, \ln \frac{0.391}{0.295} \right) = (0.456, -0.738, 0.282)$$

$$\hat{g}(\boldsymbol{X}_3^C) = (0.212 \cdot 0.386 \cdot 0.402)^{\frac{1}{3}} = 0.32$$

$$clr(\boldsymbol{X}_3^C) = clr(0.212, 0.386, 0.402) = \left(\ln \frac{0.212}{0.32}, \ln \frac{0.386}{0.32}, \ln \frac{0.402}{0.32} \right) = (-0.411, 0.186, 0.225)$$

Por lo que,

$$\boldsymbol{Z} = \begin{pmatrix} clr(\boldsymbol{X}_1^C) \\ clr(\boldsymbol{X}_2^C) \\ clr(\boldsymbol{X}_3^C) \end{pmatrix} = \begin{pmatrix} -0.045 & 0.552 & -0.507 \\ 0.456 & -0.738 & 0.282 \\ -0.411 & 0.186 & 0.225 \end{pmatrix}$$

Ahora debemos expresar esta matriz, $\boldsymbol{Z}$, como combinación lineal de:

$$\boldsymbol{Z} = \boldsymbol{U}\begin{pmatrix} {\lambda_1}^{\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & {\lambda_2}^{\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & {\lambda_s}^{\frac{1}{2}} \end{pmatrix}\boldsymbol{V}'$$

Recordemos que:

1. Las raíces cuadradas de los s valores propios positivos $\lambda_1, \dots, \lambda_s$ de $\boldsymbol{ZZ}'$ o $\boldsymbol{Z}'\boldsymbol{Z}$, con $s \leq \min\{(D-1), n\}$ siendo el rango de $\boldsymbol{Z}$ y estando $\lambda_1, \dots, \lambda_s$ en orden descendente de magnitud.
2. La matriz de vectores propios de $\boldsymbol{ZZ}'$, $\boldsymbol{U}$, con dimensión (*n, s*).
3. La matriz de vectores propios de $\boldsymbol{Z}'\boldsymbol{Z}, \boldsymbol{V}$, con dimensión (*D, s*).

<u>Cálculos</u>:

1. Matriz de valores propios positivos:

   Matriz $\boldsymbol{ZZ}'$:

$$\begin{pmatrix} -0.045 & 0.552 & -0.507 \\ 0.456 & -0.738 & 0.282 \\ -0.411 & 0.186 & 0.225 \end{pmatrix}\begin{pmatrix} -0.045 & 0.456 & -0.411 \\ 0.552 & -0.738 & 0.186 \\ -0.507 & 0.282 & 0.225 \end{pmatrix} = \begin{pmatrix} 0.564 & -0.571 & 0.007 \\ -0.571 & 0.833 & -0.261 \\ 0.007 & -0.261 & 0.254 \end{pmatrix}$$

   Calculamos los valores propios de dicha matriz:

$$\det(\boldsymbol{ZZ}' - \lambda I) = \det\begin{pmatrix} 0.564-\lambda & -0.571 & 0.007 \\ -0.571 & 0.833-\lambda & -0.261 \\ 0.007 & -0.261 & 0.254-\lambda \end{pmatrix} = -\lambda^3 + 1.651\lambda^2 - 0.43\lambda + 0.00014$$

   Calculamos los valores propios igualando la ecuación a cero:

$$-\lambda^3 + 1.651\lambda^2 - 0.43\lambda + 0.00014 = 0$$

   Valores propios ordenados descendentemente:

$$\lambda_1 = 1.327 \quad \lambda_2 = 0.324 \quad \lambda_3 = 0.00033$$

   Por lo que,

$$\begin{pmatrix} {\lambda_1}^{\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & {\lambda_2}^{\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & {\lambda_s}^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} 1.327^{\frac{1}{2}} & 0 & 0 \\ 0 & 0.324^{\frac{1}{2}} & 0 \\ 0 & 0 & 0.00033^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} 1.152 & 0 & 0 \\ 0 & 0.5695 & 0 \\ 0 & 0 & 0.01816 \end{pmatrix}$$

2. Matriz $\boldsymbol{U}$, matriz de vectores propios de $\boldsymbol{ZZ}'$:

$$\boldsymbol{U} = \begin{pmatrix} -0.589 & -0.565 & 0.577 \\ 0.784 & -0.228 & 0.577 \\ -0.195 & 0.793 & 0.577 \end{pmatrix}$$

3. Matriz $\boldsymbol{V}$, matriz de vectores propios de $\boldsymbol{Z'Z}$:

   Matriz $\boldsymbol{Z'Z}$:

$$\begin{pmatrix} -0.045 & 0.456 & -0.411 \\ 0.552 & -0.738 & 0.186 \\ -0.507 & 0.282 & 0.225 \end{pmatrix}\begin{pmatrix} -0.045 & 0.552 & -0.507 \\ 0.456 & -0.738 & 0.282 \\ -0.411 & 0.186 & 0.225 \end{pmatrix} = \begin{pmatrix} 0.379 & -0.438 & 0.059 \\ -0.438 & 0.884 & -0.446 \\ 0.059 & -0.446 & 0.387 \end{pmatrix}$$

$$\boldsymbol{V} = \begin{pmatrix} 0.403 & -0.709 & 0.577 \\ -0.816 & 0.006 & 0.577 \\ 0.413 & 0.705 & 0.577 \end{pmatrix}, \quad \boldsymbol{V'} = \begin{pmatrix} 0.403 & -0.816 & 0.413 \\ -0.709 & 0.006 & 0.705 \\ 0.577 & 0.577 & 0.577 \end{pmatrix}$$

Así que tenemos que la matriz $\boldsymbol{Z}$ se descompone en:

$$\boldsymbol{Z} = \begin{pmatrix} -0.589 & -0.565 & 0.577 \\ 0.784 & -0.228 & 0.577 \\ -0.195 & 0.793 & 0.577 \end{pmatrix}\begin{pmatrix} 1.152 & 0 & 0 \\ 0 & 0.5695 & 0 \\ 0 & 0 & 0.01816 \end{pmatrix}\begin{pmatrix} 0.403 & -0.816 & 0.413 \\ -0.709 & 0.006 & 0.705 \\ 0.577 & 0.577 & 0.577 \end{pmatrix}$$

Ahora procedemos a realizar la mejor aproximación de rango 2 de la matriz $\boldsymbol{Z}$, $\boldsymbol{Y}$, que se obtiene tomando los dos valores propios superiores de modo que:

$$\boldsymbol{Y} = \begin{pmatrix} u_{11} & u_{12} \\ \cdots & \cdots \\ u_{n1} & u_{n2} \end{pmatrix}\begin{pmatrix} {\lambda_1}^{\frac{1}{2}} & 0 \\ 0 & {\lambda_2}^{\frac{1}{2}} \end{pmatrix}\begin{pmatrix} v_{11} & \cdots & v_{D1} \\ v_{12} & \cdots & u_{D2} \end{pmatrix}$$

Así que,

$$diag = \begin{pmatrix} 1.327^{\frac{1}{2}} & 0 \\ 0 & 0.324^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} 1.152 & 0 \\ 0 & 0.5695 \end{pmatrix}$$

$$\boldsymbol{U} = \begin{pmatrix} -0.589 & -0.565 \\ 0.784 & -0.228 \\ -0.195 & 0.793 \end{pmatrix}$$

$$\boldsymbol{V} = \begin{pmatrix} 0.403 & -0.816 & 0.413 \\ -0.709 & 0.006 & 0.705 \end{pmatrix}$$

$$\boldsymbol{Y} = \begin{pmatrix} -0.589 & -0.565 \\ 0.784 & -0.228 \\ -0.195 & 0.793 \end{pmatrix}\begin{pmatrix} 1.152 & 0 \\ 0 & 0.5695 \end{pmatrix}\begin{pmatrix} 0.403 & -0.816 & 0.413 \\ -0.709 & 0.006 & 0.705 \end{pmatrix}$$

Podemos calcular que cantidad de la varianza total explicada por $\boldsymbol{X}$, la totvar[$\boldsymbol{X}$] retiene la aproximación $\boldsymbol{Y}$ con la siguiente fórmula:

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{s} \lambda_i} = \frac{1.327 + 0.324}{1.327 + 0.324 + 0.00033} \approx 1$$

Para el Biplot debemos expresar esta matriz $\boldsymbol{Y}$ como producto de otras, $\boldsymbol{G}$ y $\boldsymbol{H}$, de dimensiones (n, 2) y (D, 2) respectivamente. Dichas matrices se calculan tal que:

$$\mathbf{Y} = \begin{pmatrix} \sqrt{n-1}(\sqrt{\lambda_1})^{\alpha}\, u_{11} & \sqrt{n-1}\,(\sqrt{\lambda_2})^{\alpha} u_{12} \\ \dots & \dots \\ \sqrt{n-1}(\sqrt{\lambda_1})^{\alpha}\, u_{n1} & \sqrt{n-1}\,(\sqrt{\lambda_2})^{\alpha} u_{n2} \end{pmatrix} \begin{pmatrix} \frac{(\sqrt{\lambda_1})^{1-\alpha}\, v_{11}}{\sqrt{n-1}} & \dots & \frac{(\sqrt{\lambda_1})^{1-\alpha}\, v_{1D}}{\sqrt{n-1}} \\ \frac{(\sqrt{\lambda_2})^{1-\alpha}\, v_{21}}{\sqrt{n-1}} & \dots & \frac{(\sqrt{\lambda_2})^{1-\alpha}\, v_{2D}}{\sqrt{n-1}} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \\ \dots \\ \boldsymbol{g}_n \end{pmatrix} (\boldsymbol{h}_1 \quad \boldsymbol{h}_2 \quad \dots \quad \boldsymbol{h}_D)$$

Siendo α una constante que toma valor entre 0 y 1. Para obtener $\boldsymbol{Y}$ vamos a realizar los cálculos con $\alpha = 0.5$ de modo que las raíces cuadradas de los valores singulares se dividan por igual entre los vectores singulares izquierdo y derecho. Al usar ese valor para α estamos realizando un HJ-Biplot, que con sus propiedades consigue que:

- Los marcadores fila y los columna se puedan representar en el mismo sistema de referencia.
- Las calidades de representación de filas y columnas son las mismas

Si se utilizase un $\alpha = 0$ estaríamos realizando un GH-Biplot que entre otras cosas consigue una mejor calidad en la representación de las variables o columnas.

Por el contrario, si utilizásemos un $\alpha = 1$ estaríamos realizando un JK-Biplot que consigue una mejor representación de los individuos o filas.

Realizando las cuentas con $\alpha = 0.5$ nuestra matriz quedaría tal que:

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \\ \dots \\ \boldsymbol{g}_n \end{pmatrix} = \begin{pmatrix} \sqrt{3-1}\left(\sqrt{1.327}\right)^{0.5}(-0.589) & \sqrt{3-1}\left(\sqrt{0.5695}\right)^{0.5}(-0.565) \\ \sqrt{3-1}\left(\sqrt{1.327}\right)^{0.5} 0.784 & \sqrt{3-1}\left(\sqrt{0.5695}\right)^{0.5}(-0.228) \\ \sqrt{3-1}\left(\sqrt{1.327}\right)^{0.5}(-0.195) & \sqrt{3-1}\left(\sqrt{0.5695}\right)^{0.5} 0.793 \end{pmatrix}$$

$$= \begin{pmatrix} -0.863 & -0.695 \\ 1.149 & -0.28 \\ -0.286 & 0.974 \end{pmatrix}$$

$$\boldsymbol{H} = (\boldsymbol{h}_1 \quad \boldsymbol{h}_2 \quad \dots \quad \boldsymbol{h}_D)$$

$$= \begin{pmatrix} \frac{\left(\sqrt{1.327}\right)^{0.5} 0.403}{\sqrt{3-1}} & \frac{\left(\sqrt{1.327}\right)^{0.5}(-0.816)}{\sqrt{3-1}} & \frac{\left(\sqrt{1.327}\right)^{0.5} 0.413}{\sqrt{3-1}} \\ \frac{\left(\sqrt{0.5695}\right)^{0.5}(-0.709)}{\sqrt{3-1}} & \frac{\left(\sqrt{0.5695}\right)^{0.5} 0.006}{\sqrt{3-1}} & \frac{\left(\sqrt{0.5695}\right)^{0.5} 0.705}{\sqrt{3-1}} \end{pmatrix}$$

$$= \begin{pmatrix} 0.295 & -0.52 & 0.355 \\ -0.502 & 0.003 & 0.355 \end{pmatrix}$$

Las matrices $\boldsymbol{G}$ y $\boldsymbol{H}$ se utilizarán para representar el Biplot según lo explicado:

- Se representará un origen, *O*, que corresponderá con el centro del conjunto de datos composicionales.
- Habrá un vértice con posición en $\boldsymbol{h}_i$ para cada una de las *D* partes.
- Habrá un punto con posición en $\boldsymbol{g}_i$ para cada una de las *n* muestras.

Visualización

Veamos realmente cómo se ve un Biplot. Para ello vamos a utilizar un conjunto nuevo de datos aportados por (*CoDaWeb*, s. f.) y expuestos en el apartado Anexo. Además, se utilizará el programa R con los paquetes compositions, HardyWeinberg y calibrate. Para evitar una gran cantidad de cálculos utilizamos funciones de estos los paquetes mencionados.

No se va a realizar el Biplot de los datos composicionales utilizados en los cálculos anteriores ya que son escasos y no permiten hacer una buena visualización o interpretación.

En primer lugar cargamos los datos y los definimos como dataframe. Los datos se han obtenido del paquete "compositions" y aportan información de datos geoquímicos del Jura suizo.

Como podemos observar en la tercera línea de código, los datos tienen 359 muestras (o datos composicionales) y 11 variables que con la siguiente línea vemos sus nombres. Las cuatro primeras son de clasificación por lo que no se deben seleccionar para formar una parte de la matriz de datos composicionales. Por ello para crear la matriz X se seleccionan las 7 últimas columnas y después calculamos el cierre de la matriz $\boldsymbol{X}$, $\boldsymbol{X.com}$.

```
data(juraset)
juraset <- as.data.frame(juraset)
dim(juraset)
## [1] 359 11
colnames(juraset)
## [1] "X"   "Y"   "Rock"   "Land"   "Cd"   "Cu"   "Pb"   "Co"   "Cr"   "Ni"   "Zn"
X <- as.matrix(juraset[5:11])
X.com <- acomp(X)
```

Después debemos transformar el conjunto de datos aplicando la función log de modo que obtenemos los logaritmos naturales de todas las partes de todo el conjunto de datos composicionales.

Además, como se ha explicado anteriormente, centraremos los datos y los estandarizaremos pero no en vez de por filas (dato composicional) por variables así que debemos transponer primero, aplicar la estandarización y volver a transponer para volver a tener un dato composicional por fila.

Por último, aplicaremos la función "princomp" para realizar el Análisis de Componentes Principales sobre los datos y obtener la mejor aproximación de rango 2 y poder visualizar el Biplot.

```
lX <- log(X)
Xclr <- t(scale(t(lX), center=TRUE, scale=FALSE))
pca.results <- princomp(Xclr,cor=FALSE)
```

Una vez realizadas estas transformaciones procedemos a plotear el Biplot.

```
biplot(pca.results, var.axes=TRUE),col=c("#3C87B9","black"),cex=c(2,0),
       main="       ",xlab="  ", ylab="  ",xlabs = rep(".", 359))
abline(h=0, col="black", lty=4)
abline(v=0, col="black", lty=4)
```

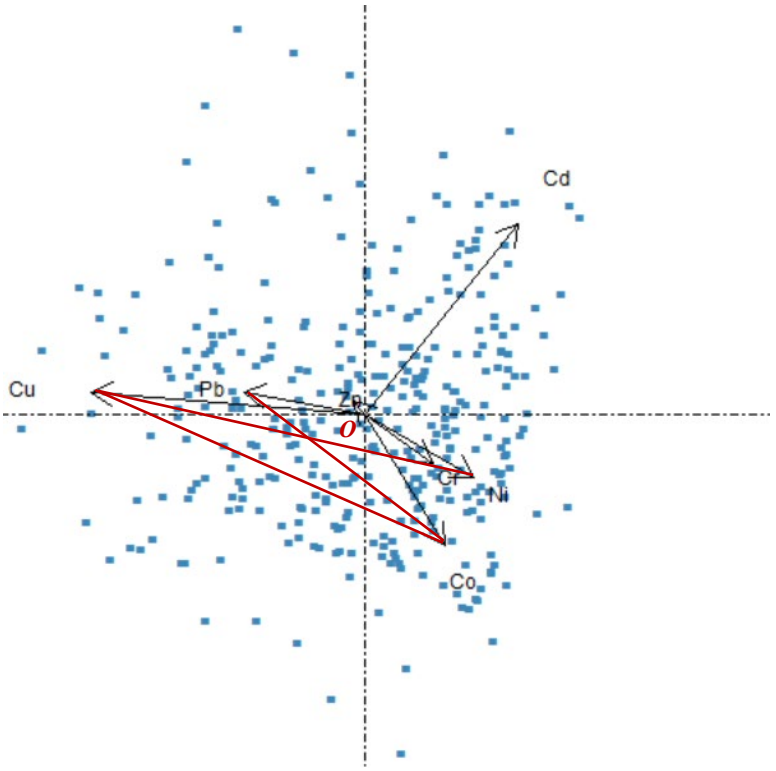

Gráfico 5. Biplot de datos composicionales

Expliquemos visualmente las partes que hemos definido anteriormente:

- Un origen *O* que se encuentra en el centro del conjunto de datos composicionales y del cuál salen todos los rayos.
- Un vértice por cada una de las *D* partes. Estos vértices consisten en un punto que resulta el extremo de los rayos que los unen al origen. En el *Gráfico 5* vemos 7 vértices: Cu, Pb, Zn, Cd, Cr, Ni y Co.
- Un punto para cada una de las *n* muestras, en el *Gráfico 5* hay representados con puntos los 36 datos composicionales.
- Los rayos del Biplot están representados en negro y unen el origen con cada uno de los vértices. Por ello, hay representados 7 rayos que corresponden a cada una de las variables.
- Por último, se ha representado en color rojo algunos de los enlaces o link, que corresponden con la unión entre dos vértices.

Teniendo en cuenta la serie de propiedades mencionadas en el apartado de Interpretación podemos decir que:

1. La longitud de los rayos corresponde a la varianza de las relaciones logarítmicas simples entre elementos individuales. En este caso todos los rayos con mayor longitud son Cu, Co y Cd. Además, el rayo correspondiente con Ni es muy corto, lo que indica que esta pobremente representado.

   Cabe destacar que es importante también observar la longitud de los enlaces. Estos corresponden con la desviación estándar de los correspondientes log-cocientes. Observando el *Gráfico 5* vemos que tienen gran variabilidad $\frac{Cu}{Co}, \frac{Cu}{Cd}, \frac{Co}{Cd}$, que coinciden con los vértices que presentaban mayores longitudes de rayos.

2. En el caso de que los algunos links generen un ángulo, el coseno de este indicaría la correlación entre los log-cocientes de las partes que formen dichos links.

   Si se diese el caso de que se forma un ángulo recto entonces se espera una correlación de 0 entre los log-cocientes de las partes, es decir, no existiría relación entre dichos ratios.

   En este caso, si representásemos todos los links veríamos que alguno se cruza como: $\frac{Pb}{Co}, \frac{Cu}{Ni}$. Sin embargo, ninguno lo hace generando un ángulo recto.

4. No hay presencia de vértices que sean coincidentes por lo que no hay partes composicionales que se puedan considerar redundantes o no aporten ninguna información.

5. Tampoco hay presencia de vértices colineales.

CONCLUSIONES

Como se podrá intuir tras la lectura de este trabajo, poder manejar de forma correcta los datos composicionales es fundamental. Permite una correcta interpretación de estos y la posibilidad de analizarlos llegando a conclusiones válidas en las diversas áreas donde se pueden generar este tipo de datos, geología, química, áreas de la salud, etc.

Con lo que se ha explicado de los datos composicionales, durante los últimos 60 años se ha ampliado mucho los conocimientos sobre ellos, generando las bases para su estructuración en forma de espacio vectorial. En este trabajo se ha comprobado dicha estructuración, comenzando por definir en el símplex unas operaciones específicas similares a las que tenemos en $\mathbb{R}^D$: la operación interna de perturbación y la operación externa de potenciación. Con ello le hemos dotado de estructura de cuerpo. También se ha demostrado que es posible definir la distancia de Aitchison, lo que le dota además de esa estructura de espacio vectorial que posee.

Por otra parte, se han definido las transformaciones log-cocientes y se ha comprobado que es más costoso encontrar transformaciones que no sean muy complejas y que sí sean isométricas, para que las distancias euclídeas que tenemos en $\mathbb{R}^D$sean trasladables a distancias en el símplex. De las trasformaciones mencionadas, la transformación clr es la más completa, sin embargo, otra muy completa y frecuentemente utilizada es la transformación ilr que no se ha comentado en este trabajo por la complejidad y extensión que conlleva.

Ahora bien, desde una perspectiva estadística, es muy importante poder tratar los datos composicionales de manera correcta. Por las propiedades del símplex no es posible aplicar directamente los conceptos básicos de estadística multivariante como la media de un vector o la varianza de un conjunto de datos ya que, al pasar a usar datos composicionales hablamos de muestras de vectores, de manera que ahora se considerarán otras variables como similares. Algunas de estas se han definido en el trabajo como el centro de la muestra, que representa lo que representaría la media muestral en unas variables habituales.

Aunque no se presenta en este trabajo, como extensión es necesario continuar el estudio de los datos composicionales. Esto se debe a que hay multitud de técnicas como Análisis de Componentes Principales o el Análisis de clusters que no se conocen en el símplex y que, para aplicarlos actualmente, es necesario trasladar los datos composicionales a $\mathbb{R}^D$, aplicar los análisis y por transformación por anti imagen observar dichos resultados en el símplex para poder llegar a conclusiones.

Aun con la gran evolución del entendimiento sobre los datos composicionales, hay multitud de cosas así como problemas sobre ellos que todavía queda por estudiar y extender. Debido a la longitud de este trabajo, hay multitud de aspectos sobre los datos composicionales que no se han podido mencionar o detallar.

BIBLIOGRAFÍA

Aitchison, J. (1981). A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, *13*(2), 175-189. https://doi.org/10.1007/BF01031393

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Mathematical Geology*, *13*, 175-189.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, *70*(1), 57-65.

Aitchison, J. (1984). The statistical analysis of geochemical compositions. *Journal of the International Association for Mathematical Geology*, *16*(6), 531-564. https://doi.org/10.1007/BF01029316

Aitchison, J. (1986). *A Concise Guide to Compositional Data Analysis*.

Aitchison, J. (1992). The triangle in statistcs. *The Art of Statistical Science.*, *8*, 89-104.

Aitchison, J. (1994). Principles of compositional data analysis. En *Multivariate Analysis and Its Applications* (Vol. 24, pp. 73-81). Institute of Mathematical Statistics. https://doi.org/10.1214/lnms/1215463786

Aitchison, J., & Pawlowsky-Glahn, V. (1997). *The one-hour course in compositional data analysis or compositional data analysis is simple. 97*, 3-35.

Aitchison, J., & Shen, S. M. (1980). Logistic-Normal Distributions: Some Properties and Uses. *Biometrika*, *67*(2), 261-272. https://doi.org/10.2307/2335470

Aitchison, J., & Shen, S. M. (1984). Measurement error in compositional data. *Journal of the International Association for Mathematical Geology*, *16*(6), 637-650. https://doi.org/10.1007/BF01029322

Albaladejo, J. P., Martin Fernández, J. A., & García, J. G. (s. f.). *MODELIZACIÓN Y ANÁLISIS DE DATOS SOBRE PROPORCIONES* (p. 20).

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research*, *65*(12), 4185-4193. https://doi.org/10.1029/JZ065i012p04185

CoDaWeb. (s. f.). Recuperado 2 de mayo de 2022, de http://www.compositionaldata.com/
Datos Composicionales: Vol. Capitulo 3. (s. f.).

Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of Parts and Their Balances in Compositional Data Analysis. *Mathematical Geology*, *37*(7), 795-828. https://doi.org/10.1007/s11004-005-7381-9

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barcelo-Vidal, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, *35*(3), 22.

Fišerová, E., & Hron, K. (2011). On the Interpretation of Orthonormal Coordinates for Compositional Data. *Mathematical Geosciences*, *43*(4), 455-468. https://doi.org/10.1007/s11004-011-9333-x

Greenacre, M. (2019). *Compositional Data Analysis in Practice*.

Krzanowski, W. J. (1988). MISSING VALUE IMPUTATION IN MULTIVARIATE DATA USING THE SINGULAR VALUE DECOMPOSITION OF A MATRIX. *Biometrical Letters*, *25*(1,2), 31-39.

Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2000). *Zero replacement in compositional data sets. Studies in Classification, Data Analysis, and Knowledge Organization*. 155-160.

Martín-Fernandez, J. A., Barcelo-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology*, *35*(3), 26.

Martín-Fernández, J. A., Bren, M., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (1999). A measure of difference for compositional data based on measures of divergence. *Lippard er al.*, 211-216.

Mateu-Figueras, G., Martín-Fernandez, J. A., Pawlowsky-Glahn, V., & Barceló-Vidal, C. (2003). *EL PROBLEMA DEL ANÁLISIS ESTADÍSTICO DE DATOS COMPOSICIONALES*. 9.

McAlister, D. (1879). The Law of the Geometric Mean. *Proceedings of the Royal Society of London*, *29*, 367-376.

Pawlowsky-Glahn, V., & Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, *15*(5), 384-398. https://doi.org/10.1007/s004770100077

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2011). *Lecture Notes on Compositional Data Analysis*. 108.

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*, *60*(359-367), 489-498. https://doi.org/10.1098/rspl.1896.0076

Tolosana-Delgado, R. (2011). Guía para el análisis espacial de datos composicionales. *Boletín Geológico y Minero*, *122*(4), 469-482.

ANEXO

BASES DE DATOS

- Portada:
  Base de Datos "ArcticLake" proporcionada por el paquete "compositions" de R.

- *Gráfico 2, 3, 4*:
  Base de Datos que recoge datos del grupo sanguíneo aportada por (*CoDaWeb*, s. f.) y publicada en el Departamento de Estadística e Investigación Operativa de la Universitat Politècnica de Catalunya. http://www-eio.upc.es/~jan/IBC/BloodMN.csv

- *Gráfico 5*:
  Base de Datos "juraset" proporcionada por el paquete "compositions" de R.

LÍNEAS DE CÓDIGO

GRÁFICOS

- Portada:

```
data(ArcticLake)
colnames(ArcticLake)

## [1] "sand"  "silt"  "clay"  "depth"

Bc<-acomp(ArcticLake[,c("silt","sand","depth")])
plot(Bc,center=TRUE,labels = c(" "," "," "))
isoPortionLines(by=0.2,at=seq(0,1,by=0.1),
                parts=1:3,total=1,labs=FALSE,lines=TRUE,unit="",
                col="#9E9E9E")
plot(Bc,add=TRUE,pch=19,col="#A9CCE3")
```

- Media geométrica y media normal, *Gráfico 2*:

```
X <- read.csv("http://www-eio.upc.es/~jan/IBC/BloodMN.csv",sep=";")
#DIAGRAMA DE DATOS COMPOSICIONALES
Xc <- acomp(X[,c("MM","NN","MN")])
plot(Xc)
isoPortionLines(by=0.2,at=seq(0,1,by=0.1),
                parts=1:3,total=1,labs=FALSE,lines=TRUE,unit="",
                col="#9E9E9E")
plot(Xc,add=TRUE,pch=19,col="#A9CCE3")


#MEDIA GEOMETRICA
gm <- mean (Xc)
gm
##          MM          NN          MN
## "0.3487108" "0.1736861" "0.4776032"
plot(gm,add=TRUE,pch=19,col="#D32F2F")

#MEDIA NORMAL
vector1 <- as.vector(X['MM'])
mean1<-mean(vector1)
```

```
vector2 <- as.vector(X[,'NN'])
mean2<-mean(vector2)

vector3 <- as.vector(X[,'MN'])
class(vector3)
mean3<-mean(vector3)

mean<-c (mean1, mean2, mean3)
mean
## [1] 216.9184 132.4490 315.3673

Xmean <- acomp(mean)
Xmean
## [1] "0.3263232" "0.1992509" "0.4744259"


plot(Xmean,add=TRUE,pch=19,col="#2A3339")
```

- Datos composicionales centralizados y estandarizados:

```
X <- read.csv("http://www-eio.upc.es/~jan/IBC/BloodMN.csv",sep=";")


#DIAGRAMA DATOS COMPOSICIONALES CENTRALIZADOS
plot(Xc,center=TRUE)
plot(Xc,center=TRUE,labels = c("C(MM)","C(NN)","C(MN)"),add=TRUE,pch=19,co
l="#A9CCE3")

isoPortionLines(by=0.2,at=seq(0,1,by=0.1),
                parts=1:3,total=1,labs=FALSE,lines=TRUE,unit="",col="#9E9E
9E")
plot(Xc,add=TRUE,pch=19,col="#A9CCE3")

#Media geométrica
gm <- mean(Xc,center=TRUE)
plot(gm,add=TRUE,pch=19,col="#D32F2F")



#DIAGRAMA DATOS COMPOSICIONALES ESTANDARIZADOS

plot(Xc,scale = TRUE,add = TRUE,pch=19,col="#A9CCE3")

#Media geométrica
Xdatos <- X[,c("MM","NN","MN")]
Xdatos<-as.matrix(Xdatos)


#Matriz de Variación
X.com <- acomp(Xdatos)
TT <- variation(X.com)
#Matriz de Variación Normalizada
NTT<-(1/2)*TT


#Varianza Total
sumatorio<-TT[1,1]+TT[1,2]+TT[1,3]+TT[2,1]+TT[2,2]+TT[2,3]+TT[3,1]+TT[3,2]
```

```
+TT[3,3]
totvar<-(1/6)*sumatorio


XE<-X.com^(1/sqrt(totvar))


gm <- mean(XE)
plot(gm,add=TRUE,pch=19,col="#D32F2F")
```

CÁLCULOS NUMÉRICOS

- Ejemplos numéricos media geométrica, matriz de variación, varianza total, centralización y estandarización:

```
#CREACIÓN DE MATRIZ DE DATOS
a<-c(1,2,1,3,1,4,1,2,3)
X<-matrix(a, nrow = 3, ncol = 3)
X<-t(X)
X<-as.data.frame(X,row.names = NULL)
rownames(X)<-c("x","y","z")

#Datos composicionales
Xc <- acomp(X[,])
Xc

#Media Geométrica
gm <- mean(Xc)
gm

#Matriz de Variación
X.com <- acomp(X)
TT <- variation(X.com)
TT

#Matriz de Variación Normalizada

NTT<-(1/2)*TT
NTT

#Varianza Total

sumatorio<-TT[1,1]+TT[1,2]+TT[1,3]+TT[2,1]+TT[2,2]+TT[2,3]+TT[3,1]+TT[3,2]
+TT[3,3]
totvar<-(1/6)*sumatorio
totvar
```

```
#CENTRALIZACIÓN
XC<-perturbe(Xc,gm^(-1))
XC
```

```
#Comprobación
gm <- mean(XC)
gm



#ESTANDARIZACIÓN
XE<-XC^(1/sqrt(totvar))

XE1<-acomp(XE[1,])
XE2<-acomp(XE[2,])
XE3<-acomp(XE[3,])

XE<-matrix(c(XE1,XE2,XE3),nrow=3,ncol=3)
XE<-t(XE)
XE

#Comprobación
#matriz de variación de los datos estandarizados
x12<-c(log(XE[1,1]/XE[1,2]),log(XE[2,1]/XE[2,2]),log(XE[3,1]/XE[3,2]))
t12<-var(x12)

x13<-c(log(XE[1,1]/XE[1,3]),log(XE[2,1]/XE[2,3]),log(XE[3,1]/XE[3,3]))
t13<-var(x13)

x21<-c(log(XE[1,2]/XE[1,1]),log(XE[2,2]/XE[2,1]),log(XE[3,2]/XE[3,1]))
t21<-var(x21)

x23<-c(log(XE[1,2]/XE[1,3]),log(XE[2,2]/XE[2,3]),log(XE[3,2]/XE[3,3]))
t23<-var(x23)

x31<-c(log(XE[1,3]/XE[1,1]),log(XE[2,3]/XE[2,1]),log(XE[3,3]/XE[3,1]))
t31<-var(x31)

x32<-c(log(XE[1,3]/XE[1,2]),log(XE[2,3]/XE[2,2]),log(XE[3,3]/XE[3,2]))
t32<-var(x32)

TTE<-c(0,t21,t31,t12,0,t32,t13,t23,0)
TTE<-matrix(TTE, nrow=3, ncol=3)
TTE

#cálculo de la variación total
sumatorio<-TTE[1,1]+TTE[1,2]+TTE[1,3]+TTE[2,1]+TTE[2,2]+TTE[2,3]+TTE[3,1]+
TTE[3,2]+TTE[3,3]
totvar<-(1/6)*sumatorio
totvar
```

- Ejemplo numérico Biplot:

```
#CREACIÓN DE MATRIZ DE DATOS
a<-c(1,2,1,3,1,4,1,2,3)
X<-matrix(a, nrow = 3, ncol = 3)
X<-t(X)
X<-as.data.frame(X,row.names = NULL)
rownames(X)<-c("x","y","z")
```

```
#Datos composicionales
Xc <- acomp(X[,])

#Media geométrica
gm <- mean(Xc)
gm

#Datos composicionales centralizados
XC<-perturbe(Xc,gm^(-1))
XC

#MATRIZ Z
Z <- clr(XC)
Z

#MATRIZ ZZT
ZZT<-tcrossprod(Z, Z)
ZZT

#Valores propios de ZZT
vp_ZZT<-eigen(ZZT)$values
vp_ZZT

#MATRIZ ZTZ
ZTZ<-crossprod(Z, Z)
ZTZ

#DESCOMPOSICIÓN DE VALORES PROPIOS
XC.svd <- svd(Z)

diag<-XC.svd$d
diag<-diag(diag)
diag

U<-XC.svd$u
U

V<-XC.svd$v
V

t(V)

#CREACIÓN DE Y
diagY<-matrix(c(diag[1,1],0,0,diag[2,2]),
              nrow=2,ncol=2)
diagY

UY<-matrix(c(U[1,1],U[2,1],U[3,1],
             U[1,2],U[2,2],U[3,2]),nrow=3,ncol=2)
UY

VY<-matrix(c(V[1,1],V[2,1],V[1,2],
             V[2,2],V[1,3],V[2,3]),nrow=2,ncol=3)
VY

#VARIANZA RETENIDA POR Y
(diagY[1,1]^2+diagY[2,2]^2)/(diag[1,1]^2+diag[2,2]^2+diag[3,3]^2)
```

```
#DESCOMPOSICION Y
alpha<-0.5
vp<-c(diagY[1,1],diagY[2,2])

G<-matrix(c(sqrt(3-1)*((sqrt(vp[1]))^alpha)*UY[1,1],sqrt(3-1)*(sqrt(vp[1])
)^alpha*UY[2,1],
            sqrt(3-1)*(sqrt(vp[1]))^alpha*UY[3,1],sqrt(3-1)*(sqrt(vp[2]))^
alpha*UY[1,2],
            sqrt(3-1)*(sqrt(vp[2]))^alpha*UY[2,2],sqrt(3-1)*(sqrt(vp[2]))^
alpha*UY[3,2]),nrow=3,ncol=2)


G

H<-matrix(c((((sqrt(vp[1]))^(1-alpha))*VY[1,1])/sqrt(3-1),
            (((sqrt(vp[2]))^(1-alpha))*VY[2,1])/sqrt(3-1),
            (((sqrt(vp[1]))^(1-alpha))*VY[1,2])/sqrt(3-1),
            (((sqrt(vp[2]))^(1-alpha))*VY[2,2])/sqrt(3-1),
            (((sqrt(vp[2]))^(1-alpha))*VY[1,3])/sqrt(3-1),
            (((sqrt(vp[2]))^(1-alpha))*VY[2,3])/sqrt(3-1)),nrow=2,ncol=3)

H
```

ABSTRACT

Compositional data consists of non-negative vectors with the property that their values sum to a constant. Due to this of this special feature, it is necessary to handle and interpret compositional data differently from how vectors without such a constant sum constraint are treated.

The key to compositional data lies in the fact that the geometry of the sample space on which a vector of proportions is defined is different from the classical Euclidean geometry of $\mathbb{R}^D$. Therefore, the commonly used multivariate techniques based on this geometry are not directly applicable.

Definition. A compositional data with D parts of a constant sum k, is a vector with D non-negative components that provide relative information and whose sample space is the simplex S^D, defined by

$$S^D(k) = \left\{ \boldsymbol{x} \in \mathbb{R}^D / \boldsymbol{x} = (x_1, x_2, \dots, x_D) : x_i \geq 0; \sum_{i=1}^{D} x_i = k \right\}$$

Definition. Closure is defined as the application on a compositional data that assigns it the value of the constant sum it possesses. Assuming a vector $\boldsymbol{x} = (x_1, \dots, x_D) \in \mathbb{R}^{D+}$ the closure of $\boldsymbol{x}$ at $k > 0$ consists of

$$C(\boldsymbol{x}) = \frac{k\boldsymbol{x}}{(x_1 + x_2 + \cdots + x_D)} = \frac{k\boldsymbol{x}}{\sum_{i=1}^{D} x_i}$$

As compositional data provides information only on the relative magnitudes of the parts, not their absolute values, the total is not of interest. This allows a compositional data to be expressed in terms of proportions of the components, known as ratios. When these ratios are not used in handling compositional data, negative correlations arise between parts without such correlation.

Many studies have shown the importance of handling and interpreting compositional data. It is necessary to apply specific methods to avoid drawing erroneous conclusions in interpretations. For this purpose, there are different ways of dealing with this type of data.

On the one hand, there is a permanence perspective that is maintained by performing the analysis of compositional data within its own space restricted by the constant sum, the Simplex.

The simplest structure of a simplex is a triangle, containing a compositional data with three parts $\boldsymbol{x} = (x_1, x_2, x_3)$, where each one corresponds to the point that is distant x_1, x_2 y x_3 respectively of the opposite sides to the vertices 1, 2 y 3. Note that, as all the parts add up to a constant, one of the parts is a function of the rest and therefore S^D has one dimension minus $(D-1)$.

The Simplex has its own characteristics, operations and measures that define it as a vector space: internal operation of perturbation, external operation of powering and the simplicial metric.

Internal operation of perturbation:

Perturbation, $(S^D, \oplus)$ consists of an internal operation $(S^D \oplus S^D \xrightarrow{\oplus} S^D)$ such that the group is abelian, and the operation is equivalent to addition in the vector space.

Definition. Given two compositions of D parts $\mathbf{x}, \mathbf{y} \in S^D$, perturbation is defined as:

$$\boldsymbol{x} \oplus \boldsymbol{y} = C[x_1 y_1, \dots, x_D y_D] = \frac{[x_1 y_1, \dots, x_D y_D]}{(x_1 y_1 + \cdots + x_D y_D)}$$

External operation of powering:

Powering, $(S^D, \odot)$ consists of an external operation $(\mathbb{R}\odot S^D \xrightarrow{\odot} S^D)$ such that a is a scalar, $a \in \mathbb{R}$, and the operation is equivalent to the product by a scalar in the vector space.

Definition. Given a compositional data of D parts $\boldsymbol{x} \in S^D$, and a scalar $a \in \mathbb{R}$, powering is defined as:

$$a\odot\boldsymbol{x} = C(x_1^a, \dots, x_D^a) = \frac{(x_1^a, \dots, x_D^a)}{(x_1^a + \cdots + x_D^a)}$$

Simplicial metric:

- Internal product:
  Definition. Internal product is defined as an application that to any pair of compositions $\boldsymbol{x}, \boldsymbol{y} \in S^D$ is associated with a number, i.e., $S^D \times S^D \xrightarrow{<\ ,\ >} \mathbb{R}$

$$<\boldsymbol{x}, \boldsymbol{y}>_a = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\ln\left(\frac{x_i}{x_j}\right)\ln\left(\frac{y_i}{y_j}\right) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j\neq i}^{D}\ln\left(\frac{x_i}{x_j}\right)\ln\left(\frac{y_i}{y_j}\right)$$

- Norm:
  In a non-Euclidean space, the shortest path between two points is not necessarily a straight line, so the same properties of the vector norm are used to extract the conditions that must be satisfied in the norm in any vector space.

  Definition. Norm, $||\boldsymbol{x}||_a$, of a compositional data $\boldsymbol{x} \in S^D$ is define as:

$$||\boldsymbol{x}||_a = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\frac{x_i}{x_j}\right)^2}$$

- Aitchison distance:
  Measures the distance between two compositions if x and y are vectors.
  Definition. Given $\boldsymbol{x}, \boldsymbol{y} \in S^D$, the Aitchison distance is defined as:

$$d_a(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a$$

  being $\boldsymbol{x} \ominus \boldsymbol{y} = \boldsymbol{x}\oplus\boldsymbol{y}^{-1} = C\left(\frac{x_1}{y_1}, \dots, \frac{x_D}{y_D}\right)$
  So that,

$$d_a(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{y_i}{y_j}\right)\right)^2} = \sqrt{\sum_{i=1}^{D}\left(\ln\left(\frac{x_i}{g(x)}\right) - \ln\left(\frac{y_i}{g(y)}\right)\right)^2}$$

With these concepts, the simplex contains a measurement space structure,$(S^D, \mathbb{R}, \oplus, \odot)$ and as for any vector space, generating vectors, bases and linear dependence are essential and this is the same for the simplex.

In the simplex, an orthonormal base $\vec{\boldsymbol{b}}$ is a set of elements that are mutually orthogonal and normal. The idea is that each compositional datum $\boldsymbol{x} \in S^D$ could be expressed as a "linear combination" of the base of the simplex such that:

$$\boldsymbol{x} = (x_1, \dots, x_D) = \oplus_{i=1}^{D} u_i\vec{\boldsymbol{b}}_i$$

The equivalent of a "linear combination" being a "power-perturbation combination" as follows where the $\vec{\boldsymbol{b}}$ are compositions considered as generators:

$$\boldsymbol{x} = \left(u_1 \odot \vec{\boldsymbol{b}}_1\right) \oplus \ldots \oplus \left(u_C \odot \vec{\boldsymbol{b}}_C\right)$$

When this subspace is the totality of the unitary simplex, the $\vec{\boldsymbol{b}}$ form a basis. In general, the basis should be chosen so that the generators are "linearly independent". In the simplex, $\vec{\boldsymbol{b}}_1, \ldots, \vec{\boldsymbol{b}}_C$ are linearly independent if and only if

$$(u_1 \odot \beta_1) \oplus \ldots \oplus (u_C \odot \beta_C) = e \Leftrightarrow u_1 = \cdots = u_c = 0$$

For a space S^D with $D-1$ dimensions, a linearly independent basis has $D-1$ generators, with the most important being those that form an orthonormal basis. Suppose the generators $\beta_1, \ldots, \beta_{D-1}$ have measure 1 such that $\left|\left|\beta_i\right|\right| = 1\ (i = 1, \ldots, D-1)$ and are orthogonal such that $< \beta_i, \beta_j >= 0\ \ (i \neq j)$.

On the other hand, there is also a non-permanence perspective, which consists of applying transformations to the compositional data that allow them to be taken to a multidimensional vector space. $\mathbb{R}^D$. This is possible since we have two vector spaces of dimension $D-1$: on the one hand, the simplex $(S^D, \mathbb{R}, \oplus, \odot)$ and on the other the space $(\mathbb{R}^{D-1}, \mathbb{R}, +, \cdot)$, so a bijective application is generated between them that preserves the operations of both, $S^D \overset{f}{\rightarrow} \mathbb{R}^{D-1}$, and that it can be inverted.

For compositional data, log-ratio transformations are used. These are those that send a point of the simplex to the logarithm of a quotient involving the parts of the composition. Considering the properties of logarithms, quotients that are compared multiplicatively are compared additively.

There are a multitude of transformations, each with its own characteristics and advantages, but the most widely used are those created by Aitchison, alr and clr. By applying the transformations, they allow the use of standard unrestricted multivariate statistics applied to the transformed compositional data.

Additive log-ratio transformation, alr

Definition. The additive log-ratio transformation (alr) consists in a bijective function of $S^D \overset{alr}{\longrightarrow} \mathbb{R}^{D-1}$, defined as:

$$\boldsymbol{x}' = alr(\boldsymbol{x}) = \left(ln\,\frac{x_1}{x_s}, ln\,\frac{x_2}{x_s}, \ldots, ln\,\frac{x_{D-1}}{x_s}\right) = (ln(x_1) - ln(x_s), \ldots, ln(x_{D-1}) - ln(x_s))$$

where $\boldsymbol{x} \in S^D$, $\boldsymbol{x}' \in \mathbb{R}^{D-1}$ and x_s can be any of the parts of the compositional data.

The inverse $\mathbb{R}^{D-1} \xrightarrow{alr^{-1}} S^D$ is defined by:

$$\boldsymbol{x} = alr^{-1}(\boldsymbol{x}') = C(exp(x'_1), exp(x'_2), \ldots, \exp(x'_{D-1}), 1)$$

This transformation verifies the perturbation as well as the powering, i.e.,

$$\boldsymbol{x} \oplus \boldsymbol{y} \overset{alr}{\longrightarrow} alr(\boldsymbol{x}) + alr(\boldsymbol{y}) \ \ ; \ \ \alpha \odot \boldsymbol{x} \overset{alr}{\longrightarrow} \alpha \cdot alr(\boldsymbol{x})$$

facilitating calculations and interpretation in practice. However, it is not an isometric transformation, i.e., the angles and distances in the simplex cannot be associated with angles and distances in real space.

Centered log-ratio transformation, clr

Definition. Centered log-ratio transformation (clr) $S^D \overset{clr}{\rightarrow} U^{D-1}$, where U^{D-1} is a hyperplane of $\mathbb{R}^D$, $\boldsymbol{x} \in S^D$, $\boldsymbol{z} \in \mathbb{R}^D$ is defined as:

$$\boldsymbol{z} = clr(\boldsymbol{x}) = \left(\ln\frac{x_1}{g(\boldsymbol{x})}, \ln\frac{x_2}{g(\boldsymbol{x})}, \dots, \ln\frac{x_D}{g(\boldsymbol{x})}\right) = (\ln x_1 - \ln g(\boldsymbol{x}), \dots, \ln x_D - \ln g(\boldsymbol{x}))$$

where $g(\boldsymbol{x})$ is the geometric mean of the D parts of $\boldsymbol{x}$

$$g(\boldsymbol{x}) = \left[\prod_{j=1}^{D} x_j\right]^{1/D}$$

The inverse of this transformation is:

$$\boldsymbol{x} = clr^{-1}(\boldsymbol{z}) = C(\exp(z_1), \dots, \exp(z_D))$$

In contrast to the alr, the clr is isometric, i.e., it maintains the angles and therefore the distances when moving from simplex to real space. Therefore, the perturbation, the powering and the distances are maintained such that:

$$\boldsymbol{x} \oplus \boldsymbol{y} \overset{clr}{\rightarrow} clr(\boldsymbol{x}) + clr(\boldsymbol{y}) \quad ; \quad \alpha \odot \boldsymbol{x} \overset{clr}{\rightarrow} \alpha \cdot clr(\boldsymbol{x}) \quad ; \quad d_a^2(\boldsymbol{x}, \boldsymbol{y}) \overset{clr}{\rightarrow} d^2(clr(\boldsymbol{x}), clr(\boldsymbol{y}))$$

Once a measurement space has been created that allows working with compositional data, statistical concepts can be transported so that they can be applied to the compositional data itself in an appropriate way. The purpose of doing this is to be able to study the compositional data set and to be able to perform exploratory analyses on it.

Statistical concepts must be applied on a data set, so that in the simplex is defined as a compositional data matrix. Mathematically, a set of compositional data is represented in the form of a matrix $\boldsymbol{X}(n \times D)$, with n rows where each row is a compositional data and D columns which are the parts. The closure constraint shall maintain that the sum of each row is equal to the constant k (generally 1):

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \dots \\ \boldsymbol{x}_n \end{pmatrix} \in \mathbb{R}^{n\,x\,D}, \quad \boldsymbol{x}_i \in S^D \quad i = 1, \dots, n$$

$$\sum_{j=1}^{D} x_{ij} = k$$

Before going on to analyze the compositional dataset, three steps must first be taken:

1. Calculate the appropriate descriptive statistics for the simplex space. These are the geometric mean, the variance matrix, and the total variance.
2. Centralization and standardization of the compositional dataset.
3. Visualize the biplot corresponding to the data to analyze patterns.

In addition, the compositional data set must be checked for errors, outliers (outliers with respect to a given distribution), or the presence of zeros.

1. Descriptive Statistics

- Geometric Mean

  A measure of central tendency is the geometric mean which, when applied to a matrix of compositional data, is expressed as:

$$\hat{g}(\boldsymbol{X}) = C(\hat{g}_1, \dots, \hat{g}_D)$$

$$\hat{g}_i = \left[\prod_{j=1}^{n} x_{ji}\right]^{1/n} \qquad i = 1, \dots, D$$

- Variation Matrix

  A measure of dispersion in a normal vector space to represent the variability of a data set with respect to its mean is the variance.

  This concept can be translated to simplex space and is known as the variance matrix:

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & \dots & t_{1D} \\ \dots & \dots & \dots \\ t_{D1} & \dots & t_{DD} \end{pmatrix} \in \mathbb{R}^{D\,x\,D}, \quad t_{ij} = Var\left(\ln\left(\frac{x_{ki}}{x_{kj}}\right)\right) = Var\big(\ln(x_{ki}) - \ln(x_{kj})\big), \quad k = 1 \dots n$$

- Total Variance

  A measure of overall dispersion is the total variance.

$$\text{totvar}[\mathbf{X}] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} Var\left(\ln\left(\frac{x_i}{x_j}\right)\right) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D} t_{ij}$$

2. Centralization and Standardization

The process of centralizing a data set in real space consists of a linear transformation that converts the scores of vectors or variables into standard deviation form so that the mean of each vector equals zero. To centralize a compositional data set, each row of the data matrix, $x_i = (x_{i1}, \dots, x_{iD})$ is perturbed by the inverse of the geometric mean, i.e.:

$$\boldsymbol{X}^C = \boldsymbol{X} \oplus \hat{g}^{-1}(\boldsymbol{X}) = \boldsymbol{X} \ominus \hat{g}(\boldsymbol{X}) = \boldsymbol{X} \ominus C(\hat{g}_1, \dots, \hat{g}_D)$$

Standardization is the process of adjusting or adapting values or data to resemble a common model in order to provide easier access and processing. A centralised compositional dataset can be standardized by performing a boosting with its total variance. In other words:

$$\boldsymbol{X}^E = \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \odot \boldsymbol{X}^C = C\begin{pmatrix} X_{11}^C \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} & \dots & X_{1D}^C \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \\ \vdots & \ddots & \vdots \\ X_{n1}^C \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} & \dots & X_{nD}^C \frac{1}{\sqrt{totvar[\boldsymbol{X}]}} \end{pmatrix}$$

3. Biplot

Aitchison adapted the Biplot to use in the simplex by applying it to a compositional data set and using it as an exploratory tool. It is an exploratory plot that allows simultaneous representation of the rows and columns of any matrix using a rank 2 approximation.

In the following we will theoretically develop the construction of a Biplot in the simplex.

First, the compositional data matrix must be centralized with *n* rows and *D* columns, $\boldsymbol{X}(n \times D)$. Next, we will create the matrix $\boldsymbol{Z}$ obtaining the coefficients of the clr coordinates of the matrix $\boldsymbol{X}$, and we decompose $\boldsymbol{Z}$ so that it looks like:

$$\boldsymbol{Z} = \boldsymbol{U}\begin{pmatrix} \lambda_1^{\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & \lambda_2^{\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_s^{\frac{1}{2}} \end{pmatrix}\boldsymbol{V}'$$

Being:

- Square roots of positive eigenvalues $\lambda_1, \dots, \lambda_s$ of $\boldsymbol{ZZ'}$ or $\boldsymbol{Z'Z}$, with $s \leq \min\{(D-1), n\}$ being the range of $\boldsymbol{Z}$ and with $\lambda_1, \dots, \lambda_s$ in descending order of magnitude.
- The eigenvalue matrix of $\boldsymbol{ZZ'}$, $\boldsymbol{U}$, with dimension (*n, s*).
- The eigenvalue matrix of $\boldsymbol{Z'Z}, \boldsymbol{V}$, with dimension (*D, s*).

For the Biplot, two dimensions are usually retained. The rank 2 approximation, $\boldsymbol{Y}$, is obtained by replacing the singular values of the positions greater than those desired by zeros, remembering that in the diagonal matrix the singular values are ordered in descending order. So it follows that:

$$\boldsymbol{Y} = \begin{pmatrix} u_{11} & u_{12} \\ \dots & \dots \\ u_{n1} & u_{n2} \end{pmatrix}\begin{pmatrix} \lambda_1^{1/2} & 0 \\ 0 & \lambda_2^{1/2} \end{pmatrix}\begin{pmatrix} v_{11} & \dots & v_{D1} \\ v_{12} & \dots & u_{D2} \end{pmatrix}$$

But we must express the matrix $\boldsymbol{Y}$ such as:

$$\mathbf{Y} = \begin{pmatrix} \sqrt{n-1}(\sqrt{\lambda_1})^{\alpha}\, u_{11} & \sqrt{n-1}\,(\sqrt{\lambda_2})^{\alpha} u_{12} \\ \dots & \dots \\ \sqrt{n-1}(\sqrt{\lambda_1})^{\alpha}\, u_{n1} & \sqrt{n-1}\,(\sqrt{\lambda_2})^{\alpha} u_{n2} \end{pmatrix}\begin{pmatrix} \frac{(\sqrt{\lambda_1})^{1-\alpha}\, v_{11}}{\sqrt{n-1}} & \dots & \frac{(\sqrt{\lambda_1})^{1-\alpha}\, v_{1D}}{\sqrt{n-1}} \\ \frac{(\sqrt{\lambda_2})^{1-\alpha}\, v_{21}}{\sqrt{n-1}} & \dots & \frac{(\sqrt{\lambda_2})^{1-\alpha}\, v_{2D}}{\sqrt{n-1}} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \\ \dots \\ \boldsymbol{g}_n \end{pmatrix}(\boldsymbol{h}_1 \quad \boldsymbol{h}_2 \quad \dots \quad \boldsymbol{h}_D)$$

The vectors $(\boldsymbol{g}_1, \dots, \boldsymbol{g}_n)$ are called row markers, while the vectors $(\boldsymbol{h}_1, \dots, \boldsymbol{h}_D)$ are called column markers.

The Biplot manages to represent all the vectors in one plane and can be superimposed to visualize the relationship between the samples and the parts.

Given what we have learned over the last 60 years our knowledge of compositional data has been greatly extended, generating the basis for its structuring in the form of a vector space. In this work, this structuring has been proved, starting by defining specific operations on the simplex similar to those we have in $\mathbb{R}^D$, this provides the simplex with a body structure. It has also been shown that it is possible to define the Aitchison distance, which also gives the simplex the structure of a vector space.

On the other hand, the log-coefficient transformations have been defined and we have shown that it is more difficult to find transformations which are not very complex and are isometric. Of the transformations mentioned, the clr transformation is the most complete, however, another very complete and frequently used transformation is the ilr transformation, which has not been discussed in this work due to its complexity.

From a statistical perspective, due to the properties of the simplex it is not possible to directly apply the basic concepts of multivariate statistics such as the mean of a vector or the variance of a set of data. Some of these statistical concepts have been defined in the paper as the sample centre, which represents what the sample mean would represent for typical variables.

Despite the great evolution of understanding surrounding compositional data, there are many points and issues about it that still need to be studied and expanded upon. Due to the length of this paper, there are many aspects of compositional data that could not be mentioned or analysed.