Butlletí de la Societat Catalana de Matemàtiques
Vol. 26, núm. 1, 2011. Pàg. 5–28. DOI: 10.2436/20.2002.01.32

De les dades composicionals a una geometria euclidiana sobre el símplex

Carles Barceló i Vidal

La geometria és l'art de pensar bé i dibuixar malament.

J. H. Poincaré

Resum: La reflexió sobre la naturalesa de les dades composicionals i sobre la metodologia estadística específica per a la seva anàlisi condueix a la construcció de l'espai de les composicions i a la seva estructuració com un espai vectorial euclidià, del qual el símplex n'és l'espai suport. S'il·lustren sobre el diagrama ternari alguns dels elements més característics d'aquesta geometria.

Paraules clau: dades composicionals, espai de les composicions, pertorbació, símplex.

Classificació MSC2010: 51M05, 62-07.

1 Introducció

D'acord amb [2], les *dades composicionals* són vectors[1] $\boldsymbol{x} = (x_1, \ldots, x_D)'$ d'observacions les components dels quals són valors no negatius que representen proporcions respecte d'un total. Això fa que aquestes components estiguin sotmeses a la restricció

$$x_1 + \cdots + x_D = 1 \,. \tag{1}$$

Els condicionaments matemàtics derivats del fet de treballar posteriorment amb les logràtios de les components d'aquests vectors d'observacions, obliga que les components x_i de les dades composicionals siguin estrictament positives. Això fa que l'espai mostral d'aquest tipus de dades sigui el símplex obert de $\mathbb{R}^D$, és a dir

$$S^D = \{\boldsymbol{x} = (x_1, \ldots, x_D)' : \; x_1 > 0, \ldots, x_D > 0 \,;\; x_1 + \cdots + x_D = 1\} \,. \tag{2}$$

1 En aquest article els vectors es consideren matrius d'una sola columna (vectors-columna). S'utilitza el símbol $'$ «prima» per a indicar la transposada d'una matriu.

Per a la dimensió $D = 3$, les dades del símplex S^3 es representen sobre els coneguts *diagrames ternaris* (figura 1). Aquest tipus de dades les trobem en la majoria de disciplines quan analitzem dades que són proporcions. Així, les trobem a la geologia, quan es detalla la composició química d'un mineral. A la demografia, quan es donen els percentatges de població en els diferents trams d'edat. A l'economia, quan s'especifica percentualment la distribució dels recursos d'una empresa entre els diferents departaments. I a les matemàtiques, quan es treballa amb la distribució de probabilitat multinomial.

Un exemple interessant procedent de la genètica de poblacions el trobem en la llei coneguda com a llei d'equilibri de Hardy-Weinberg, formulada fa més de cent anys. De manera molt simplificada podem dir que aquesta llei postula que la composició genètica d'una població roman en equilibri mentre no actuï la selecció natural ni cap altre factor, i no es produeixi cap mutació. D'aquesta manera, després d'una generació d'aparellaments a l'atzar, hom esperaria que les proporcions x_{AA}, x_{AB} i x_{BB} dels tres possibles genotips AA, AB i BB —procedents d'un gen autosòmic amb dos al·lels A i B— s'aproximessin cap a p^2, $2pq$ i q^2, respectivament, on p és la proporció de l'al·lel A present en la població, i $q = 1 - p$ la de l'al·lel B. Això equival a assegurar que la dada composicional $(x_{\mathrm{AA}}, x_{\mathrm{AB}}, x_{\mathrm{BB}})'$ hauria d'estar situada «molt a prop» de la corba del símplex S^3 d'equació $\log x_{\mathrm{AA}} + \log x_{\mathrm{AB}} - 2\log x_{\mathrm{BB}} = -2\log 2$ (figura 1). Així, a partir de les dades procedents de [5], hem representat a la figura 1 la composició dels genotips MM, MN i NN que determinen els tres grups sanguinis M, N i MN dels humans (aquesta és una classificació semblant a la més coneguda dels grups sanguinis A, B, AB i O) observada en quaranta-nou grups ètnics d'arreu del món. A la vista d'aquest gràfic ens podríem preguntar si aquestes observacions s'ajusten al que estableix la llei de Hardy-Weinberg.

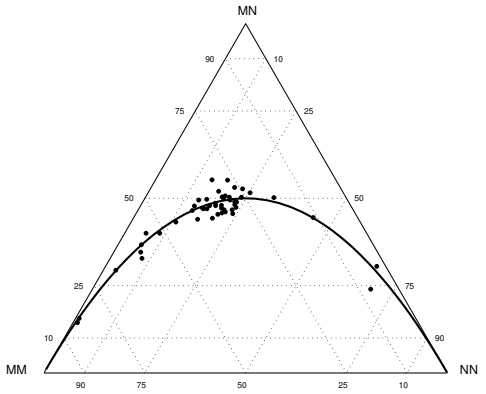


FIGURA 1: Composició dels genotips sanguinis MM, MN i NN observats en quaranta-nou grups ètnics [5]. La corba dibuixada, d'equació $\log x_{\mathrm{MM}} + \log x_{\mathrm{MN}} - 2\log x_{\mathrm{NN}} = -2\log 2$, correspon a la llei de l'equilibri de Hardy-Weinberg.

La restricció (1) que caracteritza les dades composicionals complica l'anàlisi estadística d'aquest tipus de dades. Pensem, per exemple, que el fet de modificar una component qualsevol d'un vector composicional, provoca necessàriament canvis en una o més de les altres components. Això obliga a replantejar-se el concepte d'independència estocàstica quan es treballa amb vectors aleatoris composicionals. Igualment, el coeficient de correlació entre dues components x_i i x_j d'un conjunt de dades composicionals no pot ser interpretat en la forma habitual. Tal com ja advertia Pearson [14] a final del segle XIX, el fet que les components tinguin el mateix denominador ja que representen proporcions respecte d'un mateix total —és a dir $x_i = w_i/(w_1 + \cdots + w_D)$ i $x_j = w_j/(w_1 + \cdots + w_D)$, on $w_1, \ldots, w_D$ són les components del vector $\boldsymbol{w}$ de $\mathbb{R}^D_+$ a partir de les quals es calculen els valors relatius— introdueix inevitablement una «falsa» o «espúria» correlació entre x_i i x_j. Aquestes són només algunes de les moltes dificultats que impedeixen aplicar l'anàlisi estadística estàndard quan es treballa amb dades composicionals o, si més no, interpretar-ne els resultats en la forma habitual. Aitchison [2], en el capítol 3, fa un recull exhaustiu d'aquestes dificultats. Durant molt de temps es va aplicar erròniament la metodologia estadística estàndard en l'anàlisi de les dades composicionals, alhora que en ambients científics vinculats sobretot a la geologia, per exemple a [6, 7, 8], es discutia sobre la manera d'interpretar correctament els resultats de les anàlisis estadístiques estàndard. Cal esperar fins a [1, 2] per a disposar d'una metodologia específica per a l'anàlisi d'aquest tipus de dades, que a partir d'ara denominarem genèricament amb la sigla CODA (de l'anglès *compositional data analysis*). De fet, el missatge del professor Aitchison és ben senzill: «si una dada $\boldsymbol{x} = (x_1, \ldots, x_D)'$ és composicional l'única informació rellevant que ens proporciona està continguda en les ràtios x_i/x_j de les seves components, i no en els seus valors individuals». És per això que tota la metodologia CODA està basada en les ràtios x_i/x_j o, millor dit, en les logràtios $\log(x_i/x_j)$. En essència, la metodologia CODA es basa a aplicar sobre les dades composicionals diverses transformacions definides a partir de les logràtios de les components i, tot seguit, aplicar la metodologia estadística estàndard sobre les dades logràtio transformades. Posteriorment als treballs d'Aitchison es va demostrar que aquesta metodologia CODA basada en transformacions logràtio sobre el símplex es podia explicar equivalentment des d'una òptica més «matemàtica» a partir de la seva estructuració com un espai vectorial euclidià [4, 9].

Alguns autors, en comptes de *dades composicionals*, utilitzen la terminologia *dades tancades*, en referència clara a la restricció (1). Al nostre entendre aquesta és una terminologia equívoca com també ho és en certa manera la definició donada inicialment, que obliga que la suma de les components sigui constant igual a u. Suposem, per exemple, que tenim un conjunt de dades composicionals $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})'$ $(i = 1, \ldots, n;\ D \geq 3)$ de suma constant igual a 1 i que, per conveniència, només ens interessa treballar amb les dues primeres components d'aquestes dades, és a dir $\boldsymbol{x}_i^* = (x_{i1}, x_{i2})'$ $(i = 1, \ldots, n)$. La suma de les dues components d'aquests vectors $\boldsymbol{x}_i^*$ deixarà de ser constant, però no per això deixaran de ser vectors composicionals atès que les seves com-

ponents continuen essent proporcions respecte d'un total. Per tant, el caràcter composicional d'uns vectors de dades no deriva necessàriament de la restricció (1), que no sempre es compleix, sinó de la mateixa naturalesa de les dades. En definitiva, una observació $(x_1,\ldots,x_D)'$ la considerarem *composicional* si ella i el vector d'observacions $(kx_1,\ldots,kx_D)'$, amb $k>0$, ens aporten la mateixa informació. És per aquest motiu que la restricció (1) es podria substituir per la restricció $x_1+\cdots+x_D=k$, per a qualsevol valor $k>0$. El valor d'aquesta constant és irrellevant ja que és igual a 1 si es treballa amb proporcions, a 100 si es treballa amb percentatges, a 10^6 si es treballa en parts per milió o a qualsevol altre valor depenent de la tipologia de la variable que s'estigui mesurant. Amb aquesta concepció menys restrictiva del que per a nosaltres són les dades composicionals, en la secció 2 definim les D-composicions com a classes d'equivalència de vectors de $\mathbb{R}^D_+$. D'aquesta manera el símplex S^D passa a ser un més dels possibles espais suport sobre els quals es poden representar les D-composicions. En la secció 3 veiem com la transformació logarítmica ens porta de manera natural a definir sobre $\mathbb{R}^D$ una relació d'equivalència, en correspondència amb la definida anteriorment sobre $\mathbb{R}^D_+$ a l'hora d'introduir les D-composicions. Tot seguit, definim la transformació logràtio centrada que aplica l'espai C^D de les D-composicions en un subespai de dimensió $D-1$ de l'espai real $\mathbb{R}^D$. A les seccions 4 i 5, amb ajuda d'aquesta transformació i de la seva inversa, estructurem C^D com un espai vectorial euclidià, estructura que es pot traslladar a qualsevol dels espais suport, en particular al símplex S^D. La secció 6 la dediquem fonamentalment a visualitzar sobre el símplex S^3 alguns dels conceptes afins i mètrics definits sobre l'espai de les composicions en les seccions anteriors, sense pretendre en cap moment fer-ne un desenvolupament exhaustiu. Aquesta concepció molt més general de les composicions com a classes d'equivalència de vectors de $\mathbb{R}^D_+$ —prescindint de la necessitat d'haver-se de restringir al símplex S^D— és, doncs, l'aportació que fa aquest article a la metodologia CODA, en la línia que s'havia iniciat fa uns anys [3, 4] i que no havia tingut continuïtat.

2 L'espai de les composicions

2.1 Classes d'equivalència composicionals

Habitualment, les mesures quantitatives realitzades sobre una mateixa unitat d'una població (o mostra) es materialitzen en un conjunt ordenat de D valors positius. És per això que convindrem a anomenar *D-observació* a qualsevol $(D\times 1)$-vector real $\boldsymbol{w}=(w_1,\ldots,w_D)'$ que tingui totes les seves components estrictament positives. Per tant, el conjunt de tots aquests vectors és l'octant positiu $\mathbb{R}^D_+$ de l'espai. Si el nostre interès no rau en els valors absoluts $w_1,\ldots,w_D$ de les mesures sinó en els seus valors relatius w_i/w_j, resultarà que les D-observacions $(w_1,\ldots,w_D)'$ i $(kw_1,\ldots,kw_D)'$, amb $k>0$, ens proporcionen la mateixa informació. Això ens porta a la definició següent.

DEFINICIÓ 2.1. Direm que dues D-observacions $\boldsymbol{w}, \boldsymbol{w}^* \in \mathbb{R}_+^D$ són *composicionalment equivalents* (o C-equivalents), i escriurem $\boldsymbol{w} \sim \boldsymbol{w}^*$, sempre que existeixi una constant de proporcionalitat $k > 0$ tal que $\boldsymbol{w} = k\boldsymbol{w}^*$. Aquesta relació classifica els vectors de $\mathbb{R}_+^D$ en classes d'equivalència que anomenarem *D-composicions*. La D-composició generada per una observació $\boldsymbol{w} \in \mathbb{R}_+^D$ la representarem amb $\underline{\boldsymbol{w}}$. Per tant,

$$\underline{\boldsymbol{w}} = \{k\boldsymbol{w} : k \in \mathbb{R}_+\}\,.$$

El conjunt C^D de les D-composicions, és a dir l'espai quocient $\mathbb{R}_+^D/\sim$, l'anomenarem *espai de les composicions*. Simbolitzarem amb ccl (de l'anglès, *compositional closure*) l'aplicació que fa correspondre a cada D-observació $\boldsymbol{w}$ la seva D-composició $\underline{\boldsymbol{w}}$, és a dir

$$\begin{array}{rccl} \text{ccl}: & \mathbb{R}_+^D & \longrightarrow & C^D \\ & \boldsymbol{w} & \longmapsto & \text{ccl}\,\boldsymbol{w} = \underline{\boldsymbol{w}}\,. \end{array}$$

És clar que les D-composicions es poden interpretar geomètricament com semirectes des de l'origen de coordenades de $\mathbb{R}_+^D$ (figura 2).[2]

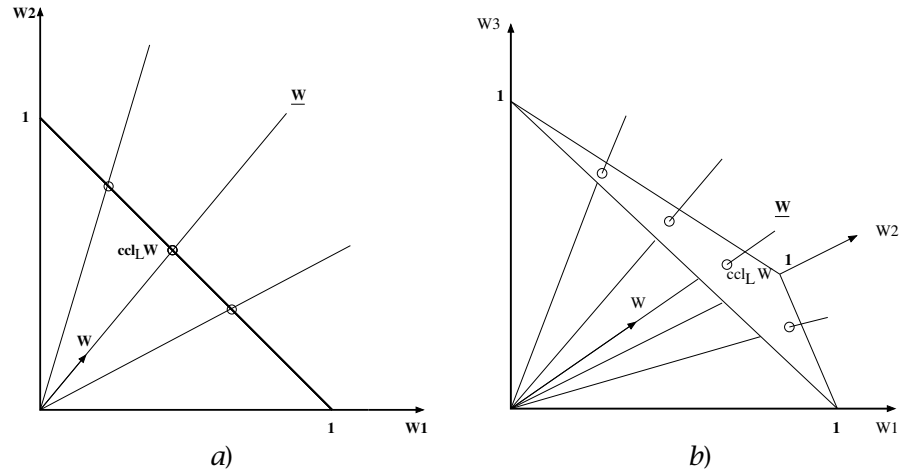


FIGURA 2: Les composicions són semirectes des de l'origen de $\mathbb{R}_+^D$ que tenen el seu representant «lineal» sobre el símplex S^D. *a*) Cas $D = 2$. *b*) Cas $D = 3$.

2.2 Criteris de selecció de representants de les composicions

Per a especificar una composició $\underline{\boldsymbol{w}}$ n'hi ha prou identificant un vector d'observacions qualsevol que pertanyi a $\underline{\boldsymbol{w}}$. Establirem diferents criteris per a seleccionar els representants de les composicions de C^D.

2 Per motius tècnics, en les figures s'utilitza una tipografia diferent de la del text, i els subíndexos es converteixen en afixos.

DEFINICIÓ 2.2. El *criteri lineal* selecciona de cada D-composició $\underline{\boldsymbol{w}}$ la D-observació $\boldsymbol{w}^* = \boldsymbol{w}/\sum_{j=1}^{D} w_j$. Observeu que $\sum_{i=1}^{D} w_i^* = 1$. Convindrem a simbolitzar aquesta observació amb $\mathrm{ccl}_{\mathrm{L}}\,\boldsymbol{w}$. El conjunt de totes aquestes observacions associades a les D-composicions no és més que el símplex definit a (2).

Resulta evident que si $\boldsymbol{w} \sim \boldsymbol{w}^*$, aleshores $\mathrm{ccl}_{\mathrm{L}}\,\boldsymbol{w} = \mathrm{ccl}_{\mathrm{L}}\,\boldsymbol{w}^*$. Per tant, el representant sobre el símplex $\mathcal{S}^D$ d'una composició $\underline{\boldsymbol{w}} \in \mathcal{S}^D$ està unívocament determinat. Geomètricament, l'observació $\mathrm{ccl}_{\mathrm{L}}\,\boldsymbol{w}$ sobre el símplex associada a la composició $\underline{\boldsymbol{w}}$ és la intersecció de la semirecta que representa $\underline{\boldsymbol{w}}$ amb l'hiperplà de $\mathbb{R}^D$ definit per l'equació $w_1 + \cdots + w_D = 1$ (figura 2). Per a la dimensió $D = 3$, la representació de les composicions sobre el símplex S^3 són els coneguts *diagrames ternaris*, molt utilitzats en l'àmbit de la geologia i altres disciplines científiques que treballen amb dades composicionals. Aquests diagrames es realitzen sobre un triangle equilàter d'alçada unitat de manera que cada punt P interior al triangle $W_1W_2W_3$ representa la composició $\underline{\boldsymbol{w}}$ associada a $(w_1, w_2, w_3)' \in S^3$, on w_1, w_2 i w_3 són, respectivament, les distàncies de P als costats W_2W_3, W_1W_3 i W_1W_2 del triangle (figura 3*a*). Anàlogament, el símplex S^4 queda representat per l'interior d'un tetraedre regular d'alçada unitat (figura 3*b*).

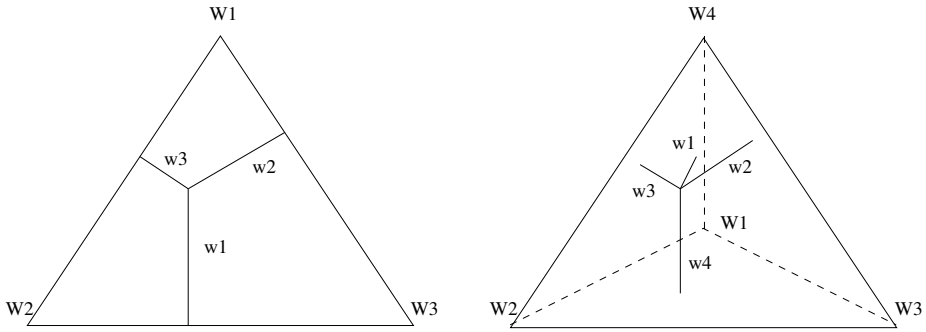


FIGURA 3: Coordenades d'un punt en el símplex. *a*) Símplex S^3. *b*) Símplex S^4.

DEFINICIÓ 2.3. El *criteri esfèric* selecciona de cada D-composició $\underline{\boldsymbol{w}}$ la D-observació $\boldsymbol{w}^* = \boldsymbol{w}/\|\boldsymbol{w}\|$. Observeu que $\|\boldsymbol{w}^*\| = 1$. Convindrem a simbolitzar aquesta observació amb $\mathrm{ccl}_{\mathrm{E}}\,\boldsymbol{w}$.

DEFINICIÓ 2.4. El *criteri hiperbòlic* selecciona de cada D-composició $\underline{\boldsymbol{w}}$ la D-observació $\boldsymbol{w}^* = \boldsymbol{w}/g(\boldsymbol{w})$, on $g(\boldsymbol{w}) = (\prod_{i=1}^{D} w_i)^{1/D}$ és la mitjana geomètrica de les components de $\boldsymbol{w}$. D'aquesta manera es compleix que $\prod_{i=1}^{D} w_i^* = 1$. Convindrem a simbolitzar aquesta observació amb $\mathrm{ccl}_{\mathrm{H}}\,\boldsymbol{w}$.

Com abans, si $\boldsymbol{w} \sim \boldsymbol{w}^*$, aleshores $\mathrm{ccl}_{\mathrm{E}}\,\boldsymbol{w} = \mathrm{ccl}_{\mathrm{E}}\,\boldsymbol{w}^*$ i també $\mathrm{ccl}_{\mathrm{H}}\,\boldsymbol{w} = \mathrm{ccl}_{\mathrm{H}}\,\boldsymbol{w}^*$. Geomètricament, $\mathrm{ccl}_{\mathrm{E}}\,\boldsymbol{w}$ i $\mathrm{ccl}_{\mathrm{H}}\,\boldsymbol{w}$ són, respectivament, la intersecció de la semirecta associada a $\underline{\boldsymbol{w}}$ amb $\mathbb{E}_+^D$ —l'octant estrictament positiu de l'esfera de radi unitat de $\mathbb{R}^D$ centrada a l'origen (figura 4*a*)— i amb la superfície hiperbòlica $\mathbb{H}_+^D$ de $\mathbb{R}_+^D$ definida implícitament per l'equació $\prod_{i=1}^D w_i = 1$ (figura 4*b*).

És clar que hi ha molts altres criteris per a seleccionar de forma unívoca les observacions que representen les composicions de C^D però els tres que hem presentat són els d'interpretació geomètrica més fàcil. Convindrem a interpretar els operadors $\mathrm{ccl}_{\mathrm{L}}, \mathrm{ccl}_{\mathrm{E}}$ i $\mathrm{ccl}_{\mathrm{H}}$ com a aplicacions de C^D (o de $\mathbb{R}_+^D$, si és el cas) en S^D, $\mathbb{E}_+^D$ i $\mathbb{H}_+^D$, respectivament.

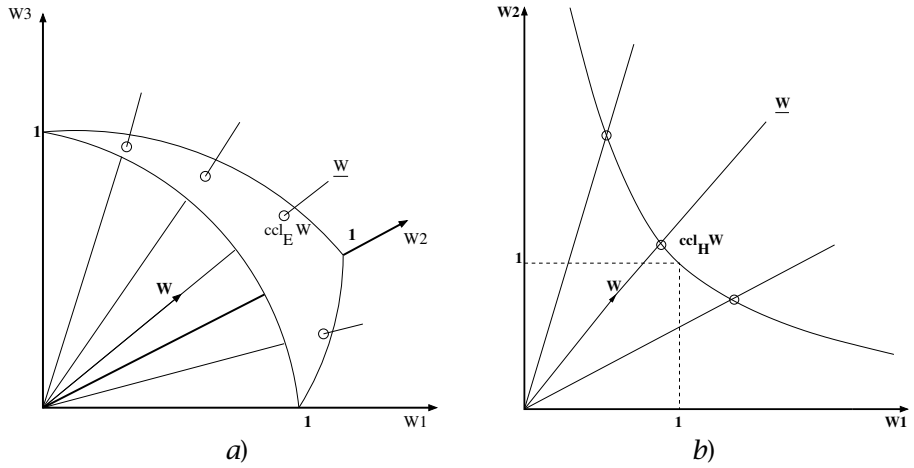


FIGURA 4: Criteris de selecció. *a*) Esfèric ($D = 3$). *b*) Hiperbòlic ($D = 2$).

3 La transformació logarítmica sobre l'espai de les composicions

3.1 Un espai quocient a $\mathbb{R}^D$

La transformació logarítmica de $\mathbb{R}_+^D$ a $\mathbb{R}^D$ ens suggereix definir sobre $\mathbb{R}^D$ una relació d'equivalència en correspondència amb la relació d'equivalència composicional definida abans sobre $\mathbb{R}_+^D$ (vegeu la definició 2.1), derivat del fet que $\boldsymbol{w} \sim \boldsymbol{w}^*$ sii $\log \boldsymbol{w} - \log \boldsymbol{w}^*$ és múltiple del vector $\mathbf{1}_D = (1, \ldots, 1)' \in \mathbb{R}^D$.

DEFINICIÓ 3.1. Direm que dos vectors $\boldsymbol{z}$ i $\boldsymbol{z}^*$ de $\mathbb{R}^D$ són *$\mathcal{L}$-equivalents*, i escriurem $\boldsymbol{z} \equiv \boldsymbol{z}^*$, si existeix una constant λ tal que $\boldsymbol{z} = \boldsymbol{z}^* + \lambda \mathbf{1}_D$. Si considerem el subespai $U = \{\lambda \mathbf{1}_D : \lambda \in \mathbb{R}\}$ de $\mathbb{R}^D$, l'anterior relació d'equivalència es pot reescriure de manera equivalent com

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

Simbolitzarem amb $\boldsymbol{z} + U$ la classe d'equivalència generada pel vector $\boldsymbol{z}$ de $\mathbb{R}^D$. El conjunt de totes les classes d'equivalència, és a dir l'espai quocient $\mathbb{R}^D/U$, el simbolitzarem amb $\mathcal{L}^D$. Simbolitzarem amb ucl l'aplicació que fa correspondre a cada vector $\boldsymbol{z}$ de $\mathbb{R}^D$ la seva classe d'equivalència $\boldsymbol{z} + U$, és a dir

$$\begin{array}{rccl} \text{ucl:} & \mathbb{R}^D & \longrightarrow & \mathcal{L}^D \\ & \boldsymbol{z} & \longmapsto & \text{ucl}\,\boldsymbol{z} = \boldsymbol{z} + U\,. \end{array}$$

En la figura 5 podem veure com les classes d'equivalència $\boldsymbol{z} + U$ es poden interpretar geomètricament com rectes paral·leles al vector $\mathbf{1}_D$. Totes aquestes rectes són ortogonals a l'hiperplà (i subespai) $V = \{\boldsymbol{z} \in \mathbb{R}^D : \boldsymbol{z}'\mathbf{1}_D = 0\}$ que passa per l'origen de $\mathbb{R}^D$ i és ortogonal a $\mathbf{1}_D$. Això fa que tots els vectors d'una mateixa classe $\boldsymbol{z} + U$ tinguin la mateixa projecció ortogonal sobre V. Escollirem precisament aquesta projecció comuna com a representant «canònic» de la classe i convindrem a simbolitzar-la amb $\text{ucl}_V\,\boldsymbol{z}$. Es fàcil veure com

$$\text{ucl}_V\,\boldsymbol{z} = \boldsymbol{z} - \frac{\sum_{j=1}^{D} z_j}{D}\mathbf{1}_D = \mathbf{H}_D\boldsymbol{z}, \tag{3}$$

essent $\mathbf{H}_D$ la coneguda *matriu de centralització* d'ordre $D \times D$. Aquesta matriu és igual a $\mathbf{I}_D - D^{-1}\mathbf{J}_D$, essent $\mathbf{I}_D$ la matriu identitat d'ordre $D \times D$, i $\mathbf{J}_D = \mathbf{1}_D\mathbf{1}_D'$.

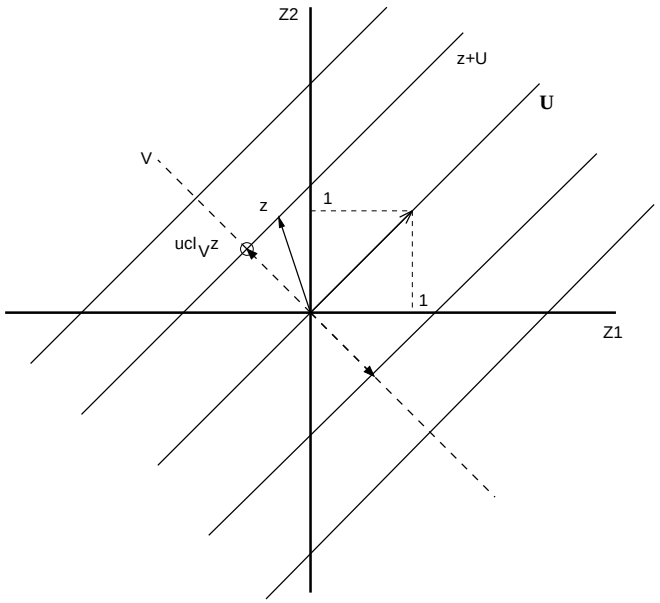


FIGURA 5: Interpretació geomètrica de les classes d'equivalència $\boldsymbol{z} + U$ de $\mathcal{L}^2 = \mathbb{R}^2/U$.

3.2 La transformació logarítmica entre els espais quocients $\mathcal{C}^D$ i $\mathcal{L}^D$

La transformació logarítmica de $\mathbb{R}_+^D$ a $\mathbb{R}^D$ és compatible amb les respectives relacions d'equivalència que hem definit en els dos espais. És a dir, es compleix que

$$\boldsymbol{w} \sim \boldsymbol{w}^*\,,\ \text{a } \mathbb{R}_+^D \quad \Longleftrightarrow \quad \log \boldsymbol{w} \equiv \log \boldsymbol{w}^*\,,\ \text{a } \mathbb{R}^D\,.$$

Això ens permet estendre la transformació logarítmica als respectius espais quocients C^D i $\mathcal{L}^D$. Simbolitzarem amb logc aquesta transformació, és a dir

$$\begin{array}{rccl} \log\text{c}\colon & C^D & \longrightarrow & \mathcal{L}^D \\ & \underline{\boldsymbol{w}} & \longmapsto & \log\text{c}\,\underline{\boldsymbol{w}} = \log \boldsymbol{w} + U\,. \end{array}$$

És immediat comprovar que logc és una aplicació bijectiva. Simbolitzarem amb expc la transformació inversa $\log\text{c}^{-1}$, la qual serà donada per

$$\begin{array}{rccl} \exp\text{c}\colon & \mathcal{L}^D & \longrightarrow & C^D \\ & \boldsymbol{z} + U & \longmapsto & \exp\text{c}(\boldsymbol{z} + U) = \operatorname{ccl}(\exp \boldsymbol{z})\,. \end{array}$$

Fent la composició de les transformacions logc i ucl_V, obtenim una transformació $\operatorname{ucl}_V \circ \log\text{c}$ de C^D en el subespai V de $\mathbb{R}^D$, és a dir

$$\underline{\boldsymbol{w}} \overset{\log\text{c}}{\longrightarrow} \log \boldsymbol{w} + U \overset{\operatorname{ucl}_V}{\longrightarrow} \mathbf{H}_D \log \boldsymbol{w} = \log \frac{\boldsymbol{w}}{g(\boldsymbol{w})}\,.$$

Això ens condueix a la definició següent.

DEFINICIÓ 3.2. Anomenarem *transformació logràtio centrada*, que simbolitzarem amb clr, la funció

$$\begin{array}{rccl} \operatorname{clr}\colon & C^D & \longrightarrow & V \\ & \underline{\boldsymbol{w}} & \longmapsto & \operatorname{clr}\underline{\boldsymbol{w}} = \log \frac{\boldsymbol{w}}{g(\boldsymbol{w})}\,. \end{array}$$

Aquesta funció es pot expressar matricialment com $\operatorname{clr}\underline{\boldsymbol{w}} = \mathbf{H}_D \log \boldsymbol{w}$. La transformació inversa clr^{-1} és donada per

$$\begin{array}{rccl} \operatorname{clr}^{-1}\colon & V & \longrightarrow & C^D \\ & \boldsymbol{z} & \longmapsto & \operatorname{clr}^{-1}\boldsymbol{z} = \operatorname{ccl}(\exp \boldsymbol{z})\,. \end{array}$$

Aquesta transformació logràtio centrada podríem també definir-la de manera semblant sobre S^D, $\mathbb{E}_+^D$ o $\mathbb{H}_+^D$. Així, per exemple, la composició $\operatorname{clr} \circ \operatorname{ccl}_L^{-1}$ és la trasformació logràtio centrada definida sobre el símplex S^D que ja trobem a [2].

4 L'espai afí de les composicions

4.1 L'espai vectorial de les composicions

El fet que el subconjunt U de $\mathbb{R}^D$, introduït a la definició 3.1, sigui un subespai de $\mathbb{R}^D$ ens permet estructurar l'espai quocient $\mathcal{L}^D = \mathbb{R}^D/U$ com un espai vectorial real. Així la suma de dues classes $\boldsymbol{z} + U$ i $\boldsymbol{z}^* + U$ serà donada per

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

i el producte d'una classe $\boldsymbol{z} + U$ per un escalar $\lambda \in \mathbb{R}$ per

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

Aquestes dues operacions estructuren $\mathcal{L}^D$ com un espai vectorial real de dimensió $D-1$, atès que U és un subespai de dimensió 1 de $\mathbb{R}^D$.

La correspondència biunívoca entre els espais quocients C^D i $\mathcal{L}^D$, donada per l'aplicació logc i la seva inversa expc, permet definir sobre C^D una estructura d'espai vectorial real isomorfa a la de $\mathcal{L}^D$.

DEFINICIÓ 4.1. En correspondència amb la suma a $\mathcal{L}^D$ definim l'operació *pertorbació* sobre C^D, que simbolitzarem amb $\oplus$, com a

$$\underline{\boldsymbol{w}} \oplus \underline{\boldsymbol{w}^*} = \operatorname{expc}\left(\operatorname{logc}\underline{\boldsymbol{w}} + \operatorname{logc}\underline{\boldsymbol{w}^*}\right) = \operatorname{ccl}(w_1 w_1^*, \ldots, w_D w_D^*)' \quad (\underline{\boldsymbol{w}}, \underline{\boldsymbol{w}^*} \in C^D)\,.$$

Igualment, en correspondència amb el producte per un escalar a $\mathcal{L}^D$, definim l'operació *producte extern* sobre C^D, que simbolitzarem amb $\odot$, com a

$$\lambda \odot \underline{\boldsymbol{w}} = \operatorname{expc}\left(\lambda \operatorname{logc}\underline{\boldsymbol{w}}\right) = \operatorname{ccl}(w_1^\lambda, \ldots, w_D^\lambda)' \quad (\underline{\boldsymbol{w}} \in C^D)\ (\lambda \in \mathbb{R})\,.$$

D'aquesta manera $(C^D, \oplus, \odot)$ queda estructurat com un espai vectorial real de dimensió $D-1$ isomorf a l'espai quocient $\mathcal{L}^D/U$. En particular, $(C^D, \oplus)$ és un grup commutatiu en el qual la composició $\underline{\mathbf{1}_D} = \operatorname{ccl}(1, \ldots, 1)'$ és l'element neutre, essent $\operatorname{ccl}(1/w_1, \ldots, 1/w_D)'$ l'element invers de la composició $\underline{\boldsymbol{w}} = \operatorname{ccl}(w_1, \ldots, w_D)'$, que convindrem a simbolitzar amb $\underline{\boldsymbol{w}}^{-1}$. Igualment, la «diferència» entre dues composicions $\underline{\boldsymbol{w}} = \operatorname{ccl}(w_1, \ldots, w_D)'$ i $\underline{\boldsymbol{w}^*} = \operatorname{ccl}(w_1^*, \ldots, w_D^*)'$, que simbolitzarem amb $\underline{\boldsymbol{w}} \ominus \underline{\boldsymbol{w}^*}$, serà donada per la composició $\operatorname{ccl}(w_1/w_1^*, \ldots, w_D/w_D^*)'$.

Les operacions $\oplus$ i $\odot$ que acabem de definir sobre l'espai C^D de les composicions es poden traslladar a qualsevol dels espais suport S^D, $\mathbb{E}_+^D$ o $\mathbb{H}_+^D$. Així, per exemple, les operacions pertorbació i producte extern sobre el símplex S^D seran donades per

$$\boldsymbol{x} \oplus \boldsymbol{x}^* = \operatorname{ccl}_{\mathrm{L}}(x_1 x_1^*, \ldots, x_D x_D^*)'\,, \quad \lambda \odot \boldsymbol{x} = \operatorname{ccl}_{\mathrm{L}}(x_1^\lambda, \ldots, x_D^\lambda)'$$

per a qualssevol $\boldsymbol{x}, \boldsymbol{x}^* \in S^D$, i $\lambda \in \mathbb{R}$.

4.2 El grup de les pertorbacions

El fet que $(C^D, \oplus, \odot)$ sigui un espai vectorial real ens permet interpretar també C^D com un espai afí de dimensió $D-1$ que té $(C^D, \oplus)$ com a grup de transformacions. D'aquesta manera, una composició $\boldsymbol{p} \in C^D$ la podem també concebre com una transformació sobre C^D, que anomenarem igualment *pertorbació*, de manera que a cada composició $\boldsymbol{x}$ li fa correspondre la composició $\boldsymbol{p} \oplus \boldsymbol{x}$. Això fa que el conjunt de totes les pertorbacions sobre C^D sigui un grup commutatiu isomorf a $(C^D, \oplus)$. D'aquesta manera tindrem que:

a) la composició de dues pertorbacions $\boldsymbol{p}_1$ i $\boldsymbol{p}_2$ no és res més que la pertorbació associada a $\boldsymbol{p}_1 \oplus \boldsymbol{p}_2$;

b) la pertorbació associada a la composició $\underline{\mathbf{1}_D}$ és la pertorbació identitat;
c) per a cada pertorbació $\boldsymbol{p}$, existeix la pertorbació inversa $\boldsymbol{p}^{-1}$;
d) donades dues composicions qualssevol

$$\underline{\boldsymbol{w}} = \mathrm{ccl}(w_1, \ldots, w_D)' \quad \text{i} \quad \underline{\boldsymbol{w}^*} = \mathrm{ccl}(w_1^*, \ldots, w_D^*)' ,$$

existeix una única pertorbació que transforma $\underline{\boldsymbol{w}}$ en $\underline{\boldsymbol{w}^*}$. Aquesta pertorbació no és altra que l'associada a la composició

$$\underline{\boldsymbol{w}^*} \ominus \underline{\boldsymbol{w}} = \mathrm{ccl}(w_1^*/w_1, \ldots, w_D^*/w_D)' .$$

És a dir, les pertorbacions tenen a C^D el mateix paper que les translacions a l'espai real.

El fet de suposar que el grup de les pertorbacions és el grup «natural» de les transformacions que operen sobre l'espai C^D de les composicions és la pedra angular de la metodologia CODA. De fet la denominació *pertorbació* per a designar l'operació $\oplus$ sobre el símplex S^D ja la trobem a [2], tot i que Aitchison la introdueix com a concepte i no com una operació interna sobre el símplex S^D. Amb tot això estem acceptant que la mesura de la «diferència» entre dues composicions $\underline{\boldsymbol{w}} = \mathrm{ccl}(w_1, \ldots, w_D)'$ i $\underline{\boldsymbol{w}^*} = \mathrm{ccl}(w_1^*, \ldots, w_D^*)'$ l'hem de basar en les ràtios w_j^*/w_j de les seves components en comptes de basar-la en les seves diferències $w_j^* - w_j$.

5 L'espai vectorial euclidià de les composicions

5.1 L'espai euclidià $\mathcal{L}^D$

Abans hem interpretat les classes d'equivalència de l'espai quocient $\mathcal{L}^D$ com a rectes paral·leles al vector $\mathbf{1}_D$ de $\mathbb{R}^D$. Això ens porta a definir de manera «natural» la distància entre dues classes $\boldsymbol{z} + U$ i $\boldsymbol{z}^* + U$ de $\mathcal{L}^D$ com la distància euclidiana entre les rectes associades. El fet que aquesta distància coincideixi amb la norma del vector diferència $\mathrm{ucl}_V \boldsymbol{z}^* - \mathrm{ucl}_V \boldsymbol{z}$, on $\mathrm{ucl}_V \boldsymbol{z}$ i $\mathrm{ucl}_V \boldsymbol{z}^*$ són les interseccions d'aquestes dues rectes amb l'hiperplà V per l'origen de $\mathbb{R}^D$ ortogonal a $\mathbf{1}_D$ (figura 6), ens permet definir-la a partir d'un producte escalar sobre $\mathcal{L}^D$.

DEFINICIÓ 5.1. El *producte escalar* $\langle\ ,\ \rangle_{\mathcal{L}}$ de dues classes $\boldsymbol{z} + U$ i $\boldsymbol{z}^* + U$ de $\mathcal{L}^D$ es defineix igual al producte escalar ordinari dels vectors $\mathrm{ucl}_V \boldsymbol{z}$ i $\mathrm{ucl}_V \boldsymbol{z}^*$ de $V \subset \mathbb{R}^D$, és a dir

$$\langle \boldsymbol{z} + U, \boldsymbol{z}^* + U \rangle_{\mathcal{L}} = \langle \mathrm{ucl}_V \boldsymbol{z}, \mathrm{ucl}_V \boldsymbol{z}^* \rangle .$$

PROPOSICIÓ 5.2. *Es compleix que*

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

per a qualssevol $\boldsymbol{z} + U$ *i* $\boldsymbol{z}^* + U$ *de* $\mathcal{L}^D$.

PROVA. N'hi ha prou amb tenir en compte la igualtat (3) i les propietats de la matriu de centralització $\mathbf{H}_D$. □

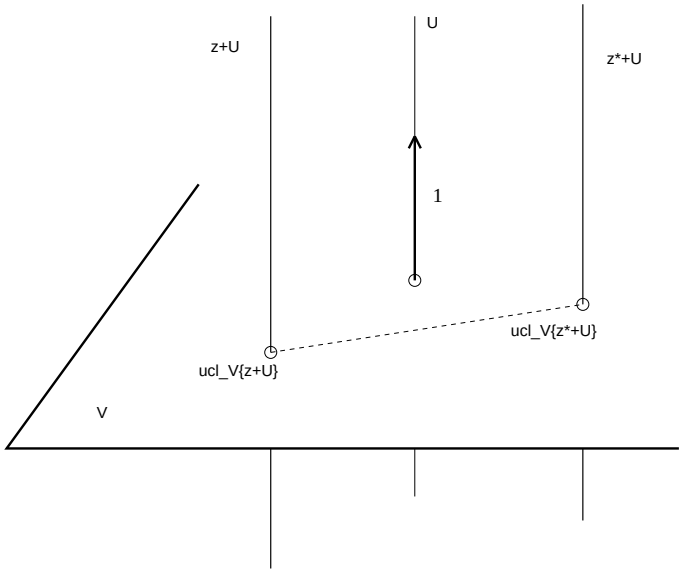


FIGURA 6: Distància entre dues classes $\boldsymbol{z}+U$ i $\boldsymbol{z}^*+U$ de $\mathcal{L}^3=\mathbb{R}^3/U$.

A partir del producte escalar $\langle\ ,\ \rangle_{\mathcal{L}}$ en resulten immediatament les expressions de la norma i la distància a l'espai $\mathcal{L}^D$. La $\mathcal{L}$-norma d'una classe d'equivalència $\boldsymbol{z}+U\in\mathcal{L}^D$ serà donada per

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

i la $\mathcal{L}$-distància entre $\boldsymbol{z}+U$ i $\boldsymbol{z}^*+U$ per

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

$$=\left[\sum_{j=1}^{D}(z_j^*-z_j)^2-\frac{1}{D}\left(\sum_{j=1}^{D}(z_j^*-z_j)\right)^2\right]^{1/2}.$$

D'aquesta manera l'espai quocient $\mathcal{L}^D$ queda estructurat com un espai vectorial euclidià.

5.2 L'espai euclidià de les composicions

La transformació bijectiva logc de C^D a $\mathcal{L}^D$ permet traspassar a C^D l'estructura euclidiana que acabem de definir sobre $\mathcal{L}^D$.

DEFINICIÓ 5.3. El *producte escalar composicional* de dues composicions $\underline{\boldsymbol{w}}$ i $\underline{\boldsymbol{w}^*}$ de C^D es defineix igual al producte escalar de $\operatorname{logc}\underline{\boldsymbol{w}}$ i $\operatorname{logc}\underline{\boldsymbol{w}^*}$ a $\mathcal{L}^D$, és a dir

$$\langle \underline{\boldsymbol{w}}, \underline{\boldsymbol{w}^*} \rangle_C = \langle \log \boldsymbol{w} + U \,,\, \log \boldsymbol{w}^* + U \rangle_{\mathcal{L}} \,.$$

Anàlogament, la *C-norma* d'una composició $\underline{\boldsymbol{w}}$ de C^D es defineix igual a la $\mathcal{L}$-norma de $\operatorname{logc}\underline{\boldsymbol{w}}$ a $\mathcal{L}^D$, és a dir

$$\| \underline{\boldsymbol{w}} \|_C = \| \log \boldsymbol{w} + U \|_{\mathcal{L}} \,.$$

I finalment, la *distància composicional* entre dues composicions $\underline{\boldsymbol{w}}$ i $\underline{\boldsymbol{w}^*}$ de C^D es defineix igual a la $\mathcal{L}$-distància entre $\operatorname{logc}\underline{\boldsymbol{w}}$ i $\operatorname{logc}\underline{\boldsymbol{w}^*}$ a $\mathcal{L}^D$, és a dir

$$d_C(\underline{\boldsymbol{w}}, \underline{\boldsymbol{w}^*}) = d_{\mathcal{L}}(\log \boldsymbol{w} + U \,,\, \log \boldsymbol{w}^* + U) \,.$$

PROPOSICIÓ 5.4. *Per a qualssevol $\underline{\boldsymbol{w}}$ i $\underline{\boldsymbol{w}^*}$ de C^D es compleix que*

$$\begin{aligned}
\langle \underline{\boldsymbol{w}}, \underline{\boldsymbol{w}^*} \rangle_C &= (\log \boldsymbol{w})' \mathbf{H}_D \log \boldsymbol{w}^* = \\
&= \sum_{j=1}^{D} \log w_j \log w_j^* - \frac{1}{D} \left(\sum_{j=1}^{D} \log w_j \right) \left(\sum_{j=1}^{D} \log w_j^* \right) , \\
\| \underline{\boldsymbol{w}} \|_C &= \left[(\log \boldsymbol{w})' \mathbf{H}_D \log \boldsymbol{w} \right]^{1/2} = \\
&= \left[\sum_{j=1}^{D} (\log w_j)^2 - \frac{1}{D} \left(\sum_{j=1}^{D} \log w_j \right)^2 \right]^{1/2} , \\
d_C(\underline{\boldsymbol{w}}, \underline{\boldsymbol{w}^*}) &= \left[(\log \boldsymbol{w}^* - \log \boldsymbol{w})' \mathbf{H}_D (\log \boldsymbol{w}^* - \log \boldsymbol{w}) \right]^{1/2} = \\
&= \left[\sum_{j=1}^{D} \left(\log \frac{w_j^*}{w_j} \right)^2 - \frac{1}{D} \left(\sum_{j=1}^{D} \log \frac{w_j^*}{w_j} \right)^2 \right]^{1/2} .
\end{aligned}$$

El fet que $\operatorname{clr}\underline{\boldsymbol{w}} = \mathbf{H}_D \log \boldsymbol{w} = \log(\boldsymbol{w}/g(\boldsymbol{w}))$ permet simplificar les expressions anteriors.

PROPOSICIÓ 5.5. *Per a qualssevol $\underline{\boldsymbol{w}}$ i $\underline{\boldsymbol{w}^*}$ de C^D es compleix que*

$$\langle \underline{\boldsymbol{w}}, \underline{\boldsymbol{w}^*} \rangle_C = \langle \operatorname{clr}\underline{\boldsymbol{w}}, \operatorname{clr}\underline{\boldsymbol{w}^*} \rangle = \sum_{j=1}^{D} \log \frac{w_j}{g(\boldsymbol{w})} \log \frac{w_j^*}{g(\boldsymbol{w}^*)} \,,$$

$$\| \underline{\boldsymbol{w}} \|_C = \| \operatorname{clr}\underline{\boldsymbol{w}} \| = \left[\sum_{j=1}^{D} \left(\log \frac{w_j}{g(\boldsymbol{w})} \right)^2 \right]^{1/2} ,$$

$$d_C(\underline{\boldsymbol{w}}, \underline{\boldsymbol{w}^*}) = d(\operatorname{clr}\underline{\boldsymbol{w}}, \operatorname{clr}\underline{\boldsymbol{w}^*}) = \left[\sum_{j=1}^{D} \left(\log \frac{w_j^*}{g(\boldsymbol{w}^*)} - \log \frac{w_j}{g(\boldsymbol{w})} \right)^2 \right]^{1/2} .$$

Per tant, el càlcul efectiu del producte escalar composicional, la C-norma i la C-distància a l'espai $\mathcal{C}^D$ es redueix al càlcul estàndard d'aquests mateixos valors sobre els vectors de $V \subset \mathbb{R}^D$ que s'obtenen en aplicar la transformació logràtio centrada (clr) a les composicions implicades en el càlcul. En definitiva, tant la transformació logc com la clr són isometries de l'espai vectorial euclidià $\mathcal{C}^D$ en l'espai quocient $\mathcal{L}^D$ i en el subespai $V \subset \mathbb{R}^D$, respectivament.

La terminologia pròpia dels espais vectorials euclidians és, per tant, aplicable a l'espai $\mathcal{C}^D$ de les composicions. Així, per exemple, parlarem de composicions C-ortogonals, de composicions C-unitàries, etc. A més, la distància composicional sobre $\mathcal{C}^D$ complirà les propietats habituals d'una distància en relació amb les operacions $\oplus$ i $\odot$ definides sobre l'espai de les composicions.

PROPOSICIÓ 5.6. *Per a qualssevol* $\underline{\boldsymbol{w}}, \underline{\boldsymbol{w}}_1, \underline{\boldsymbol{w}}_2 \in \mathcal{C}^D$ *i* $\lambda \in \mathbb{R}$ *es compleix que*

$$d_C(\underline{\boldsymbol{w}} \oplus \underline{\boldsymbol{w}}_1, \underline{\boldsymbol{w}} \oplus \underline{\boldsymbol{w}}_2) = d_C(\underline{\boldsymbol{w}}_1, \underline{\boldsymbol{w}}_2)$$

$$d_C(\lambda \odot \underline{\boldsymbol{w}}_1, \lambda \odot \underline{\boldsymbol{w}}_2) = |\lambda|\, d_C(\underline{\boldsymbol{w}}_1, \underline{\boldsymbol{w}}_2)\,.$$

Finalment remarquem que l'estructura mètrica composicional que acabem de definir sobre l'espai quocient de les composicions es pot traslladar fil per randa a qualsevol dels espais suport $\mathcal{S}^D$, $\mathbb{E}_+^D$ o $\mathbb{H}_+^D$ associats a $\mathcal{C}^D$.

6 Algunes interpretacions de la geometria euclidiana a $\mathcal{C}^D$

6.1 Varietats lineals de $\mathcal{C}^D$

Si l'espai vectorial $\mathcal{C}^D$ l'interpretem alhora com un espai afí, una referència afí $\mathcal{R}$ de $\mathcal{C}^D$ quedarà determinada per una composició $\boldsymbol{o}$ (origen de la referència) i per una base $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{D-1}$ de l'espai vectorial $(\mathcal{C}^D, \oplus, \odot)$. D'aquesta manera, les coordenades afins d'una composició $\boldsymbol{c}$ qualsevol seran els valors reals $\lambda_1, \ldots, \lambda_{D-1}$ tals que

$$\boldsymbol{c} \ominus \boldsymbol{o} = (\lambda_1 \odot \boldsymbol{e}_1) \oplus \ldots \oplus (\lambda_{D-1} \odot \boldsymbol{e}_{D-1})\,.$$

Però el fet que una composició sigui sempre la clausura composicional d'un vector $(w_1, \ldots, w_D)'$ de $\mathbb{R}_+^D$, farà que, mentre no sigui imprescindible utilitzar les coordenades afins, preferim treballar amb les «pseudocoordenades» $(w_1, \ldots, w_D)'$ que identifiquen la composició, amb el benentès que aquestes no són úniques ja que estan determinades a menys d'una constant multiplicativa $k > 0$.

Si concebem el subespai $V = \{\boldsymbol{z} \in \mathbb{R}^D : \boldsymbol{z}'\mathbf{1}_D = 0\}$, definit en la secció 3.1, com un espai afí de $\mathbb{R}^D$, resultarà que les seves varietats lineals són les interseccions de l'hiperplà $z_1 + \cdots + z_D = 0$ que defineix V amb les varietats lineals de $\mathbb{R}^D$. D'aquesta manera, per exemple, els hiperplans afins de V quedaran definits implícitament per un sistema lineal de dues equacions, una d'aquestes igual a l'equació $z_1 + \cdots + z_D = 0$ i l'altra a una equació lineal qualsevol $a_1 z_1 + \cdots + a_D z_D = c$, amb l'única condició que el vector $(a_1, \ldots, a_D)'$ no sigui

múltiple del vector $\mathbf{1}_D$. El fet que les varietats lineals de C^D, que anomenarem *C-varietats*, estiguin en correspondència biunívoca —via la transformació clr— amb les varietats lineals de V, ens permet enunciar la proposició següent.

PROPOSICIÓ 6.1. *Les C-varietats de dimensió $D-2$ (C-*hiperplans*) de C^D queden definides implícitament com el lloc geomètric de les composicions $\underline{\boldsymbol{w}} = \operatorname{ccl}(w_1,\dots,w_D)'$ de C^D que verifiquen una equació del tipus*

$$a_1 \log w_1 + \cdots + a_D \log w_D = c\ , \tag{4}$$

on $a_1,\dots,a_D$ i c són constants reals tals que $a_1+\cdots+a_D=0$, amb no totes les a_j iguals a 0.

PROVA. Sigui $\mathcal{E}$ un C-hiperplà de C^D. Aleshores $\operatorname{clr}\mathcal{E}$ serà també un hiperplà de V. Per tant, els elements $\boldsymbol{z}=(z_1,\dots,z_D)'$ de $\operatorname{clr}\mathcal{E}$ satisfaran una equació lineal del tipus $a_1^* z_1+\cdots+a_D^* z_D=c$, amb no tots els coeficients a_j^* iguals entre si. Com que els elements de $\operatorname{clr}\mathcal{E}$ són del tipus $(\log(w_1/g(\boldsymbol{w}),\dots,\log(w_D/g(\boldsymbol{w}))'$, amb $\boldsymbol{w}=(w_1,\dots,w_D)'\in\mathbb{R}_+^D$, les composicions $\underline{\boldsymbol{w}}$ de $\mathcal{E}$ compliran l'equació

$$a_1^* \log(w_1/g(\boldsymbol{w})) + \cdots + a_D^* \log(w_D/g(\boldsymbol{w})) = c\ .$$

Desenvolupant el primer terme d'aquesta igualtat arribem a l'equació

$$a_1 \log w_1 + \cdots + a_D \log w_D = c\,, \quad \text{essent}\ \ a_j = a_j^* - \frac{1}{D}\sum_{i=1}^{D} a_i^* \quad (j=1,\dots,D)\,.$$

Observem com es compleix que $a_1+\cdots+a_D=0$ i que no tots els a_j són iguals a 0 ja que no tots els coeficients a_j^* són iguals entre si.

Inversament, sigui $\mathcal{E}$ un subconjunt de C^D, els elements del qual satisfan l'equació (4) de l'enunciat. Aleshores, el fet que $a_1+\cdots+a_D=0$ permet reescriure aquesta equació com $a_1\log(w_1/g(\boldsymbol{w}))+\cdots+a_D\log(w_D/g(\boldsymbol{w}))=c$. Per tant, les coordenades dels elements de $\operatorname{clr}\mathcal{E}$ satisfan una equació lineal a $\mathbb{R}^D$. A més, com que $\operatorname{clr}\mathcal{E}\subset V$, podem concloure que $\operatorname{clr}\mathcal{E}$ és un hiperplà de V i, per tant, $\mathcal{E}$ un C-hiperplà de C^D. □

L'equació (4) es pot escriure de manera més abreujada com $\boldsymbol{a}'\log\boldsymbol{w}=c$, on $\boldsymbol{a}=(a_1,\dots,a_D)'$ és un vector no nul de $\mathbb{R}^D$ tal que $\boldsymbol{a}'\mathbf{1}_D=0$. Igual que ocorre a $\mathbb{R}^D$, el C-hiperplà de C^D definit per l'equació (4) té com a subespai director el definit implícitament per l'equació $a_1\log w_1+\cdots+a_D\log w_D=0$, essent $\operatorname{ccl}(\exp a_1,\dots,\exp a_D)'$ una composició C-ortogonal a aquest subespai. Vegem una altra manera de definir implícitament els C-hiperplans.

PROPOSICIÓ 6.2. *L'equació implícita del C-hiperplà que passa per la composició $\underline{\boldsymbol{c}}$ i és C-ortogonal a la direcció donada per la composició $\underline{\boldsymbol{v}}$ és donada per*

$$(\log\boldsymbol{v})'\mathbf{H}_D(\log\boldsymbol{w}-\log\boldsymbol{c})=0 \qquad (\underline{\boldsymbol{w}}\in C^D)\,. \tag{5}$$

PROVA. N'hi ha prou amb desenvolupar el producte matricial de l'esquerra de (5), tot tenint en compte que $\mathbf{H}_D \log \boldsymbol{w} = \log(\boldsymbol{w}/g(\boldsymbol{w}))$, per a qualsevol $\boldsymbol{w} \in \mathbb{R}^D_+$. □

La proposició 6.1 es pot generalitzar fàcilment a C-varietats lineals de dimensió qualsevol.

PROPOSICIÓ 6.3. *En general, les C-varietats queden definides implícitament per sistemes del tipus*

$$\boldsymbol{a}_i' \log \boldsymbol{w} = c_i \quad (\underline{\boldsymbol{w}} \in C^D) \qquad (i = 1, \ldots, m) \,,$$

essent $\mathbf{A} = [a_{ij} : i = 1, \ldots, m;\ j = 1, \ldots, D]$ una matriu no nul·la de dimensió $m \times D$ tal que $\mathbf{A}\mathbf{1}_D = \mathbf{0}_m$. La dimensió de la C-varietat lineal és igual a $D - r - 1$, on r és el rang de la matriu $\mathbf{A}$.

També resulta immediata la caracterització del C-paral·lelisme i la C-ortogonalitat dels hiperplans de C^D a partir de les seves equacions implícites.

PROPOSICIÓ 6.4. *Siguin $\mathcal{H}_1$ i $\mathcal{H}_2$ els C-hiperplans definits, respectivament, de forma implícita per les equacions*

$$\boldsymbol{a}_i' \log \boldsymbol{w} = c_i \quad (\underline{\boldsymbol{w}} \in C^D) \qquad (i = 1, 2) \,,$$

amb $\boldsymbol{a}_i \in \mathbb{R}^D$ no nuls i $\boldsymbol{a}_i' \mathbf{1}_D = 0$, per $i = 1, 2$. Aleshores,

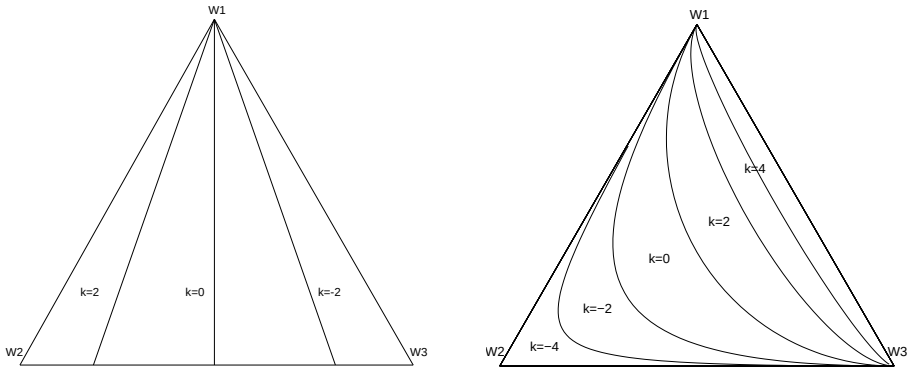
- *$\mathcal{H}_1$ i $\mathcal{H}_2$ són C-paral·lels sii $\boldsymbol{a}_1$ i $\boldsymbol{a}_2$ són paral·lels a $\mathbb{R}^D$, és a dir sii $\boldsymbol{a}_1$ és múltiple de $\boldsymbol{a}_2$.*
- *$\mathcal{H}_1$ i $\mathcal{H}_2$ són C-ortogonals sii $\boldsymbol{a}_1$ i $\boldsymbol{a}_2$ són ortogonals a $\mathbb{R}^D$, és a dir sii $\boldsymbol{a}_1' \boldsymbol{a}_2 = 0$.*

Les figures 7 i 8 ens mostren exemples de famílies de rectes C-paral·leles i C-ortogonals en el símplex S^3. A partir d'aquests gràfics resulta evident el fet que les imatges gràfiques que tenim de «recta», «paral·lelisme» i «ortogonalitat» procedents de l'espai real $\mathbb{R}^D$ no són vàlides a l'espai C^D de les composicions, malgrat ser ambdós espais mètrics euclidians. Així, per exemple, com que les C-rectes de la figura 7*b*) són geodèsiques del símplex S^3 respecte de la mètrica composicional que hem definit en la secció 5.2, resulta clar que el C-camí més curt entre dos punts del símplex no és el segment rectilini entès en la forma «estàndard». Naturalment, però, si apliquéssim la transformació logràtio centrada (clr) a totes aquestes rectes de S^3 representades en les figures 7 i 8 obtindríem imatges «estàndard» de rectes paral·leles i ortogonals contingudes en el pla $z_1 + z_2 + z_3 = 0$ de $\mathbb{R}^3$.

6.2 Boles de C^D

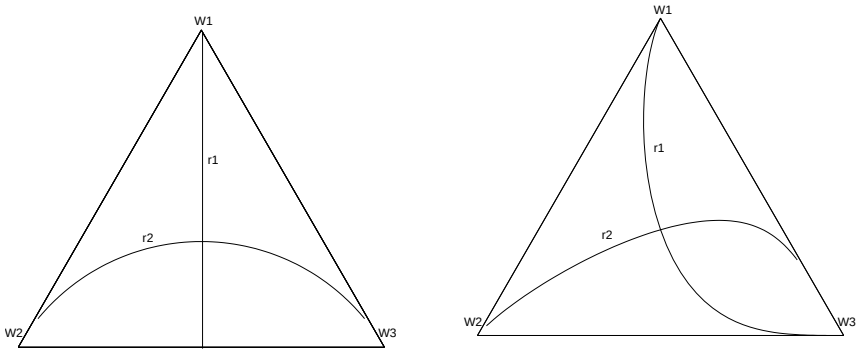
La C-bola $\mathcal{B}(\boldsymbol{a}; r)$ de radi $r > 0$ i centre la composició $\boldsymbol{a} \in C^D$ és el conjunt de totes les composicions $\boldsymbol{c} \in C^D$ tals que $d_C(\boldsymbol{c}, \boldsymbol{a}) \leq r$, és a dir

$$\mathcal{B}(\boldsymbol{a}; r) = \{\boldsymbol{c} \in C^D : \| \boldsymbol{c} \ominus \boldsymbol{a} \|_C \leq r\} \,.$$



a) *b*)

FIGURA 7: Rectes C-paral·leles a S^3. *a*) $\log w_2 - \log w_3 = k$, per a $k = -2, 0, 2$. *b*) $\log w_1 - 2\log w_2 + \log w_3 = k$, per a $k = -4, -2, 0, 2, 4$.



a) *b*)

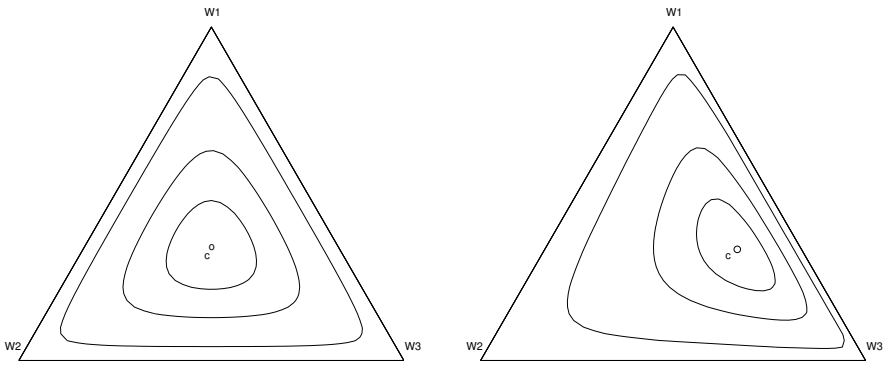
FIGURA 8: Rectes C-ortogonals a S^3. *a*) $r_1 : w_2 - w_3 = 0$; $r_2 : -2\log w_1 + \log w_2 + \log w_3 = 0$. *b*) $r_1 : \log w_1 - 3\log w_2 + 2\log w_3 = 0$; $r_2 : 5\log w_1 - \log w_2 - 4\log w_3 = 0$.

Aquesta C-bola es pot obtenir també aplicant la pertorbació $\boldsymbol{a}$ a la C-bola del mateix radi centrada a $\underline{\mathbf{1}_D}$, és a dir

$$\mathcal{B}(\boldsymbol{a}; r) = \boldsymbol{a} \oplus \mathcal{B}(\underline{\mathbf{1}_D}; r) = \left\{ \boldsymbol{a} \oplus \boldsymbol{c} : \ \boldsymbol{c} \in \mathcal{B}(\underline{\mathbf{1}_D}; r) \right\}.$$

La figura 9 ens mostra les gràfiques d'unes quantes C-circumferències representades sobre S^3. Igual com passava amb les C-rectes, els perfils d'aquestes circumferències composicionals no tenen res a veure amb els perfils «estàndard» d'aquestes figures. Fixem-nos com la proximitat a la frontera del símplex S^3 provoca distorsions en els perfils. Això és pel fet que la C-distància entre

dos punts molt «propers» entre si (en el sentit «estàndard» del terme) situats gairebé tocant la frontera del triangle és molt més gran que la C-distància de dos punts amb la mateixa «proximitat» situats en la zona central del símplex.



a) *b*)

FIGURA 9: C-circumferències a S^3 de radis $r = 1/2, 1, 2$. *a*) Centre: $\boldsymbol{c} = (1/3, 1/3, 1/3)'$. *b*) Centre: $\boldsymbol{c} = (2/6, 1/6, 3/6)'$.

6.3 Centrat d'un conjunt de dades composicionals

Si volem estudiar gràficament les posicions relatives d'un conjunt de punts de $\mathbb{R}^3$ (de $\mathbb{R}$ o de $\mathbb{R}^2$) que estan molt allunyats de l'origen de coordenades tothom té clar que podem moure aquests punts tot aplicant-los una translació que els apropi a l'origen de coordenades. Com que es tracta d'una isometria sabem que les distàncies entre els punts romandran inalterables. És clar que això mateix ho podem fer a l'espai de les composicions perquè es tracta d'un espai vectorial euclidià. Això és el que tractem en aquesta secció.

La definició següent és l'equivalent composicional, en relació amb les operacions $\oplus$ i $\odot$ de C^D, del concepte estàndard de *centre* d'un conjunt de punts de $\mathbb{R}^D$, en relació amb les operacions suma i producte.

DEFINICIÓ 6.5. Sigui un conjunt qualsevol de n composicions $\underline{\boldsymbol{w}_1}, \ldots, \underline{\boldsymbol{w}_n}$ de C^D. El *C-centre* d'aquest conjunt és la composició $\boldsymbol{g}$ definida per

$$\boldsymbol{g} = (\frac{1}{n} \odot \underline{\boldsymbol{w}_1}) \oplus \ldots \oplus (\frac{1}{n} \odot \underline{\boldsymbol{w}_n}) = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \underline{\boldsymbol{w}_i} \,. \tag{6}$$

PROPOSICIÓ 6.6. *Si* $\underline{\boldsymbol{w}_i} = \mathrm{ccl}(w_{i1}, \ldots, w_{iD})' \in C^D$ $(i = 1, \ldots, n)$*, aleshores el C-centre* $\boldsymbol{g}$ *del conjunt de composicions* $\underline{\boldsymbol{w}_1}, \ldots, \underline{\boldsymbol{w}_n}$ *és igual a*

$$\boldsymbol{g} = \mathrm{ccl}\left(\left(\prod_{i=1}^{n} w_{i1}\right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} w_{iD}\right)^{1/n}\right)' \,. \tag{7}$$

PROVA. N'hi ha prou amb desenvolupar el segon membre de la igualtat (6) tenint en compte les definicions de les operacions $\oplus$ i $\odot$ introduïdes en la secció 4.1. □

D'acord amb l'equació (7), la j-èsima component de l'observació associada al C-centre $\boldsymbol{g}$ és la mitjana geomètrica de les n components j-èsimes $w_{1j}, \ldots, w_{nj}$ de les observacions $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$. Gràficament, igual que en la geometria de l'espai, el C-centre d'un conjunt de composicions estarà situat en la «zona central» del núvol de punts que el representa. En canvi, si calculéssim el centre a partir de la mitjana aritmètica de cadascuna de les components de les composicions del conjunt, ens podríem trobar de vegades que el punt resultant no estigués situat en la «zona central» del núvol (vegeu la figura 10).

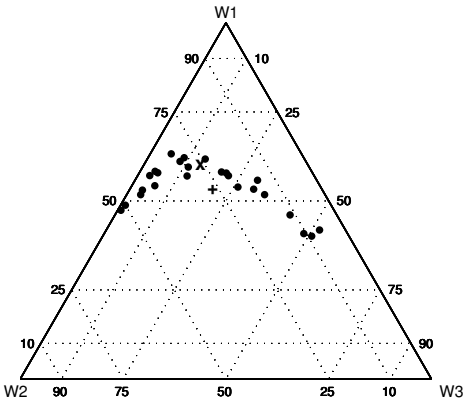
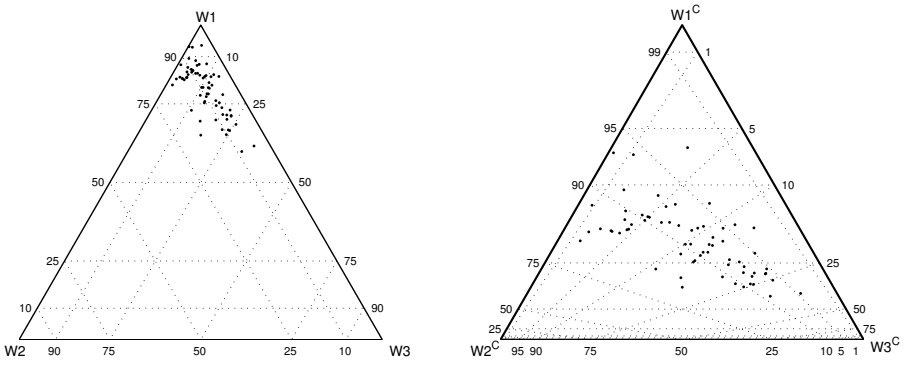


FIGURA 10: Centre d'un conjunt d'observacions a S^3 (exemple procedent de [2]). $\times$ indica la posició del C-centre; $+$ indica la posició del «centre estàndard» (mitjana aritmètica de les components de les observacions).

Si ens fixem en les observacions representades en el símplex S^3 de la figura 11*a*) observarem que estan majoritàriament situades a la part superior del diagrama ternari, molt a prop del vèrtex W_1, pel fet de ser la primera component molt més gran que les altres dues. Per tant, el núvol de punts està força allunyat de l'origen de S^3, és a dir del baricentre $(1/3, 1/3, 1/3)'$. Podem centrar-lo aplicant una perturbació $\boldsymbol{p}$ a les dades del núvol. Així, si prenem $\boldsymbol{p} = \boldsymbol{g}^{-1}$, on $\boldsymbol{g}$ és el C-centre del conjunt de les dades, el «núvol perturbat» tindrà l'origen $(1/3, 1/3, 1/3)'$ com a nou C-centre (figura 11*b*). Direm que hem *C-centrat* les observacions originals. Si comparem les figures 11*a*) i 11*b*) és clar que les distàncies «estàndard» entre les observacions originals i les observacions C-centrades no es conserven. Les que sí que es conserven són les C-distàncies. Això fa que, si $\underline{\boldsymbol{w}_1}, \ldots, \underline{\boldsymbol{w}_n}$ són les observacions originals i $\underline{\boldsymbol{w}_1^*}, \ldots, \underline{\boldsymbol{w}_n^*}$ les corresponents observacions C-centrades, les ràtios d'una

mateixa component de dues observacions qualssevol es conservin, és a dir

$$w_{ij}/w_{lj} = w^*_{ij}/w^*_{lj} \quad (j = 1,\ldots,D) \qquad (i,l = 1,\ldots,n)\ .$$



a) *b*)

FIGURA 11: C-centrat de les dades composicionals a S^3. *a*) Observacions originals. *b*) Les mateixes observacions centrades a l'origen $(1/3, 1/3, 1/3)'$.

7 Discussió

La problemàtica al voltant de l'anàlisi estadística d'una determinada tipologia de dades —com són les dades composicionals— ens ha conduït fins a la construcció d'una estructura algebraica i geomètrica sobre el símplex, l'espai suport d'aquestes dades. D'aquesta manera s'aconsegueix donar una coherència estructural a tota la metodologia CODA. És cert que alguns estadístics més aplicats no veuen la necessitat de disposar d'aquesta estructura i es limiten simplement a aplicar la metodologia estadística estàndard sobre les dades composicionals logràtio transformades. Però és innegable que disposar d'aquesta base estructural de la metodologia CODA permet comprendre-la molt millor i aplicar-la correctament. Hem de tenir en compte que molts dels procediments que es fan servir habitualment en l'anàlisi estadística de dades —com són, entre molts altres, la regressió múltiple, l'ajust per mínims quadrats, l'anàlisi en components principals i les tècniques de classificació— estan estretament lligats a la mètrica definida sobre l'espai mostral de les dades. Per tant, és imprescindible tenir sempre clara la mètrica que s'està utilitzant i sobre quin espai mostral està definida. Disposar de l'espai C^D de les composicions estructurat com a espai vectorial euclidià ens permet destriar els procediments estadístics que són compatibles amb aquesta estructura a l'hora d'analitzar un conjunt de dades composicionals. Així, per exemple, l'anàlisi de les bases

de l'espai vectorial $(S^D, \oplus, \odot)$ i de la classificació o no d'aquestes bases com a ortonormals respecte de la C-mètrica definida sobre S^D, porta a la introducció d'altres transformacions logràtio sobre el símplex [10]. Alhora posa de relleu el caràcter no isomètric de la transformació logràtio additiva (alr) sobre la qual es fonamenta en gran part la metodologia CODA introduïda a [2]. La transformació alr, que envia el vector $\boldsymbol{x} = (x_1, \ldots, x_D)'$ de S^D al vector $\boldsymbol{y} = (\log(x_1/x_D), \ldots, \log(x_{D-1}/x_D))'$ de $\mathbb{R}^{D-1}$, tot i ser un isomorfisme entre S^D i $\mathbb{R}^{D-1}$ no és una isometria, si sobre el símplex considerem la mètrica composicional definida en la secció 5.2. I això fa, per exemple, que sigui erroni realitzar una classificació en grups d'un conjunt de composicions aplicant la distància ordinària sobre les dades alr-transformades.

Aquesta estructuració del símplex ha permès també avançar cap a altres direccions més allunyades de l'anàlisi estadística i més properes a l'àlgebra, al càlcul diferencial o a la probabilitat. Des d'una perspectiva estrictament matemàtica és lògic, per exemple, que ens preguntem com es poden caracteritzar les aplicacions lineals en les quals un dels dos espais vectorials (o ambdós) és el símplex $(S^D, \odot, \oplus)$; o com són les funcions diferenciables de S^D en $\mathbb{R}^m$, de $\mathbb{R}^m$ en S^D o de S^D en S^M; o com són les funcions integrables definides sobre S^D; etc. Aquests i altres temes, alguns inicialment abordats a [4], s'han publicat recentment a la monografia [13].

El fet d'introduir la relació d'equivalència entre dades composicionals i treballar sobre l'espai quocient C^D permet ampliar encara més la metodologia CODA. L'estructura d'espai vectorial euclidià sobre C^D permet traspassar-la, no només al símplex S^D, sinó també a qualsevol espai suport de C^D, com ara la superfície esfèrica $\mathbb{E}_+^D$ o la hiperbòlica $\mathbb{H}_+^D$ introduïdes en la secció 2.2. Això permetria formalment parlar, per exemple, de C-varietats ortogonals o paral·leles sobre aquestes superfícies, tot i que resulta difícil imaginar-se gràficament com són aquestes varietats en el cas $D = 3$.

La metodologia CODA té també el seus detractors, sobretot en l'àmbit de les ciències aplicades, com ara la geologia [15, 16, 17]. La majoria de crítiques consisteixen a no acceptar la necessitat d'aplicar transformacions a les dades composicionals quan els mètodes estàndard d'anàlisi estadística aplicats sobre les dades originals «funcionen» sense problemes. Aquest pragmatisme porta sovint l'investigador aplicat a resultats que, tot i semblar versemblants, són incorrectes. Massa sovint s'ignora un principi bàsic en l'anàlisi estadística com és que cal tenir sempre en compte quin és l'espai mostral suport de les dades que s'estan analitzant. En el cas de les dades composicionals és clar que l'espai mostral no és l'espai real sinó el símplex. És cert que quan les dades estan situades en la zona central del símplex, els resultats numèrics derivats d'una anàlisi estàndard de les dades i els d'una anàlisi basada en la metodologia CODA presenten diferències petites. En canvi, quan l'anàlisi estàndard s'aplica a conjunts de dades que estan situats a prop de la frontera del símplex, ens podem trobar amb resultats inversemblants, com ara que els intervals de confiança de les estimacions caiguin fora del símplex. Allò que sovint costa d'acceptar als investigadors que es limiten a aplicar la metodologia

CODA sense voler entendre mínimament quins són els seus fonaments, és tot el que fa referència a la distància composicional. Així, per exemple, suposem que tenim les següents quatre composicions de S^3: $A_1 = (0{,}300, 0{,}200, 0{,}500)'$, $A_2 = (0{,}200, 0{,}300, 0{,}500)'$, $B_1 = (0{,}980, 0{,}010, 0{,}010)'$ i $B_2 = (0{,}970, 0{,}002, 0{,}028)'$. Resulta que la C-distància entre A_1 i A_2 és igual a 0,57, mentre que entre B_1 i B_2 és igual a 1,88. Certament, resulta difícil que la nostra mentalitat euclidiana —que tendeix d'entrada a comparar els valors numèrics basant-se en diferències i no en ràtios— accepti que la distància entre les composicions B_1 i B_2 és aproximadament tres vegades més gran que la distància entre A_1 i A_2. Cal dir, però, que aquesta no acceptació espontània d'una realitat numèrica matemàticament fonamentada desapareixeria si puguéssim representar (o imaginar) les nostres dades composicionals sobre la superfície hiperbòlica $\mathbb{H}^3_+$ en comptes del símplex S^3.

Finalment, no es pot ignorar que una dificultat important de la metodologia CODA és la impossibilitat d'aplicar les transformacions logràtio a les composicions que tenen alguna de les seves components igual a zero. En cas que el valor nul d'una component es pugui interpretar com un valor inferior a un determinat límit de detecció, té sentit reemplaçar el zero per un valor positiu molt petit utilitzant les tècniques de substitució que es fan servir habitualment en el tractament de dades mancants, tot mantenint la coherència composicional de les dades [11, 12]. D'altra banda, quan el valor nul representa efectivament un zero absolut l'estratègia que cal seguir es fonamenta en el fet de suposar que aquests zeros caracteritzen determinades subpoblacions les quals han de ser modelitzades mitjançant models condicionals. Tanmateix la solució del tractament composicional d'aquest darrer cas de zeros resta encara oberta a l'espera d'una metodologia completa que sigui compatible amb la metodologia CODA que s'ha presentat en aquest article.

Agraïments

Vull agrair als professors A. Reventós i J. A. Martín els seus consells i comentaris a l'hora d'escriure aquest article. Els resultats que s'hi recullen formen part del projecte de recerca CODA-RSS finançat pel Ministeri de Ciència i Tecnologia d'Espanya (Ref. MTM2009-13272) i l'Agència de Gestió d'Ajuts Universitaris i de Recerca de la Generalitat de Catalunya (Ref. 2009SGR424).

Referències

[1] AITCHISON, J. «The statistical analysis of compositional data». *J. R. Stat. Soc. B*, 44 (2) (1982), 139–177.

[2] AITCHISON, J. *The statistical analysis of compositional data*. Londres; Nova York: Chapman & Hall, 1986. [Reimpressió: Blackburn Press, 2003]

[3] BARCELÓ-VIDAL, C. «Fundamentación matemática del análisis de datos composicionales». Rep. Tèc. IMA 00-02-RR. Univ. de Girona, Dept. d'Informàtica i Matemàtica Aplicada, 2000.

[4] BARCELÓ-VIDAL, C.; MARTÍN-FERNÁNDEZ, J. «Mathematical foundations of compositional data analysis». A: *Proceedings of IAMG'01. The sixth annual conference of the International Association for Mathematical Geology*, 2001, 20. [CD–ROM]

[5] BOYD, W. C. *Genetics and the races of man: an introduction to modern physical anthropology*. Boston: Little, Brown & Co., 1950.

[6] CHAYES, F. «On correlation between variables of constant sum». *J. Geophys. Res.*, 65 (12) (1960), 4185–4193.

[7] CHAYES, F. «Numerical correlation and petrographic variation». *Jour. Geology*, 70 (4) (1962), 440–452.

[8] CHAYES, F. «Detecting nonrandom associations between proportions by tests of remaining-space variables». *Math. Geol.*, 15 (1) (1983), 197–206.

[9] EGOZCUE, J.; PAWLOWSKY-GLAHN, V. «Simplicial geometry for compositional data». A: *Compositional data analysis: from theory to practice*. Londres: The Geological Society, 2006, 145–159.

[10] EGOZCUE, J.; PAWLOWSKY-GLAHN, V.; MATEU-FIGUERAS, G.; BARCELÓ-VIDAL, C. «Isometric logratio transformations for compositional data analysis». *Math. Geol.*, 35 (3) (2003), 279–300.

[11] MARTÍN-FERNÁNDEZ, J.; BARCELÓ-VIDAL, C.; PAWLOWSKY-GLAHN, V. «Dealing with zeros and missing values in compositional data sets». *Math. Geol.*, 35 (3) (2003), 231–251.

[12] PALAREA-ALBALADEJO, J.; MARTÍN-FERNÁNDEZ, J. «A modified EM alr-algorithm for replacing rounded zeros in compositional data sets». *Computers & Geosciences*, 34 (8) (2008), 902–917.

[13] PAWLOWSKY-GLAHN, V.; BUCCIANTI, A. (ed.). *Compositional data analysis. Theory and applications*. Londres: Wiley, 2011.

[14] PEARSON, K. «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.» A: *Proceedings of the Royal Society of London*, 60 (1897), 489–502.

[15] TANGRI, D.; WRIGHT, R. «Multivariate analysis of compositional data: Applied comparison favour standard principal components analysis over Aitchison's loglinear contrast method». *Archaeometry*, 35 (1) (1993), 103–112.

[16] WHITTEN, E. H. T. «Open and closed compositional data in petrology». *Math. Geol.*, 27 (1995), 789–806.

[17] WORONOW, A. «The elusive benefits of logratios». A: CIMNE (ed.). *Proceedings of IAMG'97. The third annual conference of the International Association for Mathematical Geology*, 1997, 97–101.

DEPARTAMENT D'INFORMÀTICA I MATEMÀTICA APLICADA
UNIVERSITAT DE GIRONA
CAMPUS DE MONTILIVI, 17071 GIRONA
carles.barcelo@udg.edu