Models de distribució sobre el símplex

TESI DOCTORAL

codirigida per la Dra. Vera Pawlowsky Glahn i

pel Dr. Carles Barceló i Vidal

Programa de doctorat: Matemàtica Aplicada - U.P.C.

Glòria Mateu i Figueras

Barcelona, 2003

Prefaci

L'anàlisi estadística de dades composicionals ha estat i continua essent una font de reflexions científiques des de que el 1879 Karl Pearson va posar de manifest els problemes derivats de l'aplicació dels mètodes estadístics clàssics sobre aquest tipus de dades. Les dades composicionals són vectors les components dels quals representen proporcions respecte d'un total i, per tant, estan sotmesos a la restricció que la suma de les seves components és una constant. L'espai natural per a vectors de proporcions amb D components és el símplex $\mathcal{S}^D$. En l'àmbit de la modelització, ens trobem amb una gran dificultat: no coneixem en aquest espai prou classes de distribucions que permetin modelitzar adequadament la majoria dels conjunts de dades composicionals.

En els anys 80, J. Aitchison presenta una metodologia específica per treballar amb dades composicionals que hem anomenat metodologia MOVE, ja que es basa en la tècnica de les transformacions. Aquesta tècnica, tot i que en un principi no fou ben acceptada, s'ha anat aplicant en treballs d'investigació d'àmbits molt diferents. En el tema específic de la modelització de composicions aleatòries, Aitchison utilitza la transformació logquocient additiva per projectar les composicions a l'espai real i, posteriorment, les modela amb una distribució normal multivariant. És d'aquesta manera com introdueix la classe de les distribucions normals logístiques additives sobre l'espai del símplex. Tot i les bones propietats algebraiques que presenta aquesta classe de distribucions, ens trobem amb dues dificultats: el model normal no pot modelitzar alguns conjunts de dades transformades, especialment quan presenten una certa asimetria. Per altra banda, aquesta família de distribucions no és tancada respecte de l'amalgama (o suma) de components. Es coneixen, però, diferents distribucions alternatives en el símplex. Podem anomenar, entre altres, la coneguda família de distribucions de Dirichlet, les seves generalitzacions -com per exemple la distribució de Connor-Mosimann i la

distribució de Liouville-, la família de distribucions que es deriven d'aplicar les transformacions Box-Cox, o bé les distribucions d'Aitchison. Tanmateix aquestes distribucions no ens solucionen els problemes del model normal logístic additiu.

En els inicis d'aquest treball de recerca, ens vàrem marcar com a objectiu principal el desenvolupament d'una nova família de distribucions multivariants sobre el símplex que permetés modelitzar conjunts de dades composicionals quan la distribució normal logística additiva fos insuficient. Utilitzant la teoria de les transformacions d'Aitchison i la distribució normal asimètrica de A. Azzalini, hem definit una nova família de distribucions que hem anomenat normal asimètrica logística additiva. Aquesta família és especialment indicada per modelitzar conjunts de dades composicionals quan la seva transformació, a l'espai real, presenta un biaix moderat. Consegüentment, aquesta família ens aporta la solució a una de les principals dificultats de la família normal logística additiva. Estudiant amb més detall aquest nou model, hem comprovat que presenta unes bones propietats algebraiques, molt similars a les que presenta el model normal logístic additiu.

Un altre dels objectius que es varen plantejar inicialment era l'estudi empíric de la distribució de l'amalgama de components amb distribució normal logística additiva. Mitjançant simulacions, hem pogut il·lustrar l'efecte que tenen els paràmetres de la distribució normal logística additiva inicial en la distribució de l'amalgama i hem pogut comprovar que, en certs casos, el model normal asimètric proporciona un bon ajust per al logquocient de l'amalgama.

Una eina útil en l'estudi de la modelització de vectors aleatoris són els tests de bondat d'ajust. Aquests tests permeten decidir si una distribució és adequada per modelitzar un cert conjunt de dades. Malauradament, no és gens freqüent trobar a la literatura tests de bondat d'ajust aplicables a la distribució normal asimètrica. Per aquesta raó, ens vàrem plantejar un tercer objectiu: desenvolupar un test de bondat d'ajust per a la distribució normal asimètrica i realitzar un estudi de potència utilitzant diverses distribucions alternatives. La metodologia que hem escollit és la de R.B. D'Agostino i M.A. Stephens que consisteix en mesurar, a partir de diferents estadístics, la diferència entre la funció de distribució empírica (calculada mitjançant la mostra) i la funció de distribució teòrica (la normal asimètrica).

Paral·lelament al nostre estudi, ha sorgit l'estructura d'espai vectorial euclidià del símplex que, a banda de donar coherència a la teoria d'Aitchison, ens ha suggerit una nova metodologia. Es tracta de considerar el símplex com un espai vectorial i utilitzar les seves pròpies

operacions, producte escalar i distància. L'hem anomenada metodologia STAY ja que no es basa en les transformacions. L'aplicació d'aquesta nova perspectiva té conseqüències importants, ja que obliga a reescriure i reformular els conceptes i les tècniques estadístiques en funció de les operacions de $\mathcal{S}^D$. No obstant això, hem pogut comprovar que la idea proposada és equivalent a utilitzar les operacions pròpies de l'espai real i les tècniques estadístiques estàndards sobre els coeficients dels elements respecte d'una base ortonormal del símplex. Si bé en determinades situacions aquesta nova metodologia dóna resultats totalment equivalents als obtinguts amb la tècnica de les transformacions, en altres aporta canvis importants. Per exemple, ha permès expressar directament sobre el símplex elements bàsics de l'estadística clàssica com el concepte d'esperança o de variància, i reformular criteris d'optimització (biaix i variància d'un estimador). Com a conseqüència natural d'aquests resultats, ens hem plantejat un nou objectiu: definir models paramètrics sobre els coeficients de les composicions aleatòries respecte d'una base ortonormal. Utilitzant la funció de densitat com a funció d'aquests coeficients, hem definit el model normal i el model normal asimètric a $\mathcal{S}^D$.

La memòria que es presenta recull els resultats que s'han obtingut en l'estudi dels objectius que hem descrit. Hem optat per estructurar-la en cinc capítols.

En el primer, fem un recordatori de conceptes generals bàsics que tindran un paper fonamental en el nostre estudi. Introduïm així el concepte d'espai vectorial euclidià, nocions de teoria de la probabilitat i de variable aleatòria. Reservem un apartat per a l'estudi de la distribució normal asimètrica, tant en el cas univariant, com en el cas multivariant.

En el segon capítol, introduïm els conceptes més importants sobre dades composicionals que utilitzarem posteriorment en l'estudi dels models paramètrics segons les dues metodologies. Per aquesta raó, a banda de veure l'estructura d'espai vectorial, recordem també les transformacions logístiques del símplex a l'espai real, així com certs aspectes generals referents als elements de tendència central i de dispersió.

Dediquem el tercer capítol a presentar les famílies de distribucions sobre el símplex definides mitjançant les transformacions. Fem especial èmfasi en la distribució normal logística additiva introduïda per Aitchison el 1982. Utilitzant simulacions, estudiem la problemàtica de l'amalgama de dues components i proposem el model normal asimètric com una aproximació a la distribució del logquocient d'una amalgama. La part principal d'aquest capítol és la introducció de la distribució normal asimètrica logística additiva i el desenvolupament de

les seves propietats algebraiques en relació a les operacions pertorbació, potència, permutació i subcomposició. Realitzem també un estudi de la distribució del vector aleatori positiu, la composició associada al qual té una distribució normal asimètrica logística additiva. Això comporta la definició del model lognormal asimètric sobre l'espai real. Finalment, dediquem un apartat a la distribució de Dirichlet, veiem alguna de les seves generalitzacions, així com altres distribucions definides a $\mathcal{S}^D$ mitjançant transformacions.

Reservem el quart capítol pels models paramètrics sobre $\mathcal{S}^D$ definits segons la metodologia STAY, és a dir, definits sobre les coordenades d'una composició aleatòria en una base ortonormal de $\mathcal{S}^D$. Com a exemple senzill i il·lustratiu d'aquesta metodologia, iniciem el capítol amb la distribució normal a $\mathbb{R}^+$, definida sobre les coordenades respecte d'una base unitària. Donem la definició del model, estudiem les seves propietats i realitzem una comparació amb el model lognormal clàssic definit mitjançant la transformació logarítmica. A continuació, definim els models normal i normal asimètric a $\mathcal{S}^D$ sobre les coordenades d'una composició aleatòria respecte d'una base ortonormal, desenvolupem les seves propietats i calculem els seus elements característics.

El cinquè capítol conté exclusivament aspectes relacionats amb proves de bondat d'ajust per validar els models construïts segons les dues metodologies. L'originalitat d'aquest capítol es troba bàsicament en els contrastos de bondat d'ajust per al model normal asimètric univariant. Aquests, juntament amb els contrastos de normalitat univariant, permetran validar els models descrits en els capítols anteriors.

Finalment, a l'epíleg, presentem les conclusions derivades del nostre estudi i donem una relació de diferents línies de recerca desenvolupables en un futur. Completem aquesta memòria amb un llistat de les referències bibliogràfiques més importants que s'han consultat al llarg de la investigació.

Agraïments

Primer de tot vull donar les gràcies a la doctora Vera Pawlowsky i al doctor Carles Barceló per la seva tasca com a directors d'aquesta tesi, per les seves oportunes orientacions i per les revisions que han realitzat d'aquest treball. El seu ajut i suport han estat imprescindibles per arribar a la fi d'aquest treball.

I would also like to thank doctor John Aitchison for his important suggestions on my work and for his encouraging support at all time. Al doctor Juan José Egozcue per les seves llargues discussions i el seu assessorament durant aquests anys. Al doctor Pere Puig pels seus consells en relació als tests de bondat d'ajust. Thanks to doctor Adelchi Azzalini for his advice on the skew-normal distribution.

Vull agrair als meus companys del Departament d'Informàtica i Matemàtica Aplicada de la Universitat de Girona la seva companyonia, els seus consells, el suport i els ànims que m'han donat. Esther, David, Narcís, Martin, Pepus, Mei, Marta, Santi, Raimon, Jordi Ripoll, Jordi Poch, Anna, Joan, Àngel, Jaume, Marc, Maria i tots els altres, gràcies per fer-me sentir a gust al vostre costat.

Per acabar, però essent els primers en el meu pensament, gràcies a en Quim, als meus pares, a la Carmen i en Quel i a tota la família per estar al meu costat en tot moment i per entendre com d'important era per mi realitzar aquest treball. Sense la vostra comprensió i paciència no hagués estat possible acabar aquesta tesi.

Girona 18 de juny de 2003

Índex

Capítol 1

Conceptes preliminars generals

Abans d'abordar la temàtica central d'aquest treball d'investigació, hem cregut convenient recordar tots aquells conceptes generals en relació a l'estructura de l'espai sobre el qual definim famílies de distribucions paramètriques. Ens referim a elements bàsics d'un espai vectorial euclidià i a conceptes de teoria de la mesura com, per exemple, variable aleatòria, mesura de probabilitat, funció de densitat, esperança i variància entre d'altres. Dediquem doncs els dos primers apartats d'aquest capítol a presentar les eines sobre les quals se sustenten els capítols posteriors.

L'estructura algebraica de l'espai d'observacions té un paper clau en aquest treball de recerca. En el primer apartat d'aquest capítol introduïm una notació general per als elements de l'espai, les operacions, el producte escalar i la distància entre d'altres. Esmentem també certs aspectes elementals en relació a les coordenades dels vectors respecte d'una base, donat que els utilitzem posteriorment per definir lleis de probabilitat en espais vectorials diferents a $\mathbb{R}$ o $\mathbb{R}^D$.

El segon apartat d'aquest capítol està dedicat exclusivament a presentar certes nocions bàsiques de teoria de la mesura i de variable aleatòria. És ben sabut que la teoria de la mesura es defineix sobre un conjunt qualsevol i que una variable aleatòria no és més que una funció mesurable entre dos espais mesurables qualssevol. Habitualment treballem amb variables o amb vectors aleatoris que tenen imatge a l'espai real. Per aquesta raó, trobem un gran nombre de tècniques i conceptes que han estat desenvolupats utilitzant l'estructura algebraica pròpia de $\mathbb{R}^D$. No obstant això, a la pràctica tenim també variables i vectors

aleatoris definits en espais amb una estructura diferent a la dels reals. En aquests casos, cal anar alerta i utilitzar tan sols els elements i les propietats generals que es compleixen en qualsevol espai de mesura. Per evitar aquestes dificultats, proposem una metodologia alternativa: treballar amb les coordenades respecte d'una base ortonormal de l'espai i aplicar-hi tota l'anàlisi real estàndard, ja que podem identificar aquestes coordenades amb vectors de $\mathbb{R}$ o $\mathbb{R}^D$. Aquesta proposta serà tan sols aplicable quan l'espai on tinguem definida la nostra variable aleatòria tingui estructura d'espai euclidià. És important esmentar que en treballar amb les coordenades dels vectors respecte d'una base podem obtenir resultats de difícil interpretació. Per aquesta raó, sovint cal fer el pas invers i tornar a expressar els resultats en funció dels elements originals de l'espai.

Acabem aquest capítol amb un apartat introductori a la distribució normal asimètrica. Veiem que es tracta d'una generalització del model normal clàssic amb un paràmetre de forma que regula la asimetria de la distribució. Recordem amb detall la definició i les propietats d'aquesta família de distribucions i analitzem la seva problemàtica entorn de l'estimació dels paràmetres.

1.1 Estructura d'espai euclidià. Notació

En aquest apartat agrupem la notació i algunes propietats referents a l'estructura d'un espai euclidià de dimensió finita. No pretenem presentar conceptes molt formalitzats ni amb gran dosi de rigor, tan sols procurem establir les eines imprescindibles i el llenguatge que utilitzarem al llarg d'aquest treball de recerca. Podem trobar un estudi detallat i una construcció rigorosa d'espai euclidià a qualsevol llibre d'àlgebra lineal com per exemple Castellet i Llerena (1990), Xambó (1977) o bé Lang (1971).

És àmpliament conegut que un espai vectorial E sobre un cos K és un conjunt no buit amb dues operacions: una operació interna que li dóna les propietats de grup commutatiu, i una operació externa amb els elements del cos. Suposarem al llarg d'aquesta memòria que treballem sobre el cos commutatiu K=$\mathbb{R}$.

Anomenarem *vectors* als elements de E. Aquests seran denotats amb lletra minúscula i amb negreta excepte en el cas d'un espai vectorial de dimensió 1. Anomenarem *escalars* als elements del cos $\mathbb{R}$. Habitualment els denotarem amb les lletres minúscules de l'alfabet grec,

tot i que en certs casos recorrerem també a l'alfabet llatí. La notació que utilitzarem per designar un espai vectorial amb les dues operacions serà $(\mathrm{E}, \oplus, \otimes)$. Els símbols $\oplus$ i $\otimes$ indicaran respectivament les operacions interna i externa de l'espai, excepte en els casos en què tinguem una notació específica. Així, per exemple, quan treballem a l'espai real ens referirem a la suma i al producte per escalars amb la simbologia habitual. També, sobre la recta real positiva, utilitzarem la notació habitual per indicar el producte i l'operació potència. Observem que els símbols $\oplus$ i $\otimes$ emfatitzen l'analogia amb les operacions estàndards de suma de vectors i producte per escalars que tenim a l'espai $\mathbb{R}^D$.

Donat un espai vectorial $(\mathrm{E}, \oplus, \otimes)$ de dimensió D, indicarem amb $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$ una base de l'espai. També utilitzarem aquesta notació per referir-nos a un sistema de generadors. Sabem que qualsevol vector $\mathbf{v} \in \mathrm{E}$ s'expressa de manera única com a combinació lineal d'elements d'una base, és a dir, existeixen uns únics escalars $\beta_1, \beta_2, \ldots, \beta_D \in \mathbb{R}$ de manera que $\mathbf{v} = (\beta_1 \otimes \mathbf{v}_1) \oplus (\beta_2 \otimes \mathbf{v}_2) \oplus \cdots \oplus (\beta_D \otimes \mathbf{v}_D)$. Aquests escalars reben el nom de *coeficients* o *coordenades* de $\mathbf{v}$ respecte de la base $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$. Ens referirem també al vector $\mathbf{v}$ indicant tan sols les coordenades respecte de la base, és a dir, $(\beta_1, \beta_2, \ldots, \beta_D)'$. D'aquesta manera convertim el vector de E en un vector de l'espai $\mathbb{R}^D$. El gran avantatge de treballar amb les coordenades respecte d'una base és que podem utilitzar les operacions suma i producte per escalars habituals de l'espai $\mathbb{R}^D$. Així per exemple, donats $\mathbf{u}, \mathbf{v} \in \mathrm{E}$ amb coordenades $(\alpha_1, \alpha_2, \ldots, \alpha_D)'$ i $(\beta_1, \beta_2, \ldots, \beta_D)'$ en la base $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$, l'operació interna $\mathbf{u} \oplus \mathbf{v}$ es tradueix a una suma de vectors de $\mathbb{R}^D$, més concretament, les coordenades del vector $\mathbf{u} \oplus \mathbf{v}$ respecte de la base es calculen mitjançant l'operació

$$(\alpha_1, \alpha_2, \ldots, \alpha_D)' + (\beta_1, \beta_2, \ldots, \beta_D)' = (\alpha_1 + \beta_1, \alpha_2 + \beta_2, \ldots, \alpha_D + \beta_D)'. \tag{1.1}$$

De manera similar, donat $a \in \mathbb{R}$, podem comprovar que les coordenades del vector $a \otimes \mathbf{u}$ es poden calcular mitjançant el producte per escalars habitual de $\mathbb{R}^D$, és a dir,

$$a(\alpha_1, \alpha_2, \ldots, \alpha_D)' = (a\alpha_1, a\alpha_2, \ldots, a\alpha_D)'. \tag{1.2}$$

Quan sobre un espai vectorial definim un producte escalar, obtenim una nova estructura algebraica anomenada *espai vectorial euclidià*. Donats dos vectors $\mathbf{u}$ i $\mathbf{v}$ de l'espai E, indicarem amb $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathrm{E}}$ el seu producte escalar. Direm que $\mathbf{u}$ i $\mathbf{v}$ són ortogonals si $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathrm{E}} = 0$ i direm que $\mathbf{u}$ és unitari si $\langle \mathbf{u}, \mathbf{u} \rangle_{\mathrm{E}} = 1$. S'utilitzaran diversos subíndexs per indicar explícitament el

producte escalar que es considera. Reservem el subíndex "eu" per al producte escalar ordinari de l'espai real. Per a qualsevol $\mathbf{u} \in \mathrm{E}$, la determinació positiva de l'arrel quadrada $\sqrt{\langle \mathbf{u}, \mathbf{u} \rangle_{\mathrm{E}}}$ és una norma a E. En aquest cas, es tracta de la normal induïda pel producte escalar que denotarem com $\|\mathbf{u}\|_{\mathrm{E}}$. Com en el cas del producte escalar i per evitar confusions, utilitzarem diversos subíndexs.

Amb aquests elements, la teoria d'àlgebra lineal demostra que tot espai vectorial euclidià admet una base ortonormal de vectors, és a dir, una base amb vectors unitaris i ortogonals dos a dos. Denotarem per $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_D\}$ aquestes bases. Creiem important recordar que si treballem amb les coordenades dels vectors respecte d'una base ortonormal, podem calcular qualsevol producte escalar o qualsevol norma de vectors aplicant el producte escalar ordinari de l'espai $\mathbb{R}^D$ i la corresponent norma induïda. Així doncs, si $\mathbf{u}$ i $\mathbf{v}$ són vectors de l'espai vectorial E amb coordenades $(\alpha_1, \alpha_2, \ldots, \alpha_D)'$ i $(\beta_1, \beta_2, \ldots, \beta_D)'$ respecte de la base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_D\}$, llavors

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathrm{E}} = \langle (\alpha_1, \alpha_2, \ldots, \alpha_D)', (\beta_1, \beta_2, \ldots, \beta_D)' \rangle_{eu} = \sum_{i=1}^{D} \alpha_i \beta_i, \tag{1.3}$$

$$\|\mathbf{u}\|_{\mathrm{E}} = \|(\alpha_1, \alpha_2, \ldots, \alpha_D)'\|_{eu} = \sqrt{\sum_{i=1}^{D} \alpha_i^2}, \tag{1.4}$$

on, com hem indicat, el subíndex "eu" representa les operacions estàndards de l'espai $\mathbb{R}^D$.

Quan parlem de geometria, és usual treballar amb punts i vectors d'un espai. Normalment es diu que un vector és un objecte determinat per una parella ordenada de punts anomenats respectivament origen i extrem. Per definir matemàticament aquests conceptes cal utilitzar una altra estructura anomenada espai afí. Quan associem un espai afí i un espai vectorial euclidià, l'estructura conjunta que en resulta s'anomena *espai afí euclidià* o simplement *espai euclidià*. Per indicar els punts de l'espai afí i les operacions entre ells, utilitzarem la mateixa notació introduïda per als vectors de l'espai vectorial associat. És important esmentar el concepte de *sistema de referència afí*, objecte constituït per un punt, anomenat origen del sistema, i una base de l'espai vectorial associat, perquè ens permet treballar amb les coordenades cartesianes dels punts. Les coordenades cartesianes d'un punt $\mathbf{x}$ són les coordenades respecte d'una base de l'espai del vector que uneix l'origen del sistema de referència amb el punt $\mathbf{x}$. Denotarem aquestes coordenades igual que les coordenades d'un vector respecte d'una base.

Sobre un espai euclidià es defineix el concepte de mètrica o distància entre dos punts com la norma del vector que els uneix. Donats dos punts $\mathbf{x}$ i $\mathbf{y}$ de l'espai afí, utilitzarem $d_{\mathrm{E}}(\mathbf{x},\mathbf{y})$ per indicar la distància entre els punts $\mathbf{x}$ i $\mathbf{y}$. El subíndex E fa referència a la norma i per tant al producte escalar utilitzat en el càlcul. Hem vist que si treballem amb les coordenades d'un vector respecte d'una base ortonormal, podem utilitzar la norma euclidiana habitual. Si parlem en termes de distància aquesta propietat es pot enunciar com:

$$d_{\mathrm{E}}(\mathbf{x},\mathbf{y}) = d_{eu}((\alpha_1,\alpha_2,\ldots,\alpha_D)',(\beta_1,\beta_2,\ldots,\beta_D)') = \sqrt{\sum_{i=1}^{D}(\alpha_i-\beta_i)^2},$$

on $\mathbf{x}$ i $\mathbf{y}$ són punts amb coordenades cartesianes $(\alpha_1,\alpha_2,\ldots,\alpha_D)'$ i $(\beta_1,\beta_2,\ldots,\beta_D)'$ respecte d'un sistema de referència afí amb una base ortonormal. És a dir, treballar amb les coordenades dels punts en un sistema de referència on la base associada sigui ortonormal equival a treballar amb punts de l'espai $\mathbb{R}^D$, i per tant podem utilitzar les operacions clàssiques, entre elles, la distància euclidiana habitual.

Sobre un espai euclidià podem definir multitud de distàncies, tan sols cal definir una aplicació $d_{\mathrm{E}} : \mathrm{E}\times\mathrm{E} \longrightarrow \mathbb{R}$ que compleixi les propietats usuals d'una distància. En particular la distància entre dos punts de l'espai afí abans definida és una distància. En aquest treball d'investigació exigirem a qualsevol d_{E} unes propietats addicionals que seran conseqüència de la pròpia naturalesa de l'espai E. En aquests casos direm que la distància és compatible o coherent amb l'estructura de l'espai.

Definició 1.1 Sigui $(\mathrm{E},\oplus,\otimes)$ un espai euclidià. Direm que $d_{\mathrm{E}} : \mathrm{E}\times\mathrm{E} \longrightarrow \mathbb{R}$ és una distància *compatible* o *coherent* amb l'estructura algebraica de l'espai si és invariant per l'operació $\oplus$ i coherent amb l'operació $\otimes$, és a dir, si compleix les següents igualtats:

$$\begin{aligned} d_{\mathrm{E}}(\mathbf{x},\mathbf{y}) &= d_{\mathrm{E}}(\mathbf{x}\oplus\mathbf{u},\mathbf{y}\oplus\mathbf{u})\ \forall \mathbf{x},\mathbf{y},\mathbf{u}\in\mathrm{E}, \\ d_{\mathrm{E}}(\alpha\otimes\mathbf{x},\alpha\otimes\mathbf{y}) &= |\alpha| d_{\mathrm{E}}(\mathbf{x},\mathbf{y})\ \forall \mathbf{x},\mathbf{y}\in\mathrm{E},\ \forall\alpha\in\mathbb{R}. \end{aligned} \tag{1.5}$$

□

Entenem que el fet d'utilitzar una distància compatible amb l'estructura vectorial de l'espai suposa donar coherència a qualsevol anàlisi. Per exemple a l'espai real és habitual centrar un conjunt de dades, és a dir traslladar-lo a l'origen de coordenades. En aquests casos és

coherent que la distància entre dos punts traslladats sigui igual a la distància entre els dos punts inicials. La igualtat (1.5) ens assegura aquesta propietat ja que l'operació interna de l'espai E és equivalent a la translació de l'espai real. En particular podem observar que la distància entre dos punts de l'espai afí és invariant per translacions (vegeu Xambó, 1997). Aitchison (1992) i Martín-Fernández (2001) exigeixen aquestes propietats en el cas particular d'una distància definida sobre $\mathcal{S}^D$, espai vectorial que introduïm al capítol 2 d'aquesta tesi doctoral.

Recordem que dos espais euclidians es diuen isomorfs si existeix una isometria entre ells. Un resultat important que utilitzarem en aquesta tesi doctoral és el següent:

Propietat 1.1 Dos espais euclidians de dimensió finita són isomorfs si i només si tenen la mateixa dimensió. □

Això indica que, llevat d'isometries, existeix un únic espai euclidià de dimensió finita D, que es pot identificar amb l'espai $\mathbb{R}^D$ amb la suma de vectors, el producte per escalars del cos $\mathbb{R}$, el producte escalar ordinari i la distància euclidiana habitual. Sabem que aquesta isometria conserva les distàncies i el producte escalar, és a dir, conserva el paral·lelisme i la perpendicularitat. Al llarg d'aquest apartat hem aplicat constantment aquesta isometria ja que hem identificat els vectors de E amb les seves coordenades respecte de la base ortonormal i els punts de E amb les seves coordenades cartesianes respecte del sistema de referència afí corresponent.

1.2 Variables aleatòries E-valuades

Aquesta secció és un recordatori dels conceptes bàsics de la teoria matemàtica de la mesura i la probabilitat. No es pretén fer un resum detallat de totes les definicions, teoremes i demostracions. Es pot trobar una exposició rigorosa del tema a Ash (1972), Alabert (1996), Doob (1994) o Malliavin (1995), entre d'altres. L'objectiu és recordar els elements necessaris per poder definir en capítols posteriors lleis de probabilitat per a variables i vectors aleatoris que prenen valors a un espai diferent del real. Ens caldrà, per tant, començar aquesta secció amb la definició de mesura, d'espai de mesura, de funció mesurable i d'integral d'una funció respecte d'una mesura per tot seguit referir-nos a variables aleatòries i a lleis de probabilitat.

Definició 1.2 Sigui Ω un conjunt i $\mathcal{F}$ una σ-àlgebra de parts de Ω. Una *mesura* sobre $\mathcal{F}$ és una aplicació

$$\mu : \mathcal{F} \longrightarrow [0, +\infty],$$

tal que la imatge d'una unió numerable de conjunts de $\mathcal{F}$ disjunts dos a dos és igual a la suma de les imatges de cada conjunt. □

En aquestes condicions direm que $(\Omega, \mathcal{F})$ és un *espai mesurable* i que $(\Omega, \mathcal{F}, \mu)$ és un *espai de mesura*. Si a més es compleix que $\mu(\Omega) = 1$, llavors direm que μ és una *mesura de probabilitat* o simplement una *probabilitat* i que $(\Omega, \mathcal{F}, \mu)$ és un *espai de probabilitat*. Els elements de la σ-àlgebra $\mathcal{F}$ reben el nom de *conjunts mesurables*.

Com a exemple important de σ-àlgebra tenim la σ-àlgebra de Borel definida sobre el conjunt $\Omega = \mathbb{R}$ i denotada per $\mathcal{B}(\mathbb{R})$. Els seus elements s'anomenen *borelians* i els suposarem generats pels conjunts oberts amb la topologia habitual de $\mathbb{R}$. De manera similar es pot definir la σ-àlgebra de Borel sobre el conjunt $\mathbb{R}^D$ denotada per $\mathcal{B}(\mathbb{R}^D)$. Només cal considerar a $\mathbb{R}^D$ la topologia producte de D còpies de $\mathbb{R}$, és a dir, generada pels conjunts $A_1 \times A_2 \times \cdots \times A_D$ amb A_i obert de $\mathbb{R}$ (vegeu Kosniowski, 1989).

Per introduir una mesura sobre $\mathcal{F}$, no cal especificar-la sobre cada conjunt de la σ-àlgebra. La teoria de la mesura ens proporciona estructures auxiliars i teoremes de determinació de mesures que indiquen els conjunts de la σ-àlgebra sobre els quals cal fixar la mesura de manera que aquesta quedi totalment determinada.

Per exemple, es demostra que sobre l'espai mesurable $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ és suficient especificar una mesura sobre el conjunt $\{(a, b] : a < b \in \mathbb{R}\}$ i.e. sobre la col·lició d'intervals semioberts i acotats. En certs casos, podrem especificar una mesura a $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ utilitzant una funció de distribució. Recordem que una *funció de distribució* a $\mathbb{R}$ és una aplicació $F(\cdot)$ creixent i contínua per la dreta que determina una mesura μ mitjançant la fórmula

$$\mu((a, b]) = F(b) - F(a).$$

La coneguda mesura de Lebesgue, que simbolitzarem per λ, és un exemple de mesura que es pot especificar amb la funció de distribució $F(x) = x$ i per tant $\lambda((a, b]) = b - a$. La mesura de Lebesgue té una importància especial ja que és invariant per translacions. Més concretament, es pot demostrar que llevat de constants, l'única mesura invariant per translacions a l'espai

real és la mesura de Lebesgue. És per aquesta raó que l'anàlisi real utilitza la mesura de Lebesgue (Bruna, 1996).

Si es verifica que $F(+\infty) = 1$ i $F(-\infty) = 0$, llavors la mesura és una mesura de probabilitat. Com a exemple, tenim la probabilitat anomenada llei $N(0,1)$, determinada per la funció de distribució $F(x) = \int_{-\infty}^{x} (1/\sqrt{2\pi}) \exp(-y^2/2) dy$. És important destacar que no tota mesura sobre $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ es pot especificar a partir d'una funció de distribució, però sí una àmplia classe de mesures i en particular totes les mesures de probabilitat. Malauradament, no queda clara l'existència d'una funció de distribució que permeti especificar mesures de probabilitat en altres espais mesurables diferents dels reals.

Definició 1.3 Siguin $(\Omega, \mathcal{F})$ i $(\mathrm{E}, \mathcal{E})$ dos espais mesurables. Sigui $\mathrm{x} : \Omega \longrightarrow \mathrm{E}$ una aplicació. Direm que x és una *funció mesurable* respecte de $\mathcal{F}$ i $\mathcal{E}$ si i només si l'invers de tot conjunt mesurable de $\mathcal{E}$ és un conjunt mesurable de $\mathcal{F}$, i.e. $\forall A \in \mathcal{E},\ \mathrm{x}^{-1}(A) \in \mathcal{F}$. □

Si a l'espai $(\Omega, \mathcal{F})$ hi tenim definida una mesura de probabilitat, aleshores la funció mesurable $\mathrm{x} : \Omega \longrightarrow \mathrm{E}$ s'anomena *variable aleatòria* E-*valuada*. Habitualment treballem amb variables aleatòries $\mathbb{R}$-valuades, les quals anomenem simplement *variables aleatòries*. En el cas que E sigui multidimensional, l'aplicació $\mathbf{x}$ rep el nom de *vector aleatori*. Distingirem les variables dels vectors aleatoris denotant aquests últims amb negreta.

Donada una funció mesurable real, $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$, amb espai de mesura $(\Omega, \mathcal{F}, \mu)$, podem construir la integral de la funció x respecte de la mesura μ. Per indicar aquesta integral utilitzarem les notacions

$$\int_{\Omega} \mathrm{x} d\mu, \qquad \text{o} \qquad \int_{\Omega} \mathrm{x}(\omega) d\mu(\omega).$$

El procés de construcció d'aquesta integral es basa en una idea original de Lebesgue. La teoria abstracta de la integral es construeix en un espai de mesura $(\Omega, \mathcal{F}, \mu)$ qualsevol. Es defineix en primer lloc la integral d'una funció indicador, es segueix amb la integral d'una funció elemental i d'una funció mesurable positiva per acabar amb la definició general de la integral d'una funció mesurable qualsevol. És important destacar que en aquest procés de construcció es té en compte l'estructura algebraica de $\mathbb{R}$, si bé no es diu explícitament. Així, per exemple, es defineix una funció elemental com una funció $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$ que pren un

nombre finit de valors o, equivalentment, que es pot expressar com

$$\mathrm{x}(\omega) = \sum_{i=1}^{n} a_i 1_{A_i}(w)$$

per a certs $a_1, a_2, \ldots, a_n \in \mathbb{R}$, $A_1, A_2, \ldots, A_n \in \mathcal{F}$, i on 1_{A_i} és la funció indicador del conjunt A_i. Observem que en l'expressió de la funció elemental hi apareixen les operacions suma i producte per escalars habituals de l'espai $\mathbb{R}$. Per a aquests tipus de funcions es defineix la integral de x respecte de μ com

$$\int_{\Omega} \mathrm{x} d\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$

La idea intuïtiva d'integral d'una funció respecte d'una mesura és la de l'àrea que queda sota la corba. En el cas de funcions elementals, podem interpretar-ho com la suma de les àrees de n rectangles. L'altura de cada rectangle ve donada per la constant a_i i la base per $\mu(A_i)$, valor que en el cas $\Omega = \mathbb{R}$ i $\mu = \lambda$ correspon a la longitud de la base del rectangle. Per a funcions més generals, podem seguir utilitzant la idea intuïtiva d'integral com a àrea. Existeixen nombroses referències amb la definició rigorosa de la integral d'una funció real respecte d'una mesura; podem citar entre altres Ash (1972), Weir (1974), Doob (1994), Malliavin (1995), Bruna (1996) o bé Alabert (1996). No obstant això, no s'ha trobat en la bibliografia consultada la construcció i la interpretació de la integral de funcions mesurables $\mathrm{x} : \Omega \longrightarrow \mathrm{E}$ amb $(\mathrm{E}, \mathcal{E})$ espai mesurable qualsevol ni tampoc en el cas més restrictiu en què E és un espai euclidià real. Insistim en aquesta qüestió perquè el càlcul integral és una eina essencial en el càlcul de probabilitats.

Presentem a continuació procediments per construir noves mesures a partir d'altres de conegudes. Aquestes eines, si bé estan definides sobre espais mesurables qualssevol, són especialment útils i efectives en el càlcul de probabilitats a l'espai $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Veurem també com es deriven resultats que indiquen com s'integra respecte d'aquestes noves mesures si sabem integrar respecte de les originals.

Definició 1.4 Sigui $(\Omega, \mathcal{F}, \mu)$ un espai de mesura, $(\mathrm{E}, \mathcal{E})$ un espai mesurable i $\mathrm{x} : \Omega \longrightarrow \mathrm{E}$ una funció mesurable. Anomenem *mesura imatge* de μ per x a la mesura sobre $(\mathrm{E}, \mathcal{E})$ que es denota per μ_{x} i es defineix com $\mu_{\mathrm{x}}(A) = \mu(\mathrm{x}^{-1}(A))\ \forall A \in \mathcal{E}$. □

Si a l'espai $(\Omega, \mathcal{F})$ hi tenim definida una mesura de probabilitat, p, llavors la mesura imatge de p per la variable aleatòria x és també una probabilitat que s'anomena *llei de la variable*

x o bé *distribució de la variable* x, i es denota per p_{x}. Recordem que, en el cas d'una variable aleatòria real, $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$, aquesta probabilitat o llei de la variable es pot especificar mitjançant una funció de distribució $F_{\mathrm{x}}(\cdot)$ definida com

$$F_{\mathrm{x}}(x) = p_{\mathrm{x}}((-\infty, x]).$$

Teorema 1.1 (Teorema de la mesura imatge). Sigui $(\Omega, \mathcal{F}, \mu)$ un espai de mesura, $(\mathrm{E}, \mathcal{E})$ un espai mesurable, $\mathrm{x} : \Omega \longrightarrow \mathrm{E}$ una funció mesurable i μ_{x} la mesura imatge de μ per x. Sigui $g : \mathrm{E} \longrightarrow \mathbb{R}$ una funció mesurable. Aleshores,

$$\int_{\mathrm{E}} g d\mu_{\mathrm{x}} = \int_{\Omega} (g \circ \mathrm{x}) d\mu.$$

□

És a dir, la integral de la funció mesurable g respecte de la mesura imatge de μ per x és equivalent a la integral de la funció mesurable $g \circ \mathrm{x}$ respecte de la mesura μ. Observem que aquest resultat és la generalització del teorema del canvi de variable per a integrals sobre $\mathbb{R}$.

Definició 1.5 Sigui $(\Omega, \mathcal{F}, \mu)$ un espai de mesura. Sigui $f : \Omega \longrightarrow [0, +\infty]$ una funció mesurable. Definim sobre $\mathcal{F}$ una altra mesura ν com

$$\nu(A) = \int_A f d\mu \qquad \forall A \in \mathcal{F}.$$

En aquesta situació, diem que f és la *densitat de ν respecte de μ* i s'escriu $f = d\nu/d\mu$. També es diu que f és la *derivada de Radon-Nikodým* de ν respecte de μ. □

Aquesta definició ens permet especificar mesures sobre un espai mitjançant les funcions de densitat. No obstant això, caldrà tenir definida la integral de qualsevol funció de l'espai respecte de la mesura original μ. La definició 1.5 té molta importància dins la teoria de probabilitats a l'espai $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, ja que gairebé la totalitat de les lleis de probabilitats conegudes es defineixen a partir de la seva funció de densitat respecte de la mesura de Lebesgue. Per aquesta raó, habitualment parlem de funcions de densitat sense esmentar la mesura μ donat que entenem que és la mesura de Lebesgue. En aquests casos, és possible obtenir la probabilitat de qualsevol conjunt mesurable de la σ-àlgebra a partir d'una integral ordinària.

En general, quan $(\Omega, \mathcal{F}, p)$ és un espai de probabilitat i $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$ una variable aleatòria, treballem amb la densitat o derivada de Radon-Nikodým de la probabilitat p_{x} respecte de

la mesura de Lebesgue. En aquests casos, és habitual anomenar-la funció de densitat de la variable x i fins i tot denotar-la com f_{x}.

Definició 1.6 Siguin μ i ν dues mesures sobre un mateix espai mesurable $(\Omega, \mathcal{F})$. Diem que ν és una mesura *absolutament contínua respecte de* μ, i escrivim $\nu << \mu$, si i només si

$$\forall A \in \mathcal{F},\ \mu(A) = 0 \Rightarrow \nu(A) = 0.$$

□

Teorema 1.2 (Teorema de Radon-Nikodým). Siguin μ i ν dues mesures sobre un espai mesurable $(\Omega, \mathcal{F})$. Aleshores són equivalents

- $\nu << \mu$,
- existeix $f : \Omega \longrightarrow [0, +\infty]$ tal que $f = d\nu/d\mu$.

□

Propietat 1.2 Siguin $(\Omega, \mathcal{F})$ un espai mesurable, $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$ una funció mesurable i μ i ν dues mesures sobre Ω tals que $\nu << \mu$ amb densitat $f = d\nu/d\mu$. Aleshores x és integrable respecte de ν si i només si $\mathrm{x} \cdot f$ és integrable respecte de μ, i en aquest cas es compleix que

$$\int_{\Omega} \mathrm{x} d\nu = \int_{\Omega} \mathrm{x} \cdot f d\mu.$$

□

La importància del teorema de Radon-Nikodým i de la propietat 1.2 que se'n deriva, està en el càlcul efectiu d'integrals a l'espai $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Si en la propietat 1.2 $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ i μ és la mesura de Lebesgue, obtenim una integral ordinària. Més en general, siguin $(\Omega, \mathcal{F}, p)$ un espai de probabilitat, $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$ una variable aleatòria i $g : \mathbb{R} \longrightarrow \mathbb{R}$ una funció mesurable i suposem que volem calcular $\int_{\mathbb{R}} (g \circ \mathrm{x}) dp$. Si coneixem la mesura imatge p_{x}, llavors el teorema 1.1 assegura la igualtat

$$\int_{\Omega} (g \circ \mathrm{x}) dp = \int_{\mathbb{R}} g dp_{\mathrm{x}}.$$

Si a més $p_{\mathrm{x}} << \lambda$, el teorema 1.2 confirma l'existència d'una densitat f i la propietat 1.2 garanteix la igualtat

$$\int_{\mathbb{R}} g d\mathrm{p}_{\mathrm{x}} = \int_{\mathbb{R}} g \cdot f d\lambda.$$

Obtenim així una integral ordinària sobre $\mathbb{R}$ que habitualment denotem com

$$\int_{\mathbb{R}} g(x)f(x)dx.$$

Tal i com hem indicat abans, en aquests casos és habitual anomenar la funció f com la densitat de la variable aleatòria x.

Definició 1.7 Sigui $(\Omega, \mathcal{F}, p)$ un espai de probabilitat. Sigui $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$ una variable aleatòria real tal que la seva integral respecte de p existeix. Aleshores la integral

$$\mathrm{E[x]} = \int_{\Omega} \mathrm{x}dp$$

s'anomena *esperança* de la variable x i es denota com E[x]. □

Gràcies al teorema de la mesura imatge podrem convertir aquesta integral en una integral sobre l'espai $\mathbb{R}$. Observem en primer lloc que

$$\int_{\Omega} \mathrm{x}dp = \int_{\Omega} (id \circ \mathrm{x})dp,$$

on id representa la funció identitat sobre la recta real. Prenem sobre $\mathbb{R}$ la mesura imatge de p per la variable x i la funció $g = id : \mathbb{R} \longrightarrow \mathbb{R}$. Aleshores el teorema de la mesura imatge assegura que

$$\int_{\Omega} (id \circ \mathrm{x})dp = \int_{\mathbb{R}} id(x)dp_{\mathrm{x}}(x) = \int_{\mathbb{R}} xdp_{\mathrm{x}}. \tag{1.6}$$

Tot seguit, si la probabilitat p_{x} resulta ser absolutament contínua respecte de la mesura de Lebesgue de $\mathbb{R}$, el teorema de Radon-Nikodým ens assegura l'existència de la funció de densitat f, de manera que

$$\int_{\mathbb{R}} xdp_{\mathrm{x}} = \int_{\mathbb{R}} xfd\lambda,$$

que habitualment denotem com

$$\int_{\mathbb{R}} xf(x)dx. \tag{1.7}$$

D'aquesta manera el càlcul de l'esperança també es redueix al càlcul d'una integral ordinària.

L'estudi de l'esperança d'una variable aleatòria real té una especial importància ja que s'interpreta com el centre de masses de la distribució. Existeixen també altres característiques associades a una variable aleatòria real que tenen un paper clau en l'estadística matemàtica: els moments i els moments centrats.

Definició 1.8 Siguin $(\Omega, \mathcal{F}, p)$ un espai de probabilitat, $\mathrm{x} : \Omega \longrightarrow \mathbb{R}$ una variable aleatòria real i $r \geq 1$. Aleshores $\mathrm{E}[\mathrm{x}^r]$ i $\mathrm{E}[(\mathrm{x}-\mathrm{E}[\mathrm{x}])^r]$, si existeixen, s'anomenen respectivament *moment d'ordre r* i *moment d'ordre r centrat* de la variable x. □

Es pot comprovar que el moment d'ordre 1 és l'esperança definida anteriorment i que el moment d'ordre 1 centrat val sempre 0. El moment d'ordre 2 centrat s'anomena també *variància* de x i la seva arrel quadrada positiva s'anomena *desviació estàndard*. Normalment es denoten per var[x] i per std[x] respectivament. Utilitzant les propietat de les integrals a l'espai real es pot demostrar que $\mathrm{var}[\mathrm{x}] = \mathrm{E}[\mathrm{x}^2] - \mathrm{E}[\mathrm{x}]^2$ (vegeu Ash, 1972).

Aquest moment d'ordre 2 centrat té també una importància especial ja que el seu valor indica la variabilitat respecte del centre de masses de la variable aleatòria. Podem, però, reinterpretar aquest moment com el valor esperat de la distància euclidiana al quadrat al voltant del centre de masses. Aquesta és la interpretació utilitzada per Pawlowsky-Glahn i Egozcue (2000,2001) per definir la variància mètrica d'una composició aleatòria.

Quan treballem amb vectors aleatoris reals $\mathbf{x} : \Omega \longrightarrow \mathbb{R}^D$, comptem amb les generalitzacions al cas multivariant de tots els elements vistos fins al moment. Tal i com hem indicat a l'inici d'aquesta secció, els elements de la σ-àlgebra de Borel $\mathcal{B}(\mathbb{R}^D)$ són $A_1 \times A_2 \times \cdots \times A_D$ amb $A_i \in \mathcal{B}(\mathbb{R})$. El teorema de la mesura producte (vegeu Alabert, 1996) ens assegura l'existència de mesures sobre l'espai $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$. Per exemple, la mesura de Lebesgue λ a l'espai real D-dimensional es defineix com la mesura producte de les mesures de Lebesgue de cada espai $\mathbb{R}$, i es denota per $\lambda = \lambda_1 \times \lambda_2 \times \cdots \times \lambda_D$, on cada λ_i representa la mesura de Lebesgue del i-èsim espai $\mathbb{R}$. En aquest cas, la mesura d'un element de la σ-àlgebra es calcula com $\lambda(A_1 \times A_2 \times \cdots \times A_D) = \lambda_1(A_1)\lambda_2(A_2)\cdots\lambda_D(A_D)$. La mesura de Lebesgue a $\mathbb{R}^D$ és l'única mesura, llevat de constants, que és invariant per translacions i per això s'utilitza en l'anàlisi real.

Sigui $(\Omega, \mathcal{F}, p)$ un espai de probabilitat i $\mathbf{x} : \Omega \longrightarrow \mathbb{R}^D$ un vector aleatori donat per $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_D)'$, es defineix la llei del vector $\mathbf{x}$, altrament anomenada *llei conjunta*, com la mesura imatge de p per $\mathbf{x}$. Aquesta llei determina totalment la llei de cada component x_i que s'anomena *llei marginal de* x_i *en* $\mathbf{x}$, ja que donat $B \in \mathcal{B}(\mathbb{R})$,

$$p_{\mathrm{x}_i}(B) = p_{\mathbf{x}}(\mathbb{R} \times \cdots \times \mathbb{R} \times B \times \mathbb{R} \times \cdots \times \mathbb{R}).$$

En el cas de vectors aleatoris reals podem especificar la llei de probabilitat conjunta mitjançant una funció de distribució $F_{\mathbf{x}}(\cdot)$, que es defineix com:

$$\begin{aligned} F_{\mathbf{x}}(x_1, x_2, \ldots, x_D) &= p_{\mathbf{x}}((-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_D]) \\ &= p(\mathrm{x}_1 \leq x_1, \mathrm{x}_2 \leq x_2, \ldots, \mathrm{x}_D \leq x_D). \end{aligned}$$

La funció de distribució d'una llei marginal es determina immediatament a partir de la funció de distribució de la llei conjunta, ja que

$$F_{\mathrm{x}_i}(x_i) = p(\mathrm{x}_1 \in \mathbb{R}, \ldots, \mathrm{x}_{i-1} \in \mathbb{R}, \mathrm{x}_i \leq x_i, \mathrm{x}_{i+1} \in \mathbb{R}, \ldots, \mathrm{x}_D \in \mathbb{R}).$$

En cas de treballar amb una llei de probabilitat absolutament contínua respecte d'una mesura $\mu = \mu_1 \times \mu_2 \times \cdots \times \mu_D$ de l'espai $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$, el teorema de Radon-Nikodým ens assegura l'existència d'una funció de densitat conjunta $f_{\mathbf{x}}$ la qual determina també les densitats de cada marginal f_{x_i} a partir de l'expressió

$$f_{\mathrm{x}_i}(x_i) = \int_{\mathbb{R}^{D-1}} f_{\mathbf{x}}(x_1, x_2, \ldots, x_D) d\mu_1(x_1) \ldots d\mu_{i-1}(x_{i-1}) d\mu_{i+1}(x_{i+1}) \ldots d\mu_D(x_D).$$

A la pràctica, es defineix la llei conjunta d'un vector aleatori mitjançant la funció de densitat respecte de la mesura de Lebesgue de l'espai $\mathbb{R}^D$. Per aquesta raó, també parlem de funció de densitat del vector sense fer referència a la mesura μ, donat que entenem que aquesta és la mesura de Lebesgue.

Quan treballem amb més d'una variable aleatòria podem introduir el concepte d'independència. Existeixen teoremes de la teoria de la mesura que caracteritzen la independència entre variables aleatòries en termes de la seva llei, distribució o densitat conjuntes. Així, donada una mesura producte $\mu = \mu_1 \times \mu_2 \times \cdots \times \mu_D$ a $(\mathbb{R}^D, (\mathcal{B}(\mathbb{R}^D))$ i donades les variables aleatòries $\mathrm{x}_i : \Omega \longrightarrow \mathbb{R}$ amb $i = 1, 2, \ldots, D$, direm que són independents si i només si es compleix una de les tres condicions equivalents:

1. La llei del vector $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_D)'$ és producte de les lleis marginals. És a dir, $p_{\mathbf{x}} = p_{\mathrm{x}_1} \times p_{\mathrm{x}_2} \times \cdots \times p_{\mathrm{x}_D}$.

2. La funció de distribució conjunta és igual al producte de les funcions de distribució de les marginals. És a dir, $F_{\mathbf{x}}(x_1, x_2, \ldots, x_D) = F_{\mathrm{x}_1}(x_1) F_{\mathrm{x}_2}(x_2) \cdots F_{\mathrm{x}_D}(x_D)$.

3. La funció de densitat de la llei conjunta respecte de la mesura μ és producte de les densitats de les lleis marginals respecte de les mesures μ_i. És a dir, $f_{\mathbf{x}}(x_1, x_2, \ldots, x_D) = f_{\mathrm{x}_1}(x_1) f_{\mathrm{x}_2}(x_2) \cdots f_{\mathrm{x}_D}(x_D)$.

L'esperança de $\mathbf{x}$ es defineix com el vector que conté l'esperança de cada marginal, és a dir, $\mathrm{E}[\mathbf{x}] = (\mathrm{E}[\mathrm{x}_1], \mathrm{E}[\mathrm{x}_2], \ldots, \mathrm{E}[\mathrm{x}_D])'$. Podem calcular també la variància de cada marginal segons l'expressió $\mathrm{E}[(\mathrm{x}_i - \mathrm{E}[\mathrm{x}_i])^2]$, per a $i = 1, 2, \ldots, D$. En cas de tenir dues variables aleatòries, x_i i x_j, es defineix un concepte addicional: el de *covariància*. La covariància és una mesura del grau de dependència lineal entre dues variables i es calcula a partir de l'expressió

$$\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j) = \mathrm{E}[(\mathrm{x}_i - \mathrm{E}[\mathrm{x}_i])(\mathrm{x}_j - \mathrm{E}[\mathrm{x}_j])] = \mathrm{E}[\mathrm{x}_i \mathrm{x}_j] - \mathrm{E}[\mathrm{x}_i]\mathrm{E}[\mathrm{x}_j].$$

Si la covariància és nul·la es diu que les variables estan incorrelacionades. Es pot demostrar que si x_i i x_j són independents llavors són també incorrelacionades però el recíproc és, en general, fals. Donat un vector aleatori $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_D)'$, podem calcular la covariància entre cada parella de components i la variància de cada component. Totes aquestes mesures es recullen en la matriu de covariàncies, la qual conté, en la diagonal, les variàncies de cada marginal i, fora la diagonal, les covariàncies entre parelles de variables ordenades per files i columnes segons la seva posició en el vector aleatori. És a dir,

$$\begin{pmatrix} \mathrm{var}[\mathrm{x}_1] & \mathrm{cov}(\mathrm{x}_1, \mathrm{x}_2) & \cdots & \mathrm{cov}(\mathrm{x}_1, \mathrm{x}_D) \\ \mathrm{cov}(\mathrm{x}_2, \mathrm{x}_1) & \mathrm{var}[\mathrm{x}_2] & \cdots & \mathrm{cov}(\mathrm{x}_2, \mathrm{x}_D) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathrm{x}_D, \mathrm{x}_1) & \mathrm{cov}(\mathrm{x}_D, \mathrm{x}_2) & \cdots & \mathrm{var}[\mathrm{x}_D] \end{pmatrix}.$$

Observem que aquesta matriu és necessàriament simètrica ja que $\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j) = \mathrm{cov}(\mathrm{x}_j, \mathrm{x}_i)$.

Ja hem indicat que habitualment treballem amb variables o vectors aleatoris reals. En aquests casos, comptem amb la funció de densitat respecte de la mesura de Lebesgue que ens determina totalment la llei de probabilitat mitjançant el simple càlcul d'integrals ordinàries. Aquesta densitat també permet calcular l'esperança de la variable segons l'expressió (1.7), la variància i altres moments d'ordre superior.

Malgrat tot, a la pràctica trobem variables i vectors aleatoris definits a un espai E diferent de l'espai real. En aquest espai, pot no tenir sentit considerar la mesura de Lebesgue. Pensem per exemple en un espai discret, on treballem amb la mesura comptadora. Ens interessarà,

però, tenir una mesura "semblant"a la mesura de Lebesgue, és a dir, una mesura adequada o coherent amb l'estructura algebraica de l'espai E. En particular voldrem que la nostra mesura sigui invariant per les translacions de l'espai E, operació indicada per $\oplus$ a l'apartat anterior. No obstant això, el fet de treballar amb una mesura diferent a la mesura de Lebesgue té conseqüències importants, sobretot en el cas de variables o vectors aleatoris continus. És possible que no sigui immediat calcular la probabilitat de qualsevol esdeveniment, l'esperança o la variància d'una variable aleatòria. Cal tenir present que moltes de les eines de càlcul s'han desenvolupat majoritàriament per al cas real i utilitzant l'estructura algebraica pròpia dels reals. Tot i així, tenim un gran nombre de definicions i teoremes, com per exemple, el teorema de la mesura imatge o el teorema de Radon-Nikodým, que estan demostrats en qualsevol espai mesurable.

La solució que cal adoptar en aquests casos consisteix en utilitzar les eines generals aplicables en qualsevol espai mesurable o, en tot cas, adaptar i redefinir certs conceptes en relació a l'estructura algebraica del nostre espai particular. No obstant això, i en el cas que el nostre espai tingui una estructura d'espai euclidià, proposem una solució més senzilla que evita aquesta redefinició de conceptes i propietats: treballar amb les coordenades respecte d'una base ortonormal de l'espai. Tal i com hem indicat a l'apartat 1.1, qualsevol espai euclidià de dimensió D és isomorf a $\mathbb{R}^D$. Per fer ús d'aquest isomorfisme només cal identificar els vectors de coordenades respecte d'una base ortonormal amb vectors de l'espai $\mathbb{R}^D$. Així doncs, donada una variable aleatòria $\mathrm{x} : \Omega \longrightarrow \mathrm{E}$, amb $(\Omega, \mathcal{F}, p)$ espai de probabilitat i $(\mathrm{E}, \mathcal{E})$ espai mesurable amb estructura d'espai euclidià, treballarem amb les coordenades de la variable x respecte d'una base ortonormal de l'espai E. Sobre aquestes coordenades podrem definir la llei de la variable i obtenir la seva funció de distribució així com la seva funció de densitat respecte de la mesura de Lebesgue. Aquestes funcions ens permetran calcular qualsevol probabilitat utilitzant les eines de càlcul pròpies de l'espai real. També podrem calcular l'esperança i tots els altres moments mitjançant les expressions habituals. Tot i això, cal anar amb compte amb els resultats obtinguts. Quan calculem elements de l'espai suport com per exemple l'esperança, la mediana o qualsevol altre percentil mitjançant les coordenades respecte d'una base ortonormal, els resultats seran també coordenades respecte de la mateixa base. Podrem, però, obtenir l'element de l'espai E mitjançant una simple combinació lineal. En resum, la nostra proposta consisteix tan sols en identificar l'espai real amb l'espai de

coordenades respecte d'una base ortonormal d'un espai euclidià de dimensió finita i sobre el cos dels reals.

Ens cal, però, fer una matisació. En certs casos el nostre espai E serà un subconjunt de $\mathbb{R}$ o $\mathbb{R}^D$ i per tant, a efectes pràctics, haurem de distingir entre dues possibles interpretacions. La primera consisteix en considerar E com a subconjunt dels reals i utilitzar l'estructura algebraica d'aquest, és a dir, les operacions suma i producte per escalars i el producte escalar, la norma i la distància euclidiana habituals. En aquesta situació definirem les lleis de probabilitat de la manera clàssica, això és, definir les lleis directament sobre E amb la densitat respecte de la mesura de Lebesgue, o bé, utilitzar la tècnica clàssica d'aplicar una transformació a la variable o vector aleatori. La segona interpretació consisteix en considerar E com un espai vectorial per ell mateix i utilitzar la seva pròpia estructura algebraica, és a dir, l'operació interna $\oplus$, l'operació externa $\otimes$, el producte escalar $\langle\ \rangle_{\mathrm{E}}$, la norma $\|\ \|_{\mathrm{E}}$ i la distància d_{E}. Serà en aquests casos quan definirem les lleis de probabilitat treballant amb les coordenades respecte d'una base ortonormal.

En el següent capítol introduïm l'espai vectorial del símplex i veiem que, tot i ser un subconjunt de $\mathbb{R}^D$, té la seva pròpia estructura algebraica. En el capítol 3 veiem lleis de probabilitat sobre el símplex definides considerant-lo com a subconjunt de $\mathbb{R}^D$. Finalment, en el capítol 4, considerem l'estructura algebraica pròpia del símplex i definim les lleis de probabilitat sobre les coordenades del vector aleatori en una base ortonormal.

1.3 Distribució normal asimètrica a l'espai real

1.3.1 Distribució normal asimètrica univariant

La distribució normal asimètrica univariant, coneguda més popularment amb la terminologia anglesa de *skew-normal*, fou introduïda i estudiada amb detall per Azzalini (1985).

Definició 1.9 Donada una variable aleatòria z, direm que té una distribució *normal asimètrica* si és contínua i la seva funció de densitat és

$$f_{\mathrm{z}}(z) = 2\phi(z)\Phi(\lambda z), \qquad \lambda, z \in \mathbb{R},$$

on ϕ i Φ són respectivament la funció de densitat i la funció de distribució d'una normal estàndard. Utilitzarem la notació $\mathrm{z} \sim \mathcal{SN}(\lambda)$. □

El paràmetre λ és el paràmetre de forma de la distribució. Té el seu domini a tota la recta real. Quan $\lambda = 0$ obtenim la densitat d'una $\mathcal{N}(0,1)$ i quan λ tendeix a $\pm\infty$ la distribució tendeix a una normal truncada en el punt 0. La asimetria de la distribució creix a mesura que el valor absolut del paràmetre λ augmenta tot i que, a partir del valor $\lambda = 20$, aquest augment és gairebé inapreciable. Per a valors de λ positius obtenim una distribució amb biaix a la dreta i per a valors de λ negatius el biaix ens apareix a l'esquerra. El paràmetre λ està estretament relacionat amb un paràmetre que anomenem δ mitjançant les expressions següents:

$$\lambda = \frac{\delta}{\sqrt{1-\delta^2}}, \qquad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}. \tag{1.8}$$

Aquest paràmetre δ varia en l'interval $(-1,1)$. Al llarg d'aquest treball de recerca, per raons pràctiques, ens referirem en alguns casos al paràmetre δ en comptes de λ.

A la figura 1.1 hem representat la funció de densitat de diverses normals asimètriques. Podem observar clarament que a mesura que el paràmetre λ augmenta, la densitat esdevé més asimètrica. Amb el mateix gràfic observem una de les propietats de les variables normals asimètriques enunciada a Azzalini (1985): si z $\sim \mathcal{SN}(\lambda)$ llavors $-$z $\sim \mathcal{SN}(-\lambda)$.

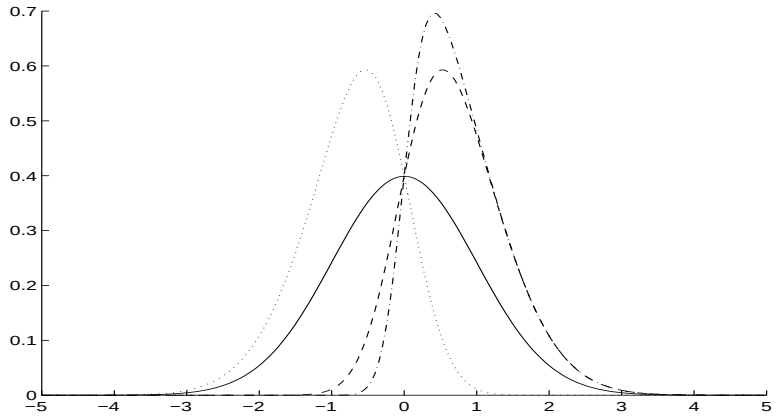


Figura 1.1: *Funcions de densitat* $\mathcal{SN}^1(0)$ (———), $\mathcal{SN}^1(2)$ (- - - -), $\mathcal{SN}^1(-2)$ (.....) *i* $\mathcal{SN}^1(4)$ (._._._.)

A continuació enunciem tres propietats (Azzalini i Dalla-Valle, 1996) que ens donen un procediment per generar mostres aleatòries. En particular veurem que podem obtenir mostres aleatòries d'una normal asimètrica a partir de mostres aleatòries de normals estàndards.

Propietat 1.3 Siguin y i w dues variables aleatòries independents amb distribució $\mathcal{N}(0,1)$.

Llavors la distribució de la variable

$$\mathrm{z} = \begin{cases} \mathrm{y} & si \quad \lambda\mathrm{y} > \mathrm{w}; \\ -\mathrm{y} & si \quad \lambda\mathrm{y} \leq \mathrm{w}; \end{cases}$$

és $\mathcal{SN}(\lambda)$. □

Propietat 1.4 Siguin y_0 i y_1 dues variables aleatòries independents amb distribució $\mathcal{N}(0,1)$. Sigui $\delta \in (-1,1)$. Llavors la variable $\mathrm{z} = \delta|\mathrm{y}_0| + (1-\delta^2)^{1/2}\mathrm{y}_1$ té una distribució $\mathcal{SN}(\lambda)$, on $\lambda = \delta/\sqrt{1-\delta^2}$. □

Propietat 1.5 Sigui $(\mathrm{x},\mathrm{y})'$ un vector aleatori normal bivariant amb marginals estandarditzades i correlació δ. Llavors la distribució de la variable

$$\mathrm{z} = \begin{cases} \mathrm{y} & si \quad \mathrm{x} > 0; \\ -\mathrm{y} & si \quad \mathrm{x} \leq 0; \end{cases}$$

és $\mathcal{SN}(\lambda)$, on $\lambda = \delta/\sqrt{1-\delta^2}$. □

Azzalini (1985) ens dóna l'expressió de la funció generatriu de moments per a una variable normal asimètrica:

$$\mathrm{M}(t) = \mathrm{E}[e^{t\mathrm{z}}] = 2\exp(t^2/2)\Phi(\delta t).$$

Amb aquesta funció podem calcular una expressió per a la mitjana, la variància i el coeficient d'asimetria γ_1:

$$\begin{aligned} \mathrm{E}[\mathrm{z}] &= \sqrt{\frac{2}{\pi}}\delta; \\ \mathrm{var}[\mathrm{z}] &= 1 - \mathrm{E}[\mathrm{z}]^2 = 1 - \frac{2}{\pi}\delta^2; \\ \gamma_1[\mathrm{z}] &= \frac{4-\pi}{2}\left(\frac{\mathrm{E}[\mathrm{z}]}{\sqrt{1-\mathrm{E}[\mathrm{z}]^2}}\right)^3. \end{aligned} \tag{1.9}$$

Es pot demostrar fàcilment que l'índex d'asimetria està acotat aproximadament dins l'interval $(-0.995, +0.995)$ i que quan $\lambda \longrightarrow \pm\infty$, o bé $\delta \longrightarrow \pm 1$, llavors γ_1 tendeix a ± 0.995. Observem doncs que la classe normal asimètrica ens proporcionarà tan sols densitats amb una asimetria moderada.

L'expressió de la funció de distribució d'una variable $\mathrm{z} \sim \mathcal{SN}(\lambda)$ és:

$$\begin{aligned} \mathrm{F}(z) = \mathrm{P}(\mathrm{z} \leq z) &= 2\int_{-\infty}^{z} \phi(t)\Phi(\lambda t)dt \\ &= 2\int_{-\infty}^{z}\int_{-\infty}^{\lambda t} \phi(t)\phi(u)dudt. \end{aligned} \tag{1.10}$$

Azzalini (1985) demostra que la funció de distribució (1.10) és igual a:

$$\mathrm{F}(z) = \Phi(z) - 2\mathrm{T}(z, \lambda),$$

on $T(z, \lambda)$ és la funció d'Owen. Aquesta funció ens dóna, per a valors positius de z i λ, la integral de la densitat bivariant normal estàndard sobre la regió limitada per les rectes $x = z, y = 0$ i $y = \lambda x$ en el pla (x, y). Les igualtats $\mathrm{T}(z, \lambda) = \mathrm{T}(-z, \lambda)$ i $-\mathrm{T}(z, \lambda) = \mathrm{T}(z, -\lambda)$ ens permeten utilitzar aquesta funció per a valors negatius de z i λ. A Young i Minder (1974) trobem la subrutina AS 76 que avalua la funció d'Owen i a Youn-Min (1985) trobem la subrutina AS 55 que proporciona una millora de la subrutina anterior. Així doncs, per calcular qualsevol probabilitat acumulada podrem aplicar algun mètode d'integració numèrica directament a l'expressió (1.10) o bé utilitzar la funció d'Owen.

Una propietat interessant de la distribució normal asimètrica és que el seu quadrat esdevé una variable χ^2_1 (Azzalini, 1985).

Propietat 1.6 Si $\mathrm{z} \sim \mathcal{SN}(\lambda)$, llavors $\mathrm{z}^2 \sim \chi^2_1$. □

A la pràctica, quan treballem amb dades reals, haurem d'incloure un paràmetre de localització i un paràmetre d'escala. Així doncs, treballarem amb la família de distribucions generades mitjançant la transformació lineal

$$\mathrm{y} = \mu + \sigma \mathrm{z}, \tag{1.11}$$

on $\mathrm{z} \sim \mathcal{SN}(\lambda)$ i $\mu, \sigma \in \mathbb{R}$, amb $\sigma > 0$. La funció de densitat per a la variable aleatòria transformada y és $2\phi(y; \mu, \sigma)\Phi(\lambda(y - \mu)/\sigma)$. Utilitzarem la notació $\mathrm{y} \sim \mathcal{SN}(\mu, \sigma, \lambda)$ per referir-nos a aquesta variable. La funció generatriu de moments, l'esperança i la variància de y són:

$$\begin{aligned} \mathrm{M}(t) &= 2\exp(t\mu + t^2\sigma^2/2)\Phi(\sigma\delta t), \\ \mathrm{E}[\mathrm{y}] &= \mu + \sigma\mathrm{E}[\mathrm{z}], \\ \mathrm{var}[\mathrm{y}] &= \sigma^2\mathrm{var}[\mathrm{z}], \end{aligned} \tag{1.12}$$

però l'índex d'asimetria resulta invariant, és a dir, $\gamma_1[\mathrm{y}] = \gamma_1[\mathrm{z}]$.

Per generar una mostra aleatòria de la variable y tan sols haurem de generar una mostra aleatòria de la variable z i transformar-la aplicant (1.11).

Donada y $\sim \mathcal{SN}(\mu, \sigma, \lambda)$ podrem calcular probabilitats acumulades aplicant un mètode d'integració numèrica a la funció de densitat o bé utilitzant la funció d'Owen segons l'expressió:

$$\mathrm{F}(y) = \mathrm{P}(\mathrm{y} \leq y) = \Phi\left(\frac{y-\mu}{\sigma}\right) - 2\mathrm{T}\left(\frac{y-\mu}{\sigma}, \lambda\right).$$

1.3.2 Distribució normal asimètrica multivariant

La versió multivariant de la distribució normal asimètrica fou introduïda per Azzalini i Dalla-Valle (1996).

Definició 1.10 Donat un vector aleatori $\mathbf{z}$ de dimensió $D \times 1$, direm que té una distribució *normal asimètrica multivariant* si és continu i la seva funció de densitat és

$$f_{\mathbf{z}}(\mathbf{z}) = 2\phi_D(\mathbf{z}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}'\mathbf{z}), \qquad \mathbf{z} \in \mathbb{R}^D, \tag{1.13}$$

on $\phi_D(.; \boldsymbol{\Omega})$ representa la funció de densitat d'un vector normal $(D \times 1)$-dimensional amb marginals estandarditzades i matriu de correlació $\boldsymbol{\Omega}$; $\Phi(.)$ és la funció de distribució d'una $\mathcal{N}(0,1)$; i $\boldsymbol{\alpha} \in \mathbb{R}^D$. Utilitzarem la notació $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$. □

En aquest cas anomenem $\boldsymbol{\alpha}$ al paràmetre de forma. Aquest vector ens indica la direcció de màxima asimetria. Igual que en el cas univariant, quan $\boldsymbol{\alpha} = \mathbf{0}$ obtenim la densitat d'una normal. Podem també relacionar $\boldsymbol{\alpha}$ amb un vector $\boldsymbol{\delta} \in \mathbb{R}^D$ mitjançant les expressions:

$$\boldsymbol{\delta} = \frac{1}{\sqrt{1+\boldsymbol{\alpha}'\boldsymbol{\Omega}\boldsymbol{\alpha}}}\boldsymbol{\Omega}\boldsymbol{\alpha}, \qquad \boldsymbol{\alpha} = \frac{1}{\sqrt{1-\boldsymbol{\delta}'\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}}}\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}. \tag{1.14}$$

Cada una de les components del vector $\boldsymbol{\delta}$ està acotada a l'interval $(-1, 1)$. En alguns casos, per raons pràctiques, ens referirem al vector $\boldsymbol{\delta}$ en comptes del vector $\boldsymbol{\alpha}$.

A continuació, detallem l'expressió de la funció generatriu de moments, el vector d'esperances i la matriu de covariàncies.

Propietat 1.7 Sigui $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$. Llavors la seva funció generatriu de moments és $M(\mathbf{t}) = \mathrm{E}[\exp(\mathbf{tz})] = 2\exp\left(\frac{1}{2}\mathbf{t}'\boldsymbol{\Omega}\mathbf{t}\right)\Phi(\boldsymbol{\delta}'\mathbf{t})$. □

Propietat 1.8 Sigui $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$. Llavors l'esperança i la matriu de covariàncies són

$$\mathrm{E}[\mathbf{z}] = \sqrt{\frac{2}{\pi}}\boldsymbol{\delta}, \qquad \mathrm{var}[\mathbf{z}] = \boldsymbol{\Omega} - \mathrm{E}[\mathbf{z}]\mathrm{E}[\mathbf{z}]' = \boldsymbol{\Omega} - \frac{2}{\pi}\boldsymbol{\delta}\boldsymbol{\delta}'.$$

□

Amb les components del vector E[$\mathbf{z}$] i la diagonal de la matriu var[$\mathbf{z}$] podem calcular, utilitzant l'expressió (1.9), el coeficient d'asimetria per a cada marginal. Azzalini i Capitanio (1999) consideren també un índex d'asimetria multivariant. En el cas de la normal asimètrica aquest índex d'asimetria esdevé:

$$\boldsymbol{\gamma}_1[\mathbf{z}] = \left(\frac{4-\pi}{2}\right)(\mathrm{E}[\mathbf{z}]'\mathrm{var}[\mathbf{z}]^{-1}\mathrm{E}[\mathbf{z}])^3.$$

El valor de $\boldsymbol{\gamma}_1$ està també acotat i el seu valor màxim és aproximadament 0.9902.

Tal i com hem fet en el cas univariant, degut a necessitats pràctiques, definim la distribució normal asimètrica multivariant amb paràmetres de localització i d'escala. Així doncs donat $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ considerem la transformació:

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\omega}\mathbf{z}, \tag{1.15}$$

on el vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_D)'$ i la matriu diagonal $\boldsymbol{\omega} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_D)$ amb $\sigma_i > 0$ $(i = 1, 2, \ldots, D)$, són respectivament els paràmetres de localització i d'escala. La funció de densitat del vector transformat $\mathbf{y}$ és

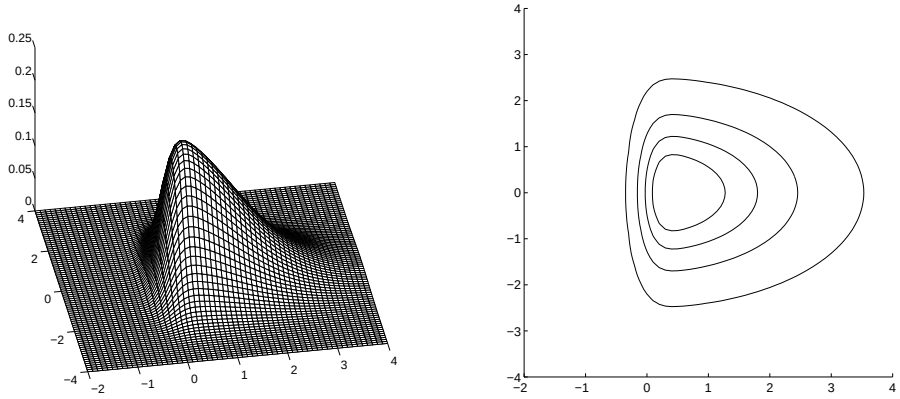
$$2\phi_D(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})), \tag{1.16}$$

on $\boldsymbol{\Sigma} = \boldsymbol{\omega}\boldsymbol{\Omega}\boldsymbol{\omega}$ és ara una matriu de covariàncies. Utilitzarem la notació $\mathbf{y} \sim \mathcal{SN}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ per referir-nos a aquest vector. La funció generatriu de moments, l'esperança i la matriu de covariàncies del vector $\mathbf{y}$ són:

$$\begin{aligned}
\mathrm{M}(t) &= 2\exp\left(t'\boldsymbol{\mu} + \frac{1}{2}t'\boldsymbol{\Sigma}t\right)\Phi(\boldsymbol{\delta}'\boldsymbol{\omega}t), \\
\mathrm{E}[\mathbf{y}] &= \boldsymbol{\mu} + \boldsymbol{\omega}\mathrm{E}[\mathbf{z}] = \boldsymbol{\mu} + \boldsymbol{\omega}\sqrt{\frac{2}{\pi}}\boldsymbol{\delta}, \\
\mathrm{var}[\mathbf{y}] &= \boldsymbol{\omega}\mathrm{var}[\mathbf{z}]\boldsymbol{\omega} = \boldsymbol{\Sigma} - \frac{2}{\pi}\boldsymbol{\omega}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\omega}.
\end{aligned}$$

A les figures 1.2(a) i 1.2(b) hem representat la funció de densitat i les respectives corbes de nivell d'un vector aleatori $\mathcal{SN}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. Podem observar que les corbes difereixen considerablement de les tradicionals el·lipses que obtenim amb un model normal bivariant.

Si generalitzem les propietats 1.4 i 1.5 obtindrem dos mètodes per generar mostres aleatòries d'un vector normal asimètric multivariant. En ambdós casos es parteix de mostres aleatòries de vectors normals multivariants.



(a) *Funció de densitat* (b) *Corbes de nivell*

Figura 1.2: $\mathcal{SN}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ *amb* $\boldsymbol{\mu} = (0,0)'$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ *i* $\boldsymbol{\alpha} = (5,0)'$

Propietat 1.9 Sigui $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_D)'$ un vector aleatori normal amb marginals estàndards i amb matriu de correlació $\boldsymbol{\Psi}$. Sigui $\mathrm{y}_0 \sim \mathcal{N}(0,1)$ una variable aleatòria independent de $\mathbf{y}$. Considerem el vector

$$\begin{pmatrix} \mathrm{y}_0 \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}_{D+1}(\mathbf{0}, \boldsymbol{\Psi}^*), \qquad \text{on} \qquad \boldsymbol{\Psi}^* = \begin{pmatrix} 1 & 0 \\ 0 & \boldsymbol{\Psi} \end{pmatrix}. \tag{1.17}$$

Considerem $\delta_1, \delta_2, \ldots, \delta_D$ valors en l'interval $(-1,1)$, i definim

$$\mathrm{z}_j = \delta_j |\mathrm{y}_0| + (1-\delta_j^2)^{1/2} \mathrm{y}_j \qquad (j = 1,2,\ldots,D). \tag{1.18}$$

Llavors la distribució del vector $\mathbf{z} = (\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_D)'$ és $\mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ on

$$\begin{aligned} \boldsymbol{\Omega} &= \boldsymbol{\Delta}(\boldsymbol{\Psi} + \boldsymbol{\lambda}\boldsymbol{\lambda}')\boldsymbol{\Delta}, \\ \boldsymbol{\alpha} &= \frac{\boldsymbol{\lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Delta}^{-1}}{(1+\boldsymbol{\lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda})^{1/2}}, \\ \boldsymbol{\Delta} &= \mathrm{diag}\left(\sqrt{1-\delta_1^2}, \sqrt{1-\delta_2^2}, \ldots, \sqrt{1-\delta_D^2}\right), \\ \boldsymbol{\lambda} &= \left(\frac{\delta_1}{\sqrt{1-\delta_1^2}}, \frac{\delta_2}{\sqrt{1-\delta_2^2}}, \ldots, \frac{\delta_D}{\sqrt{1-\delta_D^2}}\right)'. \end{aligned} \tag{1.19}$$

□

Observem que, per la propietat 1.4, cada component z_j definida segons (1.18) té una distribució normal asimètrica univariant $\mathcal{SN}(\lambda_j)$ amb $\lambda_j = \delta_j/\sqrt{1-\delta_j^2}$. Resumim tot seguit i de manera esquemàtica els punts a seguir per simular una mostra d'un vector $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$:

a. Mitjançant la matriu $\boldsymbol{\Omega}$ i el vector $\boldsymbol{\alpha}$, calcular el vector $\boldsymbol{\delta}$ a partir de (1.14).

b. Calcular la matriu $\boldsymbol{\Psi}$ amb l'expressió $\boldsymbol{\Psi} = \boldsymbol{\Delta}^{-1}(\boldsymbol{\Omega} - \boldsymbol{\delta}\boldsymbol{\delta}')\boldsymbol{\Delta}^{-1}$ que es dedueix fàcilment a partir de les expressions (1.19).

c. Generar una mostra aleatòria del vector $(\mathrm{y}_0, \mathbf{y})' \sim \mathcal{N}_{D+1}(\mathbf{0}, \boldsymbol{\Psi}^*)$ que apareix a (1.17).

d. Aplicar les transformacions (1.18) a la mostra aleatòria anterior.

Existeix també un altre mètode per generar mostres aleatòries d'una distribució $\mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ basat en la següent propietat:

Propietat 1.10 Sigui $\mathbf{x} = (\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_D)'$ un vector aleatori normal amb marginals estandarditzades i amb matriu de correlació igual a

$$\boldsymbol{\Omega}^* = \begin{pmatrix} 1 & \boldsymbol{\delta}' \\ \boldsymbol{\delta} & \boldsymbol{\Omega} \end{pmatrix}.$$

Llavors el vector

$$\mathbf{z} = \begin{cases} \mathbf{x} & \text{si } \mathrm{x}_0 > 0; \\ -\mathbf{x} & \text{altrament,} \end{cases}$$

té una distribució $\mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$, amb $\boldsymbol{\alpha} = (1/\sqrt{1-\boldsymbol{\delta}'\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}})\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}$. □

Observem que, per la propietat 1.5, la distribució de cada variable z_j donada $x_0 > 0$ és una normal asimètrica univariant $\mathcal{SN}(\lambda_j)$, on $\lambda_j = \delta_j/\sqrt{1-\delta_j^2}$.

Si volem generar mostres d'una distribució normal asimètrica amb paràmetres de localització i escala, $\mathcal{SN}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, ens cal en primer lloc generar, d'acord amb les propietats 1.9 o 1.10, una mostra d'una normal asimètrica $\mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ on ara $\boldsymbol{\Omega} = \boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1}$ i $\boldsymbol{\omega}$ és una matriu diagonal amb l'arrel quadrada de la diagonal de $\boldsymbol{\Sigma}$. Seguidament cal aplicar a la mostra obtinguda la transformació (1.15).

Un vector amb distribució normal asimètrica compleix la propietat que qualsevol subvector té també una distribució normal asimètrica. Azzalini i Capitanio (1999) ho demostren per a un vector $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ en la següent propietat:

Propietat 1.11 Sigui $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$. Considerem la partició $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)'$ de dimensions H i $D - H$, respectivament. Denotem

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \quad \text{i} \quad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix},$$

les corresponents particions de $\boldsymbol{\Omega}$ i $\boldsymbol{\alpha}$. Llavors la distribució marginal de $\mathbf{z}_1$ és $\mathcal{SN}_H(\boldsymbol{\Omega}_{11}, \overline{\boldsymbol{\alpha}}_1)$, on

$$\overline{\boldsymbol{\alpha}}_1 = \frac{\boldsymbol{\alpha}_1 + \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\alpha}_2}{\sqrt{1 + \boldsymbol{\alpha}_2'\boldsymbol{\Omega}_{1.2}\boldsymbol{\alpha}_2}} \qquad \text{i} \qquad \boldsymbol{\Omega}_{1.2} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}.$$

□

Com a cas particular, es dedueix que cada marginal d'un vector normal asimètric té també una distribució normal asimètrica. Més concretament, si $\mathbf{z} = (\mathrm{z}_1, \mathrm{z}_2, \ldots, \mathrm{z}_D)' \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ llavors $\mathrm{z}_i \sim \mathcal{SN}(\lambda_i)$ amb $\lambda_i = \delta_i/\sqrt{1 - \delta_i^2}$, on δ_i és la i-èsima component del vector $\boldsymbol{\delta}$ calculat a partir de l'expressió (1.14). La propietat 1.11, a banda d'una lleugera complicació en la notació, és també vàlida per a vectors aleatoris normals asimètrics amb paràmetres de localització i d'escala.

Una propietat que utilitzarem sovint en aquesta tesi doctoral és la propietat de les transformacions lineals. Per a un vector $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ aquesta propietat diu:

Propietat 1.12 Siguin $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ i $\mathbf{A}$ una matriu de constants de dimensió $D \times H$ de tal manera que $\mathbf{A}'\boldsymbol{\Omega}\mathbf{A}$ és una matriu de correlació. Llavors $\mathbf{z}^* = \mathbf{A}'\mathbf{z} \sim \mathcal{SN}^H(\boldsymbol{\Omega}^*, \boldsymbol{\alpha}^*)$ amb

$$\boldsymbol{\Omega}^* = \mathbf{A}'\boldsymbol{\Omega}\mathbf{A}, \qquad \boldsymbol{\alpha}^* = \frac{(\boldsymbol{\Omega}^*)^{-1}\mathbf{A}'\boldsymbol{\Omega}\boldsymbol{\alpha}}{\sqrt{1 + \boldsymbol{\alpha}'(\boldsymbol{\Omega} - \boldsymbol{\Omega}\mathbf{A}(\boldsymbol{\Omega}^*)^{-1}\mathbf{A}'\boldsymbol{\Omega})\boldsymbol{\alpha}}}.$$

□

En particular, si $\mathbf{A}$ és una matriu quadrada i no singular, tenim $\boldsymbol{\alpha}^* = \mathbf{A}^{-1}\boldsymbol{\alpha}$.

Propietat 1.13 Sigui $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$. Llavors existeix una transformació lineal $\mathbf{z}^* = \mathbf{A}'\mathbf{z}$ tal que $\mathbf{z}^* \sim \mathcal{SN}^D(\mathbf{I}_D, \boldsymbol{\alpha}^*)$ on només una component de $\boldsymbol{\alpha}^*$ és diferent de 0. □

Aquest resultat ens defineix una mena de "forma canònica" ja que les components del vector $\mathbf{z}^*$ són independents dos a dos i només una sola component "absorbeix" tota la asimetria de la distribució multivariant. Aquesta transformació lineal especial té el mateix paper que la

transformació que converteix una distribució normal multivariant en la seva forma esfèrica. En aquest cas, es compleix a més que la matriu $\mathbf{A}$ és invertible. Per tant, és possible obtenir qualsevol vector $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ aplicant la transformació convenient al vector $\mathbf{z}^*$.

Sabem que un vector aleatori multivariant té components independents si i només si la seva funció de densitat es pot expressar com a producte de les densitats de cada marginal. Per la propietat 1.11 sabem que cada marginal d'un vector aleatori amb distribució normal asimètrica té una distribució normal asimètrica univariant. Donat que $\Phi(u_1 + u_2)$ no es pot factoritzar com el producte $\Phi(u_1)\Phi(u_2)$, un vector aleatori normal asimètric $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ tindrà components independents només quan una sola component del vector $\boldsymbol{\alpha}$ sigui diferent de 0.

Propietat 1.14 Siguin $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ i $\mathbf{A}$ una matriu de constants de dimensió $D \times H$ de tal manera que $\mathbf{A}'\boldsymbol{\Omega}\mathbf{A}$ és una matriu de correlació. Aleshores, les components del vector $\mathbf{z}^* = \mathbf{A}'\mathbf{z} \sim \mathcal{SN}^H(\boldsymbol{\Omega}^*, \boldsymbol{\alpha}^*)$ són independents si i només si les dues condicions següents es compleixen simultàniament:

a. La matriu $\boldsymbol{\Omega}^*$ és diagonal.

b. El vector $\boldsymbol{\alpha}^*$ només té una component diferent de 0.

□

Podem generalitzar la propietat 1.12 per a distribucions normals asimètriques amb paràmetres de localització i d'escala. El resultat és pràcticament el mateix però la notació es complica lleugerament.

Propietat 1.15 Siguin $\mathbf{y} \sim \mathcal{SN}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ i $\mathbf{A}$ una matriu de constants de dimensió $D \times H$. Llavors $\mathbf{y}^* = \mathbf{A}'\mathbf{y} \sim \mathcal{SN}^H(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\alpha}^*)$, amb

$$\boldsymbol{\mu}^* = \mathbf{A}'\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}^* = \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}, \qquad \boldsymbol{\alpha}^* = \frac{\boldsymbol{\omega}^*(\boldsymbol{\Sigma}^*)^{-1}\mathbf{B}'\boldsymbol{\alpha}}{\sqrt{1 + \boldsymbol{\alpha}'(\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1} - \mathbf{B}(\boldsymbol{\Sigma}^*)^{-1}\mathbf{B}')\boldsymbol{\alpha}}},$$

on $\mathbf{B} = \boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\mathbf{A}$, $\boldsymbol{\omega}$ i $\boldsymbol{\omega}^*$ són matrius diagonals iguals a l'arrel quadrada de la diagonal de $\boldsymbol{\Sigma}$ i $\boldsymbol{\Sigma}^*$, respectivament. □

En particular, si $\mathbf{A}$ és una matriu quadrada i no singular, tenim $\boldsymbol{\alpha}^* = \boldsymbol{\omega}^*\mathbf{A}^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\alpha}$.

Veiem tot seguit la generalització de la propietat 1.6 al cas multivariant, que fa referència a les formes quadràtiques.

Propietat 1.16 Siguin $\mathbf{z} \sim \mathcal{SN}^D(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ i $\mathbf{B}$ una matriu quadrada simètrica i semidefinida positiva de rang p tal que $\mathbf{B}\boldsymbol{\Omega}\mathbf{B} = \mathbf{B}$. Llavors $\mathbf{z}'\boldsymbol{\Omega}\mathbf{z} \sim \chi^2_p$. □

Sigui $\mathbf{y}$ un vector aleatori d'ordre $D \times 1$, amb funció de densitat (1.16). Considerem la seva partició en dues components $\mathbf{y}_1$ i $\mathbf{y}_2$ de dimensions H i $D - H$, respectivament. Siguin

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \qquad \text{i} \qquad \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}$$

les particions corresponents dels paràmetres $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ i $\boldsymbol{\alpha}$. En la següent propietat analitzem la distribució de $\mathbf{y}_1$ condicionada per $\mathbf{y}_2$.

Propietat 1.17 L'expressió de la funció de densitat del vector $\mathbf{y}_1|\mathbf{y}_2$ és

$$\phi_H(\mathbf{y}_1; \boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{1.2}) \frac{\Phi(\boldsymbol{\alpha}_1' \boldsymbol{\omega}_1^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_{1.2}) - x_0')}{\Phi(x_0)}, \tag{1.20}$$

on

$$\begin{aligned}
\boldsymbol{\mu}_{1.2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \\
\boldsymbol{\Sigma}_{1.2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \\
x_0 &= \overline{\boldsymbol{\alpha}}_2' \boldsymbol{\omega}_2^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \\
x_0' &= (1 + \boldsymbol{\alpha}_1' \boldsymbol{\omega}_1^{-1} \boldsymbol{\Sigma}_{1.2}^{-1} \boldsymbol{\omega}_1^{-1} \boldsymbol{\alpha}_1)^{1/2} x_0, \\
\overline{\boldsymbol{\alpha}}_2 &= \frac{\boldsymbol{\alpha}_2 + \boldsymbol{\omega}_2 \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\omega}_1^{-1} \boldsymbol{\alpha}_1}{\sqrt{1 + \boldsymbol{\alpha}_1' \boldsymbol{\omega}_1^{-1} \boldsymbol{\Sigma}_{1.2}^{-1} \boldsymbol{\omega}_1^{-1} \boldsymbol{\alpha}_1}},
\end{aligned}$$

i $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2$ són matrius diagonals iguals a l'arrel quadrada de la diagonal de les matrius $\boldsymbol{\Sigma}_{11}$ i $\boldsymbol{\Sigma}_{22}$ respectivament. □

Mitjançant la densitat (1.20), observem que el vector $\mathbf{y}_1|\mathbf{y}_2$ tindrà una densitat normal asimètrica si i només si $\overline{\boldsymbol{\alpha}}_2' \boldsymbol{\omega}_2^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) = 0$. Aquesta condició és equivalent a exigir que $\overline{\boldsymbol{\alpha}}_2 = 0$, és a dir, que la component $\mathbf{y}_2$ sigui normal. Azzalini i Capitanio (1999) examinen gràficament la funció de densitat (1.20) i observen que presenta un perfil molt similar al d'una densitat normal asimètrica. Aquest resultat suggereix utilitzar una densitat normal asimètrica per aproximar la densitat (1.20). L'aproximació es calcula fàcilment igualant els tres primers moments del vector $\mathbf{y}_1|\mathbf{y}_2$ amb els tres primers moments d'un vector normal asimètric. Les equacions que en resulten permeten trobar una solució explícita, excepte en

situacions extremes, quan la distribució condicionada té un índex d'asimetria fora del rang de la normal asimètrica. Azzalini i Capitanio (1999) comproven que en la majoria dels casos aquesta aproximació obtinguda resulta ser bastant exacta i per tant afirmen que la família normal asimètrica és tancada respecte de l'operació de "condicionar".

1.3.3 Aspectes d'inferència estadística

En aquest apartat centrarem la nostra atenció en aspectes d'inferència estadística. Analitzarem en primer lloc el problema de l'estimació dels paràmetres i seguirem amb l'estudi de l'aplicació de diversos contrastos d'hipòtesi.

Estimació de paràmetres

Per trobar els estimadors dels paràmetres d'una distribució normal asimètrica utilitzarem el procediment de màxima versemblança. Donat que no és possible trobar una expressió analítica d'aquests estimadors en funció de la mostra, caldrà utilitzar procediments numèrics per trobar el màxim de la funció de versemblança. Azzalini i Capitanio (1999) discuteixen extensament els problemes derivats de l'estimació dels paràmetres. A continuació en reproduïm tan sols els aspectes més rellevants.

a. *Cas univariant*

Donada una mostra aleatòria $y_1, y_2, \ldots, y_n$ d'una variable y $\sim \mathcal{SN}(\mu, \sigma, \lambda)$, estimarem els paràmetres μ, σ i λ utilitzant el mètode de màxima versemblança. La funció de logversemblança que cal maximitzar és

$$\mathrm{l}(\mu, \sigma, \lambda) = n\ln\left(\frac{2}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2 + \sum_{i=1}^{n}\ln\left(\Phi\left(\lambda\sigma^{-1}(y_i - \mu)\right)\right).$$

No és possible trobar una expressió analítica per a l'estimador màxim versemblant en funció de la mostra, per tant, ens veiem obligats a recórrer a mètodes numèrics, com per exemple el mètode del gradient o bé l'algorisme EM, per trobar-lo. Qualsevol mètode numèric iteratiu necessita uns valors inicials; en el nostre cas prendrem els estimadors que ens proporciona el mètode dels moments. Així doncs, calcularem la mitjana ($\bar{y}$), la variància (s^2), el coeficient d'asimetria ($\widehat{\gamma}_1$) mostrals i obtindrem el valor inicial per μ, σ^2 i λ a partir de les següents relacions:

$$\lambda = \frac{\delta}{\sqrt{1-\delta^2}}, \qquad \sigma^2 = \frac{s^2}{(1 - 2\delta^2/\pi)}, \qquad \mu = \bar{y} - \sigma\delta\sqrt{\frac{2}{\pi}},$$

on $\delta = \frac{a}{\sqrt{2/\pi}\sqrt{1+a^2}}$ i $a = \left(\frac{2\widehat{\gamma_1}}{4-\pi}\right)^{1/3}$.

Seguidament podem aplicar qualsevol mètode de maximització numèrica. Ens trobem, però, amb dos problemes:

- La funció de versemblança té sempre un punt d'inflexió a $\lambda = 0$ i la matriu d'informació de Fisher corresponent esdevé singular. Per tant, si després de trobar les estimacions ens cal utilitzar la matriu d'informació de Fisher per calcular per exemple la cota de Cramér-Rao, tindrem greus problemes.

- La funció de logversemblança té un perfil molt especial que alenteix la convergència en el procés de maximització.

Per entendre millor aquests dos problemes, hem simulat una mostra de mida 100 procedent d'una $\mathcal{SN}(\mu = 0, \sigma = 1, \lambda = 5)$. En la figura 1.3(a) hem representat el perfil relatiu de la funció de logversemblança multiplicat per 2 vers el paràmetre λ. Tal i com hem esmentat, observem el punt d'inflexió a $\lambda = 0$. Cal potser recordar que el perfil de la funció de logversemblança és $\mathrm{l}^*(\lambda) = \mathrm{l}(\hat{\mu}_\lambda, \hat{\sigma}_\lambda, \lambda)$, on $\hat{\mu}_\lambda$ i $\hat{\sigma}_\lambda$ representen els valors de μ i σ que maximitzen la funció de logversemblança per a un valor fixat del paràmetre λ. Amb aquesta funció obtenim el perfil relatiu de la funció de logversemblança:

$$\mathrm{l}^*(\lambda) = \mathrm{l}(\hat{\mu}_\lambda, \hat{\sigma}_\lambda, \lambda) - \mathrm{l}(\hat{\mu}, \hat{\sigma}, \hat{\lambda})$$

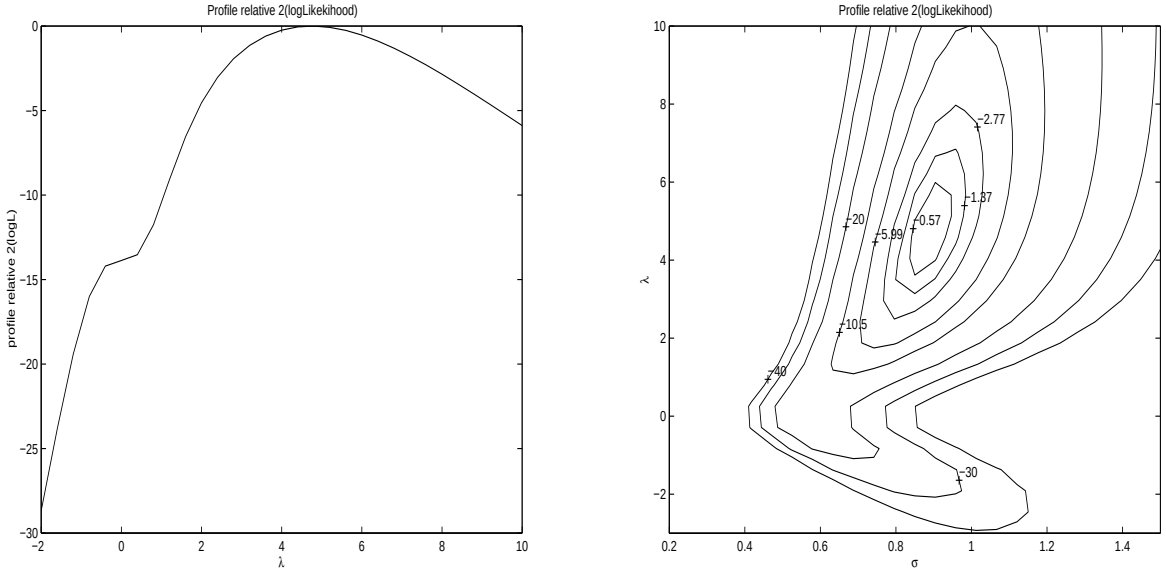
on $\hat{\mu}, \hat{\sigma}$ i $\hat{\lambda}$ són els estimadors de màxima versemblança.

En la figura 1.3(b) hem representat les corbes de nivell del perfil relatiu de la funció de logversemblança multiplicat per 2 vers els paràmetres σ i λ, és a dir, hem representat la funció

$$2\mathrm{l}^*(\sigma, \lambda) = 2(\mathrm{l}(\hat{\mu}_{\sigma,\lambda}, \sigma, \lambda) - \mathrm{l}(\hat{\mu}, \hat{\sigma}, \hat{\lambda})),$$

on $\hat{\mu}_{\sigma,\lambda}$ és el valor μ que maximitza la funció de logversemblança per a valors fixats dels paràmetres σ i λ; i $\hat{\mu}, \hat{\sigma}$ i $\hat{\lambda}$ són els estimadors de màxima versemblança. Observant aquesta figura podem veure que aquestes corbes de nivell no tenen una forma favorable per a una convergència ràpida del mètode de maximització.

Per evitar aquestes dificultats Azzalini i Capitanio (1999) proposen utilitzar una altra parametrització. Fins ara, hem considerat la parametrització directa μ, σ i λ ja que una



(a) (b)

Figura 1.3: (a) *Perfil relatiu de la funció de logversemblança multiplicat per 2 vers el paràmetre* λ *i* (b) *corbes de nivell de la mateixa funció vers els paràmetres* σ *i* λ.

variable $\mathrm{y} \sim \mathcal{SN}(\mu, \sigma, \lambda)$ es construeix com $\mathrm{y} = \mu + \sigma \mathrm{z}$ amb $\mathrm{z} \sim \mathcal{SN}(\lambda)$. Azzalini i Capitanio (1999) defineixen la parametrització centrada de la variable y. Si partim d'una variable $\mathrm{z} \sim \mathcal{SN}(\lambda)$, considerem la variable

$$\mathrm{y} = \mu^* + \sigma^* \left(\frac{\mathrm{z} - \mathrm{E}[\mathrm{z}]}{\sqrt{\mathrm{var}[\mathrm{z}]}} \right). \tag{1.21}$$

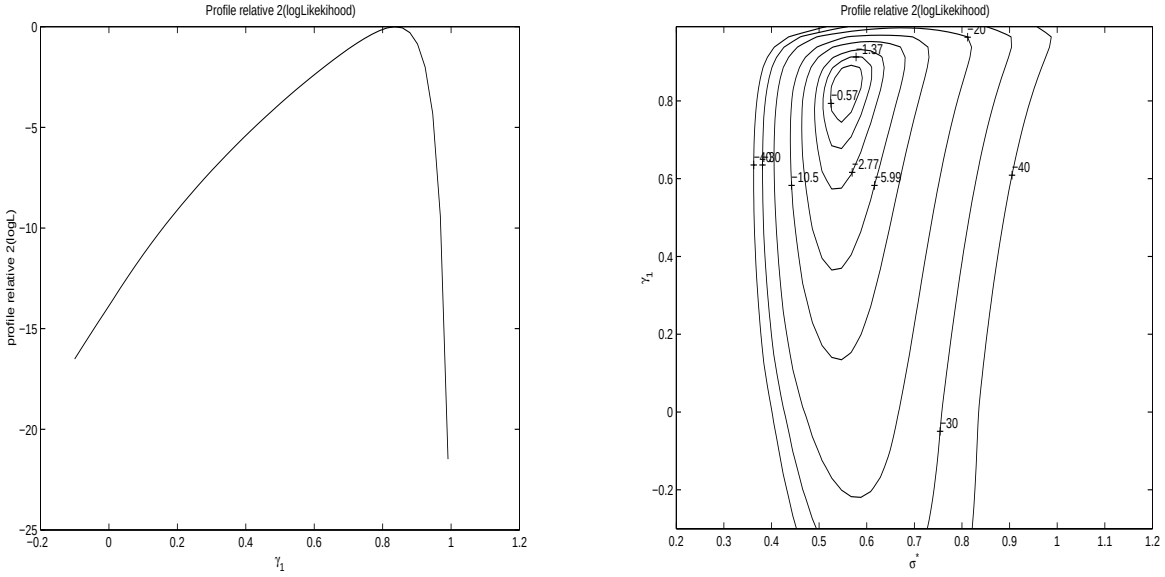
Els paràmetres anteriors μ i σ estan relacionat amb μ^* i σ^* de la següent manera:

$$\mu = \mu^* - \frac{\sigma^* \mathrm{E}[\mathrm{z}]}{\sqrt{\mathrm{var}[\mathrm{z}]}}, \qquad \sigma = \frac{\sigma^*}{\sqrt{\mathrm{var}[\mathrm{z}]}}.$$

La proposta és treballar amb els paràmetres μ^*, σ^* i γ_1 (l'índex d'asimetria habitual) en comptes dels paràmetres μ, σ i λ. Els efectes més importants que produeix aquesta nova parametrització són:

- Desaparició del punt d'inflexió i de la singularitat de la matriu d'informació de Fisher en el punt $\lambda = 0$.
- Millora considerable del perfil de la funció de logversemblança.

Podem apreciar aquestes millores a les figures 1.4(a) i 1.4(b). En la figura 1.4(a), on hem representat el perfil relatiu de la funció de logversemblança multiplicat per dos vers el paràmetre γ_1, podem observar la desaparició del punt d'inflexió. En la figura 1.4(b) tenim les corbes de nivell de la mateixa funció vers els paràmetres σ^* i γ_1. Podem observar que la forma de la funció millora considerablement i el procediment numèric convergirà al màxim més ràpidament.



(a) (b)

Figura 1.4: (a) *Perfil relatiu de la funció de logversemblança muliplicat per 2 vers el paràmetre* γ_1 *i* (b) *corbes de nivell de la mateixa funció vers els paràmetres* σ^* *i* γ_1.

Aquesta reparametrització comporta avantatges bàsicament a nivell teòric ja que evitem la singularitat de la matriu d'informació de Fisher. A la pràctica tan sols produeix una convergència més ràpida ja que la forma de la funció que cal maximitzar és molt més regular.

No obstant això, trobem un greu problema amb les dues parametritzacions. Azzalini i Capitanio (1999) disposen d'una mostra de mida 50 simulada a partir d'una normal asimètrica $\mathcal{SN}(\mu = 0, \sigma = 1, \lambda = 5)$. Aparentment aquesta mostra no presenta cap problema donat que el coeficient d'asimetria mostral γ_1 té un valor de 0.90 i per tant està dins l'interval adient per a una variable normal asimètrica. El problema el trobem amb l'estimació màxim versemblant ja que el valor de $\hat{\lambda}$ tendeix a $+\infty$, si utilitzem la primera parametrització, o el valor de $\hat{\gamma}_1$

tendeix cap al valor extrem 0.995, si utilitzem la segona. És freqüent observar aquest fenomen quan treballem amb mostres de mida petita ($n \leq 50$ aproximadament).

En aquests casos i només com a solució ad hoc Azzalini i Capitanio (1999) proposen utilitzar la següent estratègia. Si l'estimació del paràmetre γ_1 arriba, en valor absolut, al seu valor màxim, reiniciar el procés de maximització i parar quan la funció de logversemblança arribi a un valor no significativament inferior al seu màxim. A la pràctica haurem d'aplicar tan sols un test de raó de versemblances. Si obtenim el màxim en el punt $\hat{\mu_1}, \hat{\sigma_1}$ i $\hat{\lambda_1} = \infty$, tornem a començar el procés de maximització i ens parem quan el valor de la funció de logversemblança arribi a un valor $\mathrm{l}(\hat{\mu_2}, \hat{\sigma_2}, \hat{\lambda_2})$ de manera que

$$2(\mathrm{l}(\hat{\mu_2}, \hat{\sigma_2}, \hat{\lambda_2}) - \mathrm{l}(\hat{\mu_1}, \hat{\sigma_1}, \infty)) < \chi^2_{3,0.05}, \tag{1.22}$$

on $\chi^2_{3,0.05}$ és el percentil 95% d'una variable χ^2 amb tres graus de llibertat. Tot seguit prenem $\hat{\mu_2}, \hat{\sigma_2}$ i $\hat{\lambda_2}$ com a estimadors màxim versemblants. Aquesta proposta té un grau d'arbitrarietat considerable perquè podem escollir els estimadors dels paràmetres entre un gran nombre de valors. Tot i així, cal tenir present que per a valors de $\lambda > 20$ el perfil d'una densitat normal asimètrica varia de forma gairebé insignificant.

b. *Cas multivariant*

En el cas multivariant, donada una mostra aleatòria D-dimensional de mida n, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, ens cal estimar els vectors $\boldsymbol{\mu}$ i $\boldsymbol{\alpha}$, d'ordre $D \times 1$, i la matriu $\boldsymbol{\Sigma}$, d'ordre $D \times D$. La funció de logversemblança és

$$\mathrm{l}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = n\ln\left(\frac{2}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\right) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) + \sum_{i=1}^{n}\ln\left(\Phi\left(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right)\right),$$

on $\boldsymbol{\omega}$ és una matriu diagonal que conté l'arrel quadrada de la diagonal de $\boldsymbol{\Sigma}$.

Seria ideal utilitzar la reparametrització introduïda en el cas escalar però ens és impossible aplicar-la donat que no podem reparametritzar els elements fora de la diagonal de $\boldsymbol{\Sigma}$. Azzalini i Capitanio (1999) proposen utilitzar una estratègia diferent: aplicar el canvi de variable $\boldsymbol{\beta} = \boldsymbol{\omega}^{-1}\boldsymbol{\alpha}$, i trobar els estimadors màxim versemblants dels paràmetres $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ i $\boldsymbol{\beta}$ amb un mètode numèric. Amb aquest canvi de variable la funció de logversemblança esdevé

$$\mathrm{l}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = n\ln\left(\frac{2}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\right) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) + \sum_{i=1}^{n}\ln\left(\Phi\left(\boldsymbol{\beta}'(\mathbf{y}_i - \boldsymbol{\mu})\right)\right).$$

Observem que els paràmetres $\boldsymbol{\Sigma}$ i $\boldsymbol{\beta}$ apareixen separats, és a dir, els trobem en sumands diferents. Gràcies a això podrem utilitzar la propietat de factorització de la funció de log-

versemblança i el procés de maximització esdevindrà més senzill. A continuació descrivim de manera esquemàtica els punts a seguir en aquest procés de maximització. Prèviament cal calcular els valors inicials dels paràmetres amb el mètode dels moments.

Utilitzant la propietat de factorització el nostre esquema iteratiu té dues parts:

a. Fixat el valor de les estimacions $\hat{\boldsymbol{\beta}}$ i $\hat{\boldsymbol{\Sigma}}$, calcular l'estimador màxim versemblant $\hat{\boldsymbol{\mu}}$ aplicant una iteració del mètode numèric escollit.

b. Fixat el valor de l'estimador $\hat{\boldsymbol{\mu}}$, calcular l'estimador màxim versemblant $\hat{\boldsymbol{\Sigma}}$ dels dos primers sumands de la funció de logversemblança. Aquest pas és immediat ja que tan sols ens cal calcular, amb el valor $\hat{\boldsymbol{\mu}}$, la matriu de covariàncies. Paral·lelament i fixat també l'estimador $\hat{\boldsymbol{\mu}}$, trobar l'estimador màxim versemblant $\hat{\boldsymbol{\beta}}$ del tercer sumand de la funció de logversemblança aplicant una iteració del mètode numèric escollit.

Una vegada acabat aquest procés obtindrem uns estimadors $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ i $\hat{\boldsymbol{\beta}}$, i un valor per a la funció de logversemblança l. Repetirem l'esquema anterior fins que

$$\frac{\mathrm{l} - \mathrm{ant(l)}}{\mathrm{ant(l)}} < \varepsilon,$$

on ant(l) indica el valor de la funció de logversemblança en l'iterat anterior.

En aquest treball de recerca, aplicarem aquest procediment per trobar els estimadors de màxima versemblança. En particular, utilitzarem el mètode del gradient reduït generalitzat per trobar el màxim en els passos a i b de l'esquema anterior i prendrem el valor $\varepsilon = 10^{-6}$.

Podem trobar-nos també que algunes components del paràmetre de forma $\boldsymbol{\alpha}$ prenguin el valor ∞. En aquests casos Azzalini i Capitanio (1999) proposen la mateixa solució que en el cas anterior, tot i que en aquest cas haurem d'utilitzar una distribució χ^2 amb $D(D+5)/2$ graus de llibertat on D és el nombre de components del vector aleatori.

Contrastos d'hipòtesi

El model normal és un membre particular de la família normal asimètrica ja que correspon al cas $\boldsymbol{\alpha} = \mathbf{0}$. A la pràctica, sovint ens interessarà comparar el model normal asimètric ajustat amb el model normal ajustat. En aquests casos tan sols haurem de contrastar la hipòtesi nul·la $\boldsymbol{\alpha} = \mathbf{0}$ vers la hipòtesi alternativa $\boldsymbol{\alpha} \neq \mathbf{0}$, aplicant un test de raó de versemblança. Per dur a terme aquest contrast, necessitarem el màxim de la funció de logversemblança sota la hipòtesi de normalitat asimètrica, $\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}})$, i el màxim de la funció de logversemblança sota

la hipòtesi de normalitat, $\mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0})$, on $\hat{\mathbf{m}}$ i $\hat{\mathbf{S}}$ representen els estimadors màxim versemblants dels paràmetres $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ sota hipòtesi de normalitat. L'estadístic d'aquest contrast és

$$2(\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}}) - \mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0}))$$

el qual, sota la hipòtesi nul·la, segueix una distribució χ^2 amb D graus de llibertat on D és el nombre de components del vector aleatori.

Un altre aspecte interessant, sovint utilitzat en la modelització de vectors aleatoris, és la validació de l'ajust del model al conjunt de les dades. No es coneixen tests de bondat d'ajust específics per a la distribució normal asimètrica. Per aquesta raó, dediquem el capítol 5 d'aquest treball d'investigació a mostrar uns tests de bondat d'ajust que hem construït específicament per a aquesta distribució.

Capítol 2

El símplex: conceptes previs

Una composició aleatòria és un vector aleatori $\mathcal{S}^D$-valuat, és a dir, una funció mesurable amb imatge el símplex, el subconjunt de l'espai real de vectors amb components positives i de suma constant. Les seves realitzacions donen lloc a les dades composicionals, vectors amb components positives la suma de les quals és una constant. És àmpliament conegut que l'aplicació dels conceptes de teoria de la probabilitat estàndards o la utilització de tècniques estadístiques habituals dins el context de les dades composicionals poden donar lloc a resultats erronis i a greus dificultats d'interpretació. Pearson (1897) fou el primer en adonar-se del problema però va ser Aitchison (1982) qui desenvolupà una metodologia específica per a aquest tipus de dades basant-se en la tècnica de la transformació. Els resultats d'Aitchison foren publicats en diferents articles i resumits en la seva monografia Aitchison (1986). A aquesta monografia la segueixen nombrosos treballs, entre els quals podem destacar Aitchison (1992, 1997, 2001) o Aitchison et al (1998, 1999, 2000) entre d'altres. Seguint la línia iniciada per Aitchison, s'han adaptat a l'anàlisi de dades composicionals diverses tècniques estadístiques ja existents en l'anàlisi real clàssica, entre elles, la predicció d'observacions multivariants amb dependència espacial o cokrigeat (Pawlowsky, 1986, Pawlowsky-Glahn i Olea, 2003), l'anàlisi discriminant (Barceló-Vidal, 1996) o la classificació no paramètrica (Martín-Fernández, 2001). Per altra banda, aquesta metodologia ha permès ampliar les famílies de distribucions sobre el símplex. Destaquem el model normal logístic additiu (Aitchison, 1986), el model normal asimètric logístic additiu (Mateu-Figueras et al., 1998) o els models basats en la transformació Box-Cox (Barceló-Vidal, 1996).

Actualment s'ha desenvolupat la fonamentació matemàtica del símplex (Barceló-Vidal, 2000, Barceló-Vidal et al., 2001). Sabem que $\mathcal{S}^D$ té una estructura d'espai euclidià, amb unes operacions, un producte escalar i una distància diferents als elements clàssics de l'espai real (Pawlowsky-Glahn i Egozcue, 2001, Billheimer et al., 2001). Aquesta estructura ha permès reformular conceptes fonamentals dels estimadors i obtenir les seves propietats amb referència al biaix i a la variància (Pawlowsky-Glahn i Egozcue, 2001, 2002). En aquesta línia, han sorgit els primers treballs en relació a l'estudi de les tècniques de regressió lineal multivariant sobre dades composicionals (Buccianti et al., 1999, Daunis-i Estadella et al., 2002).

Un dels objectius principals d'aquest treball és l'estudi de famílies de distribucions per modelitzar dades composicionals. En el capítol anterior s'ha vist la importància de l'estructura algebraica de l'espai on tenim definides les variables aleatòries. Per aquesta raó, en aquest capítol presentem de manera detallada els conceptes bàsics referents a les dades composicionals i al seu espai mostral. Veiem en primer lloc l'estructura algebraica de $\mathcal{S}^D$ fent especial atenció a les coordenades de les composicions respecte d'una base o un sistema de generadors. Seguim amb els conceptes de subcomposició i amalgama. A continuació, presentem les transformacions logístiques ja que juguen un paper central en la definició de famílies de distribucions sobre $\mathcal{S}^D$. Finalment, estudiem les composicions aleatòries centrant-nos en els elements de tendència central i de dispersió.

2.1 Definicions bàsiques i estructura algebraica

Definició 2.1 Una *composició* $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ amb D parts és un vector amb components positives la suma de les quals val 1. El seu espai mostral natural és el *símplex* $\mathcal{S}^D$, el subconjunt de l'espai $\mathbb{R}^D$ definit com

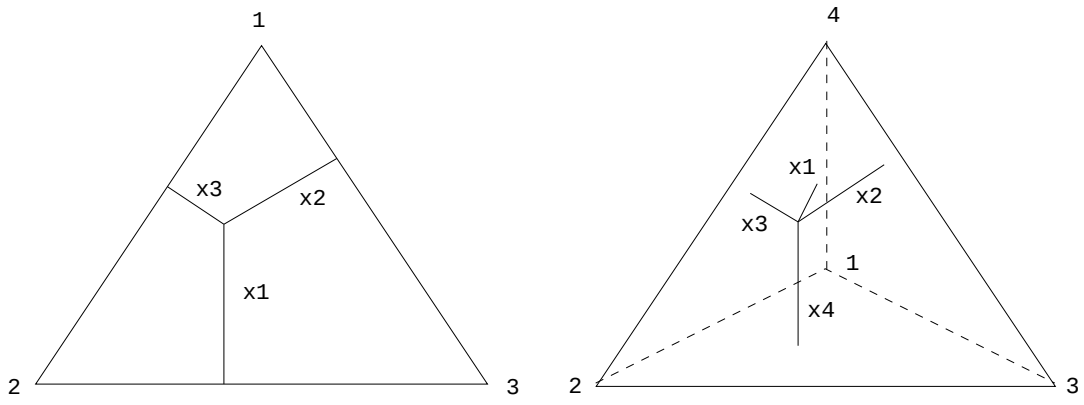
$$\mathcal{S}^D = \big\{(x_1, x_2, \ldots, x_D)' : x_1 > 0, x_2 > 0, \ldots, x_D > 0;\ \sum_{i=1}^{D} x_i = 1\big\}.$$

□

En el cas $D = 3$, el símplex $\mathcal{S}^3$ es pot representar mitjançant el *diagrama ternari*, triangle equilàter d'altura unitat (figura 2.1(a)). Existeix una correspondència biunívoca entre les composicions amb 3 parts i els punts del diagrama ternari. Cada composició $\mathbf{x} = (x_1, x_2, x_3)'$

es correspon amb el punt que dista x_1, x_2 i x_3, respectivament, dels costats oposats als vèrtexs $1, 2, 3$. En el cas $D = 4$, el símplex es representa amb un tetràedre regular d'altura unitat (figura 2.1(b)).

Barceló-Vidal (2000) i Barceló-Vidal et al. (2001) fan una revisió del símplex des d'un punt de vista matemàtic i observen que es pot interpretar com una representació de l'espai quocient on les classes d'equivalència són els conjunts $\underline{\mathbf{w}} = \{k\mathbf{w} : k > 0\}$ amb $\mathbf{w} \in \mathbb{R}_+^D$. Observem que només un element de la classe d'equivalència compleix que la suma de les seves components és igual a 1. Si escollim com a representant de la classe aquest element obtenim $\mathcal{S}^D$ com una representació d'aquest espai quocient. En aquest treball d'investigació no utilitzarem explícitament aquesta interpretació del símplex. Per aquesta raó ens caldrà definir l'operador clausura el qual assigna a cada element $\mathbf{w}$ la seva composició associada. No obstant això, les definicions i propietats que veurem es poden reescriure en termes de classes d'equivalència.



(a) (b)

Figura 2.1: (a): *Representació d'una dada composicional* $(x_1, x_2, x_3)'$ *en el símplex* $\mathcal{S}^3$*;* (b): *Representació d'una dada composicional* $(x_1, x_2, x_3, x_4)'$ *en el símplex* $\mathcal{S}^4$.

Definició 2.2 L'*operador clausura* $\mathcal{C}$ és una transformació de $\mathbb{R}_+^D$ a $\mathcal{S}^D$ que fa correspondre a cada element $\mathbf{w} \in \mathbb{R}_+^D$ la seva composició associada $\mathcal{C}(\mathbf{w})$ on

$$\mathcal{C}(\mathbf{w}) = \frac{1}{\sum_{i=1}^D w_i}\mathbf{w} = \left(\frac{w_1}{\sum_{i=1}^D w_i}, \frac{w_2}{\sum_{i=1}^D w_i}, \ldots, \frac{w_D}{\sum_{i=1}^D w_i}\right)'.$$

□

Observem que un element $\mathbf{w} \in \mathbb{R}_+^D$ determina completament una composició, però una composició determina un nombre infinit de vectors de $\mathbb{R}_+^D$. La figura 2.2 mostra la relació entre els elements de $\mathbb{R}_+^D$ i les composicions.

La restricció de la suma unitària ha estat considerada com la font de tots els problemes, donat que impedeix l'aplicació dels procediment estadístics habituals que s'utilitzen amb dades que no presenten aquesta restricció. Notem, per exemple, que el canvi en una de les parts d'una composició provoca necessàriament el canvi en com a mínim una altra de les parts. És a dir, una dada composicional $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ és, en essència, un vector de dimensió $D-1$ ja que queda completament especificat si es coneixen $D-1$ parts. En efecte, conegudes les parts $x_1, x_2, \ldots, x_{D-1}$, podem obtenir l'última component com $x_D = 1 - (\sum_{i=1}^{D-1} x_i)$. La composició $\mathbf{x}$ queda també totalment determinada si es coneixen els $D-1$ quocients x_i/x_D per a $i = 1, 2, \ldots, D-1$.

Cal remarcar que la restricció de la suma unitària es podria substituir per qualsevol altra restricció de suma constant, com per exemple $\sum_{i=1}^{D} x_i = k$. Aquesta constant k depèn només de l'escala utilitzada per expressar les dades. Observem, doncs, que els valors de les components x_i són irrellevants. Una composició amb D parts aporta tan sols informació sobre les magnituds relatives x_i/x_j $(1 \leq i, j \leq D;\ i \neq j)$ de les components que la integren. Aquest és el principi fonamental de l'anàlisi de dades composicionals que Aitchison (1997) anomena "invariància per canvi d'escala". Una conseqüència important que es dedueix d'aquest principi és que

> "*qualsevol funció aplicada sobre dades composicionals ha de poder expressar-se en termes de quocients entre les seves parts.*"

Per aquesta raó, l'investigador que vulgui aplicar qualsevol tècnica estadística a les dades composicionals interpretant-les com a informació relativa, haurà d'utilitzar quocients entre components en comptes del valor de les magnituds absolutes.

Pel que fa a qüestions de notació, quan coneguem el nombre de components d'una composició escriurem $\mathbf{x}$, però quan no estigui clar escriurem un superíndex, $\mathbf{x}^{(D)} = (x_1, x_2, \ldots, x_D)'$, que indicarà el nombre de components. Aquest superíndex ens indicarà també el nombre de components d'un subvector de $\mathbf{x}$, per exemple $\mathbf{x}^{(C)} = (x_1, x_2, \ldots, x_C)'$. Així mateix utilitzarem la notació $\mathbf{x}_{(C)} = (x_{C+1}, x_{C+2}, \ldots, x_D)'$ per representar el vector que s'obté en suprimir

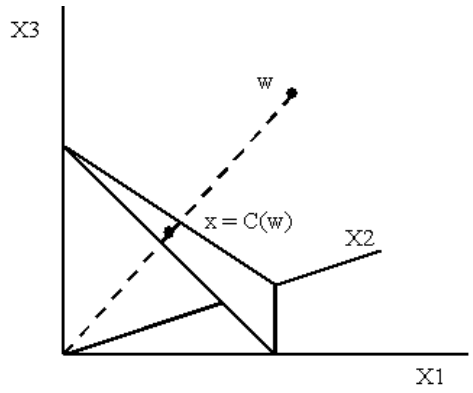


Figura 2.2: *La composició* $\mathbf{x}$ *i un dels seus vectors* $\mathbf{w}$ *associats*

les C primeres components de $\mathbf{x}$ i la notació $\mathbf{x}_{-j}$ per indicar la composició $\mathbf{x}$ sense la seva component x_j.

Sobre el símplex es defineixen dues operacions bàsiques: la pertorbació, definida per a dues composicions $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$, i la potència, definida entre una composició $\mathbf{x} \in \mathcal{S}^D$ i un escalar $\alpha \in \mathbb{R}$.

Definició 2.3 Siguin $\mathbf{x}, \mathbf{x}^*$ dues composicions amb D parts. Llavors l'operació

$$\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}\left(x_1 x_1^*, x_2 x_2^*, \ldots, x_D x_D^*\right)'$$

s'anomena *pertorbació*. □

Definició 2.4 Sigui $\mathbf{x}$ una composició amb D parts i sigui α un escalar de $\mathbb{R}$. Llavors l'operació

$$\alpha \otimes \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)'$$

s'anomena *potència*. □

Les operacions pertorbació i potència, que indiquem amb els símbols $\oplus$ i $\otimes$ respectivament, indueixen en el símplex una estructura d'espai vectorial sobre el cos $\mathbb{R}$. La pertorbació actua com a operació interna a $\mathcal{S}^D$ i li dóna les propietats de grup commutatiu; la potència actua com a operació externa respecte dels elements del cos $\mathbb{R}$ i satisfà les propietats usuals requerides en un espai vectorial. A continuació detallem aquestes propietats:

1. L'operació pertorbació $\oplus$ compleix les següents propietats:

   a. Commutativa: $\mathbf{x} \oplus \mathbf{x}^* = \mathbf{x}^* \oplus \mathbf{x} \qquad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$.

   b. Associativa: $(\mathbf{x} \oplus \mathbf{x}^*) \oplus \mathbf{x}' = \mathbf{x} \oplus (\mathbf{x}^* \oplus \mathbf{x}') \qquad \forall \mathbf{x}, \mathbf{x}^*, \mathbf{x}' \in \mathcal{S}^D$.

   c. Existència d'element neutre $\mathbf{n_e} = (1/D, \ldots, 1/D)'$ tal que $\mathbf{x} \oplus \mathbf{n_e} = \mathbf{x} \qquad \forall \mathbf{x} \in \mathcal{S}^D$.

   d. Per a cada $\mathbf{x} \in \mathcal{S}^D$ existeix un element invers $\mathbf{x}^{-1} = \mathcal{C}(1/x_1, 1/x_2, \ldots, 1/x_D)'$ que satisfà la condició $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{n_e}$.

2. L'operació potència $\otimes$ compleix les següents propietats:

   a. $\alpha \otimes (\beta \otimes \mathbf{x}) = (\alpha\beta) \otimes \mathbf{x} \qquad \forall \alpha, \beta \in \mathbb{R} \qquad \forall \mathbf{x} \in \mathcal{S}^D$.

   b. $\alpha \otimes (\mathbf{x} \oplus \mathbf{x}^*) = (\alpha \otimes \mathbf{x}) \oplus (\alpha \otimes \mathbf{x}^*) \qquad \forall \alpha \in \mathbb{R} \qquad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$.

   c. $(\alpha \otimes \mathbf{x}) \oplus (\beta \otimes \mathbf{x}) = (\alpha + \beta) \otimes \mathbf{x} \qquad \forall \alpha, \beta \in \mathbb{R} \qquad \forall \mathbf{x} \in \mathcal{S}^D$.

   d. $1 \otimes \mathbf{x} = \mathbf{x} \qquad \forall \mathbf{x} \in \mathcal{S}^D$, on 1 és l'element neutre multiplicatiu del cos $\mathbb{R}$.

La demostració d'aquestes propietats és immediata si utilitzem la definició de les dues operacions juntament amb les propietats clàssiques de la suma, el producte i la potència a l'espai real.

Aitchison (2002) defineix un producte escalar que dota el símplex d'una estructura d'espai vectorial euclidià com

$$\langle \mathbf{x}, \mathbf{x}^* \rangle_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{x_i^*}{x_j^*}. \tag{2.1}$$

Aquest producte escalar compleix les propietats habituals, és a dir, és una forma bilineal simètrica i definida positiva. La norma associada a aquest producte escalar és:

$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} = \sqrt{\frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right)^2}. \tag{2.2}$$

La distància induïda pel producte escalar que dota el símplex d'estructura d'espai mètric és l'anomenada *distància d'Aitchison* i té la següent expressió:

$$d_a(\mathbf{x}, \mathbf{x}^*) = \sqrt{\frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{x_i^*}{x_j^*} \right)^2}. \tag{2.3}$$

La distància d'Aitchison compleix les propietats usuals d'una mètrica, però també satisfà altres propietats que són conseqüència de la pròpia naturalesa de les dades composicionals.

Propietat 2.1 Siguin $\mathbf{x}, \mathbf{x}^*$ i $\mathbf{x}'$ tres dades composicionals amb D parts. La distància d'Aitchison verifica les següents propietats:

a. Invariància respecte de les pertorbacions: $d_a(\mathbf{x}, \mathbf{x}^*) = d_a(\mathbf{x}' \oplus \mathbf{x}, \mathbf{x}' \oplus \mathbf{x}^*)$.

b. Coherència respecte de l'operació potència: $d_a(\alpha \otimes \mathbf{x}, \alpha \otimes \mathbf{x}^*) = |\alpha| d_a(\mathbf{x}, \mathbf{x}^*)$, $\forall \alpha \in \mathbb{R}$.

c. Invariància respecte de les permutacions: $d_a(\mathbf{Px}, \mathbf{Px}^*) = d_a(\mathbf{x}, \mathbf{x}^*)$, $\forall \mathbf{P}$ matriu permutació de les components d'una composició.

□

Les propietat 2.1a i 2.1b permeten afirmar que la distància d_a és compatible o coherent amb l'estructura algebraica de l'espai (vegeu apartat 1.1, definició 1.1). La propietat 2.1c exigeix que la mesura de diferència entre dues composicions no depengui de l'ordre de les seves components. Aquestes han estat objecte d'estudi en diversos treballs: Aitchison (1992), Aitchison et al. (2000), Martín-Fernández et al. (1998a) o Martín-Fernández (2001). Trobem però la seva demostració explícita a Pawlowsky-Glahn i Egozcue (2002).

Una expressió equivalent a (2.3) i estudiada per Martín-Fernández et al. (1998b) és:

$$d_a(\mathbf{x}, \mathbf{x}^*) = \sqrt{\sum_{i=1}^{D} \left(\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{x_i^*}{g(\mathbf{x}^*)} \right)^2}, \tag{2.4}$$

on $g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i \right)^{1/D}$ i $g(\mathbf{x}^*) = \left(\prod_{i=1}^{D} x_i^* \right)^{1/D}$ representen les mitjanes geomètriques de les components de $\mathbf{x}$ i $\mathbf{x}^*$ respectivament.

Sabem que el símplex és un espai vectorial de dimensió $D-1$ i, per tant, qualsevol base estarà formada per $D-1$ elements (vegeu Barceló-Vidal, 2000). Hem vist que les operacions a $\mathcal{S}^D$ són $\oplus$ i $\otimes$. Tot i això, habitualment expressem les composicions amb D parts en termes de la base canònica de l'espai $\mathbb{R}^D$ i de les operacions suma i producte per escalars habituals en aquest espai, és a dir

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)' = x_1(1, 0, \ldots, 0)' + x_2(0, 1, 0, \ldots, 0)' + \cdots + x_D(0, \ldots, 0, 1)'. \tag{2.5}$$

Amb l'estructura que acabem d'introduir la base canònica de $\mathbb{R}^D$ no és una base de $\mathcal{S}^D$. Observem que els seus vectors no pertanyen ni tan sols a l'espai ja que tenen components iguals a 0. Tampoc l'expressió (2.5) representa una combinació lineal a $\mathcal{S}^D$ ja que ni la suma

ni el producte per escalars són operacions que doten al símplex d'estructura d'espai vectorial. Com a exemple d'un sistema de generadors de $\mathcal{S}^D$ tenim el conjunt $B^* = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\}$ on $\mathbf{w}_i = \mathcal{C}(1, 1, \ldots, e, \ldots, 1)'$ amb l'element e situat a la i-èsima fila. Observem que qualsevol composició $\mathbf{x}$ de $\mathcal{S}^D$ es pot escriure com

$$\mathbf{x} = (\ln x_1 \otimes \mathbf{w}_1) \oplus (\ln x_2 \otimes \mathbf{w}_2) \oplus \cdots \oplus (\ln x_D \otimes \mathbf{w}_D).$$

Una composició no s'expressa de manera única com a combinació lineal d'un sistema de generadors. Certament, la següent combinació lineal és equivalent a l'anterior

$$\mathbf{x} = \left(\ln\left(\frac{x_1}{g(\mathbf{x})}\right) \otimes \mathbf{w}_1\right) \oplus \left(\ln\left(\frac{x_2}{g(\mathbf{x})}\right) \otimes \mathbf{w}_2\right) \oplus \cdots \oplus \left(\ln\left(\frac{x_D}{g(\mathbf{x})}\right) \otimes \mathbf{w}_D\right),$$

on $g(\mathbf{x})$ representa la mitjana geomètrica de les components de $\mathbf{x}$. Per tant, direm que les coordenades de $\mathbf{x}$ respecte del sistema de generadors $B^* = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\}$ són indistintament $(\ln x_1, \ln x_2, \ldots, \ln x_D)'$ o bé

$$\left(\ln\left(\frac{x_1}{g(\mathbf{x})}\right), \ln\left(\frac{x_2}{g(\mathbf{x})}\right), \ldots, \ln\left(\frac{x_D}{g(\mathbf{x})}\right)\right)'. \tag{2.6}$$

Donat que la dimensió de $\mathcal{S}^D$ és $D - 1$, podem eliminar un dels vectors del sistema de generadors anterior i obtenir una base de $\mathcal{S}^D$. Si eliminem per exemple l'últim vector $\mathbf{w}_D$ podem expressar qualsevol composició $\mathbf{x}$ de manera única com a combinació lineal de la base $B = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$:

$$\mathbf{x} = \left(\ln\left(\frac{x_1}{x_D}\right) \otimes \mathbf{w}_1\right) \oplus \left(\ln\left(\frac{x_2}{x_D}\right) \otimes \mathbf{w}_2\right) \oplus \cdots \oplus \left(\ln\left(\frac{x_{D-1}}{x_D}\right) \otimes \mathbf{w}_{D-1}\right).$$

Així doncs direm que les coordenades de $\mathbf{x}$ respecte de la base B són

$$\left(\ln\left(\frac{x_1}{x_D}\right), \ln\left(\frac{x_2}{x_D}\right), \ldots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right)'. \tag{2.7}$$

Si en comptes d'eliminar el vector $\mathbf{w}_D$ eliminem el vector $\mathbf{w}_i$, obtenim una altra base i les components de qualsevol composició $\mathbf{x}$ respecte d'aquesta seran

$$\left(\ln\left(\frac{x_1}{x_i}\right), \ldots, \ln\left(\frac{x_{i-1}}{x_i}\right), \ln\left(\frac{x_{i+1}}{x_i}\right), \ldots, \ln\left(\frac{x_D}{x_i}\right)\right)'.$$

El producte escalar (2.1) i la norma (2.2) asseguren l'existència d'una base ortonormal. Fent uns simples càlculs podem constatar que els vectors $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$ no són ortogonals ni tampoc unitaris. No obstant això, el mètode de Gram-Schmidt aplicat sobre aquesta

base, o sobre qualsevol altre base de $\mathcal{S}^D$, ens proporciona una base ortonormal que denotarem com $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$. Si tenim en compte les propietats del producte escalar, veurem que qualsevol composició $\mathbf{x} \in \mathcal{S}^D$ es pot escriure com

$$\mathbf{x} = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a \otimes \mathbf{e}_1) \oplus (\langle \mathbf{x}, \mathbf{e}_2 \rangle_a \otimes \mathbf{e}_2) \oplus \ldots \oplus (\langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \otimes \mathbf{e}_{D-1}),$$

ja que el producte escalar $\langle \mathbf{x}, \mathbf{e}_i \rangle_a$ dóna com a resultat el mòdul de la projecció de $\mathbf{x}$ sobre $\mathbf{e}_i$, és a dir, la component i-èsima de $\mathbf{x}$ respecte de la base $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$. Direm per tant que les coordenades de $\mathbf{x}$ respecte de la base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ són:

$$(\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a)'. \tag{2.8}$$

Existeix un nombre infinit de bases ortonormals en el símplex. A Barceló-Vidal (2000) i Egozcue et al. (2001) trobem dos exemples concrets i diferents de bases ortonormals calculades amb el mètode de Gram-Schmidt.

Tal i com hem indicat en el capítol anterior, sobre les coordenades d'una composició respecte d'una base qualsevol de l'espai podem aplicar les operacions estàndards de l'espai real, és a dir, la suma de vectors i el producte d'un vector per un escalar. El resultat seran les coordenades de la composició resultant respecte de la mateixa base (vegeu apartat 1.1 expressions (1.1) i (1.2)). En el cas particular de treballar amb les coordenades respecte d'una base ortonormal, podem utilitzar a més el producte escalar ordinari i la distància euclidiana habitual. El resultat serà el mateix que aplicar el producte escalar (2.1) i la distància (2.3) sobre les composicions (vegeu apartat 1.1 expressions (1.3) i (1.4)). No podem assegurar el mateix si treballem amb coordenades respecte d'un sistema de generadors tot i que en el cas particular de les coordenades (2.6) es compleixen totes les propietats. Observem, per exemple, que la distància d'Aitchison (2.4) coincideix amb la distància euclidiana ordinària entre les components (2.6).

A l'espai vectorial $\mathcal{S}^D$ podem associar-li un espai afí. D'aquesta manera podem treballar amb punts i vectors de $\mathcal{S}^D$. Tindrem per tant sistemes de referències afins, constituïts per una base i una composició origen del sistema. És habitual prendre la composició neutra $\mathbf{n_e} = (1/D, \ldots, 1/D)'$ com a origen del sistema de referència afí.

2.2 Subcomposicions i amalgames

Donada una composició $\mathbf{x} \in \mathcal{S}^D$ ens interessarà sovint el valor de les magnituds relatives d'un subconjunt de components. Necessitem, doncs, un procediment per formar subcomposicions.

Definició 2.5 Si S és un subconjunt qualsevol de les parts $1, 2, \ldots, D$ d'una composició $\mathbf{x}$ de $\mathcal{S}^D$, i $\mathbf{x}_S$ simbolitza el subvector format per les corresponents components de $\mathbf{x}$, llavors $\mathbf{s} = \mathcal{C}(\mathbf{x}_S)$ rep el nom de *subcomposició* de les S parts de $\mathbf{x}$. □

Si el conjunt S consta de C $(2 \leq C < D)$ parts, la formació d'una subcomposició es pot considerar com una transformació de l'espai $\mathcal{S}^D$ a $\mathcal{S}^C$ la qual, a cada composició $\mathbf{x} \in \mathcal{S}^D$, li fa correspondre la composició $\mathbf{s} = \mathcal{C}(\mathbf{x}_S) \in \mathcal{S}^C$. Així doncs, podem imaginar-nos que una subcomposició és una composició en un símplex de dimensió inferior. El procés de formació d'una subcomposició té dues parts. En primer lloc cal seleccionar el subvector que ens interessa i tot seguit aplicar-li l'operador clausura. La selecció de les C components es pot dur a terme mitjançant una matriu d'ordre $C \times D$ $(2 \leq C < D)$ amb C elements iguals a 1, un a cada fila i com a màxim un a cada columna, i amb els altres $C(D-1)$ elements iguals a 0. Aquesta matriu s'anomena *matriu de selecció* i es denota $\mathbf{S}$. Així, per exemple, la matriu de selecció del subvector $\mathbf{x}_S = (x_1, x_4, x_5)'$ de la composició $\mathbf{x} = (x_1, x_2, \ldots, x_5)'$ és

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Obtenim el subvector mitjançant el producte $\mathbf{x}_S = \mathbf{S}\mathbf{x}$. Finalment, arribem a la subcomposició aplicant l'operador clausura, resultant $\mathbf{s} = \mathcal{C}(\mathbf{S}\mathbf{x})$.

Una propietat important que cal remarcar és que les subcomposicions conserven les magnituds relatives entre les seves parts.

Propietat 2.2 El quocient entre qualsevol parell de components d'una subcomposició és el mateix que el quocient entre les mateixes components en la composició inicial. □

Així doncs, quan treballem amb funcions invariants per canvi d'escala serem "subcomposicionalment coherents"(Aitchison, 1997) ja que els quocients entre les parts seleccionades romandran inalterables en la subcomposició.

La formació d'una subcomposició a partir d'una composició té una interpretació geomètrica: aquest procés es pot veure com una projecció lineal. En la figura 2.3 il·lustrem un exemple: el punt $\mathbf{P} = (p_1, p_2, p_3)$ representa una composició de 3 parts i la projectem sobre el costat 12 per obtenir $\mathbf{P'} = (p'_1, p'_2)$, que representa la subcomposició $\mathcal{C}(p_1, p_2)$.

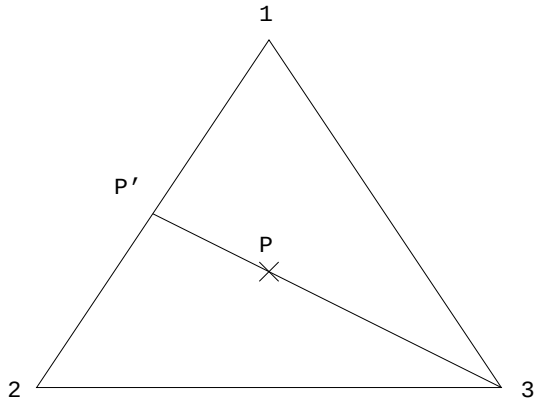


Figura 2.3: *Subcomposició $P' \in \mathcal{S}^1$ representada com una projecció lineal de $P \in \mathcal{S}^3$.*

La definició de les subcomposicions ens porta a exigir una altra condició a la distància d_a, anomenada propietat de la dominància respecte de les subcomposicions (Aitchison, 1992).

Propietat 2.3 Siguin $\mathbf{x}$ i $\mathbf{x}^*$ dues composicions amb D parts. Siguin $\mathbf{s}_x$ i $\mathbf{s}_{x^*}$ les respectives subcomposicions amb C parts. Aleshores es compleix: $d_a(\mathbf{x}, \mathbf{x}^*) \geq d_a(\mathbf{s}_x, \mathbf{s}_{x^*})$. □

En alguns casos, especialment quan les composicions tenen un gran nombre de components, ens interessarà amalgamar o sumar components per formar una nova composició "amalgamada". També ens interessarà sumar components d'una composició quan contingui una gran quantitat de parts amb un valor molt proper a zero (vegeu Martín-Fernández et al., 1997). Així doncs, donada per exemple la composició $\mathbf{x} = (x_1, x_2, \ldots, x_9)' \in \mathcal{S}^9$, el nostre interès es pot centrar en la composició $\mathbf{x}_A = (x_2 + x_3, x_1 + x_4, x_5 + x_6 + x_7, x_8 + x_9)' \in \mathcal{S}^4$.

Definició 2.6 Considerem el conjunt de les D parts d'una composició de $\mathcal{S}^D$ separat en C $(2 \leq C \leq D)$ subconjunts mútuament disjunts, i sumem les components dins de cada subconjunt. Llavors la composició amb C parts que en resulta rep el nom d'*amalgama*. □

L'amalgama de les parts d'una composició és també una transformació de $\mathcal{S}^D$ a $\mathcal{S}^C$ i es pot realitzar a partir d'una matriu d'ordre $C \times D$ amb D elements iguals a 1, cada un situat a una

columna diferent i de manera que cada fila contingui com a mínim un d'aquests elements, i amb els $(C-1)D$ elements restants iguals a 0. Anomenem aquesta matriu *matriu d'amalgama* i la denotem $\mathbf{A}$. A l'exemple anterior, la matriu de l'amalgama és igual a

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

D'aquesta manera, l'amalgama $\mathbf{x}_A = (x_2 + x_3, x_1 + x_4, x_5 + x_6 + x_7, x_8 + x_9)'$ no és més que $\mathbf{Ax}$. No tenim una interpretació geomètrica de l'operació amalgama. Aquest fet ens provoca problemes quan treballem amb variables aleatòries definides a $\mathcal{S}^D$, donat que és complicat trobar la distribució de l'amalgama $\mathbf{x}_A$ tot i conèixer la distribució de la composició aleatòria $\mathbf{x}$. En aquests casos cal recórrer a tècniques de simulació que ens aporten solucions aproximades.

2.3 Transformacions logquocients

En l'apartat anterior hem vist que l'espai mostral de les composicions és el símplex. Hem indicat també que sembla apropiat centrar l'estudi de composicions en la magnitud relativa de les components, és a dir, en els quocients entre components. Per aquesta raó, pot ser inadequat utilitzar molts dels procediments estadístics habituals que s'apliquen a l'espai $\mathbb{R}^D$ ja que aquests centren l'atenció en la magnitud absoluta de les components.

La solució introduïda per Aitchison, (1982, 1986) i utilitzada posteriorment per nombrosos autors, consisteix en transformar les composicions de $\mathcal{S}^D$ en vectors de l'espai real multivariant. D'aquesta manera, treballant amb les dades transformades, podem aplicar qualsevol tècnica estàndard sense cap mena de problema. Una vegada obtinguts els resultats tornem al símplex aplicant les eines pròpies de l'anàlisi real.

L'espai mostral dels quocients entre les parts d'una dada composicional respecte d'una d'elles és l'octant real positiu de $\mathbb{R}^{D-1}$. Si prenem els logaritmes d'aquests quocients, l'espai final esdevé tot $\mathbb{R}^{D-1}$. Així doncs, les transformacions proposades per Aitchison es basen en logaritmes de quocients. Aquesta estratègia es remunta al treball de McAlister (1879) on es desenvolupen els fonaments de la llei lognormal mitjançant el logaritme de les dades.

En aquest apartat presentem cinc transformacions: logquocient additiva, logquocient centrada, logquocient isomètrica, logquocient multiplicativa i Box-Cox, totes elles basades en quocients entre components. En primer lloc introduirem la definició habitual i en recordarem les propietats. En els tres primers casos veurem que el vector transformat coincideix amb les coordenades de la composició inicial respecte d'una base o d'un sistema de generadors de $\mathcal{S}^D$. Veurem també que les propietats d'aquestes transformacions es poden reinterpretar en termes de les coordenades de la composició.

2.3.1 Transformació logquocient additiva (alr)

Aitchison (1986) defineix la transformació logquocient additiva, anomenada alr (de l'anglès *additive logratio*) com:

Definició 2.7 Donada una composició amb D parts, la *transformació logquocient additiva* (alr) de $\mathbf{x} \in \mathcal{S}^D$ a $\mathbf{y} \in \mathbb{R}^{D-1}$ es defineix com

$$\mathbf{y} = \operatorname{alr}(\mathbf{x}) = \big(\ln(x_1/x_D), \ln(x_2/x_D), \ldots, \ln(x_{D-1}/x_D)\big)', \tag{2.9}$$

que abreujarem per

$$\mathbf{y} = \operatorname{alr}(\mathbf{x}) = \ln(\mathbf{x}_{-D}/x_D).$$

□

El jacobià d'aquesta transformació és

$$\operatorname{jac}(\mathbf{y}|\mathbf{x}^{(D-1)}) = \left(\prod_{i=1}^{D} x_i\right)^{-1}.$$

La transformació alr és bijectiva i la seva inversa és la transformació *logística additiva* (alr^{-1}), que es defineix com

$$\begin{aligned} x_i &= \frac{\exp y_i}{\sum_{j=1}^{D-1} \exp y_j + 1} \qquad (i = 1, 2, \ldots, D-1), \\ x_D &= 1 - \left(\sum_{i=1}^{D-1} x_i\right) = \frac{1}{\sum_{j=1}^{D-1} \exp y_j + 1}. \end{aligned}$$

La transformació alr és una aplicació lineal entre els espais vectorials $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$ ja que conserva ambdues operacions, la interna i l'externa:

$$\text{alr}(\mathbf{x} \oplus \mathbf{x}^*) = \text{alr}(\mathbf{x}) + \text{alr}(\mathbf{x}^*) \qquad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D, \tag{2.10}$$

$$\text{alr}(\alpha \otimes \mathbf{x}) = \alpha \text{alr}(\mathbf{x}) \qquad \forall \mathbf{x} \in \mathcal{S}^D \qquad \forall \alpha \in \mathbb{R}. \tag{2.11}$$

Un dels inconvenients de la transformació alr és la seva falta de simetria, ja que la component que figura en el denominador de cada logquocient adquireix un protagonisme especial respecte de la resta de components. Certament podríem escollir qualsevol altra component com a comú denominador. Quan es vulgui indicar que estem utilitzant la component x_j com a divisor en la transformació alr, s'indicarà amb la següent notació:

$$\text{alr}_j(\mathbf{x}) = \ln(\mathbf{x}_{-j}/x_j). \tag{2.12}$$

Les transformacions alr i alr^{-1} assignen la base canònica de $\mathbb{R}^{D-1}$ a la base no ortonormal $B = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$ de $\mathcal{S}^D$ on $\mathbf{w}_i = \mathcal{C}(1, 1, \ldots, e, \ldots, 1)'$ amb l'element e situat a la fila i-èsima. Estem, doncs, davant transformacions no isomètriques i per tant no conserven ni les distàncies, ni el producte escalar ni la norma. Això és una limitació important que caldrà tenir en compte especialment si s'aplica als vectors alr-transformats algun procediment que depengui de la mètrica.

Si recordem l'estructura d'espai vectorial de $\mathcal{S}^D$ introduïda a la secció 2.1, ens adonarem que el vector alr-transformat (2.9) coincideix amb el vector (2.7), és a dir, amb el vector de coordenades de la composició $\mathbf{x}$ respecte de la base B. Per simplificar la notació utilitzarem també l'expressió $\text{alr}(\mathbf{x})$ per indicar les coordenades de la composició $\mathbf{x}$ respecte de la base B. En aquest context la igualtat (2.10) ens diu que les coordenades respecte de la base B de la composició $\mathbf{x} \oplus \mathbf{x}^*$ són iguals a la suma habitual de les coordenades de $\mathbf{x}$ i $\mathbf{x}^*$ respecte de la mateixa base. Per altra banda la igualtat (2.11) ens diu que les coordenades de la composició $\alpha \otimes \mathbf{x}$ respecte de la base B són iguals al producte habitual entre α i el vector de coordenades de $\mathbf{x}$ respecte de la base B. Sabem que aquestes igualtats es compleixen en qualsevol espai vectorial, sempre i quan treballem amb les coordenades respecte d'una base (vegeu apartat 1.1).

El producte escalar entre dues composicions calculat utilitzant (2.1) no és igual al producte escalar ordinari de $\mathbb{R}^{D-1}$ entre les respectives components alr. La igualtat és certa en qualsevol

espai vectorial sempre i quan treballem amb coordenades respecte d'una base ortonormal però, en el nostre cas, la base B no és ortonormal. Per la mateixa raó, tampoc podrem calcular la norma d'una composició ni la distància d'Aitchison entre dues composicions de l'espai afí aplicant la norma estàndard i la distància euclidiana ordinària de $\mathbb{R}^{D-1}$ sobre les coordenades alr de les composicions. Aquest fet coincideix amb l'afirmació anterior on dèiem que la transformació alr no és isomètrica.

Si a la base B canviem qualsevol vector $\mathbf{w}_j$ pel vector $\mathbf{w}_D$, les components de qualsevol composició $\mathbf{x}$ es correspondran amb les coordenades del vector (2.12).

Així doncs, el vector $\text{alr}(\mathbf{x})$ es pot interpretar com el vector resultant d'una transformació o bé com les coordenades de la composició $\mathbf{x}$ respecte de la base B. Aquesta dualitat ens tornarà a aparèixer en els següents capítols quan introduïm models paramètrics sobre $\mathcal{S}^D$, donat que els definirem mitjançant la tècnica de les transformacions i mitjançant les coordenades respecte d'una base.

2.3.2 Transformació logquocient centrada (clr)

Aitchison (1986) defineix la transformació logquocient centrada, anomenada clr (de l'anglès *centred logratio*) com:

Definició 2.8 Donada una composició amb D parts, la *transformació logquocient centrada* (clr) de $\mathbf{x} \in \mathcal{S}^D$ a $\mathbf{z} \in \mathbb{R}^D$ es defineix com

$$\mathbf{z} = \text{clr}(\mathbf{x}) = \big(\ln(x_1/g(\mathbf{x})), \ln(x_2/g(\mathbf{x})), \ldots, \ln(x_D/g(\mathbf{x}))\big)', \tag{2.13}$$

que abreujarem per

$$\mathbf{z} = \text{clr}(\mathbf{x}) = \ln(\mathbf{x}/g(\mathbf{x})),$$

on $g(\mathbf{x})$ és la mitjana geomètrica de les D components de $\mathbf{x}$. □

En aquest cas, la transformació és simètrica entre les parts. La seva imatge és l'hiperplà V de $\mathbb{R}^D$ que passa per l'origen i és ortogonal al vector d'unitats, és a dir, $V = \text{clr}(\mathcal{S}^D) = \{\mathbf{z} \in \mathbb{R}^D; \sum_{i=1}^D z_i = 0\}$. Ens trobem, doncs, amb una nova dificultat ja que la suma de les components del vector transformat és igual a 0. Tot i això, observem que la clr és una

aplicació lineal entre $\mathcal{S}^D$ i V ja que conserva ambdues operacions, la interna i l'externa:

$$\text{clr}(\mathbf{x} \oplus \mathbf{x}^*) = \text{clr}(\mathbf{x}) + \text{clr}(\mathbf{x}^*) \qquad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D, \tag{2.14}$$

$$\text{clr}(\alpha \otimes \mathbf{x}) = \alpha \text{clr}(\mathbf{x}) \qquad \forall \mathbf{x} \in \mathcal{S}^D \qquad \forall \alpha \in \mathbb{R}. \tag{2.15}$$

La transformació clr és bijectiva entre el símplex i l'hiperplà V; la seva inversa no és altra que la *transformació logquocient centrada inversa* (clr^{-1}) que es defineix com

$$\mathbf{x} = \text{clr}^{-1}(\mathbf{z}) = \mathcal{C}(e^{z_1}, e^{z_2}, \ldots, e^{z_D})$$

Les transformacions clr i clr^{-1} són isometries entre $\mathcal{S}^D$ i V, i per tant conserven distàncies i productes escalars:

$$\langle \mathbf{x}, \mathbf{x}^* \rangle_a = \langle \text{clr}(\mathbf{x}), \text{clr}(\mathbf{x}^*) \rangle_{eu} \qquad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D, \tag{2.16}$$

$$\|\mathbf{x}\|_a = \|\text{clr}(\mathbf{x})\|_{eu} \qquad \forall \mathbf{x} \in \mathcal{S}^D,$$

$$d_a(\mathbf{x}, \mathbf{x}^*) = d_{eu}(\text{clr}(\mathbf{x}), \text{clr}(\mathbf{x}^*)) \qquad \forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D. \tag{2.17}$$

És a causa d'aquestes igualtats que Martín-Fernández et al. (1998a) defineixen la distància d'Aitchison entre dues composicions $\mathbf{x}$ i $\mathbf{x}^*$ directament com (2.17), és a dir, com la distància euclidiana entre els respectius vectors clr-transformats.

Aitchison (1986) dóna la relació entre les transformacions alr i clr:

$$\text{alr}(\mathbf{x}) = \mathbf{F}\text{clr}(\mathbf{x}), \tag{2.18}$$

essent $\mathbf{F}$ una matriu elemental especificada a la taula 2.1. La relació inversa ve donada per l'expressió

$$\text{clr}(\mathbf{x}) = \mathbf{F}^*\text{alr}(\mathbf{x}),$$

on $\mathbf{F}^*$ és la inversa generalitzada de Moore-Penrose de la matriu $\mathbf{F}$, és a dir, compleix que $\mathbf{F}\mathbf{F}^* = \mathbf{I}_{D-1}$. La seva expressió és

$$\mathbf{F}^* = \frac{1}{D} \begin{pmatrix} D-1 & -1 & \cdots & -1 \\ -1 & D-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & D-1 \\ -1 & -1 & \cdots & -1 \end{pmatrix}.$$

També es pot expressar com $\mathbf{F}^* = \mathbf{F}'(\mathbf{F}\mathbf{F}')^{-1} = \mathbf{F}'\mathbf{H}^{-1}$. Podem trobar la definició de les matrius $\mathbf{F}$ i $\mathbf{H}$ a la taula 2.1. Al llarg d'aquesta tesi doctoral, farem ús de la simbologia descrita a la taula 2.1 per denotar les matrius elementals.

Taula 2.1: *Matrius elementals.*

Notació	**Definició**	**Ordre**	**Rang**
$\mathbf{I}_k$	matriu identitat	$k \times k$	k
$\mathbf{J}_k$	matriu unitat	$k \times k$	1
$\mathbf{j}_k$	vector columna unitat	$k \times 1$	1
$\mathbf{F}_{k,k+1}$	$[\mathbf{I}_k : -\mathbf{j}_k]$	$k \times (k+1)$	k
$\mathbf{H}_k$	$\mathbf{I}_k + \mathbf{J}_k$	$k \times k$	k

Si tenim en compte l'estructura d'espai vectorial de $\mathcal{S}^D$ definida a la secció 2.1, observarem que el vector clr-transformat (2.13) coincideix amb el vector (2.6), és a dir, amb el vector de coordenades de la composició $\mathbf{x}$ respecte del sistema de generadors $B^* = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\}$ on $\mathbf{w}_i = \mathcal{C}(1, 1, \ldots, e, \ldots, 1)'$ amb l'element e situat a la i-èsima fila. Per simplificar la notació utilitzarem l'expressió $\mathrm{clr}(\mathbf{x})$ per indicar també les coordenades de la composició $\mathbf{x}$ respecte del sistema de generadors B^*. Sabem que les coordenades d'un vector respecte d'un sistema de generadors no són úniques i que en general no podem calcular el producte escalar i la distància entre composicions aplicant el producte escalar ordinari i la distància euclidiana habitual a les coordenades respecte d'un sistema de generadors. No obstant això, en el cas particular de les coordenades (2.6) es compleixen les igualtats (2.14), (2.15), (2.16) i (2.17). És a dir, podem aplicar-hi les operacions i els conceptes estàndards de l'espai real. Per altra banda, la relació (2.18) ens servirà també per transformar les coordenades en el sistema de generadors B^* en les coordenades en la base $B = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$.

Així doncs, hem vist que existeix també una dualitat en el vector $\mathrm{clr}(\mathbf{x})$, ja que es pot interpretar com el vector resultant d'una transformació sobre la composició $\mathbf{x}$ o bé com el vector de coordenades respecte del sistema de generadors B^*. Aquesta dualitat també tindrà conseqüències en la definició de models paramètrics a $\mathcal{S}^D$.

2.3.3 Transformació logquocient isomètrica (ilr)

Egozcue et al. (2003) defineixen una isometria entre els espais $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$. La motivació principal d'aquesta nova transformació és superar els defectes o inconvenients de les dues transformacions anteriors. Hem vist que la transformació alr no és una isometria. La transformació clr, tot i conservar les distàncies i el producte escalar, ens transforma les composicions en vectors d'un subespai de $\mathbb{R}^D$ amb la restricció addicional que la suma de les seves components és igual a 0. Aquests inconvenients aporten certes dificultats a l'hora d'interpretar resultats i han provocat una llarga discussió del mètode proposat per Aitchison (1986).

La transformació isomètrica sorgeix de manera natural si observem la transformació clr. La condició $\sum z_k = 0$ que satisfan les components dels vectors del subespai $V = \text{clr}(\mathcal{S}^D)$ ens indica que el vector $(1, 1, \ldots, 1)$ és ortogonal a aquest hiperplà. Si escollim una base de l'espai $\mathbb{R}^D$ formada per $D-1$ vectors ortonormals del subespai V i per un vector unitari i normal a V, és a dir, $1/\sqrt{D}(1, 1, \ldots, 1)$, i expressem els vectors clr transformats en aquesta nova base, obtindrem que la seva última component és igual a 0. A continuació, podem eliminar aquesta última component aplicant una projecció sobre l'hiperplà V. Aquest procediment, transformació clr seguida d'un canvi de base ortonormal i de la projecció ortogonal sobre el subespai V, dóna lloc a una isometria entre els espais $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$. Egozcue et al. (2003) defineixen aquesta transformació i la denoten per ilr (de l'anglès *isometric logratio*).

Definició 2.9 Donada una base ortonormal del símplex $\mathcal{S}^D$, $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$, es defineix la transformació *logquocient isomètrica* (ilr) d'una composició $\mathbf{x} \in \mathcal{S}^D$ a un vector $\mathbf{v} \in \mathbb{R}^{D-1}$ com

$$\mathbf{v} = \text{ilr}(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a)'. \tag{2.19}$$

□

La transformació isomètrica no és única, donat que en la seva definició no queda especificada la base ortonormal de $\mathcal{S}^D$ i per tant tenim la llibertat d'escollir-la.

Tal i com indica el seu nom, estem davant d'una isometria entre els espais $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$ en què la base $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ de $\mathcal{S}^D$ es correspon amb la base canònica de $\mathbb{R}^{D-1}$, ambdues ortonormals. Això implica que podem utilitzar les operacions estàndards de l'espai real, treballar amb la distància euclidiana i aplicar el producte escalar ordinari. Certament,

$\forall \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$ i $\forall \alpha \in \mathbb{R}$, es compleixen les igualtats:

$$\begin{aligned} \text{ilr}(\mathbf{x} \oplus \mathbf{x}^*) &= \text{ilr}(\mathbf{x}) + \text{ilr}(\mathbf{x}^*), \\ \text{ilr}(\alpha \otimes \mathbf{x}) &= \alpha \text{ilr}(\mathbf{x}), \\ \langle \mathbf{x}, \mathbf{x}^* \rangle_a &= \langle \text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{x}^*) \rangle_{eu}, \\ \|\mathbf{x}\|_a &= \|\text{ilr}(\mathbf{x})\|_{eu}, \\ d_a(\mathbf{x}, \mathbf{x}^*) &= d_{eu}(\text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{x}^*)). \end{aligned} \tag{2.20}$$

Egozcue et al. (2003) donen també les relacions entre les tres transformacions, alr, clr i ilr. Aquestes són:

$$\begin{aligned} &\text{clr}(\mathbf{x}) = \mathbf{U}\text{ilr}(\mathbf{x}); \quad \text{ilr}(\mathbf{x}) = \mathbf{U}'\text{clr}(\mathbf{x}); \\ &\text{alr}(\mathbf{x}) = \mathbf{F}\mathbf{U}\text{ilr}(\mathbf{x}); \quad \text{ilr}(\mathbf{x}) = (\mathbf{F}\mathbf{U})^{-1}\text{alr}(\mathbf{x}) = \mathbf{U}'\mathbf{F}^*\text{alr}(\mathbf{x}); \end{aligned} \tag{2.21}$$

on $\mathbf{F}$ és la matriu d'ordre $(D-1) \times D$ definida a la taula 2.1 i $\mathbf{U}$ és una matriu d'ordre $D \times (D-1)$ amb els vectors $\mathbf{u}_i = \text{clr}(\mathbf{e}_i)$, $i = 1, 2, \ldots, D-1$, com a columnes.

Utilitzant el fet que el jacobià de la transformació alr és igual a $\left(\prod_{i=1}^D x_i\right)^{-1}$ i la relació $\text{ilr}(\mathbf{x}) = (\mathbf{F}\mathbf{U})^{-1}\text{alr}(\mathbf{x})$, és immediat veure que el jacobià de la transformació ilr és igual a $\left(|\mathbf{F}\mathbf{U}| \prod_{i=1}^D x_i\right)^{-1}$. El producte $\mathbf{U}\mathbf{U}'$ és una matriu quadrada d'ordre D igual a la matriu

$$\frac{1}{D}\begin{pmatrix} D-1 & -1 & \cdots & -1 \\ -1 & D-1 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & D-1 \end{pmatrix},$$

la qual té dos valors propis: 0 amb multiplicitat 1, i 1 amb multiplicitat $D-1$. El subespai propi de valor propi 1 és $V = \{\mathbf{z} \in \mathbb{R}^D; \sum_{i=1}^D z_i = 0\}$. Donat que les files de la matriu $\mathbf{F}$ pertanyen al subespai V obtenim que $\mathbf{F}\mathbf{U}\mathbf{U}'\mathbf{F}' = \mathbf{F}\mathbf{F}'$. Sabem també que $\mathbf{F}\mathbf{F}'$ és igual a la matriu $\mathbf{H}_{D-1}$ (vegeu Aitchison, 1986, pàg 343) la qual té dos valors propis, 1 de multiplicitat $D-2$, i D de multiplicitat 1. En resum, $|\mathbf{F}\mathbf{U}\mathbf{U}'\mathbf{F}'| = |\mathbf{F}\mathbf{F}'| = |\mathbf{H}_{D-1}| = D$. També, i donat que $\mathbf{F}\mathbf{U}$ és una matriu quadrada, tenim que

$$|\mathbf{F}\mathbf{U}\mathbf{U}'\mathbf{F}'| = |\mathbf{F}\mathbf{U}|\,|(\mathbf{F}\mathbf{U})'| = |\mathbf{F}\mathbf{U}|^2,$$

per tant $|\mathbf{F}\mathbf{U}| = \sqrt{D}$. Així doncs, Es conclou que el jacobià de la transformació ilr és igual a

$$\text{jac}(\mathbf{v}|\mathbf{x}^{(D-1)}) = D^{-1/2}\left(\prod_{i=1}^D x_i\right)^{-1}.$$

Si recuperem l'estructura d'espai vectorial euclidià de $\mathcal{S}^D$ definida a l'apartat 2.1, ens adonarem que el vector (2.19) coincideix amb el vector (2.8), les coordenades de la composició $\mathbf{x}$ respecte de la base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ de $\mathcal{S}^D$. En cap dels dos casos anteriors tenim aquesta situació ja que el vector $\text{alr}(\mathbf{x})$ representa les coordenades respecte d'una base no ortonormal i el vector $\text{clr}(\mathbf{x})$ conté les coordenades respecte d'un sistema de generadors. Per simplificar la notació utilitzarem també l'expressió $\text{ilr}(\mathbf{x})$ per indicar les coordenades de la composició $\mathbf{x}$ respecte d'una base ortonormal. Observem que les igualtats (2.20) coincideixen amb la indicació de l'apartat 2.1 on afirmàvem que sobre les coordenades ilr podrem aplicar tota l'anàlisi real estàndard.

Observem també que les expressions (2.21) es poden interpretar com un canvi de base o un canvi en el sistema de generadors. Concretament, la relació entre les transformacions alr i ilr és un canvi de base, és a dir, la matriu $\mathbf{FU}$ és la matriu del canvi de la base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ a la base $B = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{D-1}\}$ definida anteriorment i per tant, les seves columnes estan formades per les coordenades dels vectors $\mathbf{e}_i$ en la base B, és a dir, pels vectors $\text{alr}(\mathbf{e}_i)$ $(i = 1, 2, \ldots, D-1)$.

En resum, podem veure que el vector $\text{ilr}(\mathbf{x})$ presenta la mateixa dualitat que els vectors $\text{alr}(\mathbf{x})$ i $\text{clr}(\mathbf{x})$. És a dir, es pot interpretar com el vector resultant d'una transformació o bé com les coordenades de la composició $\mathbf{x}$ respecte d'una base ortonormal.

2.3.4 Transformació logquocient multiplicativa (mlr)

Aitchison (1986) defineix també la transformació logquocient multiplicativa, anomenada mlr (de l'anglès *multiplicative logratio*).

Definició 2.10 Donada una composició amb D parts, la *transformació logquocient multiplicativa* (mlr) de $\mathbf{x} \in \mathcal{S}^D$ a $\mathbf{t} \in \mathbb{R}^{D-1}$ es defineix com $\mathbf{t} = \text{mlr}(\mathbf{x}) = (t_1, t_2, \ldots, t_{D-1})'$ amb

$$t_i = \ln\left(\frac{x_i}{1 - \sum_{j=1}^{i} x_j}\right) \qquad (i = 1, 2, \ldots, D-1).$$

□

La transformació mlr és bijectiva i la seva inversa és la transformació *logística multiplicativa* (mlr^{-1}) que es defineix com

$$\begin{aligned} x_i &= \frac{\exp t_i}{\prod_{j=1}^{i}(1+\exp t_j)} \qquad (i=1,\ldots,D-1), \\ x_D &= 1-\sum_{j=1}^{D-1} x_j = \frac{1}{\prod_{j=1}^{D-1}(1+\exp t_j)}. \end{aligned}$$

El jacobià d'aquesta transformació és

$$\mathrm{jac}(\mathbf{t}|\mathbf{x}^{(D-1)}) = \left(\prod_{j=1}^{D} x_j\right)^{-1}.$$

La transformació mlr no és una aplicació lineal entre els espais vectorials $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$ ja que no conserva cap de les dues operacions, la interna i l'externa. Certament si prenem les composicions $\mathbf{x} = (0.1, 0.5, 0.4)'$ i $\mathbf{x}^* = (0.3, 0.4, 0.3)'$ de $\mathcal{S}^3$, obtenim $\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}(0.03, 0.2, 0.12)'$, $2 \otimes \mathbf{x} = \mathcal{C}(0.01, 0.25, 0.16)'$ i les corresponents transformacions logquocients multiplicatives són

$$\begin{aligned} \mathrm{mlr}(\mathbf{x} \oplus \mathbf{x}^*) &= (-2.3671, 0.5108)', \\ \mathrm{mlr}(2 \otimes \mathbf{x}) &= (-3.7136, 0.4463)', \end{aligned}$$

valors que no coincideixen amb els termes

$$\begin{aligned} &\mathrm{mlr}(\mathbf{x}) + \mathrm{mlr}(\mathbf{x}^*) = (-2.1972, 0.2231)' + (-0.8473, 0.2877)' = (-3.0445, 0.5108)', \\ &2\mathrm{mlr}(\mathbf{x}) = 2(-2.1972, 0.2231)' = (-4.3944, 0.4462)'. \end{aligned}$$

La transformació mlr tampoc és una isometria. Si prenem les composicions $\mathbf{x}$ i $\mathbf{x}^*$ anteriors veiem que

$$\langle \mathbf{x}, \mathbf{x}^* \rangle_a = 0.1757 \qquad \text{i} \qquad \langle \mathrm{mlr}(\mathbf{x}), \mathrm{mlr}(\mathbf{x}^*) \rangle_{eu} = 1.9259.$$

És a dir, en general, donats $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$ i $\alpha \in \mathbb{R}$, tenim que

$$\mathrm{mlr}(\mathbf{x} \oplus \mathbf{x}^*) \neq \mathrm{mlr}(\mathbf{x}) + \mathrm{mlr}(\mathbf{x}^*), \tag{2.22}$$

$$\mathrm{mlr}(\alpha \otimes \mathbf{x}) \neq \alpha \mathrm{mlr}(\mathbf{x}), \tag{2.23}$$

$$\begin{aligned} \langle \mathbf{x}, \mathbf{x}^* \rangle_a &\neq \langle \mathrm{mlr}(\mathbf{x}), \mathrm{mlr}(\mathbf{x}^*) \rangle_{eu}, \\ \|\mathbf{x}\|_a &\neq \|\mathrm{mlr}(\mathbf{x})\|_{eu}, \\ d_a(\mathbf{x}, \mathbf{x}^*) &\neq d_{eu}(\mathrm{mlr}(\mathbf{x}), \mathrm{mlr}(\mathbf{x}^*)). \end{aligned}$$

Per tant, amb l'estructura d'espai vectorial de $\mathcal{S}^D$ considerada, el vector transformat no es pot interpretar com les coordenades de la composició $\mathbf{x}$ respecte d'una base de $\mathcal{S}^D$, doncs sabem que sobre les coordenades d'una composició respecte de qualsevol base podem aplicar les operacions suma i producte per escalars de l'espai real. Les desigualtats (2.22) i (2.23) ens indiquen que, en aquest cas, això no és possible.

2.3.5 Transformació Box-Cox

La transformació Box-Cox multivariant representa una generalització de la transformació logquocient additiva. Podem trobar-ne un estudi detallat a Barceló-Vidal (1996).

Definició 2.11 Donada una composició amb D parts, la *transformació Box-Cox multivariant* de paràmetre $\boldsymbol{\lambda} \in \mathbb{R}^{D-1}$ ($\mathrm{BC}^{\boldsymbol{\lambda}}$) és la transformació de $\mathbf{x} \in \mathcal{S}^D$ a $\mathbf{u} \in \mathbb{R}^{D-1}$ definida per

$$\mathbf{u} = \mathrm{BC}^{\boldsymbol{\lambda}}(\mathbf{x}) = (u_1, u_2, \ldots, u_{D-1})' \qquad \text{amb}$$

$$u_i = \begin{cases} \dfrac{(x_i/x_D)^{\lambda_i} - 1}{\lambda_i}, & \text{si } \lambda_i \neq 0, \\ \ln(x_i/x_D), & \text{si } \lambda_i = 0, \end{cases} \qquad (i = 1, 2, \ldots, D-1).$$

□

Si $\boldsymbol{\lambda} \neq 0$, la transformació inversa ve donada per

$$\mathbf{x} = \mathcal{C}\left((1+\lambda_1 u_1)^{1/\lambda_1}, (1+\lambda_2 u_2)^{1/\lambda_2}, \ldots, (1+\lambda_{D-1}u_{D-1})^{1/\lambda_{D-1}}\right)',$$

amb domini el subconjunt de $\mathbb{R}^{D-1}$ definit per

$$\left\{\mathbf{u} = (u_1, u_2, \ldots, \mathrm{u}_{D-1})' : u_1 > \frac{-1}{\lambda_1}, u_2 > \frac{-1}{\lambda_2}, \ldots, \mathrm{u}_{D-1} > \frac{-1}{\lambda_{D-1}}\right\}.$$

El jacobià d'aquesta transformació és igual a

$$\mathrm{jac}(\mathbf{u}|\mathbf{x}^{(D-1)}) = \left(\prod_{i=1}^{D} x_i\right)^{-1} \prod_{i=1}^{D-1} \left(\frac{x_i}{x_D}\right)^{\lambda_i}.$$

Aquesta família de transformacions és contínua respecte de $\boldsymbol{\lambda}$ i podem observar que quan $\boldsymbol{\lambda} \to \mathbf{0}$, les transformacions Box-Cox tendeixen a la transformació logquocient additiva.

La transformació Box-Cox de paràmetre $\boldsymbol{\lambda}$ és biunívoca, però no és una aplicació lineal entre els espais vectorials $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$ ni tampoc una transformació isomètrica. Si prenem les composicions $\mathbf{x} = (0.1, 0.5, 0.4)'$ i $\mathbf{x}^* = (0.3, 0.4, 0.4)'$ de $\mathcal{S}^3$, la transformació Box-Cox de paràmetre $\boldsymbol{\lambda} = (3, 2)'$ de les composicions $\mathbf{x} \otimes \mathbf{x}^*$ i $2 \otimes \mathbf{x}$ és igual a

$$\mathrm{BC}^{\boldsymbol{\lambda}}(\mathbf{x} \otimes \mathbf{x}^*) = (-0.4688, 1.2099)' \qquad \text{i} \qquad \mathrm{BC}^{\boldsymbol{\lambda}}(2 \otimes \mathbf{x}) = (-0.4980, 0.9382)',$$

valors que no coincideixen amb els termes

$$\mathrm{BC}^{\boldsymbol{\lambda}}(\mathbf{x}) + \mathrm{BC}^{\boldsymbol{\lambda}}(\mathbf{x}^*) = (-0.4688, 0.3177)' + (0, 0.4568)' = (-0.4688, 0.7745)',$$
$$2\mathrm{BC}^{\boldsymbol{\lambda}}(\mathbf{x}) = 2(-0.4688, 0.3177)' = (-0.9376, 0.6354)'.$$

Tampoc el producte escalar euclidià $\langle \mathrm{BC}^{\boldsymbol{\lambda}}(\mathbf{x}), \mathrm{BC}^{\boldsymbol{\lambda}}(\mathbf{x}^*) \rangle_{eu} = 0.1451$ és igual al producte escalar $\langle \mathbf{x}, \mathbf{x}^* \rangle_a = 0.1757$.

Així doncs, amb l'estructura d'espai vectorial de $\mathcal{S}^D$ considerada, el vector transformat no es pot interpretar com les coordenades de la composició $\mathbf{x}$ respecte d'una base de $\mathcal{S}^D$.

En el següent capítol veurem com definir un model paramètric sobre $\mathcal{S}^D$ a partir de les transformacions Box-Cox multivariants.

2.3.6 Comparació gràfica

Amb les figures 2.4, 2.5, 2.6 i 2.7 podrem comprovar de manera gràfica les propietats de les cinc transformacions.

En la figura 2.4(a) hem representat un quadrat amb la geometria de l'espai $\mathcal{S}^3$ de costat $\sqrt{2}$ i centrat en l'origen del símplex, la composició $(1/3, 1/3, 1/3)'$. Recordem que a $\mathcal{S}^3$ no treballem amb la mètrica euclidiana habitual i per tant no observem un quadrat igual que a l'espai real.

En els gràfics 2.4(b), 2.5(a) i 2.5(b) hem representat a l'espai $\mathbb{R}^2$ les figures que obtenim després d'aplicar les transformacions alr_1, alr_2 i alr_3 respectivament. Podem comprovar la linealitat de la transformació alr perquè conserva el paral·lelisme entre els costats. Observem també que es tracta d'una transformació no isomètrica donat que no conserva els angles ni les distàncies. Observem que els costats tenen una longitud diferent a $\sqrt{2}$. Concretament, les figures 2.4(b) i 2.5(a) tenen costats de longitud 1.506 i 2.394 aproximadament. La figura 2.5(b) té els quatre costats iguals i de longitud 2.

En la figura 2.6(a) hem representat a l'espai $\mathbb{R}^2$ el quadrat que s'obté després d'aplicar la transformació logquocient isomètrica prenent com a base ortonormal

$$\{\mathcal{C}(e^{\sqrt{1/2}}, e^{-\sqrt{1/2}}, 1)', \mathcal{C}(e^{\sqrt{1/6}}, e^{\sqrt{1/6}}, e^{-\sqrt{2/3}})'\}.$$

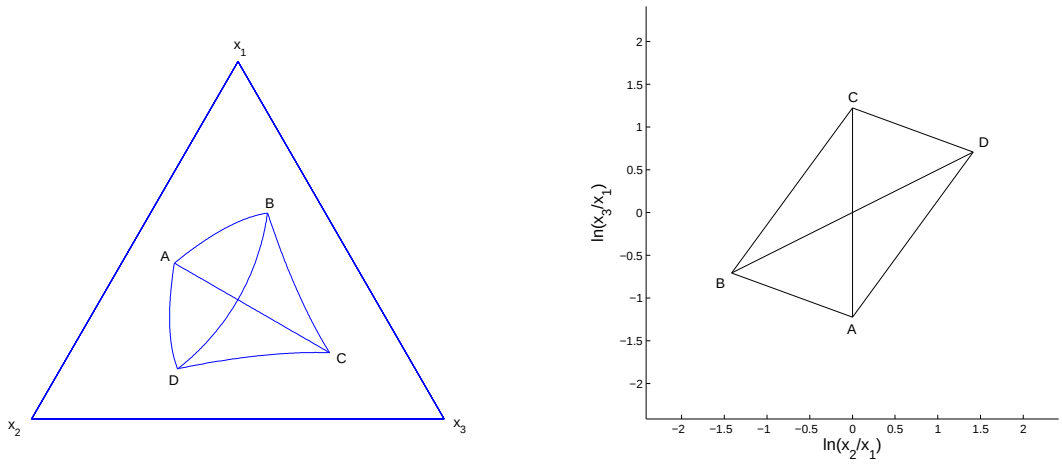
En aquest cas, l'expressió de la transformació ilr és

$$\mathbf{v} = \text{ilr}(x_1, x_2, x_3) = \left(\frac{1}{\sqrt{2}} \ln\left(\frac{x_2}{x_1}\right), \frac{1}{\sqrt{6}} \ln\left(\frac{x_1 x_2}{x_3^2}\right)\right).$$

Observem que es tracta d'una aplicació lineal i isomètrica.

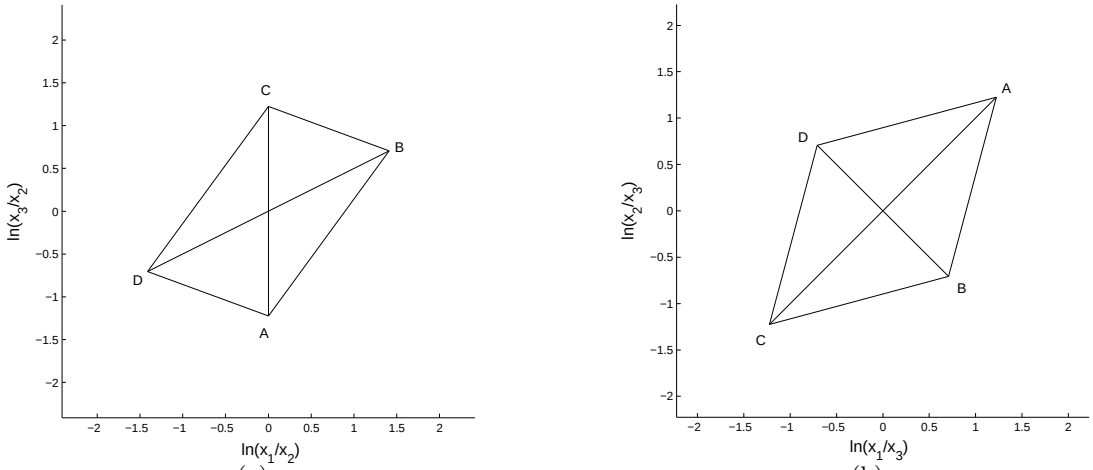
En la figura 2.6(b) hem representat el quadrat a l'espai $\mathbb{R}^3$ aplicant la transformació logquocient centrada. Tot i no poder-se apreciar molt bé en el gràfic obtenim un quadrat semblant al de la figura 2.6(a). De fet, si apliquem una projecció sobre el pla on tenim dibuixat el quadrat, observarem exactament el mateix que en la figura 2.6(a).

Finalment, la figura 2.7(a) conté el resultat d'aplicar la transformació logquocient multiplicativa i la figura 2.7(b) el resultat d'aplicar la transformació Box-Cox amb paràmetre $\boldsymbol{\lambda} = (3, 2)'$. En ambdós casos observem que els costats no són paral·lels dos a dos, senyal de la no linealitat de les dues transformacions. Tampoc conserven els angles i les distàncies, ja que, les transformacions mlr i Box-Cox no són isomètriques.



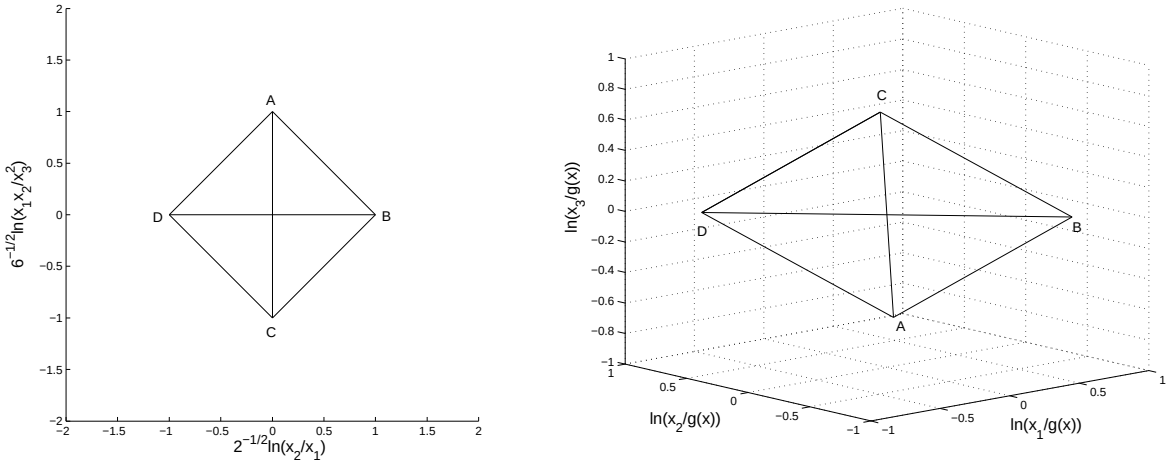
(a) (b)

Figura 2.4: (a) *Quadrat a* $\mathcal{S}^3$. (b) *Figura resultant després d'aplicar la transformació* alr_1.



(a) (b)

Figura 2.5: *Figura resultant després d'aplicar la transformació* (a) alr_2, (b) alr_3.

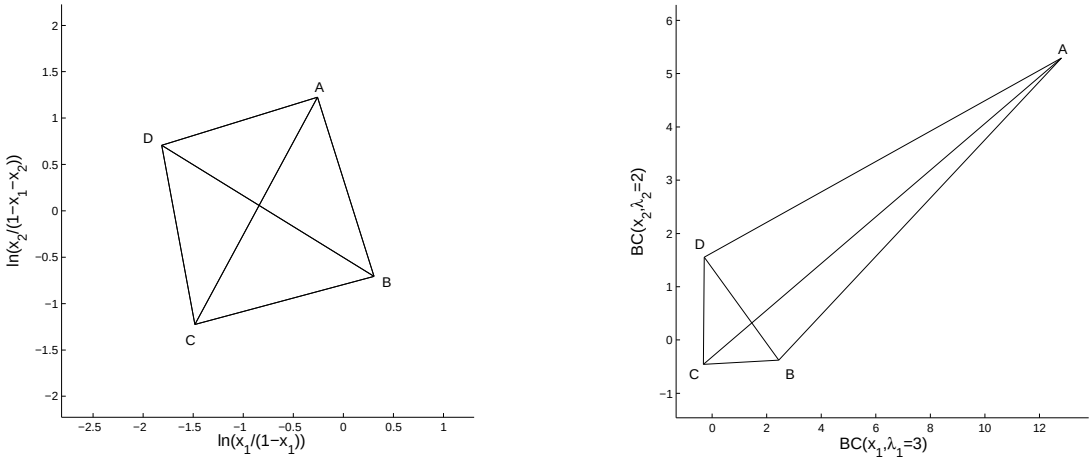


(a) (b)

Figura 2.6: (a) *Quadrat resultant després d'aplicar la transformació* ilr. (b) *Quadrat resultant després d'aplicar la transformació* clr.

2.4 Composicions aleatòries

En aquest apartat estudiarem com aplicar la teoria general de la probabilitat a les composicions aleatòries, és a dir, a vectors aleatoris $\mathcal{S}^D$-valuats. Farem una breu referència a mesures sobre el símplex i a funcions de densitat però ens centrarem bàsicament en l'estudi de l'element de tendència central i de variabilitat.



(a) (b)

Figura 2.7: (a) *Figura resultant després d'aplicar la transformació* mlr. (b)*Figura resultant després d'aplicar la transformació Box-Cox amb* $\boldsymbol{\lambda} = (3, 2)'$.

Definició 2.12 Sigui $(\Omega, \mathcal{F}, p)$ un espai de probabilitat. Una funció mesurable $\mathbf{x} : \Omega \longrightarrow \mathcal{S}^D$ s'anomena *composició aleatòria*. □

Simbolitzarem les composicions aleatòries amb la mateixa notació amb què simbolitzàvem les composicions, és a dir amb lletra minúscula i negreta $\mathbf{x}, \mathbf{x}^*, \ldots$ Distingirem, però, les dades mostrals denotant-les amb lletra majúscula $\mathbf{X}, \mathbf{X}^*, \ldots$

Donada una composició aleatòria $\mathbf{x}$, voldrem definir la seva llei de probabilitat. Com a tot espai de probabilitat podem definir una llei mitjançant una funció de densitat, és a dir, mitjançant una derivada de Radon-Nikodým de la probabilitat respecte d'una mesura.

Donat que $\mathcal{S}^D \subset \mathbb{R}^D$, podem considerar a $\mathcal{S}^D$ la mateixa estructura algebraica de l'espai real i en particular, podem prendre la mesura de Lebesgue. En aquest cas podem definir lleis de probabilitat utilitzant els procediments clàssics habituals. Per una banda podem definir les lleis directament a $\mathcal{S}^D$ amb la densitat respecte de la mesura de Lebesgue. L'altra possibilitat, utilitzada molt sovint a la pràctica, és recórrer a les transformacions. Es tracta de transformar la composició aleatòria a l'espai real multivariant, definir la funció de densitat respecte de la mesura de Lebesgue per al vector transformat i tornar a $\mathcal{S}^D$ mitjançant la transformació inversa. Aitchison i Shen (1980), Aitchison (1982, 1986), Barceló-Vidal (1996) i Mateu-Figueras et al. (1998) introdueixen lleis de probabilitat utilitzant aquesta metodologia. En aquests casos, utilitzarem la integració estàndard de l'espai real per fer efectiu el càlcul de

probabilitats o de qualsevol moment.

En l'apartat 2.1 hem vist que $\mathcal{S}^D$ té una estructura algebraica pròpia i diferent a la dels reals. En aquest context definirem les lleis de probabilitat mitjançant la derivada de Radon-Nikodým respecte d'una mesura adequada a $\mathcal{S}^D$ i diferent a la mesura de Lebesgue. No obstant això, tal i com hem indicat al capítol 1, ens sorgiran certs problemes de càlcul ja que caldrà integrar respecte d'una mesura diferent a la de Lebesgue. Com que $\mathcal{S}^D$ té estructura d'espai vectorial euclidià, solventarem aquests problemes treballant amb les coordenades de la composició aleatòria respecte d'una base ortonormal de $\mathcal{S}^D$. Sobre aquestes coordenades podrem aplicar tota l'anàlisi real estàndard; en particular, podrem definir la densitat dels coeficients respecte de la mesura de Lebesgue.

Observem que ens trobem davant tres opcions diferents. Les dues primeres basades en "sortir"del símplex, ja sigui perquè es considera tot l'espai $\mathbb{R}^D$ com a espai suport o bé perquè apliquem transformacions del símplex a l'espai real. Estudiarem amb detall aquestes opcions al capítol 3 d'aquesta tesi doctoral. La tercera opció està basada en "romandre"a $\mathcal{S}^D$ i treballar amb coordenades respecte d'una base ortonormal. Trobarem, en el capítol 4, un estudi detallat de diferents lleis de probabilitat definides segons aquesta metodologia. En qualsevol dels casos, a banda de definir diferents lleis de probabilitat, calcularem l'element de tendència central i l'element de dispersió en coherència amb la metodologia escollida.

Abans de centrar el nostre estudi en les diferents lleis de probabilitat, considerem certs aspectes generals referents als elements de tendència central i de variabilitat de les composicions aleatòries.

Si bé la definició 1.7 ens dóna l'expressió general de l'esperança d'una variable aleatòria qualsevol, la interpretació geomètrica d'aquest concepte a l'espai real ens diu que l'esperança és el valor $\mathrm{E}[\mathbf{x}]$ que minimitza l'expressió $\mathrm{E}[d_{eu}(\mathbf{x}, \mathrm{E}[\mathbf{x}])]$. Aitchison (2002) utilitza aquesta interpretació per introduir el centre d'una composició aleatòria. Tan sols cal substituir la distància euclidiana per la distància d'Aitchison (2.3).

Definició 2.13 El *centre* d'una composició aleatòria $\mathbf{x}$ és la composició $\mathrm{cen}[\mathbf{x}] \in \mathcal{S}^D$ que minimitza $\mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}[\mathbf{x}])]$. □

L'expressió d'aquest centre resulta ser $\mathrm{cen}[\mathbf{x}] = \mathcal{C}(\exp(\mathrm{E}[\ln \mathbf{x}]))$ o equivalentment

$$\mathrm{cen}[\mathbf{x}] = \mathcal{C}\left(\exp\left(\mathrm{E}\left[\ln \frac{\mathbf{x}}{g(\mathbf{x})}\right]\right)\right), \tag{2.24}$$

on $g(\mathbf{x})$ és la mitjana geomètrica de les parts de $\mathbf{x}$. Obtenim el mateix resultat interpretant el centre d'una composició aleatòria com el valor que minimitza la mesura de divergència de Kullback-Leibler (Aitchison, 1997). La igualtat (2.24) és enunciada per Aitchison (1997) però és demostrada per Pawlowsky-Glahn i Egozcue (2002), els quals demostren també la igualtat

$$\operatorname{alr}(\operatorname{cen}[\mathbf{x}]) = \operatorname{E}[\operatorname{alr}(\mathbf{x})]. \tag{2.25}$$

A partir de (2.25) i utilitzant les relacions (2.21), és immediat demostrar que

$$\operatorname{clr}(\operatorname{cen}[\mathbf{x}]) = \operatorname{E}[\operatorname{clr}(\mathbf{x})], \tag{2.26}$$

$$\operatorname{ilr}(\operatorname{cen}[\mathbf{x}]) = \operatorname{E}[\operatorname{ilr}(\mathbf{x})]. \tag{2.27}$$

Això indica que l'esperança dels vectors $\operatorname{alr}(\mathbf{x}), \operatorname{clr}(\mathbf{x})$ i $\operatorname{ilr}(\mathbf{x})$ coincideix amb la corresponent transformació de la composició $\operatorname{cen}[\mathbf{x}]$. Observem, però, que podem calcular fàcilment les esperances $\operatorname{E}[\operatorname{alr}(\mathbf{x})], \operatorname{E}[\operatorname{clr}(\mathbf{x})]$ i $\operatorname{E}[\operatorname{ilr}(\mathbf{x})]$ aplicant la definició clàssica d'esperança d'un vector aleatori real. En certs casos i per alleugerir la notació ens referirem a aquests vectors d'esperances com $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_{D-1})', \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_D)'$ i $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{D-1})'$ respectivament. Amb les relacions (2.21) entre les tres transformacions i utilitzant les propietats de l'esperança d'un vector aleatori, és immediat demostrar que

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{F}\boldsymbol{\lambda} = \mathbf{F}\mathbf{U}\boldsymbol{\xi}, \\ \boldsymbol{\lambda} &= \mathbf{F}^*\boldsymbol{\mu} = \mathbf{U}\boldsymbol{\xi}, \\ \boldsymbol{\xi} &= \mathbf{U}'\mathbf{F}^*\boldsymbol{\mu} = \mathbf{U}'\boldsymbol{\lambda}. \end{aligned} \tag{2.28}$$

En l'apartat anterior hem vist que també podem interpretar els vectors $\operatorname{alr}(\mathbf{x}), \operatorname{clr}(\mathbf{x})$ i $\operatorname{ilr}(\mathbf{x})$ com les coordenades de la composició aleatòria $\mathbf{x}$ respecte de la base B, el sistema de generadors B^* i una base ortonormal de $\mathcal{S}^D$ respectivament. De la mateixa manera, podem interpretar els vectors $\boldsymbol{\mu}, \boldsymbol{\lambda}$ i $\boldsymbol{\xi}$ com les coordenades de la composició $\operatorname{cen}[\mathbf{x}]$ respecte de la base B, el sistema de generadors B^* i una base ortonormal de $\mathcal{S}^D$, respectivament. Podem recuperar la composició $\operatorname{cen}[\mathbf{x}]$ amb una simple combinació lineal. Certament, utilitzant per exemple les coordenades $\boldsymbol{\lambda} = \operatorname{clr}(\operatorname{cen}[\mathbf{x}])$ obtenim:

$$
\begin{aligned}
&(\lambda_1 \otimes \mathbf{w}_1) \oplus \cdots \oplus (\lambda_D \otimes \mathbf{w}_D) \\
&= \left(\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_1}{g(\mathbf{x})}\right)\right] \otimes \mathbf{w}_1\right) \oplus \cdots \oplus \left(\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_D}{g(\mathbf{x})}\right)\right] \otimes \mathbf{w}_D\right) \\
&= \left(\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_1}{g(\mathbf{x})}\right)\right] \otimes (e, 1, \ldots, 1)'\right) \oplus \cdots \oplus \left(\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_D}{g(\mathbf{x})}\right)\right] \otimes (1, \ldots, 1, e)'\right) \\
&= \left(e^{\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_1}{g(\mathbf{x})}\right)\right]}, 1, \ldots, 1\right)' \oplus \cdots \oplus \left(1, \ldots, 1, e^{\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_D}{g(\mathbf{x})}\right)\right]}\right)' \\
&= \mathcal{C}\left(e^{\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_1}{g(\mathbf{x})}\right)\right]}, \ldots, e^{\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_D}{g(\mathbf{x})}\right)\right]}\right)' = \mathcal{C}\left(e^{\mathrm{E}\left[\ln\left(\frac{\mathrm{x}}{g(\mathbf{x})}\right)\right]}\right) = \mathrm{cen}[\mathbf{x}].
\end{aligned}
$$

De manera similar obtenim que

$$
\begin{aligned}
\mathrm{cen}[\mathbf{x}] &= (\xi_1 \otimes \mathbf{e}_1) \oplus \cdots \oplus (\xi_{D-1} \otimes \mathbf{e}_{D-1}) \\
&= (\mathrm{E}[\langle \mathbf{x}, \mathbf{e}_1 \rangle_a] \otimes \mathbf{e}_1) \oplus (\mathrm{E}[\langle \mathbf{x}, \mathbf{e}_2 \rangle_a] \otimes \mathbf{e}_2) \oplus \ldots \oplus (\mathrm{E}[\langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a] \otimes \mathbf{e}_{D-1}), \qquad (2.29) \\
\mathrm{cen}[\mathbf{x}] &= (\mu_1 \otimes \mathbf{w}_1) \oplus \cdots \oplus (\mu_{D-1} \otimes \mathbf{w}_{D-1}) \\
&= \left(\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_1}{\mathrm{x}_D}\right)\right] \otimes \mathbf{w}_1\right) \oplus \cdots \oplus \left(\mathrm{E}\left[\ln\left(\frac{\mathrm{x}_{D-1}}{\mathrm{x}_D}\right)\right] \otimes \mathbf{w}_{D-1}\right).
\end{aligned}
$$

Aquests resultats ens tornen a confirmar que treballar directament amb les composicions i l'estructura de $\mathcal{S}^D$ és equivalent a treballar amb les seves coordenades respecte d'una base i considerant l'estructura de l'espai real. A més, en el cas concret del centre d'una composició, podem treballar amb les coordenades alr, clr o ilr indistintament.

Pel que fa a les propietats, sabem que l'esperança d'un vector aleatori definit a l'espai real compleix que $\mathrm{E}[\mathbf{p} + \mathbf{x}] = \mathbf{p} + \mathrm{E}[\mathbf{x}]$ per a qualsevol vector de constants $\mathbf{p}$; $\mathrm{E}[\alpha\mathbf{x}] = \alpha\mathrm{E}[\mathbf{x}]$ per a qualsevol constant $\alpha \in \mathbb{R}$, i $\mathrm{E}[\mathbf{x} + \mathbf{y}] = \mathrm{E}[\mathbf{x}] + \mathrm{E}[\mathbf{y}]$ per a qualsevol parell de vectors aleatoris $\mathbf{x}$ i $\mathbf{y}$. Aitchison (2002) enuncia unes propietats anàlogues per al centre d'una composició aleatòria i Pawlowsky-Glahn i Egozcue (2002) en donen la demostració detallada.

Propietat 2.4 Siguin $\mathbf{x}$ i $\mathbf{x}^*$ dues composicions aleatòries, sigui $\mathbf{p} \in \mathcal{S}^D$ un vector de constants i $\alpha \in \mathbb{R}$ una constant qualsevol. Llavors es compleix que

a. $\mathrm{cen}[\mathbf{p} \oplus \mathbf{x}] = \mathbf{p} \oplus \mathrm{cen}[\mathbf{x}]$.

b. $\mathrm{cen}[\alpha \otimes \mathbf{x}] = \alpha\mathrm{cen}[\mathbf{x}]$.

c. $\mathrm{cen}[\mathbf{x} \oplus \mathbf{x}^*] = \mathrm{cen}[\mathbf{x}] \oplus \mathrm{cen}[\mathbf{x}^*]$.

□

Quan a l'espai real volem estimar l'esperança d'un vector aleatori a partir d'una mostra, calculem la mitjana aritmètica. Aitchison (1997) s'adona que aquest valor tan àmpliament utilitzat no és compatible amb les operacions definides al símplex i proposa la mitjana geomètrica com a element representatiu del centre d'un conjunt de dades composicionals. Així doncs, donada una mostra $\mathbf{X}$, l'estimació puntual del centre de la composició aleatòria serà la mitjana geomètrica composicional, és a dir $g(\mathbf{X}) = \mathcal{C}(g_1, g_2, \ldots, g_D)'$, on g_i és la mitjana geomètrica de la component i-èsima. Martín-Fernández (2001) demostra que $g(\mathbf{X})$ és un element compatible amb les operacions bàsiques del símplex, és a dir, $g(\mathbf{p} \oplus \mathbf{X}) = \mathbf{p} \oplus g(\mathbf{X})$, i $g(\alpha \otimes \mathbf{X}) = \alpha \otimes g(\mathbf{X})$, per a qualsevol $\mathbf{p} \in \mathcal{S}^D$ i $\alpha \in \mathbb{R}$. Pawlowsky-Glahn i Egozcue (2002) demostren, a més, que dins l'estructura considerada a $\mathcal{S}^D$ la mitjana geomètrica composicional és el millor estimador composicional lineal i no esbiaixat del centre d'una composició aleatòria.

Per altra banda, mitjançant uns simples càlculs, obtenim les igualtats:

$$\begin{aligned} \operatorname{ilr}(g(\mathbf{X})) &= \overline{\operatorname{alr}(\mathbf{X})}, \\ \operatorname{clr}(g(\mathbf{X})) &= \overline{\operatorname{clr}(\mathbf{X})}, \\ \operatorname{alr}(g(\mathbf{X})) &= \overline{\operatorname{ilr}(\mathbf{X})}, \end{aligned} \tag{2.30}$$

on $\overline{\operatorname{alr}(\mathbf{X})}$, $\overline{\operatorname{clr}(\mathbf{X})}$ i $\overline{\operatorname{ilr}(\mathbf{X})}$ representen les mitjanes aritmètiques de la mostra $\mathbf{X}$ transformada amb les transformacions alr, clr i ilr respectivament. Si ho interpretem en termes de components, les igualtats anteriors indiquen que les coordenades de la composició $g(\mathbf{X})$ coincideixen amb la mitjana aritmètica de les coordenades de la mostra $\mathbf{X}$ en les respectives bases o sistema de referència.

La relació (2.30) fou demostrada per Martín-Fernández (2001). Aquesta propietat i el fet que la mitjana geomètrica composicional sigui compatible amb les operacions pertorbació i potència permeten definir una transformació a l'espai $\mathcal{S}^D$ per centrar un conjunt de dades composicionals (vegeu Martín-Fernández et al., 1999) . Donat un conjunt de dades composicionals $\mathbf{X}$, l'operació *centrar* consisteix en pertorbar les dades amb la composició $g(\mathbf{X})^{-1}$, és a dir, calcular $\mathbf{X}^* = g(\mathbf{X})^{-1} \oplus \mathbf{X}$. El resultat és un conjunt de dades, $\mathbf{X}^*$, la mitjana geomètrica del qual és el centre del símplex $(1/D, 1/D, \ldots, 1/D)'$. Aquesta tranformació s'ha utilitzat en diversos casos pràctics (vegeu Buccianti et al.,1999, i von Eynatten et al., 2002) i resulta ser útil per veure més clarament el patró de variabilitat d'un conjunt de dades composicionals.

La variància d'una variable aleatòria a l'espai real es pot interpretar com l'esperança de la distància euclidiana al quadrat entre la variable i la seva esperança. Per aquesta raó Pawlowsky-Glahn i Egozcue (2002) suggereixen definir la variància d'una composició aleatòria com l'esperança de la distància d'Aitchison al quadrat entre la composició aleatòria i el seu centre, és a dir:

Definició 2.14 La *variància mètrica* al voltant del centre cen[$\mathbf{x}$] de la composició aleatòria $\mathbf{x}$ es defineix com

$$\text{Mvar}[\mathbf{x}] = \text{E}[d_a^2(\mathbf{x}, \text{cen}[\mathbf{x}])].$$

□

Les propietats (2.17) i (2.20) ens proporcionen dues expressions equivalents de la variància mètrica:

$$\text{Mvar}[\mathbf{x}] = \text{E}[d_{eu}^2(\text{ilr}(\mathbf{x}), \text{ilr}(\text{E}[\mathbf{x}]))] = \text{E}[d_{eu}^2(\text{clr}(\mathbf{x}), \text{clr}(\text{E}[\mathbf{x}]))]. \tag{2.31}$$

En aquest cas no serà correcte utilitzar la distància euclidiana entre els respectius vectors alr ja que la transformació alr no conserva les distàncies, o equivalentment, perquè la base B no és ortonormal.

Quan es realitza una anàlisi de components principals o es treballa amb biplots a l'espai real, s'utilitza la traça de la matriu de covariàncies del vector aleatori com una mesura de la variabilitat total. En el símplex, sabem que no té sentit treballar amb la matriu de covariàncies directes d'una composició aleatòria. Per ser coherents amb el principi d'invariància per canvis d'escala, caldria treballar amb les covariàncies entre els quocients de les parts. No obstant això, Aitchison (2002) adverteix que aquestes covariàncies són matemàticament intractables i suggereix treballar amb les covariàncies entre els logaritmes dels quocients. Per aquesta raó, defineix la variabilitat total com la traça de la matriu de covariàncies del vector $\mathbf{z} = \text{clr}(\mathbf{x})$, és a dir:

Definició 2.15 Donada una composició aleatòria $\mathbf{x}$, es defineix la *variabilitat total* com

$$\text{totvar}(\mathbf{x}) = \text{traça}(\mathbf{\Gamma}) = \frac{1}{D}\sum_{i<j} \text{var}\left[\ln\frac{x_i}{x_j}\right],$$

on $\mathbf{\Gamma} = [\gamma_{ij}] = \left[\text{cov}\left(\ln\left(\frac{\text{x}_i}{g(\mathbf{x})}\right), \ln\left(\frac{\text{x}_i}{g(\mathbf{x})}\right)\right) : i, j = 1, \ldots, D\right]$. □

La matriu $\mathbf{\Gamma}$ definida per Aitchison (1983) és una matriu d'ordre $D \times D$, simètrica i singular donat que la suma dels elements de cada fila és igual a 0.

Posteriorment Pawlowsky-Glahn i Egozcue (2002) demostren l'equivalència de les dues definicions de variància ja que obtenen la igualtat

$$\text{Mvar}[\mathbf{x}] = \text{totvar}[\mathbf{x}],$$

si bé la definició 2.14 posa de manifest la relació amb la mètrica de l'espai.

La igualtat (2.31) ens indica que podríem també definir la variabilitat total com la traça de la matriu $\mathbf{\Upsilon}$, amb

$$\mathbf{\Upsilon} = [\upsilon_{ij}] = \Big[\text{cov}\left(\langle \mathbf{x}, \mathbf{e}_i\rangle_a, \langle \mathbf{x}, \mathbf{e}_j\rangle_a\right) : i, j = 1, \ldots, D-1\Big],$$

on $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ és una base ortonormal qualsevol de $\mathcal{S}^D$. Les matrius $\mathbf{\Gamma}$ i $\mathbf{\Upsilon}$ són completament diferents. $\mathbf{\Upsilon}$ és d'ordre $(D-1) \times (D-1)$, simètrica, definida positiva i es pot interpretar com la matriu de covariàncies del vector $\mathbf{v} = \text{ilr}(\mathbf{x})$. Tot i aquestes diferències, es compleix que traça($\mathbf{\Upsilon}$) = traça($\mathbf{\Gamma}$).

Aitchison (2002) observa i posteriorment Pawlowsky-Glahn i Egozcue (2002) demostren que la mesura de dispersió Mvar o totvar compleix les mateixes propietats que la variància clàssica de l'espai real, és a dir, és una element coherent amb les operacions bàsiques definides a l'espai $\mathcal{S}^D$.

Propietat 2.5 Siguin $\mathbf{x}$ i $\mathbf{x}^*$ dues composicions aleatòries independents, $\mathbf{p} \in \mathcal{S}^D$ qualsevol composició constant, i $\alpha \in \mathbb{R}$ qualsevol constant. Llavors es compleix que

a. $\text{Mvar}[\mathbf{p} \oplus \mathbf{x}] = \text{Mvar}[\mathbf{x}]$.

b. $\text{Mvar}[\alpha \otimes \mathbf{x}] = \alpha^2 \text{Mvar}[\mathbf{x}]$.

c. $\text{Mvar}[\mathbf{x} \oplus \mathbf{x}^*] = \text{Mvar}[\mathbf{x}] + \text{Mvar}[\mathbf{x}^*]$.

□

Observem que de la propietat 2.5a es dedueix que l'operació de centrar no altera la variabilitat del conjunt de les dades.

Aitchison (1982) defineix també la matriu de covariàncies del vector $\mathbf{y} = \text{alr}(\mathbf{x})$. Tot i que aquesta matriu no ens serveix per calcular directament la variància mètrica d'una composició aleatòria, tindrà un paper important quan parlem de les famílies de distribucions a $\mathcal{S}^D$.

Definició 2.16 Donada una composició aleatòria $\mathbf{x}$ amb D parts, anomenem *matriu de covariàncies de logquocients* a la matriu

$$\mathbf{\Sigma} = [\sigma_{ij}] = \big[\mathrm{cov}\left(\ln(\mathrm{x}_i/\mathrm{x}_D), \ln(\mathrm{x}_j/\mathrm{x}_D)\right) : i, j = 1, \ldots, D-1\big].$$

□

Observem que $\mathbf{\Sigma}$ és una matriu d'ordre $(D-1) \times (D-1)$, simètrica i definida positiva. Aitchison (1986) dóna també la relació entre $\mathbf{\Sigma}$ i $\mathbf{\Gamma}$:

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{F}\mathbf{\Gamma}\mathbf{F}', \\ \mathbf{\Gamma} &= \mathbf{F}'\mathbf{H}^{-1}\mathbf{\Sigma}\mathbf{H}^{-1}\mathbf{F} = \mathbf{F}^*\mathbf{\Sigma}\mathbf{F}^{*'}, \end{aligned} \tag{2.32}$$

amb la matriu $\mathbf{F}^*$ definida a l'apartat 2.3.2 i les matrius $\mathbf{F}$ i $\mathbf{H}$ definides a la taula 2.1.

A partir de les relacions (2.21) entre les transformacions alr, clr i ilr i utilitzant les propietats de la matriu de covariàncies, podem demostrar que

$$\begin{aligned} \mathbf{\Upsilon} &= (\mathbf{U}'\mathbf{F}^*)\mathbf{\Sigma}(\mathbf{U}'\mathbf{F}^*)', \\ \mathbf{\Upsilon} &= \mathbf{U}'\mathbf{\Gamma}\mathbf{U}, \qquad (2.33) \\ \mathbf{\Sigma} &= (\mathbf{F}\mathbf{U})\mathbf{\Upsilon}(\mathbf{F}\mathbf{U})', \qquad (2.34) \\ \mathbf{\Gamma} &= \mathbf{U}\mathbf{\Upsilon}\mathbf{U}'. \end{aligned}$$

Mitjançant la relació (2.33) és gairebé immediat demostrar que traça($\mathbf{\Upsilon}$) = traça($\mathbf{\Gamma}$). Sabem que traça($\mathbf{AB}$) = traça($\mathbf{BA}$) sempre i quan les matrius $\mathbf{A}$ i $\mathbf{B}$ siguin multiplicables, per tant traça($\mathbf{\Upsilon}$) = traça($\mathbf{U}'\mathbf{\Gamma}\mathbf{U}$) = traça($\mathbf{\Gamma}\mathbf{U}\mathbf{U}'$). El producte $\mathbf{U}\mathbf{U}'$ és una matriu amb subespai propi V. Per tant obtenim $\mathbf{\Gamma}\mathbf{U}\mathbf{U}' = \mathbf{\Gamma}$ i òbviament traça($\mathbf{\Upsilon}$) = traça($\mathbf{\Gamma}$). Aitchison (1986) demostra que les traces de $\mathbf{\Gamma}$ i $\mathbf{\Sigma}$ no són iguals. Per tant tampoc ho són les traces de $\mathbf{\Upsilon}$ i $\mathbf{\Sigma}$.

No hem fet referència a la matriu de covariàncies que s'obté directament a partir de les parts d'una composició aleatòria perquè no és adequada. Pearson (1897) ja va advertir de la impossibilitat d'interpretar correctament les covariàncies i els coeficients de correlació entre les parts d'una composició aleatòria. En particular, si calculem la matriu de covariàncies habitual $\mathbf{K} = \{\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j) : i, j = 1, 2, \ldots, D\}$, observarem que

$$\sum_{j=1}^{D} \mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j) = 0 \qquad i = 1, 2, \ldots, D$$

a causa de la restricció $\sum_{i=1}^{D} \mathrm{x}_i = 1$. Sabem que $\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_i) > 0$, excepte en la situació trivial que la part x_i sigui una constant. Aquest fet provoca que necessàriament una de les altres covariàncies tingui signe negatiu, i per tant invalida la interpretació habitual de les covariàncies ja que no poden adquirir lliurement valors nuls, positius o negatius.

Capítol 3

Models paramètrics sobre $\mathcal{S}^D$ i $\mathbb{R}_+^D$. Metodologia MOVE

Dediquem aquest capítol a l'estudi de diverses lleis de probabilitat per a composicions aleatòries, és a dir, per a funcions mesurables amb imatge l'espai $\mathcal{S}^D$. De cada llei veiem la seva funció de densitat i les propietats algebraiques més importants. Analitzem també certs aspectes d'estimació de paràmetres i valorem a nivell intuïtiu l'ajust del model a unes dades. Reservem per al següent capítol el desenvolupament i l'aplicació de proves de bondat d'ajust per validar el model.

Tal i com hem indicat en els capítols anteriors, podem realitzar aquest estudi des de dues perspectives complementàries: entenent el símplex com un subconjunt de l'espai real de dimensió D, o bé considerant el símplex com espai euclidià en ell mateix. En aquest capítol recollim les lleis de probabilitat sobre $\mathcal{S}^D$ que s'obtenen utilitzant la primera perspectiva. En aquest cas, es defineixen les lleis de probabilitat aplicant principalment tècniques basades en transformacions, raó per la qual ens hi referim mitjançant l'abreviació MOVE (de l'anglès moure, transferir).

En el primer apartat, recollim certs aspectes generals que afecten al conjunt de lleis de probabilitat definides segons la metodologia MOVE. A continuació, recordem el model normal logístic additiu (aln) introduït per Aitchison (1982, 1986). Per definir aquesta llei s'utilitza la transformació logquocient additiva i el model normal multivariant de l'espai real. Veiem les seves propietats algebraiques així com la relació amb els models obtinguts mitjançant

les transformacions logquocient centrada i logquocient isomètrica. Els principals inconvenients del model aln són: la impossibilitat de modelitzar conjunts de dades composicionals la transformació de les quals presenta biaix, i la impossibilitat per descriure la distribució de l'amalgama d'una composició aleatòria. Per aquesta raó, realitzem una anàlisi empírica de la distribució de les amalgames. Es tracta d'un estudi original on es proposa una solució aproximada utilitzant la distribució normal asimètrica de l'espai real introduïda al capítol 1.

En el tercer apartat, introduïm el model normal asimètric logístic additiu (alsn). Aquest model es basa en la transformació logquocient additiva i en la distribució normal asimètrica de l'espai real multivariant (Azzalini i Dalla-Valle, 1996). El model alsn és una generalització natural del model aln, especialment indicada per modelitzar conjunts de dades quan la seva transformació a l'espai real presenta un biaix moderat. Consegüentment, aquesta família ens aporta la solució al primer dels inconvenients de la família aln. Estudiant amb més detall aquesta nova classe de distribucions, comprovem que presenta unes bones propietats algebraiques, molt similars a les del model normal logístic additiu.

Donada una distribució a $\mathcal{S}^D$, podem identificar la llei del corresponent vector aleatori de l'espai $\mathbb{R}_+^D$. En el cas del model normal asimètric logístic additiu, és necessari definir una nova distribució a $\mathbb{R}_+^D$ que anomenem distribució lognormal asimètrica. En l'apartat 3.4 d'aquest capítol introduïm aquesta distribució i donem l'expressió de la seva funció de densitat i dels seus principals moments.

Reservem un apartat per analitzar la coneguda classe de distribucions de Dirichlet amb algunes de les seves generalitzacions. Acabem el capítol amb una secció que hem anomenat "Altres distribucions"on veiem breument les famílies definides a partir de la transformació logquocient multiplicativa i de les transformacions Box-Cox multivariants (Barceló-Vidal, 1996) així com les distribucions d'Aitchison (Aitchison, 1986).

3.1 Aspectes generals

Sobre un subconjunt de l'espai real, podem introduir una llei de probabilitat de dues maneres diferents. En primer lloc, podem definir directament una llei de probabilitat mitjançant la funció de densitat de probabilitat. En segon lloc, podem utilitzar les transformacions, és a dir, aplicar a la composició aleatòria una transformació de manera que el vector resultant

prengui valors a tot l'espai real. Amb la densitat de probabilitat del vector respecte de la mesura de Lebesgue i amb el teorema del canvi de variable, obtenim finalment la densitat de la composició aleatòria. En aquests casos trobem el jacobià del canvi de variable en l'expressió de la funció de densitat.

Cal tenir en compte que $\mathcal{S}^D$ és un subconjunt d'un hiperplà de $\mathbb{R}^D$ de dimensió $D-1$. Per tant, calcularem la probabilitat d'un esdeveniment qualsevol integrant la funció de densitat respecte de la mesura de Lebesgue producte de $\mathbb{R}^{D-1}$. És a dir, per a tot esdeveniment A de $\mathcal{S}^D$, obtindrem la seva probabilitat mitjançant l'expressió

$$P(A) = \int_A f_{\mathbf{x}}(x_1, x_2, \ldots, x_{D-1}, 1 - x_1 - x_2 - \cdots - x_{D-1}) dx_1 dx_2 \cdots dx_{D-1},$$

on $f_{\mathbf{x}}$ representa la funció de densitat de probabilitat de la composició aleatòria $\mathbf{x}$.

En tots els models que estudiarem, veurem que la funció de densitat no complirà la propietat $f_{\mathbf{x}}(\mathbf{x}) = f_{\mathbf{a}+\mathbf{x}}(\mathbf{a}+\mathbf{x})$ on $f_{\mathbf{a}+\mathbf{x}}$ indica la funció de densitat de la composició aleatòria traslladada. Si analitzem el comportament de la funció de densitat respecte de les operacions pertorbació i potència que hem definit sobre $\mathcal{S}^D$ al capítol 2, veurem que si bé certes famílies seran tancades per les operacions $\oplus$ i $\otimes$ de $\mathcal{S}^D$, en cap cas es complirà la propietat

$$f_{\mathbf{x}}(\mathbf{x}) = f_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x}), \tag{3.1}$$

on ara $f_{\mathbf{a}\oplus\mathbf{x}}$ representa la funció de densitat de la composició aleatòria $\mathbf{a} \oplus \mathbf{x}$.

Pel que fa als principals moments, veurem que serà possible calcular el vector d'esperances i la matriu de covariàncies mitjançant els procediments habituals. Cal tenir en compte que aquests elements no seran directament comparables amb el centre i la variància mètrica d'una composició aleatòria ja que aquests últims s'han definit considerant l'estructura d'espai vectorial pròpia de $\mathcal{S}^D$.

3.2 Distribució normal logística additiva (aln)

3.2.1 Definició i propietats

Aitchison i Shen (1980) defineixen la distribució normal logística additiva, anomenada aln (de l'anglès *additive logistic normal*).

Definició 3.1 Una composició aleatòria $\mathbf{x}$ amb D parts té una distribució *normal logística additiva* (aln) si el vector aleatori transformat $\mathbf{y} = \text{alr}(\mathbf{x}) = \ln(\mathbf{x}_{-D}/\mathrm{x}_D)$ té una distribució $\mathcal{N}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. □

Per denotar una distribució d'aquesta classe utilitzarem la notació $\mathbf{x} \sim \mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ o més breument $\mathbf{x} \sim \mathcal{L}^D$.

Propietat 3.1 La funció de densitat d'una composició aleatòria $\mathbf{x}$ que es distribueix segons una llei $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ és

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \; & (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \left(\prod_{i=1}^D x_i\right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2}\left(\ln(\mathbf{x}_{-D}/x_D) - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\ln(\mathbf{x}_{-D}/x_D) - \boldsymbol{\mu}\right)\right] \quad (\mathbf{x} \in \mathcal{S}^D), \end{aligned}$$

o, de manera abreviada,

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \; & (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \left(\prod_{i=1}^D x_i\right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2}(\text{alr}(\mathbf{x}) - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\text{alr}(\mathbf{x}) - \boldsymbol{\mu})\right] \quad (\mathbf{x} \in \mathcal{S}^D). \end{aligned}$$

□

L'expressió d'aquesta funció de densitat sorgeix d'aplicar el teorema del canvi de variable. Certament, la definició 3.1 exigeix al vector $\text{alr}(\mathbf{x})$ una distribució normal multivariant. Observem que el vector $\boldsymbol{\mu}$ i la matriu $\boldsymbol{\Sigma}$ són el vector d'esperances i la matriu de covariàncies del vector aleatori $\text{alr}(\mathbf{x})$, introduïts també al capítol anterior. Per obtenir la densitat de la composició $\mathbf{x}$, cal multiplicar la densitat del vector $\text{alr}(\mathbf{x})$ pel terme $\left(\prod_{i=1}^D x_i\right)^{-1}$, el jacobià de la transformació introduïda a la secció 2.3.1. La densitat del model aln és la derivada de Radon-Nikodým de la probabilitat respecte a la mesura de Lebesgue. Si volem obtenir la probabilitat d'un esdeveniment qualsevol, procedirem a calcular la integral ordinària d'aquesta funció de densitat.

Donat que es tracta d'una densitat clàssica, podem calcular l'esperança de la composició aleatòria reproduint la metodologia de l'espai real, és a dir, podem obtenir l'esperança de cada component utilitzant l'expressió (1.7) (vegeu Barceló-Vidal, 1996). Aitchison (1986) observa, a més, que existeixen tots els moments d'ordre positiu, $\mathrm{E}[\prod_{i=1}^D \mathrm{x}_i^{a_i}]$ $(a_i > 0 : i = 1, 2, \ldots, D)$, però les expressions integrals no es poden reduir a una expressió simple. Tot i això, donada una distribució específica i per a valors moderats de D, podem calcular qualsevol moment utilitzant la integració hermítica. Els resultats que s'obtenen aplicant els procediments

estàndards no coincideixen amb l'expressió del centre d'una composició aleatòria (2.24) o equivalentment amb la igualtat (2.25). Certament, si fem cas d'aquesta última expressió obtindríem que $\text{cen}[\mathbf{x}] = \text{alr}^{-1}(\mu_1, \mu_2, \ldots, \mu_D)'$ ja que $\text{E}\left[\ln\left(\frac{\text{x}_i}{\text{x}_D}\right)\right] = \mu_i$.

El gran avantatge que ofereix la classe de distribucions aln és la possibilitat d'utilitzar tots els procediments estadístics que es basen en la normalitat multivariant. Precisament són les propietats de la distribució normal les que ens permeten demostrar que la família de distribucions aln és tancada per les operacions pertorbació i potència, per la formació de subcomposicions i per la permutació de les components. Trobem la demostració d'aquestes propietats a Aitchison (1986).

Propietat 3.2 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sigui $\mathbf{a} \in \mathcal{S}^D$ una composició constant i b una constant de $\mathbb{R}$. Llavors la composició $\mathbf{x}^* = \mathbf{a} \oplus (b \otimes \mathbf{x})$ es distribueix segons una llei

$$\mathcal{L}^D(\ln(\mathbf{a}_{-D}/a_D) + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma}).$$

□

Propietat 3.3 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sigui $\mathbf{u}$ un vector de $\mathbb{R}_+^D$ amb D components positives amb distribució independent de la composició $\mathbf{x}$. Llavors la composició $\mathbf{x}^* = \mathbf{u} \oplus \mathbf{x}$ es distribueix segons un model:

a. $\mathcal{L}^D(\boldsymbol{\mu} + \ln(\mathbf{u}_{-D}/\text{u}_D), \boldsymbol{\Sigma})$, si $\mathbf{u}$ és un vector de constants.

b. $\mathcal{L}^D(\boldsymbol{\mu} + \boldsymbol{\theta}, \boldsymbol{\Sigma} + \mathbf{K})$, si $\mathbf{u} \sim \mathcal{L}^D(\boldsymbol{\theta}, \mathbf{K})$.

c. $\mathcal{L}^D(\boldsymbol{\mu} + \mathbf{F}_{(D-1),D}\boldsymbol{\zeta}, \boldsymbol{\Sigma} + \mathbf{F}_{(D-1),D}\boldsymbol{\Theta}\mathbf{F}'_{(D-1),D})$, si $\mathbf{u} \sim \Lambda^D(\boldsymbol{\zeta}, \boldsymbol{\Theta})$. □

La demostració d'aquestes propietats és immediata, ja que les operacions pertorbació i potència entre composicions es corresponen amb les operacions suma i producte per un escalar entre els respectius vectors alr transformats.

Per motius que quedaran clars més endavant, volem remarcar que, tot i ser una família tancada per l'operació pertorbació, la densitat normal logística additiva no compleix la igualtat (3.1); és a dir, el valor $f_{\mathbf{x}}(\mathbf{x})$ no és igual a $f_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a}\otimes\mathbf{x})$ on $f_{\mathbf{x}}$ i $f_{\mathbf{a}\oplus\mathbf{x}}$ representen les funcions de densitat de les composicions aleatòries $\mathbf{x}$ i $\mathbf{a} \otimes \mathbf{x}$, respectivament.

Propietat 3.4 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sigui $\mathbf{s} = \mathcal{C}(\mathbf{Sx})$ la subcomposició obtinguda amb $\mathbf{S}$, matriu de selecció $C \times D$. Llavors $\mathbf{s}$ es distribueix segons una llei $\mathcal{L}^C(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$, amb

$$\boldsymbol{\mu}_S = \mathbf{Q}_S\boldsymbol{\mu} \qquad \text{i} \qquad \boldsymbol{\Sigma}_S = \mathbf{Q}_S\boldsymbol{\Sigma}\mathbf{Q}'_S,$$

on

$$\mathbf{Q}_S = \mathbf{F}_{C-1,C}\mathbf{S}\mathbf{F}'_{D-1,D}\mathbf{H}^{-1}_{D-1}$$

i les matrius $\mathbf{F}$ i $\mathbf{H}$ són les definides a la taula 2.1. □

Propietat 3.5 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sigui $\mathbf{x}_P = \mathbf{Px}$ la composició $\mathbf{x}$ amb les components reordenades per la matriu permutació $\mathbf{P}$. Llavors $\mathbf{x}_P$ es distribueix segons una llei $\mathcal{L}^D(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ amb

$$\boldsymbol{\mu}_P = \mathbf{Q}_P\boldsymbol{\mu} \qquad \text{i} \qquad \boldsymbol{\Sigma}_P = \mathbf{Q}_P\boldsymbol{\Sigma}\mathbf{Q}'_P,$$

on

$$\mathbf{Q}_P = \mathbf{F}_{D-1,D}\mathbf{P}\mathbf{F}'_{D-1,D}\mathbf{H}^{-1}_{D-1}$$

i les matrius $\mathbf{F}$ i $\mathbf{H}$ són les definides a la taula 2.1. □

La següent propietat fa referència a la distribució del vector aleatori de $\mathbb{R}^D_+$ a partir del qual obtenim distribucions normals logístiques additives.

Propietat 3.6 Sigui $\mathbf{w} \in \mathbb{R}^D_+$ un vector aleatori amb distribució lognormal $\Lambda^D(\boldsymbol{\zeta}, \boldsymbol{\Theta})$. Llavors la composició associada $\mathbf{x} = \mathcal{C}(\mathbf{w})$ es distribueix segons una llei $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, on

$$\boldsymbol{\mu} = \mathbf{F}_{D-1,D}\boldsymbol{\zeta} \qquad \text{i} \qquad \boldsymbol{\Sigma} = \mathbf{F}_{D-1,D}\boldsymbol{\Theta}\mathbf{F}'_{D-1,D}.$$

□

Recordem que la distribució lognormal es defineix també mitjançant una transformació, és a dir, la metodologia utilitzada en la seva definició és totalment coherent amb la utilitzada en la definició del model normal logístic additiu. És important destacar que una distribució lognormal amb qualsevol estructura de covariància dóna lloc a una distribució normal logística additiva.

En alguns casos, ens pot interessar conèixer la distribució d'una subcomposició $\mathbf{s}_1$ donada una altra subcomposició $\mathbf{s}_2$, és a dir, conèixer la distribució de $\mathbf{s}_1|\mathbf{s}_2$. Si la composició original és de classe $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, aquest és un problema fàcil de resoldre gràcies a les propietats de la distribució normal multivariant. Cal, però, distingir dos casos: quan $\mathbf{s}_1$ i $\mathbf{s}_2$ tenen una component en comú, i quan $\mathbf{s}_1$ i $\mathbf{s}_2$ no tenen cap component en comú. No considerarem el cas on $\mathbf{s}_1$ i $\mathbf{s}_2$ tenen més d'una component en comú perquè llavors tindríem alguns logquocients fixos la variabilitat dels quals no té sentit estudiar. La següent propietat es refereix al cas on $\mathbf{s}_1$ i $\mathbf{s}_2$ tenen una única component en comú. Suposarem, sense perdre generalitat, que aquesta part és l'última de les components.

Propietat 3.7 Sigui $\mathbf{x}$ una composició aleatòria de D parts amb una distribució $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Siguin $\mathbf{s}_1 = \mathcal{C}(\mathbf{x}^{(C)}, \mathrm{x}_D) = \mathcal{C}(\mathrm{x}_1, \ldots, \mathrm{x}_C, \mathrm{x}_D)$ i $\mathbf{s}_2 = \mathcal{C}(\mathbf{x}_{(C)}) = \mathcal{C}(\mathrm{x}_{C+1}, \mathrm{x}_{C+2}, \ldots, \mathrm{x}_D)$ dues subcomposicions amb la component x_D en comú. Llavors la distribució condicional de $\mathbf{s}_1$ donada $\mathbf{s}_2$ segueix un model $\mathcal{L}^{C+1}(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{1.2})$, amb

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \qquad \text{i} \qquad \boldsymbol{\Sigma}_{1.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21},$$

on

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

són les particions d'ordre $(C-1, D-C)$ de $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$, i $\mathbf{y}_2 = \mathrm{alr}(\mathbf{s}_2)$. □

Tenim un resultat similar quan $\mathbf{s}_1$ i $\mathbf{s}_2$ no tenen cap component en comú. En aquest cas només ens cal aplicar la propietat 3.7 seguida de la propietat 3.4.

Propietat 3.8 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Siguin $\mathbf{s}_1 = \mathcal{C}(\mathbf{x}^{(C)})$ i $\mathbf{s}_2 = \mathcal{C}(\mathbf{x}_{(C)})$ les dues subcomposicions sense cap part en comú. Llavors la distribució condicional de $\mathbf{s}_1$ donada $\mathbf{s}_2$ segueix un model $\mathcal{L}^C(\mathbf{F}\boldsymbol{\mu}_{1.2}, \mathbf{F}\boldsymbol{\Sigma}_{1.2}\mathbf{F}')$, on $\mathbf{F}$ és la matriu de la taula 2.1 d'ordre $(C-2) \times (C-1)$. □

Així, podem afirmar que la família normal logística additiva és també tancada per l'operació de condicionar.

El teorema del límit central proporciona distribucions normals a l'espai real. De la mateixa manera, podem establir un teorema del límit central que proporcioni distribucions normals

logístiques additives a l'espai del símplex. No pretenem presentar el teorema d'una forma rigorosa: el nostre objectiu és mostrar com les distribucions aln es poden generar amb un procés d'aquest tipus.

Propietat 3.9 Sigui $\mathbf{x}_n$ $(n = 0, 1, 2, \ldots)$ una seqüència de composicions aleatòries generades a partir de pertorbacions independents $\mathbf{u}_n$ $(n = 1, 2, \ldots)$ i no necessàriament amb una distribució normal logística:

$$\mathbf{x}_n = \mathbf{u}_n \oplus \mathbf{x}_{n-1} \qquad (n = 1, 2, \ldots).$$

Llavors, sota certes condicions de regularitat i per a valors de n grans, $\mathbf{x}_n$ tendeix cap a una distribució normal logística additiva. □

3.2.2 Aspectes d'inferència estadística

Per estimar els paràmetres $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ d'una composició aleatòria $\mathbf{x} \sim \mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a partir dels valors d'una mostra hem d'aplicar tan sols els procediments estàndards del model normal als logquocients. Sigui $\mathbf{X} = [x_{k,r} : k = 1, 2, \ldots, n \,; r = 1, 2, \ldots, D]$ una mostra de mida n d'una composició amb D parts. Mitjançant la transformació alr, transformem $\mathbf{X}$ en una matriu $\mathbf{Y}$ d'ordre $n \times (D-1)$:

$$\mathbf{Y} = [y_{k,i} = \ln(x_{k,i}/x_{k,D}) : k = 1, 2, \ldots, n \,; i = 1, 2, \ldots, D-1].$$

Obtenim les estimacions dels paràmetres $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ amb les expressions:

$$\hat{\mu}_i = \frac{1}{n}\sum_{k=1}^{n} y_{k,i} \qquad i = 1, \ldots, D-1,$$

$$\hat{\sigma}_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n}(y_{k,i} - \hat{\mu}_i)(y_{k,j} - \hat{\mu}_j) \qquad i, j = 1, \ldots, D-1.$$

Recordem que la mitjana mostral i la matriu de covariàncies corregides són estimadors centrats i de mínima variància dels paràmetres $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$, estretament relacionats amb els estimadors que s'obtenen a partir del mètode de màxima versemblança. En determinades situacions ens interessarà treballar directament amb els estimadors de màxima versemblança del model normal, en aquests casos utilitzarem la notació $\hat{\mathbf{m}}$ i $\hat{\mathbf{S}}$ per representar-los.

Cal tenir en compte que en aplicar la transformació logquocient additiva podem escollir la component que s'utilitza com a denominador en els logquocients. Per aquesta raó, la mostra

transformada i els estimadors calculats a partir d'aquesta dependran del denominador escollit. Tot i això, en la propietat 3.5 hem vist que la distribució aln és tancada per les permutacions i que existeix una relació lineal entre els paràmetres del model aln de la composició permutada i els paràmetres del model aln de la composició original. Utilitzant aquestes relacions podem demostrar la següent propietat:

Propietat 3.10 El valor màxim de la funció de logversemblança d'una mostra d'una composició de classe aln és invariant respecte del grup de les permutacions de les components de la composició. □

En particular, es dedueix que el valor màxim de la funció de logversemblança no depèn del denominador que s'utilitzi en la transformació logquocient.

Per validar el model normal logístic additiu amb un test de bondat d'ajust, tan sols cal aplicar un test de normalitat multivariant a les dades transformades. Existeix en la literatura una gran varietat de contrastos de normalitat multivariant. Aitchison (1986) utilitza les proves de bondat d'ajust basades en la funció de distribució empírica. Ens trobem, però, amb la dificultat addicional de la dependència del denominador escollit en la transformació logquocient additiva. Per aquesta raó Aitchison et al. (2003) proposen una adaptació d'aquests contrastos per eliminar la dependència del denominador (vegeu capítol 5).

3.2.3 Altres parametritzacions

Hem vist que la distribució normal logística additiva està definida mitjançant una transformació del símplex a l'espai real. Aitchison (1986) va escollir la transformació logquocient additiva, però es pot definir la mateixa llei de probabilitat utilitzant la transformació logquocient centrada o logquocient isomètrica donat que existeix una relació matricial entre elles. Certament, és equivalent afirmar que el vector $\mathrm{alr}(\mathbf{x})$ segueix una distribució normal que afirmar que els vectors $\mathrm{clr}(\mathbf{x})$ o $\mathrm{ilr}(\mathbf{x})$ segueixen una distribució normal (degenerada en el cas del vector $\mathrm{clr}(\mathbf{x})$).

Aitchison (1986) utilitza les relacions matricials entre els paràmetres $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\lambda}$ i $\boldsymbol{\Gamma}$ i les relacions entre les transformacions alr i clr per obtenir la densitat d'una composició aleatòria $\mathbf{x} \sim \mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ en termes dels paràmetres $\boldsymbol{\lambda}$ i $\boldsymbol{\Gamma}$ i del vector $\mathrm{clr}(\mathbf{x})$ (vegeu Aitchison, 1986,

pàg. 116). L'expressió és

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) = \quad & (2\pi)^{-(D-1)/2} D^{-1/2} (\mid \mathbf{\Gamma} \mid^{+})^{-1/2} \left(\prod_{i=1}^{D} x_i \right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2} (\mathrm{clr}(\mathbf{x}) - \boldsymbol{\lambda})' \mathbf{\Gamma}^{-} (\mathrm{clr}(\mathbf{x}) - \boldsymbol{\lambda}) \right] \qquad (\mathbf{x} \in \mathcal{S}^D), \end{aligned}$$

on $\mid \mathbf{\Gamma} \mid^{+}$ és el pseudodeterminant i $\mathbf{\Gamma}^{-}$ és la inversa generalitzada de Moore-Penrose de la matriu $\mathbf{\Gamma}$. Aitchison (1986) demostra també que $\mid \mathbf{\Gamma} \mid^{+}$ és el producte dels valors propis positius de la matriu $\mathbf{\Gamma}$.

La densitat del vector aleatori $\mathrm{clr}(\mathbf{x})$ és una densitat normal singular. Per aquesta raó és necessari l'ús de la inversa generalitzada i el pseudodeterminant de la matriu $\mathbf{\Gamma}$. Per evitar treballar amb distribucions degenerades, Aitchison utilitza una estratègia doble en els seus treballs. En la modelització de conjunts de dades composicionals amb distribucions multivariants, utilitza la transformació alr. Per altra banda, en les aplicacions no paramètriques que exigeixen simetria en el tractament de les parts d'una composició o bé hi intervenen les distàncies, utilitza la transformació clr.

Utilitzant les relacions (2.21), (2.28) i (2.34), podem escriure també l'expressió de la densitat en funció dels paràmetres $\boldsymbol{\xi}$ i $\mathbf{\Upsilon}$ i del vector $\mathrm{ilr}(\mathbf{x})$. Més concretament,

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) = \quad & (2\pi)^{-(D-1)/2} \mid (\mathbf{FU}) \mathbf{\Upsilon} (\mathbf{FU})' \mid^{-1/2} \left(\prod_{i=1}^{D} x_i \right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2} (\mathbf{FU}\mathrm{ilr}(\mathbf{x}) - \mathbf{FU}\boldsymbol{\xi})' (\mathbf{FU}\mathbf{\Upsilon}(\mathbf{FU})')^{-1} (\mathbf{FU}\mathrm{ilr}(\mathbf{x}) - \mathbf{FU}\boldsymbol{\xi}) \right] \quad (\mathbf{x} \in \mathcal{S}^D). \end{aligned}$$

Sabem que $|\mathbf{FU}\mathbf{\Upsilon}(\mathbf{FU})'| = |\mathbf{FU}|\ |\mathbf{\Upsilon}|\ |(\mathbf{FU})'| = |\mathbf{\Upsilon}|\ |\mathbf{FU}|^2$. Donat que $|\mathbf{FU}|^2 = D$ (vegeu secció 2.3.3), l'expressió de la funció de densitat d'una composició aleatòria normal logística additiva en funció dels paràmetres $\boldsymbol{\xi}$, $\mathbf{\Upsilon}$ i del vector $\mathrm{ilr}(\mathbf{x})$ és igual a

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) = \quad & (2\pi)^{-(D-1)/2} |\mathbf{\Upsilon}|^{-1/2} D^{-1/2} \left(\prod_{i=1}^{D} x_i \right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2} (\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\xi})' \mathbf{\Upsilon}^{-1} (\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\xi}) \right] \quad (\mathbf{x} \in \mathcal{S}^D). \end{aligned} \tag{3.2}$$

Observem que el terme $D^{-1/2} \left(\prod_{i=1}^{D} x_i \right)^{-1}$ es correspon amb el jacobià del canvi en la transformació ilr. És a dir, obtindríem la mateixa expressió si partíssim de la densitat del vector $\mathrm{ilr}(\mathbf{x}) \sim \mathcal{N}^{D-1}(\boldsymbol{\xi}, \mathbf{\Upsilon})$ i apliquéssim la transformació ilr inversa.

Existeixen a la pràctica un nombre considerable de conjunts de dades composicionals la distribució dels quals queda acceptablement explicada a partir d'un model normal logístic additiu. Però pecaríem d'optimistes si creguéssim que la majoria de conjunts de dades es

poden ajustar amb una distribució d'aquest tipus. Per exemple, el model normal multivariant no pot ajustar adequadament conjunts de dades transformades si presenten una certa asimetria. En l'apartat 3.3 introduïm una generalització de la distribució aln que aporta la solució a aquest problema. També, en l'apartat 3.6 veiem altres solucions possibles utilitzant les transformacions Box-Cox.

3.2.4 Amalgames de composicions amb distribució normal logística additiva

Quan treballem amb dades composicionals de grans dimensions o amb una gran quantitat de parts properes a zero, és habitual amalgamar o sumar algunes components i treballar amb la nova composició amalgamada. Coneguda la distribució de la composició inicial, serà interessant trobar la distribució de l'amalgama. Un dels principals inconvenients que presenten les distribucions aln és la dificultat per descriure la distribució d'una amalgama $\mathbf{x}_A = \mathbf{A}\mathbf{x}$, amb $\mathbf{A}$ matriu d'amalgama $C \times D$. La raó d'aquesta dificultat és la impossibilitat d'expressar el logaritme de la suma de components en termes dels logaritmes de les components.

Donada una composició $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_D)'$ amb D parts amb una distribució $\mathcal{L}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, sabem que el vector alr transformat, $(\ln(\mathrm{x}_1/\mathrm{x}_D), \ln(\mathrm{x}_2/\mathrm{x}_D), \ldots, \ln(\mathrm{x}_{D-1}/\mathrm{x}_D))'$, segueix un model normal $\mathcal{N}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ i, per tant, el vector de quocients $(\mathrm{x}_1/\mathrm{x}_D, \mathrm{x}_2/\mathrm{x}_D, \ldots, \mathrm{x}_{D-1}/\mathrm{x}_D)'$ segueix una distribució lognormal $\Lambda^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Si amalgamem les C primeres components, haurem de descriure la distribució de la composició $\mathbf{x}_A = (\sum_{i=1}^{C} \mathrm{x}_i, \mathrm{x}_{C+1}, \ldots, \mathrm{x}_D)'$, la qual després de la transformació logquocient esdevé el vector

$$\mathbf{y}_A = \left(\ln\left(\frac{\sum_{i=1}^{C} \mathrm{x}_i}{\mathrm{x}_D} \right), \ln\left(\frac{\mathrm{x}_{C+1}}{\mathrm{x}_D} \right), \ldots, \ln\left(\frac{\mathrm{x}_{D-1}}{\mathrm{x}_D} \right) \right)'.$$

Observem que l'estudi de la distribució de la primera component del vector $\mathbf{y}_A$ és equivalent a l'estudi de la distribució del logaritme de la suma de C variables lognormals. En general, el nostre problema és equivalent a estudiar la distribució de la variable $\mathrm{y} = \ln(\sum_{i=1}^{C} \mathrm{y}_i)$ on y_i $(i = 1, 2, \ldots, C)$ són variables amb distribució lognormal.

Hem iniciat la nostra recerca amb un recull bibliogràfic exhaustiu. Ens trobem davant d'un problema històric estudiat bàsicament en l'àmbit de l'enginyeria de la comunicació. Wilkinson (1934) fou el primer en abordar aquest problema en el cas particular de la suma de lognormals independents i proposà una aproximació a la distribució de y amb un model

normal. Fenton (1960) aproxima la distribució de la suma $\sum_{i=1}^{C} \mathrm{y}_i$ amb un model lognormal igualant els principals moments. Aitchison i Brown (1957) proposen dues generalitzacions aplicables a distribucions que aproximen la lognormal. La primera consisteix en representar la funció de densitat en termes de polinomis ortogonals associats a la distribució lognormal. La segona consisteix en tractar el logaritme de la variable aproximadament com una normal i representar la densitat d'aquesta amb les sèries de Gram-Charlier o Edgeworth, és a dir, en termes de polinomis d'Hermite. Posteriorment, Schleher (1977) aproxima la densitat de la suma de lognormals independents utilitzant una forma generalitzada de les sèries de Gram-Charlier i millora les aproximacions dels moments de Wilkinson i Fenton. Marlow (1967) demostra que, sota certes condicions generals, la distribució de la variable y amb y_i independents i idènticament distribuïdes és asimptòticament normal. Naus (1969) obté la funció generatriu de moments, l'esperança i la variància en el cas particular $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2)$, quan y_1 i y_2 són variables independents idènticament distribuïdes segons un model $\Lambda(0, \sigma^2)$. Posteriorment, Hamdan (1971), Naus (1973) i Crow i Shimizu (1988) presenten diferents generalitzacions del treball de Naus (1969). Un estudi de la densitat de la suma de lognormals independents utilitzant la transformada de Fourier de la funció característica és realitzat per Barakat (1976). Schwartz i Yeh (1982) presenten un procés iteratiu per avaluar la mitjana i la variància de y i validen l'aproximació amb un model normal mitjançant simulacions. Sota la suposició que la suma de lognormals té una distribució lognormal, Ho (1995) afirma que el mètode de Schleher (1977) dóna millors aproximacions quan treballem amb la suma de variables independents però l'aproximació no és adient si les variables presenten correlacions entre elles. Seguidament, aplica certes modificacions al mètode de Schwartz i Yeh (1982) i obté un mètode eficient per avaluar la mitjana i la variància del logaritme de la suma de lognormals. Més recentment, Pirinen (2000) aplica les aproximacions de Fenton-Wilkinson i Schwartz i Yeh amb les extensions introduïdes per Ho.

En definitiva, trobar la distribució exacta de la suma de variables lognormals és encara un problema obert. La distribució convergeix molt lentament a una distribució normal i exhibeix una gran asimetria. No obstant això, existeix un acord general d'aproximar la suma de lognormals independents amb un model lognormal i, per tant, el seu logaritme amb un model normal. Trobem, però, nombrosos mètodes per aproximar els paràmetres i moments de la distribució resultant.

Paral·lelament a aquesta recerca bibliogràfica, hem realitzat una primera anàlisi de la distribució de la variable y mitjançant simulacions (Mateu-Figueras et al., 2000). Ens hem restringit al cas més simple del logaritme de la suma de dues variables aleatòries lognormals, $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2)$ amb $(\mathrm{y}_1, \mathrm{y}_2)' \sim \Lambda^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. A partir d'unes simulacions preliminars, hem observat que, si bé en alguns casos una distribució normal resulta adequada per descriure la variable, en altres casos resulta desaconsellable a causa de la asimetria que presenta la distribució empírica de les dades. Aquest resultat ens ha portat a pensar en el model normal asimètric com una possible alternativa. Certament, en alguns casos el model normal asimètric proporciona un ajust raonable. Malgrat això, trobem exemples on el coeficient d'asimetria mostral és superior al valor 0.995, i per tant el model normal asimètric no resulta adequat. A continuació, i per il·lustrar aquests primers resultats, donem tres exemples gràfics.

A la figura 3.1 hem representat l'histograma de dades simulades a partir del logaritme de la suma de les components d'una mostra lognormal bivariant. El coeficient d'asimetria mostral és molt proper a 0, concretament, -0.033. A les dades hem ajustat un model normal, la densitat del qual està representada en el mateix gràfic. Si apliquem qualsevol test de bondat d'ajust per validar el model, arribem a la conclusió d'acceptar la hipòtesi de normalitat. Així doncs, si bé des d'un punt de vista teòric la variable no segueix un model normal, a efectes pràctics obtenim un ajust raonable.

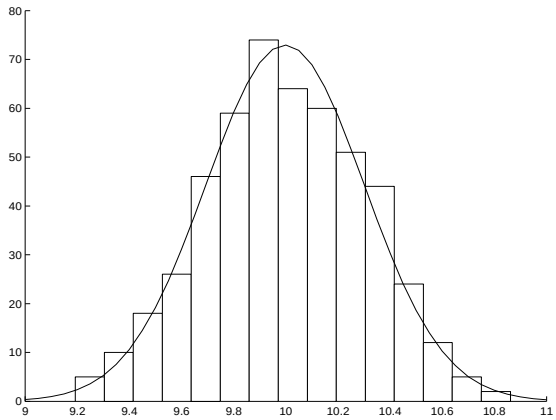
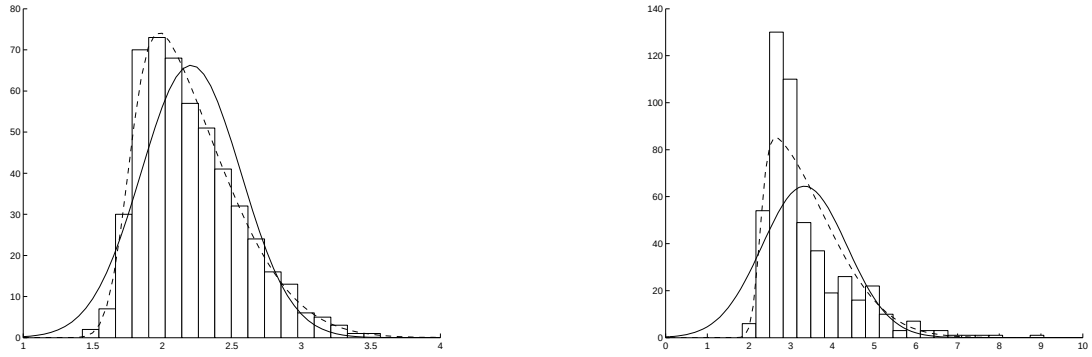


Figura 3.1: *Histograma i model normal ajustat a una mostra de la variable* $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2)$ *on* $(\mathrm{y}_1, \mathrm{y}_2)' \sim \Lambda^2((10, 0)', \begin{pmatrix} 0.1 & 0.27 \\ 0.27 & 0.9 \end{pmatrix})$.

En el gràfic 3.2(a) hem representat l'histograma d'una altra mostra simulada amb les densitats superposades dels models normal i normal asimètric ajustats. En aquest cas el coeficient d'asimetria mostral és de 0.775, i per tant el model normal no proporciona un bon ajust. Contràriament podem observar que el model normal asimètric s'ajusta millor al perfil de l'histograma. Si apliquem els contrastos de bondat d'ajust que descriurem posteriorment al capítol 5, obtenim en tots els casos un p-valor superior a 0.5, i per tant arribem a la conclusió d'acceptar el model normal asimètric. Finalment, a la figura 3.2(b) observem també l'histograma d'una mostra simulada i els models normal i normal asimètric ajustats. En aquest cas l'índex d'asimetria mostral és de 1.798 i per tant no obtenim un bon ajust en cap cas. Certament, si apliquem un test de bondat d'ajust, ambdós models resulten rebutjats.



(a) (b)

Figura 3.2: *Histograma, model normal i model normal asimètric ajustats a una mostra de la variable* $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2)$ *on* $(\mathrm{y}_1, \mathrm{y}_2)' \sim \Lambda^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *amb* (a) $\boldsymbol{\mu} = (0, 2)'$, $\boldsymbol{\Sigma} = \begin{pmatrix} 0.5 & -0.35 \\ -0.35 & 0.3 \end{pmatrix}$ *i* (b) $\boldsymbol{\mu} = (1, 1)'$ *i* $\boldsymbol{\Sigma} = \begin{pmatrix} 2.3 & -0.5 \\ -0.5 & 0.3 \end{pmatrix}$.

A la vista d'aquests resultats i per analitzar fins a quin punt té sentit la modelització de y amb una distribució normal asimètrica, hem estudiat el valor del coeficient d'asimetria, $\gamma_1[\mathrm{y}]$, utilitzant una gran quantitat de mostres de la variable $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2)$, simulades a partir d'un vector $(\mathrm{y}_1, \mathrm{y}_2)' \sim \Lambda^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ i variant el valors dels paràmetres dintre d'uns rangs fixats. Per al vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, hem considerat $\mu_1, \mu_2 \in [-20, 20]$. Per als elements de la matriu $\boldsymbol{\Sigma}$, hem considerat $\sigma_1^2, \sigma_2^2 \in [0.01, 5]$ i el coeficient de correlació $\rho = \mathrm{corr}(\ln \mathrm{y}_1, \ln \mathrm{y}_2) \in [-0.9, +0.9]$. En

cada cas hem generat mostres de mida 25000. A partir d'aquestes simulacions hem analitzat el comportament del coeficient d'asimetria mostral, $\hat{\gamma}_1$, i hem constatat heurísticament que:

1. El coeficient d'asimetria mostral és independent dels valors de μ_1 i μ_2, i depèn només del valor absolut de la seva diferència. Quan $|\mu_1 - \mu_2| > 10$, la asimetria és pràcticament nul·la i per tant, en aquests casos, l'ajust amb un model normal té sentit. Contràriament, quan $|\mu_1 - \mu_2| < 10$ la asimetria pren valors més elevats i en molts casos arriba al seu valor màxim quan $\mu_1 = \mu_2$ tot i que el seu valor està influenciat pels altres paràmetres.

2. El quocient σ_2^2/σ_1^2, influeix en la asimetria només quan $|\mu_1 - \mu_2| < 10$. En aquests casos, s'observa que $\hat{\gamma}_1$ pren els valors mínims quan $\sigma_2^2/\sigma_1^2 = 1$ i va augmentant a mesura que aquest quocient s'allunya de 1.

3. Referent al paràmetre ρ, obtenim els valors més grans del coeficient d'asimetria quan $\rho = -0.9$. Aquests van disminuint a mesura que $\rho \longrightarrow +0.9$.

Podem apreciar millor aquestes característiques observant les figures 3.3, 3.4 i 3.5, on hem representat el valor del coeficient d'asimetria de cada mostra vers la diferència $\mu_1 - \mu_2$ i el quocient σ_2^2/σ_1^2. En el gràfic 3.3 hem fixat $\rho = +0.9$; en el 3.4 hem considerat $\rho = 0$; i en la figura 3.5, $\rho = -0.9$. Es fa difícil donar un rang de valors dels paràmetres μ_1, μ_2, σ_1^2, σ_2^2 i ρ, on el coeficient d'asimetria prengui valors dins l'interval $(-0.995, +0.995)$, i per tant l'ajust amb un model normal asimètric sigui raonable. No obstant això, i en els casos considerats en les nostres simulacions, podem constatar que quan $|\mu_1 - \mu_2| < 10$ els coeficients d'asimetria resulten superiors a $+0.995$ només en els casos extrems, és a dir, quan $\rho \longrightarrow -0.9$, o quan un dels dos paràmetres σ_1^2 o σ_2^2 tendeix al seu valor màxim 5, o al seu valor mínim 0.01.

D'aquest estudi es conclou que, des d'un punt de vista pràctic, podrem intentar l'ajust amb un model normal asimètric en els casos on el coeficient d'asimetria mostral estigui dins l'interval $(-0.995, 0.995)$. El següent pas raonable és estudiar des d'un punt de vista analític si el model normal asimètric pot aproximar convenientment la distribució del logaritme de la suma de variables lognormals.

Tal i com hem indicat a la secció 1.3.2, Azzalini i Capitanio (1999) utilitzen el model normal asimètric per aproximar la distribució d'un vector aleatori condicionat. La seva estratègia per calcular els paràmetres de la densitat aproximada, consisteix en igualar els tres

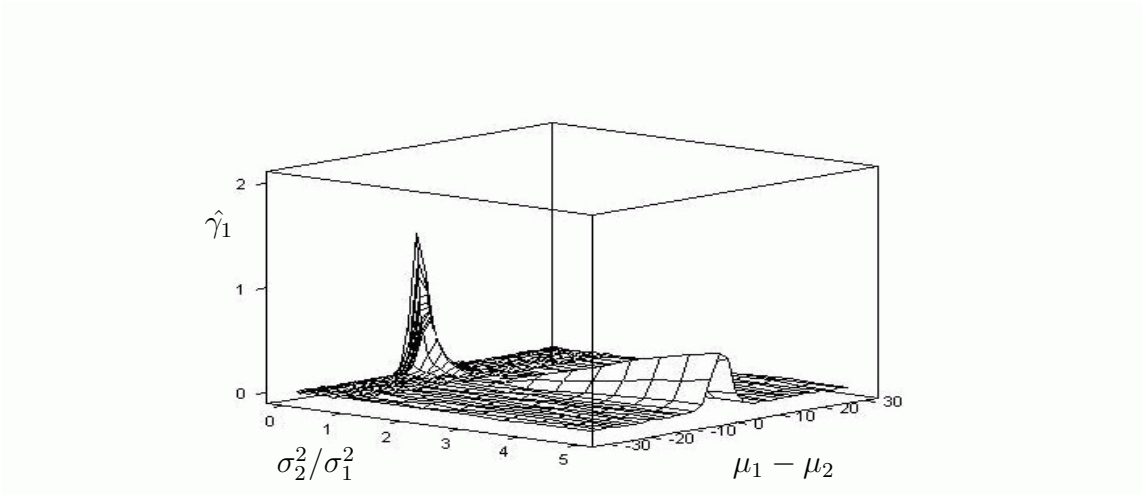


Figura 3.3: *Coeficient d'asimetria mostral de la variable* $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2)$ *generada a partir d'un vector* $(\mathrm{y}_1, \mathrm{y}_2)'$ *amb distribució lognormal amb paràmetre* $\rho = +0.9$.

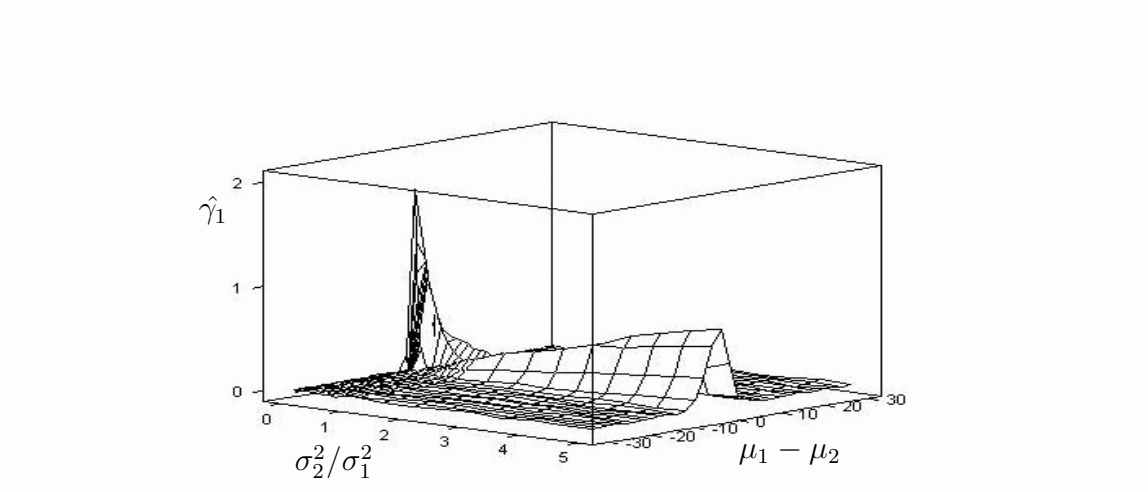


Figura 3.4: *Coeficient d'asimetria mostral de la variable* $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2)$ *generada a partir d'un vector* $(\mathrm{y}_1, \mathrm{y}_2)'$ *amb distribució lognormal amb paràmetre* $\rho = 0$.

primers moments. En el nostre cas, és difícil trobar una expressió analítica dels moments de la variable y. No obstant això, el desenvolupament de Taylor d'ordre 2 de la variable $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2) = \ln(\mathrm{e}^{\mathrm{z}_1} + \mathrm{e}^{\mathrm{z}_2})$ en el punt $(\mathrm{E}[\mathrm{z}_1], \mathrm{E}[\mathrm{z}_2])' = (\mu_1, \mu_2)' = \boldsymbol{\mu}$ permetrà calcular aproximacions analítiques per a E[y], var[y] i γ_1[y] en funció dels paràmetres $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$. Si el model normal asimètric és adequat, podrem utilitzar aquestes expressions per calcular els paràmetres de la densitat normal asimètrica teòrica que aproxima la densitat de y.

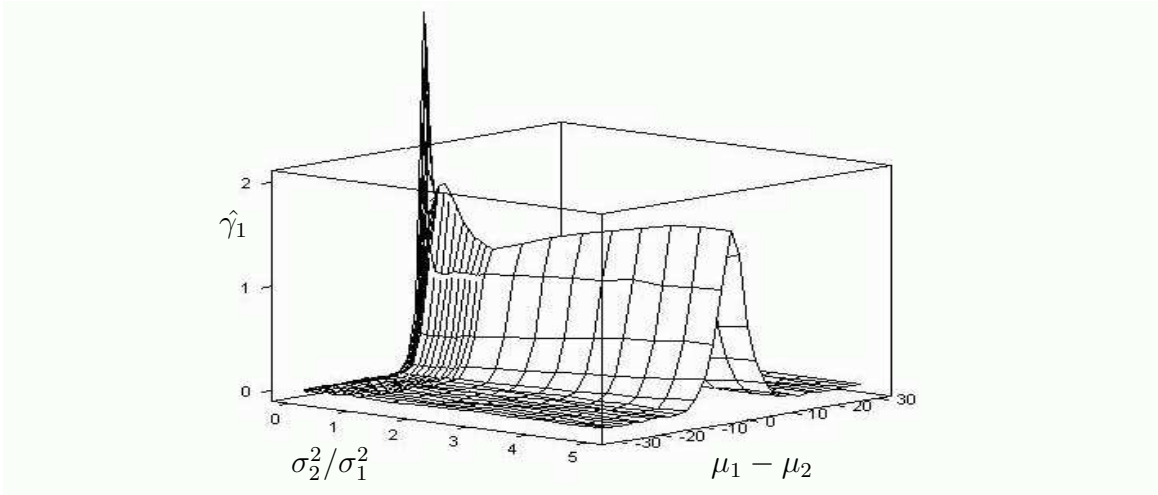


Figura 3.5: *Coeficient d'asimetria mostral de la variable* $\mathrm{y} = \ln(\mathrm{y}_1 + \mathrm{y}_2)$ *generada a partir d'un vector* $(\mathrm{y}_1, \mathrm{y}_2)'$ *amb distribució lognormal amb paràmetre* $\rho = -0.9$.

El desenvolupament de Taylor d'ordre 2 de la variable $\mathrm{y} = \ln(\mathrm{e}^{\mathrm{z}_1} + \mathrm{e}^{\mathrm{z}_2})$ en el punt $(\mathrm{E}[\mathrm{z}_1], \mathrm{E}[\mathrm{z}_2])' = \boldsymbol{\mu}' = (\mu_1, \mu_2)'$ és

$$\mathrm{y} = \ln(\mathrm{e}^{\mu_1} + \mathrm{e}^{\mu_2}) + \sum_{i=1}^{2} \kappa_i(\mathrm{z}_i - \mu_i) + \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} \kappa_i(\delta_{ij} - \kappa_j)(\mathrm{z}_i - \mu_i)(\mathrm{z}_j - \mu_j) + R_3(\mathbf{z}, \boldsymbol{\mu}),$$

on $R_3(\mathbf{z}, \boldsymbol{\mu})$ és la resta de Taylor d'ordre 3, μ_i representa la component i del vector $\boldsymbol{\mu}$, σ_{ij} la component ij de la matriu $\boldsymbol{\Sigma}$, i δ_{ij} és la delta de Kronecker, igual a 1 quan $i = j$, i igual a 0 quan $i \neq j$. Per alleugerir la notació hem utilitzat $\kappa_i = \mathrm{e}^{\mu_i}/(\mathrm{e}^{\mu_1} + \mathrm{e}^{\mu_2})$ $(i = 1, 2)$. Aquest desenvolupament permet calcular fàcilment una expressió aproximada de l'esperança, la variància i el coeficient d'asimetria de la variable y. Aquests tres coeficients identificaran la densitat normal asimètrica que caldria utilitzar per aproximar la densitat de la variable y. Aquesta metodologia fou introduïda a Aitchison i Bacon-Shone (1999) per trobar els moments de la combinació lineal convexa de diverses composicions aleatòries.

La nostra conclusió és que el model normal asimètric proporciona, en certs casos, una bona aproximació per al logaritme de la suma de variables lognormals. Observem, doncs, que hem ampliat el ventall de possibilitats contemplades fins al moment, que es restringien principalment a l'ús d'aproximacions amb un model normal.

Deixem com a línia de recerca futura l'anàlisi de la viabilitat i de la qualitat d'aquestes aproximacions i la generalització al cas $\mathrm{y} = \ln(\sum_{i=1}^{C} \mathrm{y}_i)$, amb $(\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_C)' \sim \Lambda^C(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

3.3 Distribució normal asimètrica logística additiva (alsn)

Una de les principals limitacions del model normal logístic additiu és la impossibilitat de modelitzar conjunts de dades amb una certa asimetria. En aquesta secció ens proposem generalitzar el model aln i aportar una primera solució a aquest problema. Utilitzant el model normal asimètric multivariant i la transformació alr, definim una nova família de distribucions a $\mathcal{S}^D$ que hem anomenat normal asimètrica logística additiva (alsn).

3.3.1 Definició i propietats

Mateu-Figueras et al. (1998) defineixen la distribució normal asimètrica logística additiva, que per coherència amb la distribució normal logística additiva anomenem alsn (de l'anglès *additive logistic skew-normal*).

Definició 3.2 Una composició aleatòria $\mathbf{x}$ amb D parts té una distribució *normal asimètrica logística additiva* (alsn) si el vector aleatori transformat $\mathbf{y} = \text{alr}(\mathbf{x}) = \ln(\mathbf{x}_{-D}/\text{x}_D)$ té una distribució $\mathcal{SN}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. □

Per denotar una distribució d'aquesta classe utilitzem la notació $\mathbf{x} \sim \mathcal{LS}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ o, més breument, $\mathbf{x} \sim \mathcal{LS}^D$. En aquest apartat fem servir $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ per denotar els paràmetres del model alsn però cal tenir en compte que no es corresponen amb l'esperança ni amb la matriu de covariàncies del vector $\text{alr}(\mathbf{x})$.

Propietat 3.11 La funció de densitat d'una composició aleatòria $\mathbf{x}$ que es distribueix segons una llei $\mathbf{x} \sim \mathcal{LS}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ és

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \; & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \left(\prod_{i=1}^D x_i\right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2}\left(\ln(\mathbf{x}_{-D}/x_D) - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\ln(\mathbf{x}_{-D}/x_D) - \boldsymbol{\mu}\right)\right] \\ & \times \Phi\left(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\ln(\mathbf{x}_{-D}/x_D) - \boldsymbol{\mu})\right), \end{aligned}$$

on $\mathbf{x} \in \mathcal{S}^D$ i Φ representa la funció de distribució d'una normal estàndard. De manera abreviada podem escriure

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \; & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \left(\prod_{i=1}^D x_i\right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2}(\text{alr}(\mathbf{x}) - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\text{alr}(\mathbf{x}) - \boldsymbol{\mu})\right] \\ & \times \Phi\left(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\text{alr}(\mathbf{x}) - \boldsymbol{\mu})\right). \end{aligned} \tag{3.3}$$

□

L'expressió de la densitat (3.3) sorgeix d'aplicar el teorema del canvi de variable. Si la observem amb detall, identificarem la densitat d'un vector normal asimètric multivariant multiplicada pel terme $\left(\prod_{i=1}^{D} x_i\right)^{-1}$, el jacobià de la transformació alr.

La densitat del model alsn és la derivada de Radon-Nikodým de la probabilitat respecte de la mesura de Lebesgue, i per tant obtindrem la probabilitat d'un esdeveniment qualsevol a partir d'una integral ordinària. Podríem també calcular l'esperança de la composició aleatòria $\mathbf{x}$ aplicant el procediment habitual, és a dir, utilitzant que $\mathrm{E}[\mathbf{x}] = (\mathrm{E}[\mathrm{x}_1], \mathrm{E}[\mathrm{x}_2], \ldots, \mathrm{E}[\mathrm{x}_D])'$ amb

$$\mathrm{E}[\mathrm{x}_i] = \int_{\mathcal{S}^D} x_i f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) dx_1 dx_2 \ldots dx_{D-1} \qquad i = 1, 2, \ldots, D-1.$$

Deixem com a problema obert el càlcul efectiu d'aquestes esperances. Tot i això, en cap cas el resultat serà igual al centre d'una composició aleatòria. La igualtat (2.25) ens diu que $\mathrm{cen}[\mathbf{x}] = \mathrm{alr}^{-1}(\mathrm{E}[\mathrm{alr}(\mathbf{x})])$ i sabem que $\mathrm{alr}(\mathbf{x}) \sim \mathcal{SN}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. Així doncs, amb la definició del centre d'una composició aleatòria, obtindríem que $\mathrm{cen}[\mathbf{x}] = \mathrm{alr}^{-1}(\boldsymbol{\mu} + \sqrt{2/\pi}\boldsymbol{\omega}\boldsymbol{\delta})$, on $\boldsymbol{\omega}$ és una matriu diagonal amb l'arrel quadrada de la diagonal de $\boldsymbol{\Sigma}$, i $\boldsymbol{\delta}$ és el vector relacionat amb $\boldsymbol{\alpha}$ mitjançant les expressions (1.14).

Una distribució de la classe $\mathcal{LS}^D$ queda determinada si coneixem $(D-1)(D+4)/2$ paràmetres, $D-1$ més que en el cas d'una distribució de la classe $\mathcal{L}^D$. Aquests $D-1$ paràmetres que s'afegeixen corresponen a les components del paràmetre de forma $\boldsymbol{\alpha}$.

Podem observar que quan $\boldsymbol{\alpha} = \mathbf{0}$ obtenim la funció de densitat d'un vector de classe aln. Per aquesta raó, el model alsn representa una generalització del model aln.

La propietat 1.15 de la distribució normal asimètrica ens permetrà demostrar que la família alsn és tancada per les operacions pertorbació i potència, tancada per la formació de subcomposicions així com per la permutació de les seves components.

Propietat 3.12 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{LS}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. Sigui $\mathbf{a} \in \mathcal{S}^D$ una composició constant i $b \in \mathbb{R}$ una constant. Llavors la composició aleatòria $\mathbf{x}^* = \mathbf{a} \oplus (b \otimes \mathbf{x})$ es distribueix segons una llei

$$\mathcal{LS}^D(\ln(\mathbf{a}_{-D}/a_D) + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma}, \boldsymbol{\alpha}).$$

Demostració. Siguin $\mathbf{y}$ i $\mathbf{y}^*$ els vectors aleatoris que s'obtenen després d'aplicar la transformació loquocient additiva a les composicions $\mathbf{x}$ i $\mathbf{x}^*$. A partir de la definició de $\mathbf{x}^*$ es dedueix

que la relació entre $\mathbf{y}$ i $\mathbf{y}^*$ és

$$\mathbf{y}^* = \ln\left(\frac{\mathbf{a}_{-D}}{a_D}\right) + b\mathbf{y}.$$

Sabem que $\mathbf{y} \sim \mathcal{SN}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. Aleshores, es conclou fàcilment que la distribució de $\mathbf{y}^*$ és $\mathcal{SN}^{D-1}(\ln(\mathbf{a}_{-D}/a_D) + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma}, \boldsymbol{\alpha})$. □

Si bé la família alsn és tancada per les operacions $\oplus$ i $\otimes$, la seva funció de densitat no compleix la propietat (3.1). Certament, amb uns simples càlculs es pot comprovar que el valor $f_{\mathbf{x}}(\mathbf{x})$ no és igual a $f_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x})$, on $f_{\mathbf{x}}$ i $f_{\mathbf{a}\oplus\mathbf{x}}$ representen les funcions de densitat de les composicions aleatòries $\mathbf{x}$ i $\mathbf{a} \otimes \mathbf{x}$, ambdues amb distribució normal asimètrica logística additiva.

Coneguda la distribució d'una composició, una qüestió d'interès és trobar la distribució exacta d'una subcomposició. En la següent propietat demostrem que qualsevol subcomposició d'un vector de classe alsn és també de classe alsn i donem la relació entre els respectius paràmetres.

Propietat 3.13 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{LS}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. Sigui $\mathbf{s} = \mathcal{C}(\mathbf{Sx})$ la subcomposició obtinguda a partir de la matriu de selecció $\mathbf{S}$ d'ordre $C \times D$. Llavors $\mathbf{s}$ es distribueix segons una llei $\mathcal{LS}^C(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S, \boldsymbol{\alpha}_S)$, amb

$$\boldsymbol{\mu}_S = \mathbf{Q}_S\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}_S = \mathbf{Q}_S\boldsymbol{\Sigma}\mathbf{Q}_S', \qquad \boldsymbol{\alpha}_S = \frac{\boldsymbol{\omega}_S(\boldsymbol{\Sigma}_S)^{-1}\mathbf{B}'\boldsymbol{\alpha}}{\sqrt{1 + \boldsymbol{\alpha}'(\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1} - \mathbf{B}(\boldsymbol{\Sigma}_S)^{-1}\mathbf{B}')\boldsymbol{\alpha}}},$$

on $\mathbf{Q}_S = \mathbf{F}_{(C-1)\times C}\mathbf{SF}'_{(D-1)\times D}\mathbf{H}^{-1}_{D-1}$, $\mathbf{B} = \boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\mathbf{Q}_S'$, i $\boldsymbol{\omega}_S$ i $\boldsymbol{\omega}$ són matrius diagonals amb l'arrel quadrada de les diagonals de les matrius $\boldsymbol{\Sigma}_S$ i $\boldsymbol{\Sigma}$ respectivament.

Demostració. Siguin $\mathbf{y}$ i $\mathbf{y}_S$ els vectors que s'obtenen després d'aplicar la transformació logquocient additiva a la composició $\mathbf{x}$ i a la subcomposició $\mathbf{s}$ respectivament. A partir de la definició de $\mathbf{s}$ es pot demostrar que la relació entre $\mathbf{y}_S$ i $\mathbf{y}$ és

$$\mathbf{y}_S = \mathbf{Q}_S\mathbf{y},$$

on $\mathbf{Q}_S = \mathbf{F}_{(C-1)\times C}\mathbf{SF}_{(D-1)\times D}\mathbf{H}^{-1}_{D-1}$ (Aitchison, 1986, pàg. 119). Sabem que el vector $\mathbf{y}$ segueix un model $\mathcal{SN}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. Aleshores, aplicant la propietat 1.15 de la distribució normal asimètrica es conclou que $\mathbf{s} \sim \mathcal{LS}^C(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S, \boldsymbol{\alpha}_S)$, on

$$\boldsymbol{\mu}_S = \mathbf{Q}_S\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}_S = \mathbf{Q}_S\boldsymbol{\Sigma}\mathbf{Q}_S', \qquad \boldsymbol{\alpha}_S = \frac{\boldsymbol{\omega}_S(\boldsymbol{\Sigma}_S)^{-1}\mathbf{B}'\boldsymbol{\alpha}}{\sqrt{1 + \boldsymbol{\alpha}'(\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1} - \mathbf{B}(\boldsymbol{\Sigma}_S)^{-1}\mathbf{B}')\boldsymbol{\alpha}}}.$$

□

Donada una composició aleatòria $\mathbf{x}$ de classe $\mathcal{LS}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, suposem que volem canviar d'ordre les seves components aplicant una matriu permutació. En la següent propietat demostrem que la composició resultant és de classe $\mathcal{LS}^D$ i donem els seus paràmetres en funció dels de la composició original.

Propietat 3.14 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{LS}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. Sigui $\mathbf{x}_P = \mathbf{P}\mathbf{x}$ la composició $\mathbf{x}$ amb les components reordenades per la matriu permutació $\mathbf{P}$. Llavors $\mathbf{x}_P$ es distribueix segons una llei $\mathcal{LS}^D(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P, \boldsymbol{\alpha}_P)$, amb

$$\boldsymbol{\mu}_P = \mathbf{Q}_P\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}_P = \mathbf{Q}_P\boldsymbol{\Sigma}\mathbf{Q}_P', \qquad \boldsymbol{\alpha}_P = \boldsymbol{\omega}_P(\mathbf{Q}_P')^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\alpha},$$

on $\mathbf{Q}_P = \mathbf{F}_{(D-1)\times D}\mathbf{P}\mathbf{F}'_{(D-1)\times D}\mathbf{H}_{D-1}^{-1}$, i $\boldsymbol{\omega}_P$ i $\boldsymbol{\omega}$ són matrius diagonals amb l'arrel quadrada de les diagonals de les matrius $\boldsymbol{\Sigma}_P$ i $\boldsymbol{\Sigma}$ respectivament.

Demostració. Siguin $\mathbf{y}$ i $\mathbf{y}_P$ els vectors que s'obtenen després d'aplicar la transformació logquocient additiva a les composicions $\mathbf{x}$ i $\mathbf{x}_P$. A partir de la definició de la composició $\mathbf{x}_P$ es pot demostrar que la relació entre $\mathbf{y}$ i $\mathbf{y}_P$ és

$$\mathbf{y}_P = \mathbf{Q}_P\mathbf{y}$$

on $\mathbf{Q}_P = \mathbf{F}\mathbf{P}\mathbf{F}'\mathbf{H}^{-1}$ (Aitchison, 1986, pàg. 118). Sabem que el vector $\mathbf{y}$ segueix un model $\mathcal{SN}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. Aleshores, aplicant la propietat 1.15 de la distribució normal asimètrica multivariant, es pot concloure que $\mathbf{x}_P \sim \mathcal{LS}^D(\mathbf{Q}_P\boldsymbol{\mu}, \mathbf{Q}_P\boldsymbol{\Sigma}\mathbf{Q}_P', \boldsymbol{\omega}_P(\mathbf{Q}_P')^{-1}\boldsymbol{\omega}^{-1}\boldsymbol{\alpha})$. □

Amb aquesta propietat deduïm que la distribució normal asimètrica logística additiva és independent de la component que s'utilitza com a denominador a la transformació logquocient additiva.

Un dels principals inconvenients que presenta la distribució alsn és la impossibilitat de descriure la distribució de qualsevol amalgama $\mathbf{x}_A = \mathbf{A}\mathbf{x}$ (essent $\mathbf{A}$ una matriu d'amalgama d'ordre $C \times D$) d'una composició de classe $\mathcal{LS}^D$. La raó d'aquesta dificultat és la impossibilitat d'expressar el logaritme de la suma de components en termes dels logaritmes de les components.

3.3.2 Aspectes d'inferència estadística

En aquest apartat analitzem aspectes d'inferència estadística per a la distribució normal asimètrica logística additiva. Concretament, fem referència a l'estimació dels paràmetres i

veiem certs contrastos d'hipòtesi. L'estratègia que seguim per fer aquesta anàlisi és molt senzilla: transformem les dades composicionals en observacions de l'espai real aplicant la transformació logquocient additiva i utilitzem tots els procediments estadístics multivariants que es coneixen per a la distribució normal asimètrica.

Sigui $\mathbf{X} = [x_{k,r} : k = 1, 2, \ldots, n; r = 1, 2, \ldots, D]$ una mostra de mida n d'una composició aleatòria amb D parts. Mitjançant la transformació logquocient additiva transformem la mostra a l'espai $\mathbb{R}^{D-1}$, $\mathbf{Y} = [y_{k,i} = \ln(x_{k,i}/x_{k,D}) : k = 1, 2, \ldots, n; i = 1, 2, \ldots, D-1]$. Amb aquesta mostra estimem els paràmetres $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ i $\boldsymbol{\alpha}$ del model normal asimètric aplicant el mètode de màxima versemblança (vegeu apartat 1.3.3). Denotem per $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ i $\hat{\boldsymbol{\alpha}}$ aquests estimadors. Recordem que no és possible escriure una expressió analítica dels estimadors en funció de la mostra, i per tant hem de recórrer a mètodes de maximització numèrica.

Cal tenir també en compte que en aplicar la transformació logquocient additiva podem escollir la component que s'utilitza com a denominador en els logquocients. Per tant, la mostra transformada i els estimadors calculats a partir d'ella dependran del denominador escollit. Aquest fet no és un inconvenient ja que en la propietat 3.14 hem demostrat que la família alsn és tancada per l'operació permutació de components. En la mateixa propietat hem calculat la relació entre els paràmetres del model alsn d'una composició permutada i els paràmetres de la composició original. Mitjançant aquestes relacions i fent uns simples càlculs algebraics, es dedueix la següent propietat:

Propietat 3.15 El valor del màxim de la funció de logversemblança d'una mostra d'una composició de classe alsn és invariant respecte del grup de les permutacions de les components de la composició. □

En particular, es dedueix que el valor màxim de la funció de logversemblança, $\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}})$, no depèn del denominador que s'utilitzi en la transformació logquocient additiva.

Donat que per al model alsn utilitzem mètodes numèrics per calcular els estimadors, hem fet un estudi empíric per comprovar aquesta invariància respecte del denominador. Hem partit d'un conjunt de dades reals corresponents a una subcomposició amb 3 parts, $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3)$, que recull el percentatge en pes d'òxids presents en una mostra de 332 roques procedents dels dipòsits de bauxita Halimba (Hongria). Les dades han estat amablement cedides pel Dr. G. Bárdossy de l'Acadèmia de Ciències Hongaresa.

Hem aplicat les tres transformacions alr possibles prenent cada vegada una de les parts $\mathrm{x}_1, \mathrm{x}_2$ i x_3 com a denominador, i hem calculat en cada cas els estimadors de màxima versemblança així com el valor màxim de la funció de logversemblança en cada cas. Els resultats obtinguts figuren a la taula 3.1. Podem observar com el valor $\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}})$ és pràcticament el mateix en cada cas, obtenint diferències tan sols a partir del tercer decimal. Pel que fa als estimadors dels paràmetres, podem comprovar que estan aproximadament relacionats segons les expressions de la propietat 3.14. Aquestes relacions no s'acompleixen a la perfecció, però aquest fet és degut únicament al procés numèric d'estimació.

Taula 3.1: *Estimadors i valor màxim de la funció de logversemblança per a cada una de les tres transformacions* alr.

transformació	$\mathrm{alr}_1 = \ln(\mathbf{x}_{-1}/\mathrm{x}_1)$	$\mathrm{alr}_2 = \ln(\mathbf{x}_{-2}/\mathrm{x}_2)$	$\mathrm{alr}_3 = \ln(\mathbf{x}_{-3}/\mathrm{x}_3)$
$\hat{\boldsymbol{\mu}}$	$(-1.983, -0.858)$	$(1.985, 1.127)$	$(0.858, -1.128)$
$\hat{\boldsymbol{\Sigma}}$	$\begin{pmatrix} 2.114 & -0.020 \\ -0.020 & 0.017 \end{pmatrix}$	$\begin{pmatrix} 2.109 & 2.129 \\ 2.129 & 2.125 \end{pmatrix}$	$\begin{pmatrix} 0.017 & 0.036 \\ 0.036 & 2.163 \end{pmatrix}$
$\hat{\boldsymbol{\alpha}}$	$(-4.338, 0.050)$	$(3.761, 0.565)$	$(0.339, -4.366)$
$\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}})$	-209.411	-209.412	-209.413

El model aln és un cas particular del model alsn ja que correspon al cas $\boldsymbol{\alpha} = \mathbf{0}$. A la pràctica, sovint ens interessarà comparar l'ajust per un model aln amb l'ajust per un model alsn. En aquests casos, haurem de contrastar la hipòtesi nul·la $\mathrm{H}_0 : \boldsymbol{\alpha} = \mathbf{0}$ vers la hipòtesi alternativa $\mathrm{H}_1 : \boldsymbol{\alpha} \neq \mathbf{0}$ mitjançant un test de raó de versemblança. Per realitzar aquest contrast, tan sols cal aplicar la transformació alr a la mostra i calcular el màxim de la funció de logversemblança sota la suposició d'un model alsn i aln. Denotem $\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}})$ i $\mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0})$ aquests valors on $\hat{\mathbf{m}}$ i $\hat{\mathbf{S}}$ representen els estimadors de màxima versemblança dels paràmetres d'un model aln. L'estadístic del contrast és la diferència $2(\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}}) - \mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0}))$ la qual, sota la hipòtesi nul·la, segueix una distribució χ^2 amb $D-1$ graus de llibertat, on $D-1$ és la dimensió del vector alr transformat. Observem que aquest tipus de contrast és invariant respecte del denominador escollit en la transformació alr ja que utilitzem només el valor del

màxim de la versemblança per prendre la nostra decisió.

Per validar el model normal logístic additiu caldrà aplicar un test de bondat d'ajust del model normal asimètric a les dades alr transformades. Donat que no existeixen en la literatura aquest tipus de contrastos, desenvolupem en el capítol 5 d'aquest treball d'investigació proves de bondat d'ajust per a la distribució normal asimètrica.

3.3.3 Exemples

En aquesta secció ajustem sobre tres conjunts de dades composicionals el model normal asimètric logístic additiu. Iniciem l'estudi amb la transformació de les dades de $\mathcal{S}^D$ a $\mathbb{R}^{D-1}$ mitjançant la transformació alr. Tot seguit ajustem un model normal asimètric calculant els estimadors amb el mètode de màxima versemblança. Paral·lelament ajustem un model normal multivariant i comparem els dos ajustos amb el test de raó de versemblança. Reservem per al següent capítol la validació del model amb una prova de bondat d'ajust.

1. Base de dades HALIMBA

Aquesta base de dades recull el percentatge en pes dels òxids presents en una mostra de 332 roques procedents dels dipòsits de bauxita Halimba (Hongria). Treballarem amb la subcomposició $(Al_2O_3, SiO_2, Fe_2O_3)'$.

Apliquem la transformació logquocient additiva amb la component Al_2O_3 com a denominador. Els coeficients d'asimetria de les dues marginals $\ln(SiO_2/Al_2O_3)$ i $\ln(Fe_2O_3/Al_2O_3)$ són -0.395 i -0.122 respectivament. Degut a aquesta moderada asimetria, té sentit intentar l'ajust amb un model normal asimètric bivariant. Els estimadors de màxima versemblança del model normal asimètric són

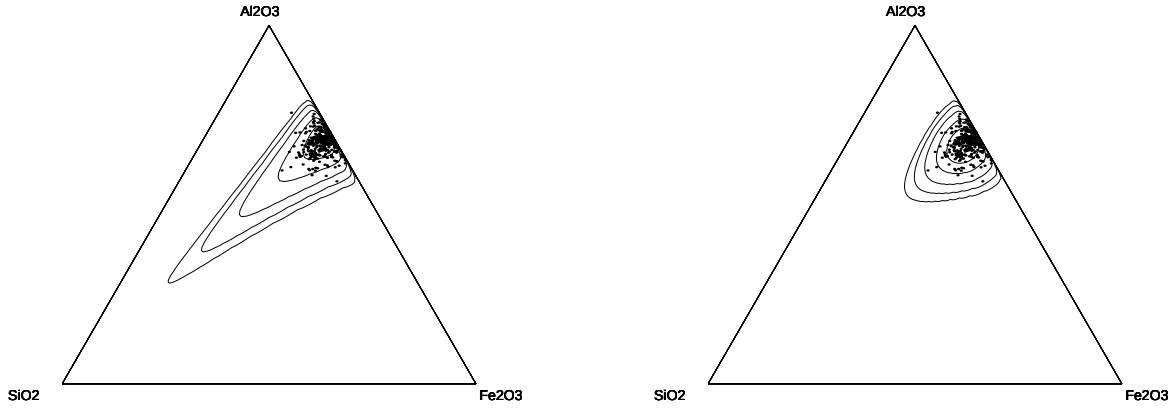
$$\hat{\boldsymbol{\mu}} = (-1.983, -0.858)', \qquad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 2.114 & -0.020 \\ -0.020 & 0.017 \end{pmatrix}, \qquad \hat{\boldsymbol{\alpha}} = (-4.338, 0.05)',$$

i el valor de la funció de logversemblança en aquest punt és $\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}}) = -209.411$. Si ajustem un model normal bivariant obtenim els següents estimadors de màxima versemblança

$$\hat{\mathbf{m}} = (-3.133, -0.846)', \qquad \hat{\mathbf{S}} = \begin{pmatrix} 0.794 & -0.006 \\ -0.006 & 0.016 \end{pmatrix},$$

i el valor de la funció de màxima versemblança és $\mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0}) = -222.32$. En les figures 3.6(a) i 3.6(b) hem representat les corbes de nivell dels dos models ajustats. De manera visual

podem observar que el model alsn recull millor la variabilitat de les dades. Si comparem numèricament els dos ajustos aplicant el test de raó de versemblança, obtenim una diferència altament significativa, $2(\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}}) - \mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0})) = 25.818 = \chi^2_2(1.000)$. Així doncs, en aquest cas, sembla més apropiat utilitzar un model alsn per ajustar aquest conjunt de dades.



(a) *Ajust amb un model* aln (b) *Ajust amb un model* alsn

Figura 3.6: *Subcomposició* $(Al_2O_3, SiO_2, Fe_2O_3)'$ *de la base de dades HALIMBA*

2. Base de dades HONGITE

Aquesta és una base de dades simulada procedent d'Aitchison (1986) que reprodueix la composició en pes de 5 minerals en una mostra de 25 roques. Considerem en aquest cas tan sols la subcomposició de les tres primeres components que anomenarem respectivament A, B i C. Aitchison (1986) va demostrar que el model aln era raonable per ajustar aquest conjunt de dades. A continuació intentarem millorar l'ajust amb un model alsn.

Utilitzant com a denominador la tercera component, apliquem la transformació alr. Els coeficients d'asimetria de les marginals $\ln(A/C)$ i $\ln(B/C)$ són 0.297 i 0.137 respectivament. Els estimadors de màxima versemblança del model normal asimètric són

$$\hat{\boldsymbol{\mu}} = (0.206, -1.097)', \qquad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 3.415 & 4.694 \\ 4.694 & 6.474 \end{pmatrix}, \qquad \hat{\boldsymbol{\alpha}} = (12.815, -9.763)',$$

i el valor de la funció de màxima versemblança en aquest punt és $\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}}) = -26.454$. Els estimadors de màxima versemblança per al model normal són

$$\hat{\mathbf{m}} = (1.6, 0.799)', \qquad \hat{\mathbf{S}} = \begin{pmatrix} 1.472 & 2.052 \\ 2.052 & 2.881 \end{pmatrix},$$

i el valor de la funció de màxima versemblança és $\mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0}) = -27.012$. Si apliquem el test de raó de versemblança per comparar numèricament els dos models, obtenim que la diferència $2(\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}}) - \mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0})) = 1.116 = \chi^2_2(0.428)$ no és significativa. Així doncs, no tenim motius suficients per rebutjar la hipòtesi de normalitat de les dades alr transformades.

3. Base de dades CONVEX

Aitchison i Bacon-Shone (1999) analitzen la distribució de combinacions lineals convexes de composicions aleatòries independents amb distribució normal logística additiva. És a dir, la distribució de la composició aleatòria $\mathbf{x} = \sum_{i=1}^{C} \pi_i \mathbf{x}_i$, on $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_C$ són C composicions independents amb distribució aln, i $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_C) \in \mathcal{S}^C$. D'un extens estudi de simulació es dedueix que en alguns casos la distribució de $\mathbf{x}$ es pot aproximar adequadament amb un model normal logístic additiu. No obstant això, existeixen casos on aquesta aproximació no és adequada degut principalment a la asimetria que presenten algunes marginals del vector $\mathrm{alr}(\mathbf{x})$. En aquestes situacions Aitchison i Bacon-Shone (1999) proposen el model normal asimètric com a alternativa.

La base de dades Convex consta de 200 observacions en el símplex $\mathcal{S}^3$. S'ha obtingut mitjançant la combinació lineal convexa $\mathbf{X} = 0.5\mathbf{X}_A + 0.5\mathbf{X}_B$, on $\mathbf{X}_A$ i $\mathbf{X}_B$ són mostres simulades de composicions aleatòries $\mathbf{x}_A$ i $\mathbf{x}_B$ independents amb distribució $\mathbf{x}_A \sim \mathcal{L}^2(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$ i $\mathbf{x}_B \sim \mathcal{L}^2(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, amb

$$\boldsymbol{\mu}_A = (0,0)', \qquad \boldsymbol{\mu}_B = (-3,3)' \qquad \text{i} \qquad \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_B = \begin{pmatrix} 0.8 & 0 \\ 0 & 1.4 \end{pmatrix}.$$

Apliquem en primer lloc la transformació logquocient additiva a la composició $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3)'$ escollint com a denominador la component x_3. Els índexs d'asimetria de les marginals del conjunt de dades transformades són -0.096 i 0.862. Aquests valors estan dins el rang de variació del coeficient d'asimetria d'una distribució normal asimètrica. Per tant, té sentit intentar l'ajust amb un model d'aquest tipus. Els estimadors de màxima versemblança són

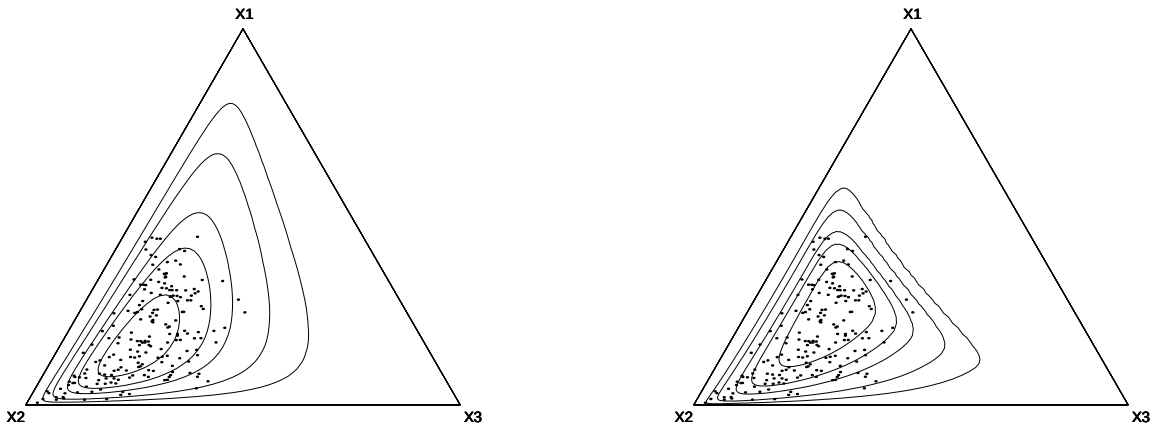
$$\hat{\boldsymbol{\mu}} = (0.581, 0.812)', \qquad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.201 & -0.306 \\ -0.306 & 0.703 \end{pmatrix}, \qquad \hat{\boldsymbol{\alpha}} = (-5.173, 5.219)',$$

i el valor del màxim de la funció de logversemblança és $\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}}) = -417.071$. Intentem

també l'ajust amb un model normal i obtenim els estimadors

$$\hat{\mathbf{m}} = (-0.114, 1.347)', \qquad \hat{\mathbf{S}} = \begin{pmatrix} 0.717 & 0.067 \\ 0.067 & 0.416 \end{pmatrix},$$

amb un valor de la funció de màxima versemblança de $\mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0}) = -445.072$. En les figures 3.7(a) i 3.7(b) hem representat les corbes de nivell dels dos models ajustats. Observem clarament que el model alsn proporciona un millor ajust al conjunt de dades. Si apliquem el test de raó de versemblança per comparar els dos models, obtenim que la diferència $2(\mathrm{l}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\alpha}}) - \mathrm{l}(\hat{\mathbf{m}}, \hat{\mathbf{S}}, \mathbf{0})) = 56.002 = \chi^2_2(1.000)$ és altament significativa i per tant rebutgem el model aln.



(a) *Ajust amb un model* aln (b) *Ajust amb un model* alsn

Figura 3.7: *Base de dades CONVEX*

3.3.4 Altres parametritzacions

Hem definit la distribució normal asimètrica logística additiva utilitzant la transformació logquocient additiva, però podem definir la mateixa llei de probabilitat utilitzant la transformació logquocient isomètrica. En la definició només cal exigir que el vector transformat $\mathrm{ilr}(\mathbf{x})$ segueixi una distribució normal asimètrica.

En la definició del model normal asimètric logístic additiu imposàvem que $\mathrm{alr}(\mathbf{x}) \sim \mathcal{SN}^{D-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. D'acord amb la igualtat $\mathrm{ilr}(\mathbf{x}) = \mathbf{U}'\mathbf{F}^*\mathrm{alr}(\mathbf{x})$ i la propietat 1.15, podem

assegurar que $\text{ilr}(\mathbf{x}) \sim \mathcal{SN}^{D-1}(\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$ amb

$$\boldsymbol{\xi} = \mathbf{U}'\mathbf{F}^*\boldsymbol{\mu}, \qquad \boldsymbol{\Upsilon} = \mathbf{U}'\mathbf{F}^*\boldsymbol{\Sigma}(\mathbf{U}'\mathbf{F}^*)', \qquad \boldsymbol{\varrho} = \boldsymbol{\upsilon}((\mathbf{U}'\mathbf{F}^*)^{-1})'\boldsymbol{\omega}^{-1}\boldsymbol{\alpha} = \boldsymbol{\upsilon}(\mathbf{FU})'\boldsymbol{\omega}^{-1}\boldsymbol{\alpha},$$

on $\boldsymbol{\upsilon}$ i $\boldsymbol{\omega}$ són matrius diagonals iguals a l'arrel quadrada de la diagonal de $\boldsymbol{\Upsilon}$ i $\boldsymbol{\Sigma}$ respectivament. És fàcil obtenir les relacions matricials inverses, és a dir,

$$\begin{aligned} &\boldsymbol{\mu} = (\mathbf{U}'\mathbf{F}^*)^{-1}\boldsymbol{\xi} = \mathbf{FU}\boldsymbol{\xi}, \quad \boldsymbol{\Sigma} = (\mathbf{U}'\mathbf{F}^*)^{-1}\boldsymbol{\Upsilon}((\mathbf{U}'\mathbf{F}^*)')^{-1} = \mathbf{FU}\boldsymbol{\Upsilon}(\mathbf{FU})', \\ &\boldsymbol{\alpha} = \boldsymbol{\omega}((\mathbf{FU})^{-1})'\boldsymbol{\upsilon}^{-1}\boldsymbol{\varrho}. \end{aligned} \tag{3.4}$$

En aquest cas hem utilitzat la notació $\boldsymbol{\xi}, \boldsymbol{\Upsilon}$ i $\boldsymbol{\varrho}$, però igual que en la definició 3.2, cal tenir en compte que els paràmetres $\boldsymbol{\xi}$ i $\boldsymbol{\Upsilon}$ no es corresponen amb l'esperança i la matriu de covariàncies del vector $\text{ilr}(\mathbf{x})$.

El jacobià del canvi en la transformació ilr és $D^{-1/2}\left(\prod_{i=1}^D x_i\right)^{-1}$, i per tant la funció de densitat de probabilitat de la composició $\mathbf{x}$ en la parametrització $\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho}$ i $\text{ilr}(\mathbf{x})$ és

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho}) = \; & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} D^{-1/2}\left(\prod_{i=1}^D x_i\right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})'\boldsymbol{\Upsilon}^{-1}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right] \\ & \times \Phi\left(\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right), \end{aligned} \tag{3.5}$$

on $\mathbf{x} \in \mathcal{S}^D$ i Φ representa la funció de distribució d'una normal estàndard.

Amb les relacions matricials anteriors podem demostrar que les densitats (3.3) i (3.5) són iguals, és a dir,

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \; & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Sigma} \mid^{-1/2}\left(\prod_{i=1}^D x_i\right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2}(\text{alr}(\mathbf{x}) - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\text{alr}(\mathbf{x}) - \boldsymbol{\mu})\right] \times \Phi\left(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\text{alr}(\mathbf{x}) - \boldsymbol{\mu})\right) \\ = \; & 2(2\pi)^{-(D-1)/2} \mid (\mathbf{FU})\boldsymbol{\Upsilon}(\mathbf{FU})' \mid^{-1/2}\left(\prod_{i=1}^D x_i\right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2}(\mathbf{FU}\text{ilr}(\mathbf{x}) - \mathbf{FU}\boldsymbol{\xi})'(\mathbf{FU}\boldsymbol{\Upsilon}(\mathbf{FU})')^{-1}(\mathbf{FU}\text{ilr}(\mathbf{x}) - \mathbf{FU}\boldsymbol{\xi})\right] \\ & \times \Phi\left(\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\mathbf{FU})^{-1}\boldsymbol{\omega}\boldsymbol{\omega}^{-1}(\mathbf{FU}\text{ilr}(\mathbf{x}) - \mathbf{FU}\boldsymbol{\xi})\right) \\ = \; & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} D^{-1/2}\left(\prod_{i=1}^D x_i\right)^{-1} \\ & \times \exp\left[-\tfrac{1}{2}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})'\boldsymbol{\Upsilon}^{-1}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right] \times \Phi\left(\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right) \\ = \; & f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho}). \end{aligned}$$

En aquest cas, no és immediat escriure la densitat en termes de la transformació logquocient centrada. Sabem que $\text{clr}(\mathbf{x}) = \mathbf{F}^*\text{alr}(\mathbf{x})$, no obstant això, no podem aplicar la propietat 1.15 de la distribució normal asimètrica ja que aquesta es demostrà suposant una transformació lineal a un espai de dimensió igual o inferior a l'inicial. És evident que la distribució del

vector clr($\mathbf{x}$) serà una distribució degenerada o singular però no podem parlar de distribució normal asimètrica singular ja que encara no ha estat definida.

3.4 Distribució lognormal asimètrica a $\mathbb{R}_+^D$

A la pràctica, obtenim una distribució normal logística additiva a $\mathcal{S}^D$ quan la distribució d'un vector aleatori de $\mathbb{R}_+^D$ és lognormal (vegeu propietat 3.6). A continuació formulem aquesta qüestió en el cas de la distribució normal asimètrica logística additiva. El problema no té una solució senzilla ja que cal definir una nova distribució a l'espai $\mathbb{R}_+^D$ que hem anomenat distribució lognormal asimètrica.

En aquest apartat definim en primer lloc la distribució lognormal asimètrica univariant i donem l'expressió de la seva funció de densitat i dels seus principals moments. Seguim amb l'estudi de la distribució lognormal asimètrica multivariant i acabem demostrant que la composició obtinguda a partir d'un vector amb distribució lognormal asimètrica té una distribució normal asimètrica logística additiva.

3.4.1 Distribució lognormal asimètrica univariant

Definició 3.3 Direm que una variable aleatòria w $\in \mathbb{R}^+$ té una distribució *lognormal asimètrica univariant* amb paràmetres μ, σ i λ quan la variable aleatòria transformada y = ln(w) tingui una distribució $\mathcal{SN}(\mu, \sigma, \lambda)$. □

Propietat 3.16 La funció de densitat d'una variable aleatòria amb distribució lognormal asimètrica amb paràmetres μ, σ i λ és

$$\mathrm{f}(w) = \begin{cases} \frac{2}{\sqrt{2\pi}\sigma w} \exp\left(-\frac{1}{2\sigma^2}(\ln w - \mu)^2\right) \Phi\left(\lambda\sigma^{-1}(\ln w - \mu)\right) & w > 0, \\ 0 & w \leq 0, \end{cases}$$

on Φ és la funció de distribució d'una normal estàndard. □

A la figura 3.8 hem representat la funció de densitat de variables lognormals asimètriques. Hem fixat els paràmetres $\mu = 0$, $\sigma = 0.3$ i hem variat el paràmetre λ. En el cas $\lambda = 0$ obtenim la funció de densitat d'una distribució lognormal. Les densitats obtingudes en els casos $\lambda = -1, 0, 2$ tenen un fort biaix cap a la dreta però podem observar com en el cas

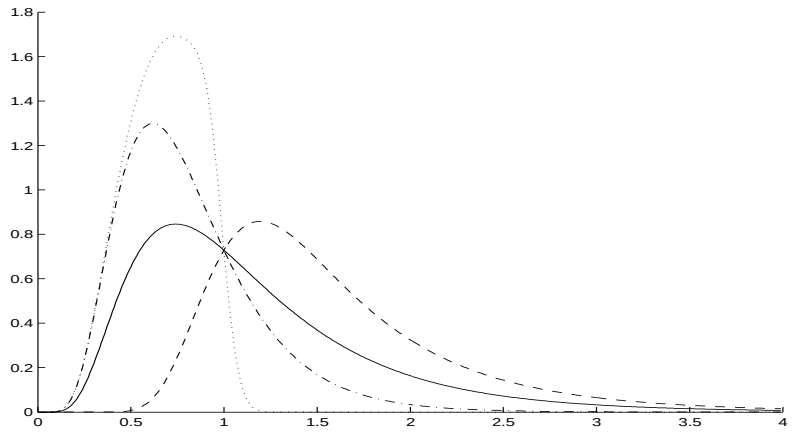


Figura 3.8: *Funcions de densitat lognormals asimètriques amb* $\mu = 0, \sigma = 0.3$ *i* $\lambda = -8$ (.....), $\lambda = -1$ (.-.-.-.), $\lambda = 0$ (——) *i* $\lambda = 2$ (- - - -).

$\lambda = -8$ tenim una lleugera asimetria cap a l'esquerra. Així doncs, ens trobem davant una generalització de la distribució lognormal que permet densitats amb un biaix negatiu.

Volem emfatitzar que aquesta densitat és totalment coherent dintre la metodologia MOVE, ja que ha estat definida mitjançant la transformació logarítmica.

Per estudiar amb més detall aquesta distribució, calculem l'expressió dels moments que ens permetran trobar una expressió per a l'esperança, la variància i el coeficient d'asimetria.

Propietat 3.17 Sigui w una variable aleatòria lognormal asimètrica amb paràmetres μ, σ i λ. Aleshores el moment d'ordre r al voltant de l'origen és

$$\mu'_r = \mathrm{E}[\mathrm{w}^r] = 2\exp(r\mu + r^2\sigma^2/2)\Phi(\sigma\delta r),$$

on $\delta = \lambda/\sqrt{1+\lambda^2}$ i Φ és la funció de distribució d'una normal estàndard.

Demostració. A partir de la definició de variable aleatòria lognormal asimètrica sabem que $\mathrm{w} = \exp(\mathrm{y})$, on y és una variable aleatòria $\mathcal{SN}(\mu, \sigma, \lambda)$. Aleshores, $\mathrm{E}[\mathrm{w}^r] = \mathrm{E}[\exp(r\mathrm{y})]$. És a dir, el moment d'ordre r al voltant de l'origen de la variable w coincideix amb el valor de la funció generatriu de moments de la variable y avaluada al punt r. Utilitzant l'expressió (1.12) obtenim que $\mathrm{E}[\mathrm{w}^r] = 2\exp(r\mu + r^2\sigma^2/2)\Phi(\sigma\delta r)$. □

A partir de la propietat 3.17 i mitjançant càlculs algebraics senzills, obtenim una expressió

analítica de l'esperança, la variància i el coeficient d'asimetria de la variable w:

$$\begin{aligned} \mathrm{E}[\mathrm{w}] &= 2\exp(\mu+\sigma^2/2)\Phi(\sigma\delta); \\ \mathrm{var}[\mathrm{w}] &= 2\exp(2\mu+\sigma^2)\left(\exp(\sigma^2)\Phi(2\sigma\delta)-2\Phi^2(\sigma\delta)\right); \\ \gamma_1[\mathrm{w}] &= \frac{\exp(3\sigma^2)\Phi(3\sigma\delta)+8\Phi^3(\sigma\delta)-6\exp(\sigma^2)\Phi(\sigma\delta)\Phi(2\sigma\delta)}{\sqrt{2}[\exp(\sigma^2)\Phi(2\sigma\delta)-2\Phi^2(\sigma\delta)]^{3/2}}. \end{aligned} \tag{3.6}$$

Observem que el coeficient d'asimetria $\gamma_1[\mathrm{w}]$ no depèn del paràmetre μ. Per analitzar amb més detall la asimetria d'una variable lognormal asimètrica, hem representat a la figura 3.9 el coeficient $\gamma_1[\mathrm{w}]$ vers λ per a diferents valors fixats de σ. Podem observar clarament un increment del valor del coeficient d'asimetria a mesura que augmentem el valor dels paràmetres σ i λ. Fixat el paràmetre σ, observem un màxim (en alguns casos local i en altres global) en el punt $\lambda = 0$. En el gràfic podem veure clarament que, en certs casos, obtenim valors negatius del coeficient d'asimetria. Això ens indica que la família lognormal asimètrica admet densitats esbiaixades per la dreta i per l'esquerra. Finalment podem observar que quan $\lambda \longrightarrow \pm\infty$ el coeficient d'asimetria tendeix a un valor que depèn del paràmetre σ. Més concretament,

$$\begin{aligned} \lim_{\lambda\to+\infty}\gamma_1[\mathrm{w}] &= \frac{\exp(3\sigma^2)\Phi(3\sigma)+8\Phi^3(\sigma)-6\exp(\sigma^2)\Phi(\sigma)\Phi(2\sigma)}{\sqrt{2}\left(\exp(\sigma^2)\Phi(2\sigma)-2\Phi^2(\sigma)\right)^{3/2}}, \\ \lim_{\lambda\to-\infty}\gamma_1[\mathrm{w}] &= \frac{\exp(3\sigma^2)\Phi(-3\sigma)+8\Phi^3(-\sigma)-6\exp(\sigma^2)\Phi(-\sigma)\Phi(-2\sigma)}{\sqrt{2}\left(\exp(\sigma^2)\Phi(-2\sigma)-2\Phi^2(-\sigma)\right)^{3/2}}. \end{aligned}$$

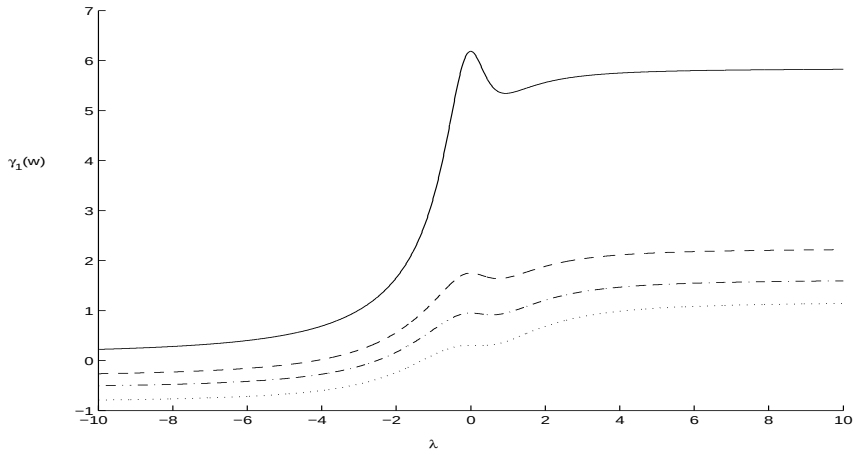


Figura 3.9: *Coeficient d'asimetria vers* λ *amb* $\sigma = 0.1$ (.....), $\sigma = 0.3$ (._._._), $\sigma = 0.5$ (- - - -) *i* $\sigma = 1$ (——).

3.4.2 Distribució lognormal asimètrica multivariant

De manera natural podem definir el model lognormal asimètric en el cas multivariant.

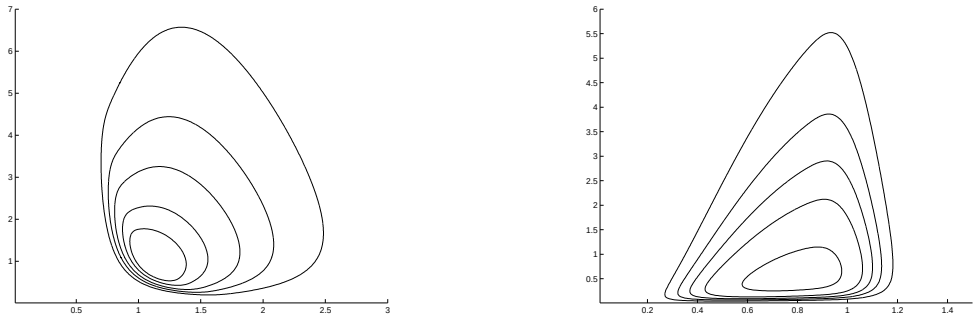
Definició 3.4 Direm que un vector aleatori $\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_D)' \in \mathbb{R}_+^D$ té una distribució *lognormal asimètrica multivariant* amb paràmetres $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ i $\boldsymbol{\alpha}$ quan el vector aleatori transformat $\mathbf{y} = \ln(\mathbf{w}) = (\ln \mathrm{w}_1, \ln \mathrm{w}_2, \ldots, \ln \mathrm{w}_D)'$ tingui una distribució $\mathcal{SN}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. □

Propietat 3.18 La funció de densitat d'un vector aleatori $\mathbf{w} \in \mathbb{R}_+^D$ amb distribució lognormal asimètrica i amb paràmetres $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ i $\boldsymbol{\alpha}$ és

$$\mathrm{f}(\mathbf{w}) = \begin{cases} \frac{2}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2} \prod_{i=1}^D w_i} \exp\left(-\frac{1}{2}(\ln\mathbf{w} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\ln\mathbf{w} - \boldsymbol{\mu})\right) \Phi\left(\boldsymbol{\alpha}\boldsymbol{\omega}^{-1}(\ln\mathbf{w} - \boldsymbol{\mu})\right) & \mathbf{w} \in \mathbb{R}_+^D, \\ 0 & \mathbf{w} \notin \mathbb{R}_+^D; \end{cases}$$

on $\boldsymbol{\omega}$ és una matriu diagonal que conté l'arrel quadrada de la diagonal de $\boldsymbol{\Sigma}$ i Φ és la funció de distribució de la normal estàndard. □

Les expressions (3.6) calculades en el cas univariant ens permetran trobar l'esperança, la variància i el coeficient d'asimetria de cada marginal.



(a) (b)

Figura 3.10: *Corbes de nivell d'una distribució lognormal asimètrica bivariant amb paràmetres* $\boldsymbol{\mu} = (0,0)', \boldsymbol{\Sigma} = \begin{pmatrix} 0.1 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}$ *i a)* $\boldsymbol{\alpha} = (5,3)'$ *i b)* $\boldsymbol{\alpha} = (-5,0)'$.

A les figures 3.10(a) i 3.10(b) hem representat les corbes de nivell de dues densitats lognormals asimètriques bivariants. En el primer gràfic observem una asimetria positiva en les dues marginals del vector $\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2)'$, concretament $\gamma_1[\mathrm{w}_1] = 1.3$ i $\gamma_1[\mathrm{w}_2] = 2.7$. En

canvi, en el segon gràfic observem un moderat biaix a l'esquerra per a la primera marginal. El valor exacte dels respectius coeficients d'asimetria és $\gamma_1[\mathrm{w}_1] = -0.3$ i $\gamma_1[\mathrm{w}_2] = 2.5$.

La funció generatriu de moments d'un vector normal asimètric multivariant permet trobar l'expressió analítica dels moments al voltant de l'origen d'un vector lognormal asimètric. La demostració és similar a la donada en el cas univariant.

Propietat 3.19 Sigui $\mathbf{w} \in \mathbb{R}_+^D$ un vector aleatori amb distribució lognormal asimètrica amb paràmetres $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ i $\boldsymbol{\alpha}$. Aleshores els moments al voltant de l'origen són

$$\mu'_{t_1 t_2 \ldots t_D} = \mathrm{E}[\mathrm{w}_1^{t_1} \mathrm{w}_2^{t_2} \ldots \mathrm{w}_D^{t_D}] = 2\exp(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})\Phi(\boldsymbol{\delta}'\boldsymbol{\omega}\mathbf{t}),$$

on $\mathbf{t}' = (t_1, t_2, \ldots, t_D)'$, $\boldsymbol{\delta} = \boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1}\boldsymbol{\alpha}/\sqrt{1 + \boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1}\boldsymbol{\alpha}}$, $\boldsymbol{\omega}$ és una matriu diagonal amb l'arrel quadrada de la diagonal de $\boldsymbol{\Sigma}$, i Φ és la funció de distribució d'una normal estàndard.

□

3.4.3 Composició associada a un vector lognormal asimètric

La següent propietat fa referència a la distribució del vector aleatori a partir de la qual obtenim una distribució de classe alsn a l'espai $\mathcal{S}^D$.

Propietat 3.20 Sigui $\mathbf{w} \in \mathbb{R}_+^D$ un vector aleatori amb distribució lognormal asimètrica amb paràmetres $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ i $\boldsymbol{\alpha}$. Aleshores la composició associada $\mathbf{x} = \mathcal{C}(\mathbf{w})$ es distribueix segons una llei $\mathcal{LS}^D(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\alpha}^*)$ on

$$\boldsymbol{\mu}^* = \mathbf{F}\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}^* = \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}' \qquad \text{i} \qquad \boldsymbol{\alpha}^* = \frac{\boldsymbol{\omega}^*(\boldsymbol{\Sigma}^*)^{-1}\mathbf{B}'\boldsymbol{\alpha}}{\sqrt{1 + \boldsymbol{\alpha}'\left(\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1} - \mathbf{B}(\boldsymbol{\Sigma}^*)^{-1}\mathbf{B}'\right)\boldsymbol{\alpha}}},$$

amb $\mathbf{B} = \boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\mathbf{F}'$ i $\boldsymbol{\omega}$ i $\boldsymbol{\omega}^*$ són matrius diagonals que contenen l'arrel quadrada de $\boldsymbol{\Sigma}$ i $\boldsymbol{\Sigma}^*$ respectivament.

Demostració. A partir de la definició de distribució lognormal asimètrica resulta que el vector $\ln \mathbf{w} \sim \mathcal{SN}^D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$. La relació entre el vector $\ln \mathbf{w}$ i el vector logquocient transformat $\mathbf{y} = \ln(\mathbf{x}_{-D}/x_D)$ ve donada per $\mathbf{y} = \mathbf{F} \ln \mathbf{w}$ (Aitchison, 1986, pàg. 117). Aplicant la propietat 1.15 de la distribució normal asimètrica, obtenim que la distribució del vector $\mathbf{F} \ln \mathbf{w}$ és $\mathcal{SN}^{D-1}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\alpha}^*)$, amb

$$\boldsymbol{\mu}^* = \mathbf{F}\boldsymbol{\mu}, \qquad \boldsymbol{\Sigma}^* = \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}' \qquad \text{i} \qquad \boldsymbol{\alpha}^* = \frac{\boldsymbol{\omega}^*(\boldsymbol{\Sigma}^*)^{-1}\mathbf{B}'\boldsymbol{\alpha}}{\sqrt{1 + \boldsymbol{\alpha}'\left(\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}\boldsymbol{\omega}^{-1} - \mathbf{B}(\boldsymbol{\Sigma}^*)^{-1}\mathbf{B}'\right)\boldsymbol{\alpha}}}.$$

□

Aquest resultat ens diu que qualsevol vector aleatori que tingui una distribució lognormal asimètrica amb qualsevol estructura de covariància dóna lloc a una composició aleatòria amb distribució normal asimètrica logística additiva.

3.5 Distribució de Dirichlet

Aitchison (1986) recorda que la distribució més coneguda sobre el símplex és la distribució de Dirichlet.

Definició 3.5 La *distribució de Dirichlet* amb paràmetre $\boldsymbol{\alpha} \in \mathbb{R}^D_+$, que simbolitzarem per $\mathcal{D}^D(\boldsymbol{\alpha})$, és la distribució sobre $\mathcal{S}^D$ que té per funció de densitat

$$f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^D \alpha_i\right)}{\prod_{i=1}^D \Gamma(\alpha_i)} \prod_{i=1}^D x_i^{\alpha_i - 1} \qquad (\mathbf{x} \in \mathcal{S}^D). \tag{3.7}$$

La *classe de les distribucions de Dirichlet* consisteix en el conjunt de totes les distribucions que s'obtenen en permetre que el vector $\boldsymbol{\alpha}$ variï sobre l'espai $\mathbb{R}^D_+$. □

La densitat (3.7) és també la derivada de Radon-Nikodým respecte de la mesura de Lebesgue. Es tracta, doncs, d'una densitat clàssica, i per tant obtindrem la probabilitat d'un esdeveniment qualsevol calculant integrals ordinàries d'aquesta funció de densitat.

Per calcular el vector d'esperances i la matriu de covariàncies s'utilitza el procediment habitual, tot obtenint

$$\begin{aligned}
&\mathrm{E}[x_i] = \alpha_i/\alpha_+, \\
&\mathrm{var}[x_i] = \alpha_i(\alpha_+ - \alpha_i)/(\alpha_+^2(\alpha_+ + 1)), \\
&\mathrm{cov}(x_i, x_j) = -\alpha_i\alpha_j/(\alpha_+^2(\alpha_+ + 1)) \qquad (i \neq j),
\end{aligned}$$

on $\alpha_+ = \sum_{i=1}^D \alpha_i$. Recordem que aquests valors són poc adients per representar el centre i la variabilitat d'una composició aleatòria si considerem l'estructura de $\mathcal{S}^D$ definida al capítol 2.

Propietat 3.21 Sigui $\mathbf{x}$ una composició aleatòria de Dirichlet $\mathcal{D}^D(\boldsymbol{\alpha})$. Sigui $\mathbf{S}$ una matriu de selecció $C \times D$ $(2 \leq C < D)$, $\mathbf{A}$ una matriu d'amalgama $C \times D$ $(2 \leq C \leq D)$, i $\mathbf{P}$ una matriu permutació $D \times D$. Llavors:

a. $\mathbf{s} \sim \mathcal{D}^C(\mathbf{S}\boldsymbol{\alpha})$, on $\mathbf{s} = \mathcal{C}(\mathbf{S}\mathbf{x})$.

b. $\mathbf{x}_A \sim \mathcal{D}^C(\mathbf{A}\boldsymbol{\alpha})$, on $\mathbf{x}_A = \mathbf{A}\mathbf{x}$.

c. $\mathbf{x}_P \sim \mathcal{D}^D(\mathbf{P}\boldsymbol{\alpha})$, on $\mathbf{x}_P = \mathbf{P}\mathbf{x}$.

□

Les propietats 3.21a, 3.21b i 3.21c posen de manifest l'elegància i la simplicitat matemàtica de la classe de distribucions de Dirichlet. Tant la permutació de components com el pas a subcomposicions o amalgames d'una composició de Dirichlet donen lloc a una altra distribució de Dirichlet amb paràmetres relacionats de manera molt simple amb els paràmetres de la composició original.

La família de distribucions de Dirichlet no és tancada per les operacions pertorbació i potència. És a dir, donades $\mathbf{x} \sim \mathcal{D}^D(\boldsymbol{\alpha})$, una composició constant $\mathbf{p} \in \mathcal{S}^D$ i $\alpha \in \mathbb{R}$, les composicions aleatòries $\mathbf{p} \oplus \mathbf{x}$ i $\alpha \otimes \mathbf{x}$ no segueixen una distribució de Dirichlet.

Propietat 3.22 Els contorns d'isoprobabilitat d'una distribució de Dirichlet $\mathcal{D}^D(\boldsymbol{\alpha})$ són convexos si $\alpha_i > 1$ $(i = 1, \ldots, D)$. □

El fet que els contorns d'isoprobabilitat de les distribucions de Dirichlet siguin sempre convexos si $\alpha_i > 1$ $(i = 1, 2, \ldots, D)$ fa que resulti difícil utilitzar-les per modelitzar conjunts de dades composicionals com els de la figura 3.11, amb un perfil marcadament còncau. Per altra banda, aquests perfils solen ser bastant habituals a la pràctica.

La distribució de Dirichlet es genera mitjançant la clausura d'un vector aleatori amb components independents, cadascuna d'elles amb distribució Gamma amb paràmetres α_i i β, és a dir, amb un paràmetre d'escala β idèntic en totes les components (vegeu Stuart i Ort, 1986). La crítica que Aitchison realitza a la distribució de Dirichlet és la seva gran rigidesa, derivada de l'exigència d'independència de les components a partir de les quals es construeix.

Propietat 3.23 Sigui $\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_D)' \in \mathbb{R}_+^D$ un vector aleatori les components del qual w_i $(i = 1, \ldots, D)$ són independents i amb una distribució Gamma igualment escalada $Ga(\alpha_i, \beta)$. Llavors la composició $\mathbf{x} = \mathcal{C}(\mathbf{w})$ es distribueix segons una llei $\mathcal{D}^D(\boldsymbol{\alpha})$. □

D'aquesta propietat s'observa que les components d'una composició de Dirichlet són gairebé independents, ja que la correlació entre elles estarà únicament motivada pel fet d'haver-les dividit totes per la suma comú $\sum_{i=1}^{D} \mathrm{w}_i$ en el procés de construcció de la composició.

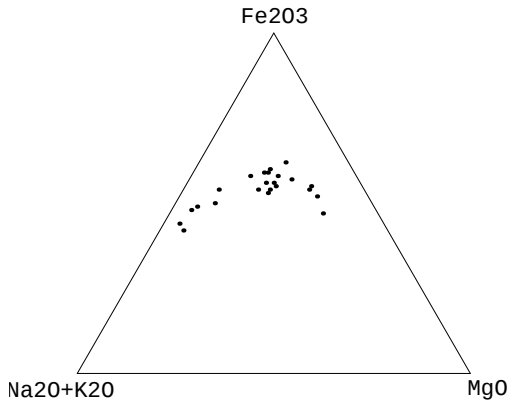


Figura 3.11: *Composició d'una mostra de 23 laves afíriques procedents de l'illa de Skye (Escòcia) adaptades per Aitchison (1986) de Thompson, Esson i Duncan (1972, Fig. 7).*

Contràriament a aquest resultat, recordem que obteníem una distribució aln o alsn a partir d'un vector de $\mathbb{R}^D_+$ amb qualsevol estructura de covariància. Aquesta estructura d'independència torna a aparèixer reflectida en la següent propietat:

Propietat 3.24 Sigui $\mathbf{x}$ una composició aleatòria de Dirichlet $\mathcal{D}^D(\boldsymbol{\alpha})$. Siguin $\mathbf{s}_1$ i $\mathbf{s}_2$ dues subcomposicions de $\mathbf{x}$ qualssevol construïdes a partir de subconjunts disjunts de components de la composició $\mathbf{x}$. Llavors es compleix que $\mathbf{s}_1$ i $\mathbf{s}_2$ són independents. □

Per estimar el paràmetre $\boldsymbol{\alpha}$ d'una composició aleatòria $\mathbf{x} \sim \mathcal{D}^D(\boldsymbol{\alpha})$ a partir dels valors d'una mostra s'utilitza el mètode de màxima versemblança. A Narayanan (1991) trobem l'algorisme AS 266 que proporciona l'estimació màxim versemblant del paràmetre $\boldsymbol{\alpha}$.

En resum, podem afirmar que la classe de distribucions de Dirichlet compleix elegants propietats en relació a les subcomposicions i a les amalgames. Així i tot, el seu àmbit d'aplicació és molt restringit a causa de les propietats 3.23 i 3.24. A la pràctica, quan treballem amb conjunts de dades, sovint no podem suposar a priori que es compleixen aquestes propietats d'independència. Aquest fet ha impulsat a molts autors a buscar generalitzacions d'aquest model amb menys estructura d'independència. Trobem a la literatura nombrosos treballs, com per exemple la distribució de Dirichlet escalada que s'obté d'eliminar a la propietat 3.23 el requeriment de components del vector $\mathbf{w}$ igualment escalades.

Definició 3.6 La *distribució de Dirichlet escalada* amb paràmetres $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}_+^D$, que simbolitzarem per $\mathcal{D}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$, és la distribució sobre $\mathcal{S}^D$ que té per funció de densitat

$$\mathrm{p}_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\Gamma\left(\sum_{i=1}^D \alpha_i\right) \prod_{i=1}^D \beta_i^{\alpha_i} x_i^{\alpha_i - 1}}{\prod_{i=1}^D \Gamma(\alpha_i) \left(\sum_{i=1}^D \beta_i x_i\right)^{\sum_{i=1}^D \alpha_i}}.$$

La *classe de distribucions de Dirichlet escalades* consisteix en el conjunt de totes les distribucions que s'obtenen en permetre que els vectors $\boldsymbol{\alpha}$ i $\boldsymbol{\beta}$ variïn sobre l'espai $\mathbb{R}_+^D$. □

Podem observar que les distribucions $\mathcal{D}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ són idèntiques a les distribucions $\mathcal{D}^D(\boldsymbol{\alpha})$ si i només si $\beta_1 = \beta_2 = \cdots = \beta_D$. Pel que fa a les propietats, aquesta generalització és una família tancada per la permutació de components i pel pas a subcomposicions, però no és tancada pel pas a amalgames. Com a propietat addicional, obtenim que les distribucions de Dirichlet escalades són tancades per l'operació pertorbació (Pawlowsky, comunicació personal).

Propietat 3.25 Sigui $\mathbf{x}$ una composició aleatòria $\mathcal{D}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Siguin $\mathbf{a} \in \mathcal{S}^D$ una composició constant, $\mathbf{S}$ una matriu de selecció $C \times D$ $(C < D)$, i $\mathbf{P}$ una matriu permutació $D \times D$. Llavors:

a. $\mathbf{x}^* \sim \mathcal{D}^D(\boldsymbol{\alpha}, \mathbf{a}^{-1}\boldsymbol{\beta})$, on $\mathbf{x}^* = \mathbf{a} \oplus \mathbf{x}$.

b. $\mathbf{s} \sim \mathcal{D}^C(\mathbf{S}\boldsymbol{\alpha}, \mathbf{S}\boldsymbol{\beta})$, on $\mathbf{s} = \mathcal{C}(\mathbf{S}\mathbf{x})$.

c. $\mathbf{x}_P \sim \mathcal{D}^D(\mathbf{P}\boldsymbol{\alpha}, \mathbf{P}\boldsymbol{\beta})$, on $\mathbf{x}_P = \mathbf{P}\mathbf{x}$.

□

La mateixa crítica que fèiem a la distribució de Dirichlet pot aplicar-se a aquesta generalització, ja que s'obté a partir de variables aleatòries independents. La distribució de Dirichlet generalitzada no aconsegueix suavitzar suficientment l'estructura d'independència, i per tant el seu àmbit d'aplicació també resulta fortament restringit.

Propietat 3.26 Sigui $\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_D)' \in \mathbb{R}_+^D$ un vector aleatori amb components w_i independents i distribuïdes segons un model $Ga(\alpha_i, \beta_i)$ $(i = 1, \ldots, D)$. Llavors la composició associada $\mathbf{x} = \mathcal{C}(\mathbf{w})$ es distribueix segons un model $\mathcal{D}^D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ i és independent de la variable $\sum_{i=1}^D \beta_i \mathrm{w}_i$, la qual es distribueix segons un model $Ga(\sum_{i=1}^D \alpha_i, 1)$. □

3.6 Altres distribucions

En aquest apartat volem tan sols esmentar altres distribucions que s'han definit sobre el símplex, l'estudi de les quals excedeix els objectius del nostre treball. Per aquest motiu, es consideren línies obertes per a futurs desenvolupaments.

La distribució de Connor-Mosimann (Connor i Mosimann, 1969) representa una altra generalització de la classe de Dirichlet. Fou definida a partir d'un vector aleatori de $\mathbb{R}_+^D$ amb components independents i amb distribució beta. Ometem aquí l'expressió de la seva funció de densitat ja que a la pràctica resulta bastant intractable i presenta encara una considerable estructura d'independència. Trobem altres generalitzacions de la distribució de Dirichlet en els treballs de Darroch i James (1974) i James i Mosimann (1980), amb les quals s'intenta reduir, també amb un èxit limitat, la independència de la classe de Dirichlet. En aquesta mateixa línia, trobem els treballs de Gupta i Richards (1987), i Rayens i Srinivasan (1994) on es fa un estudi exhaustiu de les distribucions de Liouville i de les distribucions de Liouville generalitzades. Dins la família de Liouville generalitzada, trobem densitats que admeten la dependència entre dues o més subcomposicions disjuntes d'una composició aleatòria. Tot i així, tenim una dificultat afegida, ja que aquestes densitats depenen d'una funció contínua escollida per l'usuari. Si donada una mostra de dades composicionals volem estimar els paràmetres d'una distribució de Liouville, caldrà escollir en primer lloc l'expressió de l'esmentada funció i aplicar tot seguit un mètode d'estimació dels paràmetres.

Barndorff-Nielsen i Jorgensen (1991) introdueixen un nou model sobre el símplex $\mathcal{S}^D$ utilitzant D variables aleatòries independents amb distribució inversa gaussiana generalitzada i condicionant per la seva suma. Demostren que aquest model conté la distribució de Dirichlet com un cas particular. Malgrat això, el model presenta també una forta estructura d'independència ja que una composició d'aquest tipus prové d'un vector de $\mathbb{R}_+^D$ amb components independents. Veiem doncs, que l'objectiu de trobar una classe de distribucions prou flexible i que contingui la classe de Dirichlet com a cas particular no s'ha acomplert del tot.

Pel que fa al model normal logístic additiu, hem vist que presenta unes bones propietats però en certs casos el model normal multivariant és poc adient per modelitzar conjunts de dades transformades, especialment si s'observa un cert biaix. En la secció 3.3 hem introduït una possible solució al problema mitjançant la distribució normal asimètrica logística additiva.

No obstant això, Aitchison (1986) proposa una altra possible solució utilitzant les transformacions Box-Cox. Aquesta alternativa ha estat àmpliament estudiada per Barceló-Vidal (1996). La idea és basa en trobar la tranformació Box-Cox (vegeu apartat 2.3.5) que millor ajusta les dades transformades a una distribució normal. Aquesta família de transformacions és contínua respecte de $\boldsymbol{\lambda}$ i podem observar que quan $\boldsymbol{\lambda} \to \mathbf{0}$, les transformacions Box-Cox tendeixen a la transformació logquocient additiva. Així doncs, les transformacions Box-Cox permeten ampliar la classe de les distribucions normals logístiques additives. Malauradament, les distribucions sobre el símplex que s'obtenen no tenen la simplicitat i l'elegància de les distribucions de classe $\mathcal{L}^{D-1}$, ja que existeix una gran dependència del denominador utilitzat en la transformació. Així per exemple, si $\left((\mathrm{x}_1/\mathrm{x}_3)^\lambda, (\mathrm{x}_2/\mathrm{x}_3)^\lambda\right)'$ segueix una distribució normal bivariant, llavors $(\mathrm{x}_1/\mathrm{x}_2)^\lambda$ no segueix un distribució normal. Trobem a més una altra dificultat teòrica afegida degut a que el rang de variació de la component i-èsima no és tota la recta real, sinó la semirecta $(-1/\lambda_i, +\infty)$.

Aitchison (1986, capítols 6 i 13) discuteix breument altres distribucions sobre el símplex. De la mateixa manera que introdueix la classe normal logística additiva, defineix la classe *normal logística multiplicativa*. Obtenim aquesta classe de distribucions aplicant la transformació logquocient multiplicativa (vegeu apartat 2.3.4) i modelant posteriorment amb una distribució normal multivariant. No estudiarem amb detall aquest model ja que li manquen les elegants propietats que tenen les distribucions de classe aln. En particular, aquestes distribucions depenen totalment de l'ordre en què estan disposades les parts d'una composició i, per tant, no obtenim una família tancada per la permutació de les components d'una composició aleatòria. Tot i això, cal destacar una propietat interessant: és possible descriure la distribució de l'amalgama $(\mathbf{x}^{(C)}, \mathbf{j}'\mathbf{x}_{(C)})'$.

Aitchison (1986) defineix també la classe de distribucions normals logístiques partides. Aquesta classe de distribucions es defineix mitjançant la transformació partida. La transformació partida és la composició de dues transformacions. La primera és una transformació del símplex $\mathcal{S}^D$ a l'espai producte $\mathcal{S}^2 \times \mathcal{S}^C \times \mathcal{S}^{D-C}$ que transforma una composició $\mathbf{x}$ amb D parts en el vector $(\mathbf{x}_A; \mathbf{s}_1, \mathbf{s}_2)'$, on $\mathbf{x}_A = (\mathbf{j}'\mathbf{x}^{(C)}, \mathbf{j}'\mathbf{x}_{(C)})'$, $\mathbf{s}_1 = \mathcal{C}(\mathbf{x}^{(C)})$ i $\mathbf{s}_2 = \mathcal{C}(\mathbf{x}_{(C)})$. La segona és una transformació logquocient. Aitchison (1986) proposa utilitzar la logquocient additiva o la logquocient multiplicativa. Aquestes ens transformen les subcomposicions i les amalgames en vectors de l'espai real. Quan el vector transformat es distribueix segons un model normal

multivariant llavors diem que la composició $\mathbf{x}$ té una distribució *normal logística partida*. Observem que apliquem logquocients a l'amalgama i a les subcomposicions separadament i treballem amb la suposició que el vector resultant és normal multivariant. Així doncs, podem tenir una dependència entre les subcomposicions i entre les subcomposicions i l'amalgama. Això és completament oposat a les distribucions de Dirichlet, on només era possible la independència. Podem a més generalitzar aquesta definició prenent C subcomposicions en comptes de dues.

Aitchison (1986) proposa també una generalització de la classe normal logística utilitzant la transformació logística híbrida, de $\mathbb{R}^D$ a $\mathcal{S}^D$. Aquesta transformació no és més que una mixtura entre les transformacions additiva i multiplicativa.

Finalment, Aitchison (1985) introdueix la classe de distribucions $\mathcal{A}^{D-1}(\alpha, B)$, conegudes com a *distribucions d'Aitchison*. Aquesta família de distribucions prové d'una mixtura de dues classes: la classe de distribucions de Dirichlet $\mathcal{D}^{D-1}$ i la classe normal logística additiva $\mathcal{L}^{D-1}$, ambdues incloses com a casos particulars. Aitchison (1986, capítol 13) estudia algunes propietats d'aquesta distribució però no aprofundeix en el seu estudi perquè demostra que és una distribució molt propera a les distribucions aln. Per altra banda aquest model és extremadament complicat i de difícil aplicació a la pràctica.

Capítol 4

Models paramètrics sobre $\mathcal{S}^D$. Metodologia STAY

Dediquem aquest capítol a l'anàlisi de diverses lleis de probabilitat a l'espai $\mathcal{S}^D$ segons la metodologia STAY. Es tracta d'un estudi original on es considera el símplex com un espai vectorial en ell mateix i es defineixen els models sobre les coordenades de la composició aleatòria en una base ortonormal. Per aquesta raó ens hi referim mitjançant l'abreviació STAY (en anglès romandre, restar). De cada llei estudiem l'expressió de la funció de densitat, les propietats algebraiques més importants i certs aspectes d'estimació de paràmetres.

Iniciem el capítol amb un exemple il·lustratiu d'aquesta metodologia en el cas univariant. Definim sobre la recta real positiva la distribució normal a $\mathbb{R}^+$. Habitualment es considera $\mathbb{R}^+$ com a subconjunt de $\mathbb{R}$, s'utilitza la geometria euclidiana habitual i la mesura de Lebesgue. Tot i així, sabem que $\mathbb{R}^+$ és un espai vectorial euclidià amb unes operacions, un producte escalar i una mesura propis i diferents als estàndards. Treballant amb les coordenades respecte d'una base ortonormal, definim la llei normal a $\mathbb{R}^+$, donem l'equació de la funció de densitat i l'expressió de l'esperança i la variància. Veiem que aquesta densitat es correspon amb la funció que dóna McAlister quan al 1879 introdueix la distribució lognormal. Per aquesta raó, realitzem un estudi comparatiu indicant les similituds i les diferències entre el model lognormal clàssic i el nou model normal a $\mathbb{R}^+$. Podem trobar part d'aquest treball a Mateu-Figueras et al. (2002) i a Mateu-Figueras i Pawlowsky-Glahn (2003). Tolosana Delgado et al. (2003) apliquen la distribució normal a $\mathbb{R}^+$ al krigeat de variable positives.

En el segon apartat del capítol, analitzem certs aspectes generals referents a les lleis de probabilitat sobre $\mathcal{S}^D$ definides segons la metodologia STAY. Seguidament, en el tercer i quart apartat introduïm la llei normal a $\mathcal{S}^D$ i la llei normal asimètrica a $\mathcal{S}^D$ mitjançant la funció de densitat sobre les coordenades de la composició aleatòria en una base ortonormal. Restringides al símplex, aquestes lleis de probabilitat coincideixen com a mesures de probabilitat amb les lleis que s'obtenen a partir dels models normal logístic additiu i normal asimètric logístic additiu, però les funcions de densitat corresponents no compleixen les mateixes propietats ni tenen els mateixos elements característics.

4.1 Exemple introductori: variables aleatòries normals a $\mathbb{R}^+$

Habitualment considerem $\mathbb{R}^+$ com un subconjunt de $\mathbb{R}$ i per tant, utilitzem la mateixa estructura algebraica. Això ens permet aplicar tots els procediments estadístics habituals de l'espai real. En particular, permet definir lleis de probabilitat mitjançant densitats respecte de la mesura de Lebesgue. Tot i que aquestes lleis estan definides a tota la recta real, la funció de densitat assigna el valor 0 a la semirecta $(-\infty, 0]$. En altres paraules, es considera un esdeveniment amb probabilitat nul·la. Tal i com hem indicat en capítols anteriors, una altra estratègia freqüentment utilitzada a la pràctica consisteix a recórrer a les transformacions. És a dir, aplicar a la variable una transformació de manera que el resultat sigui una variable aleatòria amb valors a tota la recta real. Seguidament, definir una funció de densitat respecte de la mesura de Lebesgue per la variable transformada. Per últim, aplicar el teorema del canvi de variable i obtenir la densitat de la variable original també respecte de la mesura de Lebesgue. En aquests casos, el jacobià del canvi de variable apareix sempre en l'expressió de la funció de densitat. La llei lognormal és un exemple típic de llei de probabilitat definida mitjançant la transformació logarítmica.

En aquest apartat de la tesi doctoral considerem $\mathbb{R}^+$ amb la seva pròpia estructura algebraica. Introduïm en primer lloc les operacions, el producte escalar i la distància que donen a $\mathbb{R}^+$ una estructura d'espai euclidià. Calculem també les coordenades d'un element respecte d'una base ortonormal. Cal dir que $\mathbb{R}^+$ té dimensió 1 i per tant, una base estarà formada per tan sols un element. Així doncs, en aquest cas és més adient utilitzar la terminologia de "base de norma 1"o "base unitària". A partir de la coordenada d'una variable respecte

d'aquesta base, definim la llei normal a $\mathbb{R}^+$, en calculem l'esperança, la variància i n'estudiem les propietats més importants. Finalment veiem que sobre la recta real positiva la llei normal a $\mathbb{R}^+$ i la llei lognormal són idèntiques, és a dir, assignen la mateixa probabilitat als esdeveniments de $\mathbb{R}^+$. No obstant això, veiem que compleixen propietats i tenen valors característics diferents.

4.1.1 Estructura algebraica de l'espai $\mathbb{R}^+$

Sabem que la recta real, amb les operacions suma i producte per escalars, té una estructura d'espai vectorial. Quan treballem a $\mathbb{R}^+$, aquesta estructura no és adequada ja que, si traslladem un vector o fem el producte per un escalar arbitrari de $\mathbb{R}$, podem obtenir resultats que no es troben a $\mathbb{R}^+$. Però, com recorden Pawlowsky-Glahn i Egozcue (2001), coneixem dues operacions que indueixen una estructura d'espai vectorial a $\mathbb{R}^+$. Donats $x, y \in \mathbb{R}^+$, l'operació interna, que té un paper anàleg a la suma de l'espai real, és el producte habitual $x \cdot y$. Donats $x \in \mathbb{R}^+$ i $\alpha \in \mathbb{R}$, l'operació externa, que té un paper anàleg al producte per escalars de l'espai real, és la potència x^α. És gairebé immediat comprovar que aquestes dues operacions indueixen a $\mathbb{R}^+$ una estructura d'espai vectorial sobre el cos $\mathbb{R}$.

Per altra banda, per a qualssevol $x, y \in \mathbb{R}^+$ podem definir un producte escalar com

$$\langle x, y \rangle_+ = \ln x \ln y,$$

que indueix la norma $\|x\|_+ = |\ln x|$, $\forall x \in \mathbb{R}^+$. Podem comprovar fàcilment que es compleixen les propietats habituals del producte escalar i de la norma.

A continuació podem introduir una estructura d'espai afí sobre $\mathbb{R}^+$ i una distància entre qualsevol parella de punts $x, y \in \mathbb{R}^+$ mitjançant l'expressió

$$d_+(x, y) = |\ln y - \ln x|. \tag{4.1}$$

Es pot comprovar com (4.1) compleix les propietats habituals de qualsevol distància i en concret les especificades a la definició 1.1. Això indica que és una distància compatible amb l'estructura vectorial de l'espai. Observem, però, que és una distància diferent a la distància euclidiana ordinària utilitzada a $\mathbb{R}$. El seu interès pràctic ja fou observat per Galton (1879) quan indicà que davant molts fenòmens socials o relatius a la vida és incorrecte afirmar que els errors de mesura per excés i per defecte calculats amb la distància euclidiana habitual es

produeixen amb la mateixa freqüència. F. Galton es refereix bàsicament a dades estrictament positives sobre les quals es compleix la llei de Fechner que, en la seva forma més simple, diu "sensació=log(estímul)". En l'argumentació que utilitza apareix un exemple concret: suposem que tenim un tint de color gris obtingut barrejant 8 porcions de blanc sobre negre. Si volem identificar aquest gris, tenim tanta probabilitat d'equivocar-nos i aparellar-lo amb un gris obtingut amb 16 porcions de blanc que amb un obtingut amb 4 porcions de blanc. En el primer cas cometem un error per excés i en el segon un error per defecte i l'experiència ens diu que ambdós errors es produeixen amb la mateixa freqüència. Malgrat això, la distància euclidiana ordinària entre els valors 8 i 16 és diferent a la distància entre els valors 4 i 8. En canvi, si prenem la distància (4.1) obtenim el mateix valor:

$$d_+(8,16) = |\ln 16 - \ln 8| = |\ln 8 - \ln 4| = d_+(4,8).$$

D'aquest raonament es conclou que no podem tractar aquest tipus de dades com si fossin elements de $\mathbb{R}$ sinó que cal considerar-les dins $\mathbb{R}^+$ amb l'estructura definida anteriorment. Pel que fa a la mesura, veiem que no té sentit considerar la mesura de Lebesgue i, com a conseqüència, serà incoherent definir lleis de probabilitat mitjançant la derivada de Radon-Nikodým respecte de la mesura de Lebesgue. Davant això i per evitar qualsevol incompatibilitat amb l'estructura de $\mathbb{R}^+$, proposem treballar amb les coordenades respecte d'una base de norma 1.

Amb les operacions, el producte escalar i la norma abans definits, la teoria d'àlgebra lineal ens assegura l'existència d'una base ortonormal. En aquest cas concret, l'espai $\mathbb{R}^+$ té dimensió 1 i per tant la base estarà formada per tan sols un vector de norma 1. Tenim únicament dues possibilitats: o bé el vector e o bé el seu invers respecte de l'operació interna definida abans. En aquesta secció i sense perdre generalitat considerarem tan sols el vector unitari e com a base unitària o canònica de $\mathbb{R}^+$.

Donat que habitualment considerem $\mathbb{R}^+$ com un subconjunt de la recta real, expressem els seus elements en termes de la base canònica de $\mathbb{R}$. Certament, qualsevol $x \in \mathbb{R}^+$ es pot escriure com $x = x1$, i per tant diem que x és la component del vector x respecte de la base 1. No obstant això, dins l'estructura definida a $\mathbb{R}^+$, el vector 1 no és ni tan sols una base sinó que es tracta de l'element neutre de l'espai, en altres paraules, un vector de norma 0 i ortogonal a qualsevol altre. Per altra banda, l'expressió $x = x1$ no representa una combinació

lineal ja que l'operació producte per escalars no és l'operació externa de $\mathbb{R}^+$. Utilitzant la base unitària de $\mathbb{R}^+$ i l'operació potència podem escriure $x = e^{\ln x}$, i per tant diem que $\ln x$ és la coordenada del vector x respecte de la base e.

En l'apartat 1.1 hem comprovat que si treballem amb les coordenades respecte d'una base ortonormal podem aplicar les operacions clàssiques de l'espai real, així com el producte escalar ordinari i la distància euclidiana habitual. Certament, en el nostre cas és immediat veure que les operacions producte i potència es corresponen amb la suma i el producte per escalars dels respectius logaritmes:

$$\begin{aligned} x \cdot y &= e^{\ln x} e^{\ln y} = e^{\ln x + \ln y} \qquad \forall x, y \in \mathbb{R}^+, \\ x^\alpha &= (e^{\ln x})^\alpha = e^{\alpha \ln x} \qquad \forall x \in \mathbb{R}^+, \forall \alpha \in \mathbb{R}. \end{aligned}$$

Observem que el resultat d'aplicar les operacions suma i producte per escalars als coeficients respecte de la base e és també un coeficient respecte de la mateixa base. Així doncs, si volem recuperar l'element de $\mathbb{R}^+$ haurem d'aplicar la operació inversa: l'exponencial. Aquest procediment s'utilitza reiteradament a la pràctica ja que es difícil interpretar un resultat en termes del seu logaritme.

També podem aplicar el producte escalar ordinari, la norma i la distància euclidiana habitual sobre els logaritmes ja que $\forall x, y \in \mathbb{R}^+$,

$$\begin{aligned} \langle x, y \rangle_+ &= \ln x \ln y = \langle \ln x, \ln y \rangle_{eu}, \\ \|x\|_+ &= |\ln x| = \sqrt{(\ln x)^2} = \|\ln x\|_{eu}, \\ d_+(x, y) &= |\ln y - \ln x| = \sqrt{(\ln y - \ln x)^2} = d_{eu}(\ln x, \ln y). \end{aligned} \tag{4.2}$$

Observem que aquests resultats no són vectors de $\mathbb{R}^+$, es tracta tant sols d'escalars i per tant no caldrà prendre exponencials per interpretar els resultats.

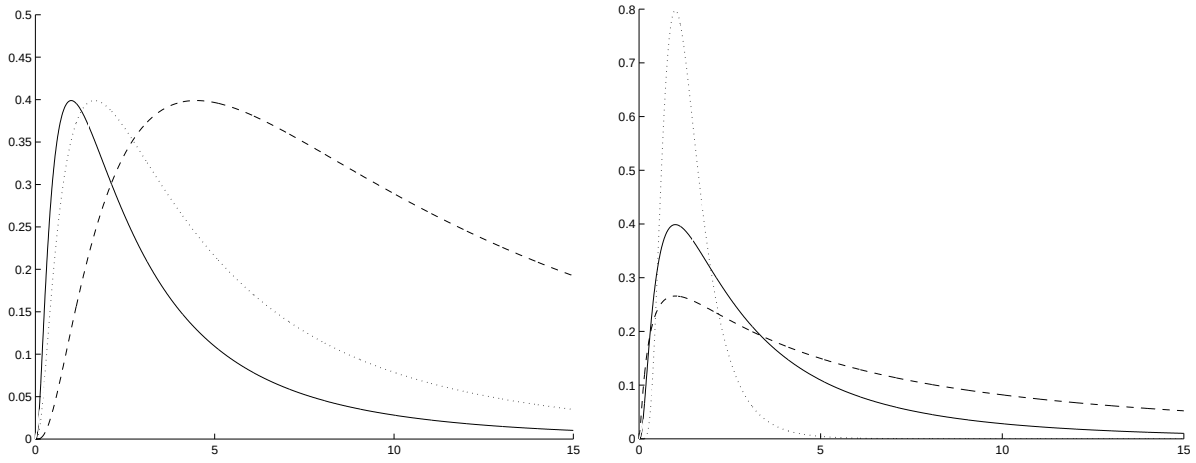
4.1.2 Distribució normal a $\mathbb{R}^+$

Utilitzant l'estructura algebraica geomètrica introduïda sobre l'espai real positiu, podem definir una llei de probabilitat que anomenarem llei normal a $\mathbb{R}^+$. La definició es du a terme mitjançant la funció de densitat de les coordenades d'una variable respecte de la base de norma 1 de l'espai $\mathbb{R}^+$.

Definició 4.1 Sigui $(\Omega, \mathcal{F}, p)$ un espai de probabilitat i $\mathrm{x} : \Omega \longrightarrow \mathbb{R}^+$ una variable aleatòria $\mathbb{R}^+$-valuada. Direm que x té una distribució *normal a* $\mathbb{R}^+$ amb paràmetres μ i σ, i ho denotarem com $\mathrm{x} \sim \mathcal{N}^+(\mu, \sigma)$, si la seva funció de densitat és

$$f^+(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right), \qquad x \in \mathbb{R}^+. \tag{4.3}$$

□



(a) (b)

Figura 4.1: *Funcions de densitat* (a) $\mathcal{N}^+(0,1)$ (——), $\mathcal{N}^+(0.5,1)$ (.....), $\mathcal{N}^+(1.5,1)$ (- - - - -), (b) $\mathcal{N}^+(0,1)$ (——), $\mathcal{N}^+(0,0.5)$ (.....) *i* $\mathcal{N}^+(0,1.5)$ (- - - -).

Aquesta funció de densitat està absolutament restringida a l'espai real positiu. Això indica que el valor 0 i els reals negatius són esdeveniments impossibles que ara no s'identifiquen amb esdeveniments de probabilitat nul·la. L'expressió (4.3) coincideix amb la funció anomenada "law of frequency"proposada per McAlister (1879) quan va introduir la llei lognormal. Tot i així, difereix del que habitualment es coneix com a funció de densitat d'una llei lognormal. Observem a més que (4.3) ens recorda a la densitat d'una variable normal clàssica de la recta real. Per aquesta raó i per les propietats que veurem tot seguit, l'hem anomenada distribució normal a $\mathbb{R}^+$. A la figura 4.1 hem representat la densitat (4.3) per a diferents valors dels paràmetres. En la figura 4.1(a) hem variat el paràmetre μ deixant fix el paràmetre σ i en la figura 4.1(b) hem variat el paràmetre σ deixant fix el paràmetre μ.

La funció (4.3) és la densitat de les coordenades de la variable aleatòria x respecte de la base e, és a dir, la derivada de Radon-Nikodým de la probabilitat respecte de la mesura

de Lebesgue. Per tant, podem tractar-la com la densitat d'una variable aleatòria real. Així doncs, per obtenir la probabilitat dins un interval (a, b) amb $0 < a < b \in \mathbb{R}^+$ qualssevol, caldrà fer el càlcul de la integral ordinària de la funció (4.3) entre els valors $\ln a$ i $\ln b$, és a dir, entre les coordenades de a i b respecte de la base unitària,

$$P(a < x < b) = \int_{\ln a}^{\ln b} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) d\lambda(\ln x). \tag{4.4}$$

Per trobar qualsevol percentil, per exemple el valor de la mediana, realitzarem el càlcul invers: buscarem el valor Md de manera que $P(\mathrm{x} < \mathrm{Md}) = 1/2$. És a dir

$$\int_{-\infty}^{\ln \mathrm{Md}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) d\lambda(\ln x) = \frac{1}{2}. \tag{4.5}$$

Aquesta expressió és la mateixa que obtenim en calcular la mediana d'una variable aleatòria normal a l'espai real i per tant sabem que $\ln \mathrm{Md} = \mu$. Aquest resultat ens indica que la coordenada de la mediana respecte de la base e és el paràmetre μ. Per tant, el valor de la mediana l'obtindrem prenent l'exponencial, és a dir, $\mathrm{Md} = e^{\mu}$.

També podem justificar aquesta densitat directament a partir de conceptes de teoria de la mesura a l'espai real positiu. Si tenim en compte els consells de Galton (1879) i la distància (4.1), veurem que la mesura adequada en el nostre espai és absolutament contínua respecte de la mesura de Lebesgue i podem definir-la a partir de la seva derivada de Radon-Nikodým respecte de λ, és a dir,

$$d\nu = \frac{1}{x} d\lambda.$$

Així, per exemple, la mesura d'un interval qualsevol (a, b) amb $0 < a < b \in \mathbb{R}^+$ es pot calcular com

$$\nu(a, b) = \int_a^b \frac{1}{x} d\lambda(x) = \ln b - \ln a = \ln \frac{b}{a}.$$

Si volem especificar una mesura de probabilitat sobre la recta real positiva podem utilitzar, entre altres eines, la funció de densitat de la probabilitat respecte d'una mesura de $\mathbb{R}^+$. Precisament la densitat (4.3) introduïda en la definició 4.1 és la derivada de Radon-Nikodým o densitat de probabilitat respecte de la mesura ν. Així doncs, segons la teoria de la mesura, podríem calcular la probabilitat dins l'interval (a, b) amb $0 < a < b$ com

$$P(a < \mathrm{x} < b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) d\nu(x).$$

És difícil realitzar aquest càlcul ja que no es tracta d'una integral ordinària. Tot i així, podem convertir-la en una integral ordinària utilitzant la densitat de la mesura ν respecte de la mesura de Lebesgue. És a dir

$$P(a < \mathrm{x} < b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) \frac{1}{x} d\lambda(x), \tag{4.6}$$

i la probabilitat que en resulta és exactament la mateixa que obteníem treballant amb els logaritmes en l'expressió (4.4). Observem també que la funció dins la integral (4.6) és igual a la funció de densitat d'una llei lognormal amb paràmetres μ i σ. De fet, és la funció que McAlister (1879) anomena "law of facility". Arribem doncs a la conclusió que la llei normal a $\mathbb{R}^+$ i la llei lognormal a $\mathbb{R}$ són la mateixa llei de probabilitat si restringim la segona a la recta real positiva. Reservem el següent apartat per comparar més a fons aquestes dues lleis de probabilitat. Centrem-nos ara en l'estudi de les propietats d'aquesta nova distribució.

Propietat 4.1 Sigui $\mathrm{x} \sim \mathcal{N}^+(\mu, \sigma)$, $a \in \mathbb{R}^+$ un vector constant i $\beta \in \mathbb{R}$ un escalar. Llavors, la variable aleatòria $\mathrm{x}^* = a \cdot \mathrm{x}^\beta$ té una distribució $\mathcal{N}^+(\ln a + \beta\mu, \beta^2\sigma)$. □

La demostració d'aquesta propietat és immediata si treballem amb les components de les variables respecte de la base ortonormal i apliquem la propietat de les transformacions lineals per a variables aleatòries reals.

Propietat 4.2 Sigui x una variable aleatòria normal a $\mathbb{R}^+$ amb funció de densitat f_{x}^+ i sigui $a \in \mathbb{R}^+$ un vector constant. Aleshores compleix la igualtat $f_{a\cdot\mathrm{x}}^+(ax) = f_{\mathrm{x}}^+(x)$ on $f_{a\cdot\mathrm{x}}^+$ representa la funció de densitat de la variable aleatòria $a \cdot \mathrm{x}$.

Demostració. Per la propietat 4.1 sabem que $a \cdot \mathrm{x} \sim \mathcal{N}^+(\ln a + \mu, \sigma)$ i per tant, a partir de l'expressió de les respectives funcions de densitat, tenim que

$$\begin{aligned} f_{a\cdot\mathrm{x}}^+(ax) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln(ax) - (\ln a + \mu)}{\sigma}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) = f_{\mathrm{x}}^+(x). \end{aligned}$$

□

Observem que la distribució normal sobre la recta real compleix una propietat totalment equivalent amb l'operació suma, l'operació interna de $\mathbb{R}$. És a dir, si $\mathrm{x} \sim \mathcal{N}(\mu, \sigma)$ amb funció

de densitat f_{x} i a $\in \mathbb{R}$, aleshores $f_{\mathrm{x}}(x) = f_{a+\mathrm{x}}(a+x)$ on $f_{a+\mathrm{x}}$ representa la funció de densitat de la variable aleatòria traslladada $a + \mathrm{x}$. No obstant això, observem que a $\mathbb{R}$ no es compleix la igualtat $f_{\mathrm{x}}(x) = f_{a\cdot\mathrm{x}}(a \cdot x)$ ni a $\mathbb{R}^+$ la igualtat $f^+_{\mathrm{x}}(x) = f^+_{a+\mathrm{x}}(a+x)$.

Propietat 4.3 Sigui $\mathrm{x} \sim \mathcal{N}^+(\mu, \sigma)$. Llavors es compleix que $\mathrm{E[x]} = \mathrm{Md(x)} = e^{\mu}$.

Demostració. Mitjançant l'expressió (4.5) hem calculat la mediana d'una variable aleatòria $\mathrm{x} \sim \mathcal{N}^+(\mu, \sigma)$ i hem obtingut el valor e^{μ}. Així doncs, només queda demostrar que E[x] té el mateix valor. Sabem que l'esperança és un element de l'espai suport. Si apliquem la definició clàssica d'esperança als logaritmes, obtindrem les coordenades de E[x] respecte de la base unitària. El càlcul per realitzar és

$$\int_{-\infty}^{+\infty} \ln x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) d\lambda(\ln x).$$

Obtenim com a resultat el paràmetre μ ja que l'expressió anterior es correspon amb l'esperança d'una variable aleatòria normal de l'espai $\mathbb{R}$. Com a últim pas i per obtenir el valor de E[x], només ens cal prendre l'exponencial. Així doncs $\mathrm{E[x]} = e^{\mu}$. □

Propietat 4.4 Si $\mathrm{x} \sim \mathcal{N}^+(\mu, \sigma)$, llavors $\mathrm{var[x]} = \sigma^2$.

Demostració. La variància d'una variable aleatòria real es defineix com $\mathrm{var[y]} = \mathrm{E[(y - E[y])^2]}$. No obstant això, podem interpretar-la com el valor esperat de la distància ordinària al quadrat al voltant del centre de masses E[y]. És a dir, $\mathrm{var[y]} = \mathrm{E}[d^2_{eu}(\mathrm{y}, \mathrm{E[y]})]$. Aquesta interpretació ja fou utilitzada per Pawlowsky-Glahn i Egozcue (2001, 2002) per definir la "variància mètrica", la variància d'un vector aleatori sobre un espai de Hilbert de dimensió finita qualsevol. En aquesta direcció i donada $\mathrm{x} \sim \mathcal{N}^+(\mu, \sigma)$, definim la seva variància com $\mathrm{var[x]} = \mathrm{E}[d^2_+(\mathrm{x}, \mathrm{E[x]})]$ ja que d_+ és una distància coherent a $\mathbb{R}^+$. L'altra possibilitat és treballar amb les coordenades respecte de la base de norma 1 i definir $\mathrm{var[x]} = \mathrm{E}[d^2_{eu}(\ln x, \mathrm{E}[\ln x])]$. Per la propietat 4.3 i per la igualtat (4.2), les dues expressions són idèntiques: $\mathrm{var[x]} = \mathrm{E}[d^2_+(\mathrm{x}, \mathrm{E[x]})] = \mathrm{E}[d^2_{eu}(\ln x, \mathrm{E}[\ln x])]$. El seu resultat és igual a σ^2 donat que la segona igualtat es correspon amb la variància d'una variable aleatòria normal a $\mathbb{R}$ amb paràmetres μ i σ. □

És important mencionar que el valor σ^2 no és un element de l'espai suport. Es tracta tan sols d'un valor numèric que descriu la dispersió de la variable, i per tant no cal prendre

l'exponencial. Un valor molt sovint utilitzat a la pràctica és l'arrel quadrada de la variància, valor que anomenem desviació estàndard.

Com a conseqüència d'aquests dos darrers resultats, l'expressió d'un interval centrat en l'esperança i de longitud $2k\sigma$ és igual a

$$(e^{\mu-k\sigma}, e^{\mu+k\sigma}). \tag{4.7}$$

Certament, podem comprovar que el punt mig de l'interval és e^{μ} i que

$$d_+(e^{\mu-k\sigma}, e^{\mu+k\sigma}) = 2k\sigma.$$

Podem arribar a la mateixa expressió prenent l'exponencial dels extrems de l'interval calculat a partir de les coordenades respecte de la base e i utilitzant la distribució normal estàndard a $\mathbb{R}$. Si busquem en la literatura, trobarem l'aplicació pràctica dels intervals (4.7). Tenim per exemple el treball de Ahrens (1954), que obté intervals de predicció a $\mathbb{R}^+$ prenent l'exponencial dels intervals calculats a partir dels logaritmes de les dades.

Mitjançant l'expressió de la funció de densitat d'una variable aleatòria normal a $\mathbb{R}^+$, podem comprovar fàcilment que els intervals (4.7) són d'isodensitat. Per tant, arribem a la conclusió que la distribució normal a $\mathbb{R}^+$ és simètrica al voltant de l'esperança e^{μ}. Aquesta simetria pot semblar paradoxal ja que, certament, no la observem en el perfil de la funció de densitat. Cal, però, pensar aquesta simetria dins $\mathbb{R}^+$ amb l'estructura que hem construït a l'apartat 4.1.1 i no euclidianament, és a dir, imaginant $\mathbb{R}^+$ com a subconjunt de $\mathbb{R}$ amb l'estructura habitual. A la figura 4.2 hem representat l'interval $(\exp(\mu-\sigma), \exp(\mu+\sigma))$ sobre una funció de densitat $\mathcal{N}^+(\mu, \sigma)$ amb $\mu = 0$ i $\sigma = 1$. Observem com efectivament en els extrems de l'interval la funció de densitat pren el mateix valor.

De la mateixa manera que hem demostrat les propietats 4.1, 4.2, 4.3 i 4.4, podem demostrar que la llei normal a $\mathbb{R}^+$ compleix exactament les mateixes propietats que la llei normal clàssica de l'espai real. Un aspecte important que es deriva d'aquest fet és que podem trobar estimadors consistents i intervals de confiança exactes per a l'esperança d'una variable aleatòria x $\sim \mathcal{N}^+(\mu, \sigma)$. Suposem $x_1, x_2, \ldots, x_n$ una mostra de la variable x. El procediment que seguirem per trobar les estimacions de l'esperança serà treballar amb les coordenades de les dades respecte de la base de norma 1, és a dir, amb els valors $y_1, y_2, \ldots, y_n$ on $y_i = \ln x_i$ per a $i = 1, 2, \ldots, n$. Sobre aquestes coordenades, aplicarem la teoria de la llei normal a

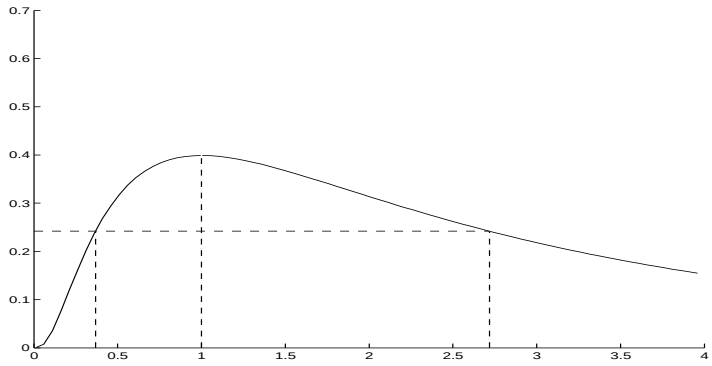


Figura 4.2: *En línia discontínua, extrems i punt mig de l'interval* $(e^{\mu-\sigma}, e^{\mu+\sigma})$ *sobre la funció de densitat* $\mathcal{N}^+(\mu = 0, \sigma = 1)$.

l'espai real, que ens diu que la mitjana aritmètica $\bar{y}$ és un estimador consistent i que l'interval $\bar{y} \pm z_{\alpha/2} S_y^*/\sqrt{n}$ és l'interval al $(1-\alpha)100\%$ de confiança per a l'esperança, amb S_y^* la desviació estàndard corregida (amb denominador $n-1$) dels logaritmes. Tot seguit, prendrem l'exponencial d'aquests termes perquè l'esperança és un element de l'espai suport $\mathbb{R}^+$. Així doncs, l'estimador consistent de l'esperança d'una variable aleatòria normal a $\mathbb{R}^+$ serà $\exp(\bar{y})$ i un interval al $(1-\alpha)100\%$ de confiança per a l'esperança serà

$$\left(e^{\bar{y} - z_{\alpha/2} \frac{S_y^*}{\sqrt{n}}}, e^{\bar{y} + z_{\alpha/2} \frac{S_y^*}{\sqrt{n}}} \right).$$

Ara bé, si ens fixem amb la justificació de la distribució lognormal que fa Galton (1879), veiem que proposa la mitjana geomètrica com a mesura del centre de masses d'un conjunt de dades estrictament positives en substitució de l'habitual mitjana aritmètica. Observem, però, que aquesta afirmació és una altra manera d'indicar el que acabem de veure, ja que la mitjana geomètrica de les dades originals es correspon amb l'exponencial de la mitjana aritmètica de les dades logtransformades:

$$\left(\prod_{i=1}^{n} \mathrm{x}_i \right)^{1/n} = e^{1/n \sum_{i=1}^{n} \ln x_i} = e^{\bar{y}}.$$

Cal anar amb compte amb les tècniques d'inferència estadística i amb les representacions gràfiques habituals de dades donat que moltes d'elles han estat desenvolupades suposant l'estructura pròpia de l'espai real. Si les apliquem directament sobre qualsevol mostra d'una variable aleatòria positiva per a la qual té més sentit l'estructura pròpia de $\mathbb{R}^+$, podem

obtenir resultats o interpretacions errònies. Per evitar qualsevol problema, ens caldrà aplicar aquestes tècniques estadístiques estàndards a les mostres logtransformades. No obstant això, en certs casos podrem modificar fàcilment el procediment estadístic o la representació gràfica i adaptar-los al nostre espai $\mathbb{R}^+$. Quan això sigui possible, podrem treballar directament amb les dades originals de l'espai real positiu. Per exemple, en representar l'histograma d'unes dades dividim el rang de valors de la variable en intervals de la mateixa amplada o distància euclidiana. Si treballem amb una variable aleatòria a $\mathbb{R}^+$ i tenim en compte l'estructura algebraica de $\mathbb{R}^+$, no té sentit representar l'histograma habitual d'una mostra de la variable ja que no té sentit calcular una distància amb l'aplicació d_{eu}. Tanmateix podem representar l'histograma habitual de la mostra logtransformada, és a dir, l'histograma dels coeficients en la base unitària de la mostra. Una altra possibilitat és modificar l'histograma habitual i adaptar-lo al nostre espai $\mathbb{R}^+$. La modificació és senzilla, només cal calcular l'amplitud dels intervals amb la distància d_+ que és coherent amb l'estructura pròpia de $\mathbb{R}^+$. En aquest cas podrem representar l'histograma de les dades originals.

4.1.3 Comparació amb la distribució lognormal

Abans hem indicat que podem interpretar $\mathbb{R}^+$ com un subconjunt de $\mathbb{R}$ i utilitzar l'estructura habitual de $\mathbb{R}$. En aquest context, recordem que una variable aleatòria positiva x té una distribució *lognormal a* $\mathbb{R}$ amb paràmetres μ i σ, si la variable $\mathrm{y} = \ln \mathrm{x}$ té una distribució normal a $\mathbb{R}$ amb mitjana μ i variància σ^2. Habitualment s'utilitza la notació $\mathrm{x} \sim \Lambda(\mu, \sigma)$ i la funció de densitat de la llei de probabilitat és

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) & x > 0, \\ 0 & x \leq 0. \end{cases} \tag{4.8}$$

Si comparem aquesta densitat amb la densitat d'una llei normal a $\mathbb{R}^+$, trobem certes diferències. En primer lloc, observem que (4.8) inclou un cas per a valors $x \leq 0$. Aquest fet pot semblar en principi paradoxal ja que la variable hauria d'estar totalment restringida a l'espai real positiu. No obstant això, aquest cas és necessari ja que estem interpretant $\mathbb{R}^+$ com a subconjunt de $\mathbb{R}$ i per tant estem considerant implícitament la variable x com una variable aleatòria real. Per donar consistència a aquest plantejament, la densitat lognormal (4.8) assigna el valor 0 a la semirecta $(-\infty, 0]$, és a dir, identifica els esdeveniments impos-

sibles amb els esdeveniments de probabilitat nul·la. L'altra diferència que podem observar comparant les densitats lognormal i normal a $\mathbb{R}^+$ és el terme $1/x$. Aquest terme és el jacobià del canvi de variable i és necessari per poder aplicar l'anàlisi real estàndard. En particular, per calcular probabilitats a partir de la integral ordinària de la funció (4.8). A la figura 4.3 hem representat les funcions de densitat d'una llei lognormal a $\mathbb{R}$ i d'una llei normal a $\mathbb{R}^+$ sobre un mateix eix. Observem com les dues densitats difereixen considerablement.

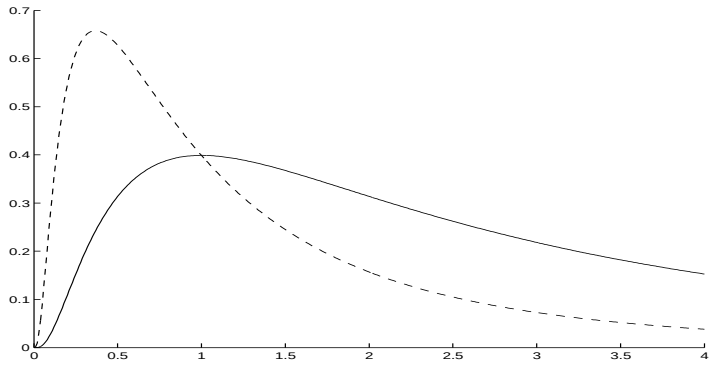


Figura 4.3: *Funcions de densitat* $\Lambda(0,1)$ (- - - -) *i* $\mathcal{N}^+(0,1)$ (——).

Tot i les diferències indicades en l'expressió de les funcions de densitat, les igualtats (4.4) i (4.6) ens indiquen que les dues lleis assignen la mateixa probabilitat als subconjunts de $\mathbb{R}^+$. En el cas de la llei normal a $\mathbb{R}^+$ treballem amb els coeficients de la variable respecte de la base unitària e. En aquesta situació, recordem que la probabilitat d'un interval (a,b) amb $0 < a < b$ qualssevol es calcula segons l'expressió (4.4); és a dir, mitjançant la integral ordinària de la densitat (4.3) entre els valors $\ln a$ i $\ln b$. En el cas de la llei lognormal a $\mathbb{R}$, estem considerant $\mathbb{R}^+$ com un subconjunt de $\mathbb{R}$ i la funció (4.8) és una densitat clàssica. Així doncs, la probabilitat dins un interval (a,b) amb $a < b$ qualssevol és igual a (4.6); és a dir, calculem la integral ordinària de la funció de densitat lognormal entre els valors a i b. La probabilitat resultant és igual en ambdós casos sempre i quan $0 < a < b$, tot i que, en el primer cas, només té sentit calcular probabilitats per a valors $0 < a < b$ mentre que en el segon ho podem fer per a valors $a < b$ sense cap altra restricció. La igualtat entre les probabilitats implica necessàriament la igualtat en tots els percentils. Així doncs, els percentils d'una variable normal a $\mathbb{R}^+$ coincideixen amb els corresponents percentils d'una

variable lognormal a $\mathbb{R}$. Recordem per exemple que la mediana d'una variable $\mathrm{x} \sim \mathcal{N}^+(\mu, \sigma)$ és igual a e^μ, valor que coincideix amb la mediana d'una variable $\mathrm{x} \sim \Lambda(\mu, \sigma)$. Arribem doncs a la conclusió que, restringides a la recta real positiva, les dues lleis de probabilitat són la mateixa.

Pel que fa a coincidències, observem que la llei lognormal a $\mathbb{R}$ compleix la propietat 4.1 de la llei normal a $\mathbb{R}^+$.

Però, sens dubte, la distribució lognormal a $\mathbb{R}$ no compleix les mateixes propietats que la distribució normal a $\mathbb{R}^+$. En primer lloc, la família lognormal no és tancada per l'operació interna de l'espai, la translació en aquest cas, ja que el resultat és la densitat lognormal amb tres paràmetres. Per altra banda no es compleix la igualtat que obteníem per la distribució normal a $\mathbb{R}^+$, és a dir, $f_{a\cdot\mathrm{x}}(ax) \neq f_{\mathrm{x}}(x)$, amb $f_{a\cdot\mathrm{x}}$ i f_{x} funcions de densitats de les variables lognormals $a \cdot \mathbf{x}$ i $\mathbf{x}$ respectivament. Tampoc el valor esperat ni la variància d'una llei lognormal coincideixen amb els respectius valors en el cas normal a $\mathbb{R}^+$. Recordem que per a una variable aleatòria $\mathrm{x} \sim \Lambda(\mu, \sigma)$ obtenim

$$\begin{aligned} \mathrm{E}[\mathrm{x}] =& e^{\mu+\sigma^2/2}, \\ \mathrm{var}[\mathrm{x}] =& e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right). \end{aligned}$$

(vegeu Aitchison i Brown, 1957 o Crow i Shimizu, 1988). Sabem, a més, que la densitat lognormal presenta una forta asimetria a la dreta. Per aquesta raó la mitjana sempre té un valor superior a la mediana. Per altra banda, si considerem a $\mathbb{R}^+$ la mateixa estructura que a l'espai real, utilitzarem l'expressió $(\mathrm{E}[\mathrm{x}] - k\mathrm{std}[\mathrm{x}], \mathrm{E}[\mathrm{x}] + k\mathrm{std}[\mathrm{x}])$ per calcular l'interval centrat a l'esperança de la variable x i de longitud $2k$ desviacions. En aquest cas la notació std[x] indica la desviació estàndard de la variable. Malgrat tot, aquests tipus d'intervals no s'utilitzen a la pràctica ja que, en certs casos l'extrem inferior pren valors negatius i l'interval perd el sentit. Per exemple, en el cas de la lognormal amb $\mu = 0$ i $\sigma = 1$, obtenim que l'interval centrat en l'esperança i de radi una desviació estàndard és igual a $(-0.512, 3.810)$. Per aquesta raó, s'utilitzen els intervals indicats en la secció anterior per al cas normal a $\mathbb{R}^+$. És a dir, donada $\mathrm{x} \sim \Lambda(\mu, \sigma)$ es pren l'interval $(e^{\mu-k\sigma}, e^{\mu+k\sigma})$ (vegeu Ahrens, 1954). Tot i ser utilitzats a la pràctica, aquests intervals no es consideren òptims ja que no són intervals d'isodensitat ni tampoc tenen longitud mínima. A la figura 4.4 hem representat l'interval $(e^{\mu-\sigma}, e^{\mu+\sigma})$ sobre una funció de densitat lognormal amb paràmetres $\mu = 0$ i $\sigma = 1$. Podem

observar clarament que en els extrems de l'interval la funció no pren els mateixos valors.

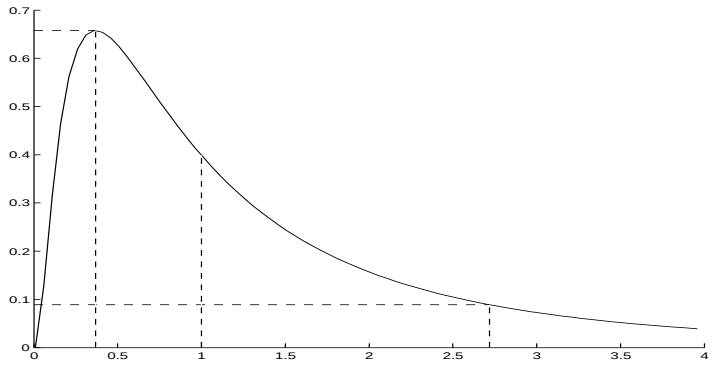


Figura 4.4: *En línia discontínua, punt e^{μ} i extrems de l'interval $(e^{\mu-\sigma}, e^{\mu+\sigma})$ sobre la funció de densitat* $\Lambda(\mu = 0, \sigma = 1)$.

L'altra gran diferència entre el model normal a $\mathbb{R}^+$ i el model lognormal a $\mathbb{R}$ fa referència als estimadors. És complicat calcular un estimador consistent i intervals de confiança exactes per a l'esperança d'una variable lognormal. Trobem en la literatura un gran nombre de procediments destinats al càlcul d'estimadors, però molts d'ells són de difícil aplicació perquè requereixen d'extenses taules o bé aporten només resultats aproximats. Donat que l'esperança d'una variable aleatòria lognormal és $\exp(\mu + \sigma^2/2)$, un estimador puntual raonable podria ser $\exp(\bar{y} + S_y^*/2)$, on $\bar{y}$ i S_y^* representen respectivament la mitjana aritmètica i la desviació estàndard corregida dels logaritmes de les dades. No obstant això, aquest estimador no és òptim donat que presenta un cert biaix. Existeixen estudis que demostren que l'estimador eficient per a la mitjana d'una variable lognormal és $\mathrm{t} = \exp(\bar{y})\gamma_n(V)$, anomenat estimador de Sichel. La funció $\gamma_n(V)$ que apareix en l'expressió depèn de la mida (n) de la mostra i de la variància logarítmica (V) i actua com a factor corrector del biaix. Diferents autors han calculat taules del terme $\gamma_n(V)$. Podem trobar-les per exemple a Clark i Harper (2000). Tot i ser un estimador eficient i àmpliament utilitzat a la pràctica (vegeu Krige, 1981 o Rendu, 1981), cal anar amb compte ja que esdevé terriblement inestable quan la variància logarítmica mostral és superior a 3. Existeixen també intervals del $(1-\alpha)100\%$ de confiança per a l'estimador t. La seva expressió és $(\mathrm{t}\Psi_{\alpha/2}, \mathrm{t}\Psi_{1-\alpha/2})$ on els termes $\Psi_{\alpha/2}$ i $\Psi_{1-\alpha/2}$ depenen de la mida de la mostra, de la variància logarítmica mostral així com del nivell de confiança $1-\alpha$. Existeixen també altres mètodes per calcular estimacions de l'esperança, però en

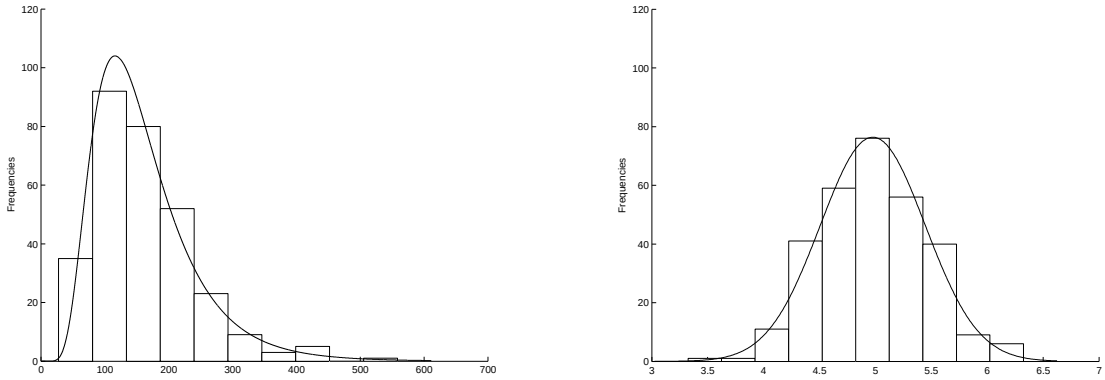
general donen lloc a estimacions no òptimes. Podem anomenar, entre d'altres, el mètode dels moments, el mètode dels quantils, el mètode de Mood, o bé un mètode gràfic, tots ells descrits a Aitchison i Brown (1957) o bé a Koch i Link (1980).

Per il·lustrar les diferències entre els estimadors obtinguts amb un model lognormal i amb un model normal a $\mathbb{R}^+$, hem simulat una mostra de 300 dades representant mides de pous de petroli expressades en milers de barrils. Davis (1986) introdueix aquesta variable com un exemple modelable sovint amb una llei lognormal. L'objectiu està en calcular una estimació puntual i una estimació al 90% de confiança per a l'esperança de la variable.

Si decidim que les diferències entre les dades són absolutes, estarem en el context de la llei lognormal. En aquest cas esperarem obtenir en mitjana 161.93 milers de barrils d'un pou de petroli. També, obtenim que l'esperança estarà compresa dins l'interval $(151.83, 174.73)$ amb probabilitat 0.9. L'estimació puntual i l'interval de confiança per a l'esperança s'han calculat utilitzant el mètode de Sichel i les taules de Clark i Harper (2000).

Si decidim que les diferències entre les dades són relatives, estarem en el context de la llei normal a $\mathbb{R}^+$ i calcularem les estimacions a partir de les coordenades en la base unitària e, és a dir, a partir dels logaritmes de les dades. L'exponencial de la mitjana aritmètica dels logaritmes és 145.04 i l'exponencial dels extrems de l'interval al 90% de confiança calculat amb els logaritmes és $(138.70, 151.68)$. Observem que en aquest últim cas obtenim valors molt més "conservadors". No obstant això, cal tenir en compte que aquests resultats no són del tot comparables. En el cas lognormal obtenim una estimació del valor $\exp(\mu + \sigma^2/2)$ mentre que en el cas normal a $\mathbb{R}^+$ obtenim una estimació del valor $\exp(\mu)$. Observem, doncs, que la valoració del mètode més correcte passa per decidir quina estructura de $\mathbb{R}^+$ considerem més apropiada, $\mathbb{R}^+$ com a subconjunt de $\mathbb{R}$ o bé $\mathbb{R}^+$ com a espai vectorial, decisió validable tan sols a partir de la realitat.

Per il·lustrar el procediment que cal seguir quan considerem l'estructura d'espai vectorial de $\mathbb{R}^+$ i per emfatitzar més les diferències entre els dos models, els hem ajustat a les dades simulades. En ambdós casos hem obtingut $\hat{\mu} = 4.977$ i $\hat{\sigma} = 0.470$ com a estimacions dels paràmetres ja que es calculen a partir de la mitjana i de la desviació estàndard de les dades logtransformades. Per fer-nos una primera idea intuïtiva de la bondat d'ajust del model a les dades, podem comparar el perfil dels histogrames amb la corba de la funció de densitat ajustada. Recordem que en el context del model lognormal, podem aplicar qualsevol tècnica



(a) (b)

Figura 4.5: *Mostra simulada mida 300.* (a) *Histograma amb la densitat lognormal ajustada.* (b) *Histograma de les dades logtransformades amb la densitat normal ajustada.*

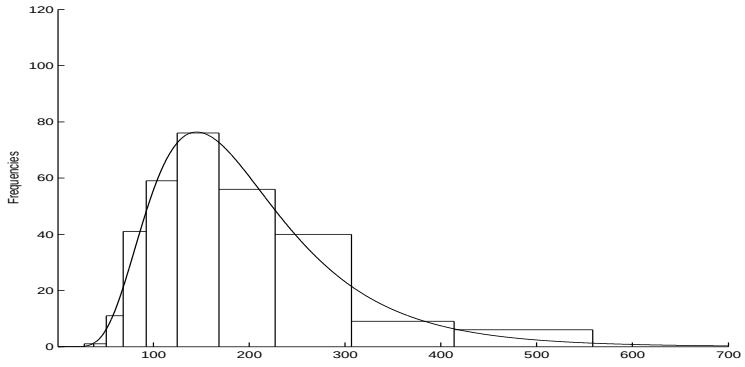
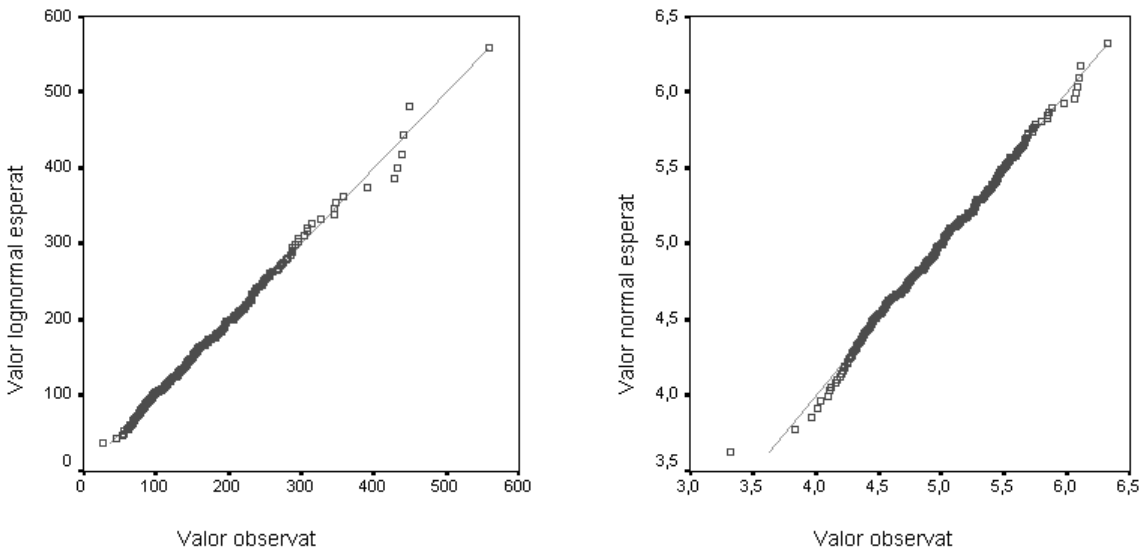


Figura 4.6: *Mostra simulada mida 300. Histograma amb la densitat normal a* $\mathbb{R}^+$ *ajustada.*

estadística estàndard en particular, podem utilitzar l'histograma clàssic de les dades. A la figura 4.5(a) hem representat l'histograma de la mostra simulada amb la densitat lognormal $\Lambda(\mu = 4.977, \sigma = 0.470)$ superposada. Observem que l'ajust és bastant raonable.

Volem remarcar que per al model normal a $\mathbb{R}^+$ no té sentit representar l'histograma habitual de les dades originals ja que en aquest gràfic es divideix l'eix OX en intervals amb la mateixa distància euclidiana d_{eu}. Si hem decidit utilitzar la distribució normal a $\mathbb{R}^+$ és perquè considerem que la distància entre dos punts és relativa, és a dir, calculable a partir de (4.2), i per tant no és coherent dibuixar intervals d'igual longitud euclidiana. La solució que adoptem

és dibuixar l'histograma habitual de les coordenades de les dades respecte de la base e. Així doncs, per valorar intuïtivament la bondat d'ajust del model normal a $\mathbb{R}^+$ a les nostres dades caldrà comparar el perfil de l'histograma de la mostra logtransformada amb la densitat normal clàssica ajustada. Mitjançant el gràfic 4.5(b) podem comprovar que l'ajust sembla també raonable. Observem, però, que aquest darrer gràfic també és vàlid per valorar intuïtivament l'ajust d'un model lognormal a les dades, ja que una variable lognormal és aquella el logaritme de la qual té una distribució normal a $\mathbb{R}$. Per aquesta raó, en ambdós casos arribem a la mateixa conclusió. Tal i com hem apuntat en l'apartat anterior podem adaptar l'histograma a l'estructura de $\mathbb{R}^+$ i treballar amb la mostra original. Tan sols cal dibuixar l'histograma però dividint l'eix OX amb intervals de la mateixa distància d_+. D'aquesta manera podem valorar intuïtivament la bondat d'ajust comparant també el perfil d'aquest histograma amb la corba de la funció de densitat normal a $\mathbb{R}^+$ ajustada (vegeu figura 4.6). La conclusió és la mateixa que l'obtinguda amb les coordenades respecte de la base unitària. Observem, doncs, que l'histograma que cal representar depèn de l'estructura de $\mathbb{R}^+$ que considerem més adequada.



(a) (b)

Figura 4.7: *Mostra simulada mida 300.* (a) *Diagrama quantil-quantil del model lognormal aplicat a les dades originals.* (b) *Diagrama quantil-quantil del model normal a* $\mathbb{R}^+$ *aplicat a les coordenades en la base* e.

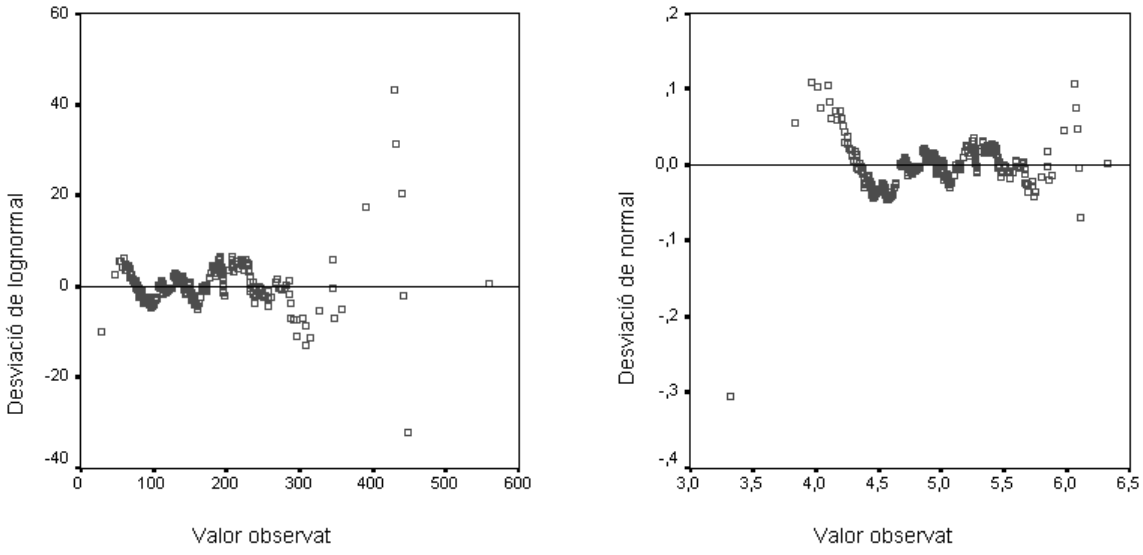
Una tècnica utilitzada sovint per valorar gràficament la bondat d'ajust d'una distribució a unes dades són els diagrames quantil-quantil, coneguts també com diagrames Q-Q. Es tracta d'un diagrama bivariant on es representen els quantils mostrals vers els quantils teòrics, és a dir, els quantils que esperaríem tenir si les dades provinguessin realment de la distribució suposada. Per trobar aquests quantils teòrics, cal primer calcular amb les dades ordenades el percentil que representa cada una. Seguidament, i mitjançant la inversa de la funció de distribució suposada, podem trobar els valors teòrics que representen el mateix percentil. Si el model és apropiat, observarem en el gràfic una tendència lineal.

En la figura 4.7(a) hem representat el gràfic Q-Q del model lognormal que proporciona directament el paquet estadístic SPSS. Observem que el núvol de punts segueix la línia central del gràfic a excepció dels quantils amb un valor més elevat, que se n'allunyen lleugerament. La conclusió final és que el model lognormal és adequat. Cal notar que valorem l'ajust del núvol de punts a la recta des d'un punt de vista euclidià, és a dir, valorem a ull si el núvol de punts es troba a prop o lluny de la recta. En la figura 4.7(b) hem representat el gràfic Q-Q corresponent al model normal a $\mathbb{R}^+$. Per ser coherents amb les recomanacions donades a la secció 4.1.2 treballem amb les coordenades respecte de la base unitària, és a dir, representem els quantils dels logaritmes de les dades vers els quantils esperats segons un model normal clàssic. Observem en el gràfic 4.7(b) que el núvol de punts també s'ajusta a la recta i arribem a la conclusió que el model normal a $\mathbb{R}^+$ és apropiat. És important remarcar que aquest segon diagrama Q-Q també seria aplicable al model lognormal. En aquest cas, és més complicat adaptar els diagrames Q-Q a l'estructura pròpia de $\mathbb{R}^+$, ja que ens caldria valorar intuïtivament l'ajust del núvol de punts a la recta amb la mètrica definida a $\mathbb{R}^+$. I això és difícil donat que els nostres ulls miren i mesuren les distàncies euclidianament.

Podem valorar gràficament la bondat d'ajust del model observant també els gràfics de la figura 4.8. Aquests gràfics s'anomenen diagrames Q-Q sense tendència; en ells hi ha representades les distàncies euclidianes ordinàries entre els punts dels gràfics Q-Q anteriors i la línia central. Per al model lognormal estem suposant $\mathbb{R}^+$ com a subconjunt de $\mathbb{R}$ amb la mateixa estructura, per tant és coherent utilitzar la distància euclidiana. En el cas del model normal a $\mathbb{R}^+$, només té sentit utilitzar la distància euclidiana entre les coordenades de les dades respecte de la base unitària. Així doncs, podem representar el diagrama Q-Q sense tendència del model normal a $\mathbb{R}^+$ si utilitzem la mostra logtransformada. No obstant això, en

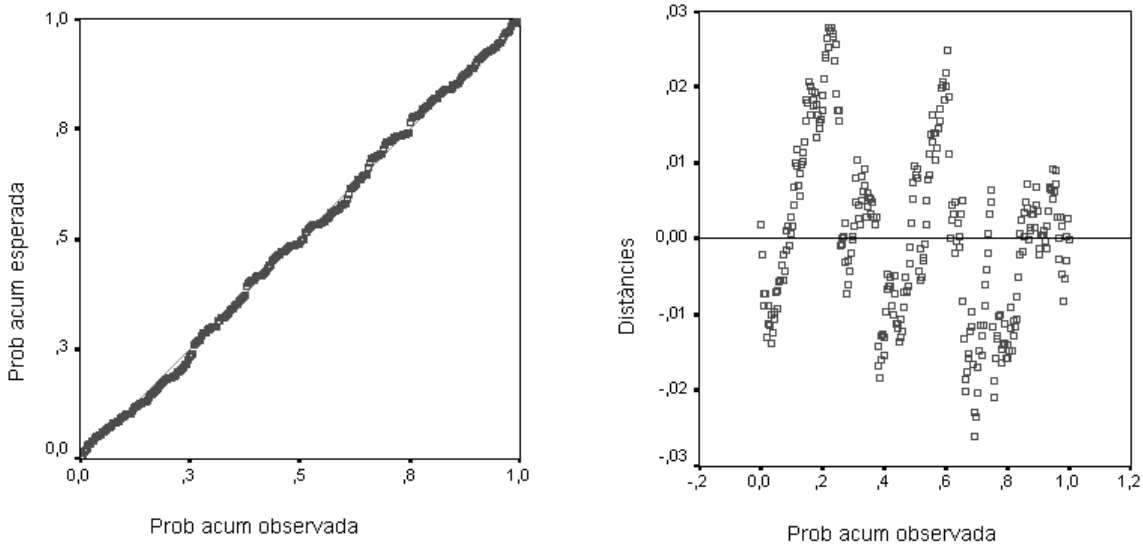
aquest cas podem adaptar fàcilment els diagrames Q-Q sense tendència a l'estructura pròpia de $\mathbb{R}^+$ per així treballar amb les dades originals. Tan sols cal representar les distàncies calculades segons l'aplicació d_+ entre els punts i la recta del diagrama Q-Q. Si realitzem aquests càlculs obtindrem un diagrama Q-Q sense tendència equivalent al diagrama 4.8(b) ja que la distància euclidiana entre coordenades en la base e és igual a la distància d_+ entre els elements de $\mathbb{R}^+$.

Una tècnica de bondat d'ajust molt semblant als diagrames Q-Q són els anomenats diagrames P-P. La idea és bàsicament la mateixa però en comptes de representar quantils es representen probabilitats acumulades, és a dir, es dibuixen els percentils empírics vers els percentils teòrics que s'obtindrien si les dades provinguessin realment de la distribució suposada. En la figura 4.9 hem representat el diagrama P-P i el diagrama P-P sense tendència de les dades originals suposant un model lognormal. Obtindríem exactament el mateix gràfic per al model normal a $\mathbb{R}^+$ ja que els dos models defineixen la mateixa llei sobre $\mathbb{R}^+$ i per tant les probabilitats calculades són exactament les mateixes.



(a) (b)

Figura 4.8: *Mostra simulada mida 300.* (a) *Diagrama quantil-quantil sense tendència del model lognormal aplicat a les dades originals.* (b) *Diagrama quantil-quantil sense tendència del model normal a* $\mathbb{R}^+$ *aplicat a les coordenades en la base e.*



(a) (b)

Figura 4.9: *Mostra simulada mida 300.* (a) *Diagrama P-P.* (b) *Diagrama P-P sense tendència.*

4.2 Distribucions a $\mathcal{S}^D$. Aspectes generals

La definició general de derivada de Radon-Nikodým d'una probabilitat és valida en qualsevol espai de mesura i, en particular, sobre $\mathcal{S}^D$. Així doncs, serà correcte definir una llei de probabilitat mitjançant la seva funció de densitat respecte d'una mesura, sempre i quan aquesta mesura sigui adequada sobre el símplex. Malgrat això, poden aparèixer certes dificultats que afecten bàsicament al càlcul de probabilitats i dels elements que requereixen del càlcul integral. La raó és senzilla: només sabem fer el càlcul efectiu d'integrals de funcions reals respecte de la mesura de Lebesgue (i de qualsevol altra mesura absolutament contínua respecte d'ella). Certament, donada una composició aleatòria $\mathbf{x}$ amb densitat de probabilitat $f(\cdot)$ respecte d'una mesura ν de $\mathcal{S}^D$, podem escriure la probabilitat d'un esdeveniment A qualsevol com

$$P(A) = \int_A f(\mathbf{x}) d\nu(\mathbf{x}),$$

però no sabem fer efectiu aquest càlcul.

Com que $\mathcal{S}^D$ té una estructura d'espai vectorial amb un producte escalar, podem evitar aquestes dificultats utilitzant la propietat 1.1 que assegura un isomorfisme entre $\mathcal{S}^D$ i $\mathbb{R}^{D-1}$.

Per fer ús d'aquest isomorfisme només cal identificar qualsevol element de l'espai amb les seves coordenades respecte d'una base ortonormal. Així doncs, la metodologia que aplicarem per definir lleis de probabilitat és la mateixa que la de l'apartat anterior però en el cas multivariant. És a dir, introduirem la funció de densitat de les coordenades de la composició aleatòria respecte d'una base ortonormal de $\mathcal{S}^D$. Tal i com hem vist al capítol 2, podem tractar aquestes coordenades com elements de l'espai $\mathbb{R}^{D-1}$ i aplicar tota la teoria estàndard. En particular, podem definir la funció de densitat habitual, és a dir, la funció de densitat de les coordenades respecte de la mesura de Lebesgue de $\mathbb{R}^{D-1}$. Aquesta funció ens permetrà obtenir la probabilitat d'un esdeveniment A calculant la integral ordinària sobre les coordenades del conjunt A respecte de la mateixa base ortonormal. És a dir, si $f^*(\cdot)$ és la funció de densitat de probabilitat de les coordenades, podem calcular la probabilitat de l'esdeveniment $A \subseteq \mathcal{S}^D$ com

$$P(A) = \int_{A^*} f^*(v_1, v_2, \ldots, v_{D-1}) dv_1 dv_2 \ldots dv_{D-1},$$

on A^* i $(v_1, v_2, \ldots, v_{D-1})$ representen les coordenades respecte d'una base ortonormal de $\mathcal{S}^D$ que caracteritzen al conjunt A i la composició $\mathbf{x}$.

Cal tenir en compte que si utilitzem aquesta metodologia per calcular qualsevol element de l'espai suport $\mathcal{S}^D$, obtindrem les coordenades d'aquest element respecte de la base ortonormal. Tot seguit, podrem recuperar la composició original a partir d'una simple combinació lineal a $\mathcal{S}^D$.

En els dos apartats següents definim lleis de probabilitat sobre $\mathcal{S}^D$ indicant l'expressió de les funcions f^*. En ambdós casos, les famílies són tancades per les operacions pertorbació i potència. Les seves funcions de densitat dels coeficients compleixen, a més, la igualtat

$$f^*_{\mathbf{x}}(\mathbf{x}) = f^*_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x}), \tag{4.9}$$

on $f^*_{\mathbf{x}}$ i $f^*_{\mathbf{a}\oplus\mathbf{x}}$ representen les densitats de les composicions aleatòries $\mathbf{x}$ i $\mathbf{a}\oplus\mathbf{x}$ respectivament, amb $\mathbf{a}$ composició constant. Aquesta propietat ens informa que podem imaginar la funció de densitat d'una composició aleatòria pertorbada com el resultat d'aplicar la mateixa pertorbació a la corba de la densitat de la composició aleatòria inicial. Això té conseqüències importants quan fem estadística a $\mathcal{S}^D$ ja que sovint apliquem l'operació de centrar una composició (vegeu Martín-Fernández et al., 1999).

En cada cas és possible realitzar el càlcul de l'esperança i la matriu de covariàncies utilitzant també les coordenades respecte d'una base ortonormal i els procediments estàndards de l'espai real. Els resultats que s'obtenen són coherents amb el centre i la variància mètrica d'una composició aleatòria que s'han definit a l'apartat 2.4.

4.3 Distribució normal a $\mathcal{S}^D$

4.3.1 Definició i propietats

Definició 4.2 Sigui $(\Omega, \mathcal{F}, p)$ un espai de probabilitat i $\mathbf{x} : \Omega \longrightarrow \mathcal{S}^D$ una composició aleatòria. Direm que $\mathbf{x}$ té una distribució *normal a* $\mathcal{S}^D$ amb paràmetres $\boldsymbol{\xi}$ i $\boldsymbol{\Upsilon}$, si la funció de densitat de les coordenades en una base ortonormal de $\mathcal{S}^D$ és

$$f^*_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\xi}, \boldsymbol{\Upsilon}) = (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)\right], \qquad (4.10)$$

on $\text{ilr}(\mathbf{x})$ representa el vector de coordenades de $\mathbf{x}$ respecte de la base ortonormal escollida. □

Escriurem $\mathbf{x} \sim \mathcal{N}^D_{\mathcal{S}}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ per indicar que la composició $\mathbf{x}$ segueix un model normal a $\mathcal{S}^D$. El subíndex $\mathcal{S}$ indica que es tracta d'un model en el símplex i el superíndex D mostra el nombre de parts de la composició aleatòria. Hem utilitzat la notació $\boldsymbol{\xi}$ i $\boldsymbol{\Upsilon}$ per als paràmetres del model ja que es corresponen amb el vector d'esperances i la matriu de covariàncies del vector $\text{ilr}(\mathbf{x})$.

Volem insistir que la densitat (4.10) és la densitat dels coeficients o coordenades de $\mathbf{x}$ respecte d'una base ortonormal, i per tant és la derivada de Radon-Nikodým de la probabilitat respecte de la mesura de Lebesgue en l'espai $\mathbb{R}^{D-1}$ de coeficients. Això permet calcular la probabilitat d'un esdeveniment qualsevol a partir d'una integral ordinària. És a dir, si suposem A un esdeveniment de $\mathcal{S}^D$, llavors la seva probabilitat serà

$$P(A) = \int_{A^*} (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)\right] d\lambda(\text{ilr}(\mathbf{x})),$$

on A^* representa les coordenades del conjunt A respecte de la base ortonormal considerada.

Un aspecte important que volem remarcar és que la llei normal a $\mathcal{S}^D$ coincideix, sobre $\mathcal{S}^D$, amb la llei normal logística additiva introduïda a l'apartat 3.2 i definida mitjançant

transformacions. Recordem que la densitat (3.2) és la densitat de probabilitat del model aln en la parametrització $\boldsymbol{\xi}$ i $\boldsymbol{\Upsilon}$. En aquest cas es considera $x_D = 1 - (\sum_{i=1}^{D-1} x_i)$ i el vector $\text{ilr}(\mathbf{x})$ és el resultat d'aplicar la transformació logquocient isomètrica a la composició $\mathbf{x}$. És per aquesta raó que observem el jacobià de la transformació en la funció de densitat. Recordem també que la densitat (3.2) és la derivada de Radon-Nikodým de la probabilitat respecte de la mesura de Lebesgue a l'espai $\mathbb{R}^{D-1}$, l'espai imatge per la transformació. Per tant, podem calcular la probabilitat de l'esdeveniment $A \subset \mathcal{S}^D \subset \mathbb{R}^D$ com

$$P(A) = \int_A \frac{D^{-1/2}\left(\prod_{i=1}^D x_i\right)^{-1}}{(2\pi)^{(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{1/2}} \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)\right] dx_1 dx_2 \cdots dx_{D-1}. \tag{4.11}$$

Donat que es tracta també d'una integral ordinària, podem fer efectiu el seu càlcul aplicant la metodologia estàndard.

Al llarg d'aquest treball d'investigació, hem denotat per $\text{ilr}(\mathbf{x})$ el vector de coordenades de la composició $\mathbf{x}$ respecte de la base ortonormal ja que la seva expressió coincideix amb el vector resultant d'aplicar la transformació logquocient isomètrica a la mateixa composició. Per demostrar que la llei normal logística additiva i la llei normal a $\mathcal{S}^D$ coincideixen sobre $\mathcal{S}^D$, és important que distingim clarament entre el vector de coordenades respecte de la base i el vector resultant d'aplicar la transformació. Per aquesta raó i per evitar possibles confusions, denotarem les coordenades respecte de la base ortonormal de la composició $\mathbf{x}$ com $\mathbf{v} = (v_1, v_2, \ldots, v_{D-1})'$. D'aquesta manera podem escriure la probabilitat de l'esdeveniment A segons el model normal a $\mathcal{S}^D$ com

$$P(A) = \int_{A^*} (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{v} - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\mathbf{v} - \boldsymbol{\xi}\right)\right] dv_1 dv_2 \cdots dv_{D-1}, \tag{4.12}$$

on A^* representa les coordenades de A en la base ortonormal.

En ambdós casos, les expressions (4.11) i (4.12) són integrals de funcions reals respecte de la mesura de Lebesgue i podem utilitzar els procediments estàndards del càlcul integral a l'espai real. En particular, podem aplicar el teorema del canvi de variable. Prenem doncs, l'expressió (4.12) i apliquem el canvi $\mathbf{v} = \text{ilr}(\mathbf{x})$, el jacobià del qual és $D^{-1/2}\left(\prod_{i=1}^D x_i\right)^{-1}$. El teorema del canvi de variable assegura la igualtat

$$P(A) = \int_{A^*} (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{v} - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\mathbf{v} - \boldsymbol{\xi}\right)\right] dv_1 dv_2 \cdots dv_{D-1}$$

$$= \int_{\text{ilr}^{-1}(A^*)} \frac{D^{-1/2}\left(\prod_{i=1}^{D} x_i\right)^{-1}}{(2\pi)^{(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{1/2}} \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)\right] dx_1 dx_2 \cdots dx_{D-1}.$$

De la mateixa manera que les coordenades de $\mathbf{x}$ respecte d'una base ortonormal coincideixen amb el vector resultant d'aplicar la transformació logquocient isomètrica a $\mathbf{x}$, les coordenades de l'esdeveniment A també coincideixen amb l'esdeveniment resultant d'aplicar-li la transformació ilr, i per tant $\text{ilr}^{-1}(A^*) = A$. Obtenim, doncs, que les lleis de probabilitat definides amb els models normal a $\mathcal{S}^D$ i normal logístic additiu són la mateixa sobre $\mathcal{S}^D$. No obstant això, veurem que difereixen en les propietats i en els valors característics.

Podem observar que la densitat (4.10) es correspon amb la densitat d'un vector normal multivariant a l'espai real. Per aquesta raó, l'hem anomenada llei normal a $\mathcal{S}^D$. A continuació, veurem que compleix les mateixes propietats que un model normal clàssic.

Per coherència amb tot el treball d'investigació, tornem a recuperar la notació $\text{ilr}(\mathbf{x})$ per indicar les coordenades de $\mathbf{x}$ respecte de la base ortonormal de $\mathcal{S}^D$.

Propietat 4.5 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$. Sigui $\mathbf{a} \in \mathcal{S}^D$ una composició constant i b un escalar del cos $\mathbb{R}$. Llavors la composició $\mathbf{x}^* = \mathbf{a} \oplus (b \otimes \mathbf{x})$ es distribueix segons una llei $\mathcal{N}_{\mathcal{S}}^{D}(\text{ilr}(\mathbf{a}) + b\boldsymbol{\xi}, b^2\boldsymbol{\Upsilon})$.

Demostració. A partir de la definició de $\mathbf{x}^*$ i de les propietats dels coeficients en una base ortonormal, es dedueix que $\text{ilr}(\mathbf{x}^*) = \text{ilr}(\mathbf{a}) + b\text{ilr}(\mathbf{x})$. Així doncs, les coordenades de la composició $\mathbf{x}^*$ s'obtenen mitjançant una transformació lineal de les coordenades de la composició $\mathbf{x}$. Podem tractar la funció de densitat de les coordenades de $\mathbf{x}$ com una densitat a l'espai real, i per tant obtindrem la densitat de $\mathbf{x}^*$ mitjançant un simple canvi de variable. També, i donat que l'expressió de la densitat de les coordenades de $\mathbf{x}$ coincideix amb una densitat normal clàssica, podem aplicar la propietat de les transformacions lineals del model normal a l'espai real. En ambdós casos, obtenim que la densitat de les coordenades de la composició $\mathbf{x}^*$ és

$$\begin{aligned} f^*_{\mathbf{x}^*}(\mathbf{x}^*) = {} & (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} b^{-1} \\ & \times \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}^*) - (\text{ilr}(\mathbf{a}) - b\boldsymbol{\xi})\right)' b^{-2}\boldsymbol{\Upsilon}^{-1}\left(\text{ilr}(\mathbf{x}) - (\text{ilr}(\mathbf{a}) - b\boldsymbol{\xi})\right)\right], \end{aligned}$$

i, per tant, es conclou que $\mathbf{x}^* \sim \mathcal{N}_{\mathcal{S}}^{D}(\text{ilr}(\mathbf{a}) + b\boldsymbol{\xi}, b^2\boldsymbol{\Upsilon})$. □

Observem que els paràmetres de la distribució de $\mathbf{x}^*$ són $\text{ilr}(\mathbf{a}) + b\boldsymbol{\xi}$ i $b^2\boldsymbol{\Upsilon}$, valors que es corresponen amb l'esperança i la matriu de covariàncies del vector de coordenades $\text{ilr}(\mathbf{x}^*)$ ja que

$$\begin{aligned}\text{E}[\text{ilr}(\mathbf{x}^*)] &= \text{E}[\text{ilr}(\mathbf{a}) + b\text{ilr}(\mathbf{x})] = \text{ilr}(\mathbf{a}) + b\text{E}[\text{ilr}(\mathbf{x})] = \text{ilr}(\mathbf{a}) + b\boldsymbol{\xi},\\ \text{var}[\text{ilr}(\mathbf{x}^*)] &= \text{var}[\text{ilr}(\mathbf{a}) + b\text{ilr}(\mathbf{x})] = b^2\text{var}[\text{ilr}(\mathbf{x})] = b^2\boldsymbol{\Upsilon}.\end{aligned}$$

Propietat 4.6 Sigui $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ i $\mathbf{a} \in \mathcal{S}^D$ una composició constant. Aleshores es compleix la igualtat $f^*_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x}) = f^*_{\mathbf{x}}(\mathbf{x})$, on $f^*_{\mathbf{a}\oplus\mathbf{x}}$ i $f^*_{\mathbf{x}}$ representen les funcions de densitat de les composicions aleatòries $\mathbf{x}$ i $\mathbf{a} \oplus \mathbf{x}$ respectivament.

Demostració. Per la propietat anterior sabem que $\mathbf{a} \oplus \mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^D(\text{ilr}(\mathbf{a}) + \boldsymbol{\xi}, \boldsymbol{\Upsilon})$ i per tant, a partir de l'expressió de les respectives funcions de densitat, tenim que

$$\begin{aligned}f^*_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x}) &= (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2}\\ &\quad \times \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{a} \oplus \mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi})\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{a} \oplus \mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi})\right)\right]\\ &= (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2}\\ &\quad \times \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{a}) + \text{ilr}(\mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi})\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{a}) + \text{ilr}(\mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi})\right)\right]\\ &= (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)\right] = f^*_{\mathbf{x}}(\mathbf{x}),\end{aligned}$$

tal i com s'indicava a (4.9). □

Propietat 4.7 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$. Sigui $\mathbf{x}_P = \mathbf{P}\mathbf{x}$ la composició $\mathbf{x}$ amb les components reordenades per la matriu permutació $\mathbf{P}$. Llavors $\mathbf{x}_P$ es distribueix segons un model $\mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\xi}_P, \boldsymbol{\Upsilon}_P)$ amb

$$\boldsymbol{\xi}_P = \mathbf{U}'\mathbf{P}\mathbf{U}\boldsymbol{\xi} \qquad \text{i} \qquad \boldsymbol{\Upsilon}_P = (\mathbf{U}'\mathbf{P}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}'\mathbf{P}\mathbf{U})',$$

on $\mathbf{U}$ és la matriu d'ordre $D \times (D-1)$ les columnes de la qual són les coordenades clr dels vectors d'una base ortonormal de $\mathcal{S}^D$.

Demostració. Per obtenir la distribució de la composició aleatòria $\mathbf{x}_P$ en funció de la distribució de $\mathbf{x}$, és necessari trobar la relació matricial entre les coordenades ilr de les dues composicions $\mathbf{x}_P$ i $\mathbf{x}$. Si treballem amb les coordenades clr, és a dir, amb les coordenades de les composicions respecte del sistema de referència B^*, és immediat veure que $\text{clr}(\mathbf{x}_P) = \mathbf{P}\text{clr}(\mathbf{x})$.

L'expressió (2.21) ens dóna la relació entre les coordenades clr i ilr. Així doncs, si $\text{ilr}(\mathbf{x})$ i $\text{ilr}(\mathbf{x}_P)$ representen les coordenades de les composicions $\mathbf{x}$ i $\mathbf{x}_P$ respecte de la base ortonormal, és clar que la relació entre elles és $\text{ilr}(\mathbf{x}_P) = (\mathbf{U}'\mathbf{P}\mathbf{U})\,\text{ilr}(\mathbf{x})$. Aplicant el teorema del canvi de variable o la propietat de les transformacions lineals de la llei normal a l'espai real, obtenim que la densitat de les coordenades de la composició $\mathbf{x}_P$ és igual a la densitat d'un model $\mathcal{N}_{\mathcal{S}}^D(\mathbf{U}'\mathbf{P}\mathbf{U}\boldsymbol{\xi}, (\mathbf{U}'\mathbf{P}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}'\mathbf{P}\mathbf{U})')$. □

Observem una altra vegada que els paràmetres del model resultant coincideixen amb l'esperança i la matriu de covariàncies del vector de coordenades $\text{ilr}(\mathbf{x}_P)$.

Propietat 4.8 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$. Sigui $\mathbf{s} = \mathcal{C}(\mathbf{S}\mathbf{x})$ una subcomposició amb parts seleccionades mitjançant la matriu $\mathbf{S}$, d'ordre $C \times D$. Llavors la subcomposició $\mathbf{s}$ es distribueix segons una llei $\mathcal{N}_{\mathcal{S}}^C(\boldsymbol{\xi}_S, \boldsymbol{\Upsilon}_S)$ amb

$$\boldsymbol{\xi}_S = \mathbf{U}^{*\prime}\mathbf{S}\mathbf{U}\boldsymbol{\xi} \qquad \text{i} \qquad \boldsymbol{\Upsilon}_S = (\mathbf{U}^{*\prime}\mathbf{S}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}^{*\prime}\mathbf{S}\mathbf{U})',$$

on $\mathbf{U}$ és la matriu d'ordre $D \times (D-1)$ les columnes de la qual són les coordenades clr d'una base ortonormal de $\mathcal{S}^D$ i $\mathbf{U}^*$ és la matriu d'odre $C \times (C-1)$ amb columnes les coordenades clr d'una base ortonormal de $\mathcal{S}^C$.

Demostració. Sabem que el vector de coordenades respecte de la base B de $\mathcal{S}^C$ de la composició $\mathbf{s}$ i el vector de coordenades respecte de la base B de $\mathcal{S}^D$ de la composició $\mathbf{x}$ coincideixen amb els seus vectors alr transformats. Aquestes coordenades compleixen la relació que dóna Aitchison (1986, pàg. 119) i per tant, $\text{alr}(\mathbf{s}) = \left(\mathbf{F}_{C-1,C}\mathbf{S}\mathbf{F}^*_{D,D-1}\right)\text{alr}(\mathbf{x})$, on $\mathbf{F}$ i $\mathbf{F}^*$ són les matrius descrites al capítol 2, els subíndex de les quals indiquen les seves respectives dimensions. Seguidament, utilitzant les relacions (2.21) entre els vectors de coordenades alr i ilr, obtenim que $\text{ilr}(\mathbf{s}) = \left(\mathbf{U}^{*\prime}\mathbf{F}^*_{C,C-1}\mathbf{F}_{C-1,C}\mathbf{S}\mathbf{F}^*_{D,D-1}\mathbf{F}_{D-1,D}\mathbf{U}\right)\text{ilr}(\mathbf{x})$, on les columnes de $\mathbf{U}^*$ contenen les coordenades clr de la base ortonormal considerada a $\mathcal{S}^C$, i les columnes de $\mathbf{U}$ les coordenades clr de la base ortonormal considerada a $\mathcal{S}^D$. Es pot comprovar fàcilment que els productes $\mathbf{F}^*_{C,C-1}\mathbf{F}_{C-1,C}$ i $\mathbf{F}^*_{D,D-1}\mathbf{F}_{D-1,D}$ donen lloc a dues matrius quadrades amb subespais propis de valor propi 1 que són $V_1 = \{(z_1, z_2, \ldots, z_C)' \in \mathbb{R}^C; \sum_{i=1}^C z_i = 0\}$ i $V_2 = \{(z_1, z_2, \ldots, z_D)' \in \mathbb{R}^D; \sum_{i=1}^D z_i = 0\}$, respectivament. Es compleix per tant, que $\mathbf{U}^{*\prime}\left(\mathbf{F}^*_{C,C-1}\mathbf{F}_{C-1,C}\right) = \mathbf{U}^{*\prime}$ i $\left(\mathbf{F}^*_{D,D-1}\mathbf{F}_{D-1,D}\right)\mathbf{U} = \mathbf{U}$. Així doncs, la relació entre les coordenades ilr de la subcomposició $\mathbf{s}$ i la composició $\mathbf{x}$ és $\text{ilr}(\mathbf{s}) = \left(\mathbf{U}^{*\prime}\mathbf{S}\mathbf{U}\right)\text{ilr}(\mathbf{x})$. Donada la

densitat del vector ilr($\mathbf{x}$), obtenim la densitat del vector ilr($\mathbf{s}$) aplicant el teorema del canvi de variable o la propietat de les transformacions lineals de la llei normal a l'espai real. L'expressió d'aquesta densitat és idèntica a la densitat d'una llei $\mathcal{N}_{\mathcal{S}}^{C}\left(\mathbf{U}^{*\prime}\mathbf{P}\mathbf{U}\boldsymbol{\xi}, (\mathbf{U}^{*\prime}\mathbf{P}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}^{*\prime}\mathbf{P}\mathbf{U})'\right)$. □

Així doncs, les propietats de la distribució normal de l'espai real ens han permès demostrar que la família normal a $\mathcal{S}^D$ és tancada per les operacions pertorbació, potència, permutació i pel pas a subcomposicions. Hem vist també que la seva funció de densitat compleix la igualtat (4.9). No obstant això, no és possible descriure la distribució de qualsevol amalgama $\mathbf{x}_A$ en termes de la distribució de la composició $\mathbf{x}$. La raó d'aquesta dificultat és la impossibilitat d'obtenir una relació matricial entre les coordenades en una base ortonormal de les composicions $\mathbf{x}_A$ i de $\mathbf{x}$.

Propietat 4.9 Sigui $\mathbf{x} \in \mathcal{S}^D$ una composició aleatòria amb distribució $\mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ i sigui $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{D-1})'$. Llavors $\mathrm{E}[\mathbf{x}] = (\xi_1 \otimes \mathbf{e}_1) \oplus (\xi_2 \otimes \mathbf{e}_2) \oplus \ldots \oplus (\xi_{D-1} \otimes \mathbf{e}_{D-1})$ on $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ representa una base ortonormal de $\mathcal{S}^D$.

Demostració. L'esperança de qualsevol objecte aleatori és un element de l'espai suport. Si apliquem la definició estàndard d'esperança d'un vector aleatori a les coordenades de $\mathbf{x}$ respecte de la base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ utilitzant la densitat (4.10), obtindrem les coordenades de $\mathrm{E}[\mathbf{x}]$ respecte de la mateixa base. Si realitzem aquest càlcul utilitzant els mètodes d'integració estàndards, en resulta el paràmetre $\boldsymbol{\xi}$. Finalment, obtindrem la composició $\mathrm{E}[\mathbf{x}]$ mitjançant la combinació lineal a $\mathcal{S}^D$ resultant $(\xi_1 \otimes \mathbf{e}_1) \oplus (\xi_2 \otimes \mathbf{e}_2) \oplus \ldots \oplus (\xi_{D-1} \otimes \mathbf{e}_{D-1})$. □

Recordem que per definició, la composició cen[$\mathbf{x}$] introduïda al capítol 2, minimitza l'esperança $\mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}[\mathbf{x}])]$. En aquest cas, observem que la composició $\mathrm{E}[\mathbf{x}]$ obtinguda coincideix amb el centre cen[$\mathbf{x}$] d'una composició aleatòria (vegeu igualtat (2.29)). Això és una diferència essencial entre la llei normal logística additiva i la llei normal a $\mathcal{S}^D$. Recordem que per als models definits segons la metodologia MOVE no obteníem en cap cas, la igualtat entre $\mathrm{E}[\mathbf{x}]$ i cen[$\mathbf{x}$].

Propietat 4.10 Sigui $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$. Llavors una mesura de dispersió al voltant de l'esperança ve donada per $\mathrm{Mvar}[\mathbf{x}] = \mathrm{traça}(\boldsymbol{\Upsilon})$.

Demostració. La variància mètrica es defineix com $\text{Mvar}[\mathbf{x}] = \text{E}[d_a^2(\mathbf{x}, \text{cen}[\mathbf{x}])]$. Donada $\mathbf{x} \sim \mathcal{N}_{\mathcal{S}}^D(\boldsymbol{\xi}, \boldsymbol{\Upsilon})$ sabem per la propietat 4.9 que $\text{cen}[\mathbf{x}] = \text{E}[\mathbf{x}]$. Per tant, la variància mètrica serà una mesura de dispersió al voltant de l'esperança igual a $\text{Mvar}[\mathbf{x}] = \text{E}[d_a^2(\mathbf{x}, \text{E}[\mathbf{x}])]$. Sabem que la distància d_a entre dos elements és igual a la distància euclidiana d_{eu} entre les coordenades dels elements respecte d'una base ortonormal. Per tant, podem escriure $\text{Mvar}[\mathbf{x}] = \text{E}[d_{eu}^2(\text{ilr}(\mathbf{x}), \text{E}[\text{ilr}(\mathbf{x})])]$, valor que coincideix amb $\text{traça}(\boldsymbol{\Upsilon})$. □

Tal i com hem indicat en el capítol 2, no és coherent utilitzar les covariàncies o les correlacions entre dues components qualssevol de la composició original. Per aquesta raó, calcularem sempre covariàncies i correlacions de les components respecte de la base ortonormal. En el cas d'un model normal a $\mathcal{S}^D$ les covariàncies entre parelles de components coincideixen amb els elements de fora la diagonal de la matriu $\boldsymbol{\Upsilon}$.

4.3.2 Aspectes d'inferència estadística

Per estimar els paràmetres $\boldsymbol{\xi}$ i $\boldsymbol{\Upsilon}$ del model normal a $\mathcal{S}^D$ a partir dels valors d'una mostra, tan sols cal aplicar els procediments estàndards a les coordenades de la mostra respecte d'una base ortonormal. Així doncs, donada una mostra $\mathbf{X}$ de la composició $\mathbf{x}$, les estimacions puntuals de $\boldsymbol{\xi}$ i $\boldsymbol{\Upsilon}$ seran el vector de mitjanes mostrals i la matriu de covariàncies mostrals corregides de les seves coordenades ilr:

$$\hat{\boldsymbol{\xi}} = \overline{\text{ilr}(\mathbf{X})} \qquad \widehat{\boldsymbol{\Upsilon}} = \text{cov}(\text{ilr}(\mathbf{X})).$$

Amb aquests valors podem calcular també les estimacions puntuals per a $\text{E}[\mathbf{x}]$ i $\text{Mvar}[\mathbf{x}]$ ja que

$$\widehat{\text{E}[\mathbf{x}]} = (\hat{\xi}_1 \otimes \mathbf{e}_1) \oplus (\hat{\xi}_2 \otimes \mathbf{e}_2) \oplus \cdots \oplus (\hat{\xi}_{D-1} \otimes \mathbf{e}_{D-1}),$$

$$\widehat{\text{Mvar}[\mathbf{X}]} = \text{traça}(\widehat{\boldsymbol{\Upsilon}}).$$

De la mateixa manera que hem demostrat les propietats 4.5, 4.7 i 4.8, podem veure que la llei normal a $\mathcal{S}^D$ compleix exactament les mateixes propietats que la llei normal clàssica a l'espai real. Una conseqüència important és que els estimadors de l'esperança i la variabilitat són consistents i de mínima variància.

Per validar el model normal a $\mathcal{S}^D$ amb una prova de bondat d'ajust, tan sols cal aplicar un test de normalitat multivariant a les coordenades respecte de la base ortonormal.

4.3.3 Altres parametritzacions

La distribució normal logística additiva fou definida utilitzant la transformació logquocient additiva. Al capítol 3, hem vist que obteníem la mateixa llei de probabilitat utilitzant les transformacions logquocient centrada o logquocient isomètrica. Degut a la igualtat que existeix entre les coordenades d'una composició respecte d'una base ortonormal, el sistema de generadors B^* i la base B, i els vectors resultants d'aplicar les transformacions ilr, clr i alr a la composició $\mathbf{x}$, és natural preguntar-nos si obtenim la mateixa llei de probabilitat treballant amb les coordenades respecte del sistema B^* o la base B. Per treballar amb més comoditat utilitzarem la notació $\mathbf{v}$ per referir-nos a les coordenades de $\mathbf{x}$ respecte d'una base ortonormal i $\mathbf{y}$ per a les coordenades respecte de la base B.

Donat un esdeveniment A de $\mathcal{S}^D$, sabem que

$$P(A) = \int_{A^*} (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{v}-\boldsymbol{\xi})' \boldsymbol{\Upsilon}^{-1}(\mathbf{v}-\boldsymbol{\xi})\right] d\mathbf{v},$$

on A^* representa les coordenades de A en la base ortonormal. Sabem que $\mathbf{FU}$ és la matriu del canvi a la base B. Aplicant a l'expressió anterior el canvi de variable $\mathbf{v} = (\mathbf{FU})^{-1}\mathbf{y}$, obtenim

$$\begin{aligned} P(A) &= \int_{\mathbf{FU}A^*} \frac{(2\pi)^{-(D-1)/2}}{\mid \boldsymbol{\Upsilon} \mid^{1/2}} \exp\left[-\frac{1}{2}\left((\mathbf{FU})^{-1}\mathbf{y}-\boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1}\left((\mathbf{FU})^{-1}\mathbf{y}-\boldsymbol{\xi}\right)\right] \frac{1}{|\mathbf{FU}|} d\mathbf{y} \\ &= \int_{\mathbf{FU}A^*} \frac{(2\pi)^{-(D-1)/2}}{\mid \boldsymbol{\Upsilon} \mid^{1/2} |\mathbf{FU}|} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{FU}\boldsymbol{\xi})' ((\mathbf{FU})^{-1})' \boldsymbol{\Upsilon}^{-1} (\mathbf{FU})^{-1}(\mathbf{y}-\mathbf{FU}\boldsymbol{\xi})\right] d\mathbf{y} \\ &= \int_{\mathbf{FU}A^*} (2\pi)^{-(D-1)/2} \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right] d\mathbf{y}, \end{aligned} \tag{4.13}$$

on $\boldsymbol{\mu} = \mathbf{FU}\boldsymbol{\xi}$ i $\boldsymbol{\Sigma} = \mathbf{FU}\boldsymbol{\Upsilon}(\mathbf{FU})'$. En el capítol 2 hem introduït el vector $\boldsymbol{\mu}$ i la matriu $\boldsymbol{\Sigma}$ com el vector d'esperances i la matriu de covariàncies del vector $\mathbf{y} = \text{alr}(\mathbf{x})$, és a dir, com el vector d'esperances i la matriu de covariàncies de les coordenades de $\mathbf{x}$ respecte la base B. La funció dins la integral (4.13) representa la densitat de probabilitat del vector $\mathbf{y}$. La seva expressió coincideix amb la densitat d'un model normal a l'espai real i per tant es conclou que $\boldsymbol{\mu} = \text{E}[\mathbf{y}]$ i $\boldsymbol{\Sigma} = \text{var}[\mathbf{y}]$.

Observem també que $\mathbf{FU}A^*$ són les coordenades de l'esdeveniment A respecte de la base B. Així doncs, si utilitzem la densitat de probabilitat de les coordenades respecte de la base B, les probabilitats no varien i conseqüentment, la llei de probabilitat és exactament la mateixa.

Recordem però, que la base B no és ortonormal, i per tant la distància euclídiana entre les coordenades alr no és igual a la distància d'Aitchison entre les respectives composicions. Per aquesta raó, serà incorrecte utilitzar la densitat dels coeficients en la base B en els procediments on hi intervinguin distàncies o productes escalars. Per exemple si apliquem el procediment habitual en el càlcul de Mvar[$\mathbf{x}$] utilitzant les coordenades respecte de la base B, obtindrem que Mvar[$\mathbf{x}$] = E[$d^2_{eu}(\mathbf{y}, \mathrm{E}[\mathbf{y}])$] = traça($\mathbf{\Sigma}$), valor que no coincideix amb traça($\boldsymbol{\Upsilon}$) calculada amb la densitat dels coeficients en la base ortonormal.

De forma similar, podem treballar amb la densitat dels coeficients respecte del sistema de generadors B^*. No aconsellem el seu ús ja que, si bé les coordenades clr conserven les distàncies, obtenim la dificultat addicional de treballar amb una distribució degenerada.

En resum, podem definir el model normal a $\mathcal{S}^D$ mitjançant la funció de densitat de les coordenades alr, clr o ilr. En els tres casos, obtenim la mateixa llei de probabilitat. Malgrat això, recomanem utilitzar sempre la densitat de les coordenades respecte d'una base ortonormal ja que així evitem treballar amb una distribució degenerada i assegurem la conservació de les distàncies.

4.4 Distribució normal asimètrica a $\mathcal{S}^D$

4.4.1 Definició i propietats

Definició 4.3 Sigui $(\Omega, \mathcal{F}, p)$ un espai de probabilitat i $\mathbf{x} : \Omega \longrightarrow \mathcal{S}^D$ una composició aleatòria. Direm que $\mathbf{x}$ té una distribució *normal asimètrica a* $\mathcal{S}^D$ amb paràmetres $\boldsymbol{\xi}$, $\boldsymbol{\Upsilon}$ i $\boldsymbol{\varrho}$, si la funció de densitat dels coeficients en una base ortonormal de $\mathcal{S}^D$ és

$$f^*_{\mathbf{x}}(\mathbf{x}) = 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}\left(\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)\right] \times \Phi\left[\boldsymbol{\varrho}' \boldsymbol{\upsilon}^{-1} (\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right], \tag{4.14}$$

on Φ representa la funció de distribució d'una normal estàndard i ilr($\mathbf{x}$) el vector de coordenades de $\mathbf{x}$ respecte de la base ortonormal escollida. □

Escriurem $\mathbf{x} \sim \mathcal{SN}^D_{\mathcal{S}}(\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$ per indicar que la composició $\mathbf{x}$ segueix un model normal asimètric a $\mathcal{S}^D$. El subíndex $\mathcal{S}$ indica que es tracta d'un model en el símplex i el superíndex D mostra el nombre de parts de la composició aleatòria. Utilitzem la notació $\boldsymbol{\xi}$, $\boldsymbol{\Upsilon}$ i $\boldsymbol{\varrho}$ per

representar els paràmetres del model. Al igual que amb el model normal asimètric logístic additiu, fem servir $\boldsymbol{\xi}$ i $\boldsymbol{\Upsilon}$ per denotar els paràmetres però cal tenir en compte que no es corresponen amb l'esperança ni la matriu de covariàncies del vector ilr($\mathbf{x}$), denotats per $\boldsymbol{\xi}$ i $\boldsymbol{\Upsilon}$ en capítols anteriors.

La densitat (4.14) està completament restringida al símplex. Es tracta de la derivada de Radon-Nikodým de la probabilitat respecte de la mesura de Lebesgue a l'espai $\mathbb{R}^{D-1}$ de coeficients. Per aquesta raó, podem calcular la probabilitat d'un esdeveniment $A \subseteq \mathcal{S}^D$ qualsevol amb la integral ordinària

$$P(A) = \int_{A^*} 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[\mathbf{M}\right] \Phi\left[\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right] d\lambda(\text{ilr}(\mathbf{x})),$$

on

$$\mathbf{M} = -\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right),$$

i A^* representa les coordenades de l'esdeveniment A en la base ortonormal considerada.

De la mateixa manera que la llei normal a $\mathcal{S}^D$ coincideix amb la llei normal logística additiva, podem veure que la llei normal asimètrica a $\mathcal{S}^D$ coincideix amb la llei normal asimètrica logística additiva definida a l'apartat 3.3. La densitat del model alsn en la parametrització $\boldsymbol{\xi}, \boldsymbol{\Upsilon}$ i $\boldsymbol{\varrho}$ de l'expressió (3.5) representa la derivada de Radon-Nikodým de la probabilitat respecte de la mesura de Lebesgue a l'espai $\mathbb{R}^{D-1}$. Per obtenir la densitat (3.5) hem considerat com a variables aleatòries les $D-1$ primeres components de la composició $\mathbf{x}$ ja que l'última queda fixada, és a dir, $x_D = 1 - (\sum_{i=1}^{D-1} x_i)$. Per altra banda, el vector ilr($\mathbf{x}$) s'interpreta com el vector resultant d'aplicar la transformació logquocient isomètrica a la composició $\mathbf{x}$ i com a conseqüència, apareix el jacobià del canvi en l'expressió de la densitat (3.5). En aquesta situació podem calcular la probabilitat de l'esdeveniment A com la integral ordinària

$$P(A) = \int_A \frac{2D^{-1/2}\left(\prod_{i=1}^{D} x_i\right)^{-1}}{(2\pi)^{(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{1/2}} \exp\left[\mathbf{M}\right] \Phi\left[\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right] dx_1 dx_2 \cdots dx_{D-1}, \qquad (4.15)$$

on

$$\mathbf{M} = -\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1} \left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right).$$

Per demostrar que les dues lleis de probabilitat són en realitat la mateixa, és important distingir el vector de coordenades respecte de la base ortonormal, del vector resultant d'aplicar la transformació ilr, ambdós denotats com ilr($\mathbf{x}$) al capítol 2. Utilitzem en aquest cas la

notació $\mathbf{v} = (v_1, v_2, \ldots, v_{D-1})'$ per referir-nos a les coordenades respecte de la base ortonormal. Així doncs, segons un model normal asimètric a $\mathcal{S}^D$ podem escriure la probabilitat de l'esdeveniment A com

$$P(A) = \int_{A^*} 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[\mathbf{M}^*\right] \Phi\left[\boldsymbol{\varrho}' \boldsymbol{\upsilon}^{-1}(\mathbf{v} - \boldsymbol{\xi})\right] dv_1 dv_2 \cdots dv_{D-1}, \quad (4.16)$$

on

$$\mathbf{M}^* = -\frac{1}{2} (\mathbf{v} - \boldsymbol{\xi})' \boldsymbol{\Upsilon}^{-1} (\mathbf{v} - \boldsymbol{\xi}),$$

i A^* representa les coordenades de A respecte de la mateixa base ortonormal.

Observem que les expressions (4.15) i (4.16) són integrals de funcions reals respecte de la mesura de Lebesgue. A partir d'aquí, podem utilitzar els procediments estàndards del càlcul integral a l'espai real. Al igual que hem fet a l'apartat anterior, prenem l'expressió (4.16) i apliquem la transformació $\mathbf{v} = \text{ilr}(\mathbf{x})$, el jacobià de la qual és $D^{-1/2} \left(\prod_{i=1}^{D} x_i\right)^{-1}$. El teorema del canvi de variable a l'espai real assegura la igualtat

$$\begin{aligned} P(A) &= \int_{A^*} 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[\mathbf{M}^*\right] \Phi\left[\boldsymbol{\varrho}' \boldsymbol{\upsilon}^{-1}(\mathbf{v} - \boldsymbol{\xi})\right] dv_1 dv_2 \cdots dv_{D-1} \\ &= \int_{\text{ilr}^{-1}(A^*)} \frac{2D^{-1/2} \left(\prod_{i=1}^{D} x_i\right)^{-1}}{(2\pi)^{(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{1/2}} \exp\left[\mathbf{M}\right] \Phi\left[\boldsymbol{\varrho}' \boldsymbol{\upsilon}^{-1}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right] dx_1 dx_2 \cdots dx_{D-1}, \end{aligned}$$

on

$$\begin{aligned} \mathbf{M}^* &= -\frac{1}{2} (\mathbf{v} - \boldsymbol{\xi})' \boldsymbol{\Upsilon}^{-1} (\mathbf{v} - \boldsymbol{\xi}), \\ \mathbf{M} &= -\frac{1}{2} (\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})' \boldsymbol{\Upsilon}^{-1} (\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}). \end{aligned}$$

Observem com efectivament el darrer terme d'aquesta igualtat coincideix amb l'expressió (4.15) ja que $\text{ilr}^{-1}(A^*) = A$. Arribem doncs a la conclusió que les dues lleis de probabilitat són la mateixa sobre $\mathcal{S}^D$. Tot i així veurem que difereixen en les propietats i en els valors característics.

Per altra banda, podem observar que la densitat (4.14) és molt similar a la densitat normal asimètrica multivariant de l'espai real. Aquest és el motiu pel qual l'hem anomenat model normal asimètric a $\mathcal{S}^D$. Seguidament enunciem i demostrem les propietats principals. Per coherència amb tot el treball d'investigació, tornem a utilitzar la notació $\text{ilr}(\mathbf{x})$ per referir-nos a les coordenades de $\mathbf{x}$ respecte d'una base ortonormal.

Propietat 4.11 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{SN}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$. Siguin $\mathbf{a} \in \mathcal{S}^D$ una composició constant i b un escalar del cos $\mathbb{R}$. Llavors la composició $\mathbf{x}^* = \mathbf{a} \oplus (b \otimes \mathbf{x})$ es distribueix segons una llei $\mathcal{SN}_{\mathcal{S}}^{D}(\text{ilr}(\mathbf{a}) + b\boldsymbol{\xi}, b^2\boldsymbol{\Upsilon}, \boldsymbol{\varrho})$.

Demostració. A partir de la definició de $\mathbf{x}^*$ i de les propietats dels coeficients en una base ortonormal, es dedueix que $\text{ilr}(\mathbf{x}^*) = \text{ilr}(\mathbf{a}) + b\text{ilr}(\mathbf{x})$. És a dir, les coordenades en la base ortonormal de la composició $\mathbf{x}^*$ s'obtenen a partir d'una transformació lineal de les coordenades de $\mathbf{x}$. Aplicant el teorema del canvi de variable o senzillament utilitzant la propietat de les transformacions lineals que compleix la distribució normal asimètrica a l'espai real, obtenim que la densitat de les coordenades $\text{ilr}(\mathbf{x}^*)$ és

$$\begin{aligned} f^*_{\mathbf{x}^*}(\mathbf{x}^*) = {} & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} b^{-1} \\ & \times \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}^*) - (\text{ilr}(\mathbf{a}) + b\boldsymbol{\xi})\right)' b^{-2}\boldsymbol{\Upsilon}^{-1}\left(\text{ilr}(\mathbf{x}^*) - (\text{ilr}(\mathbf{a}) + b\boldsymbol{\xi})\right)\right] \\ & \times \Phi\left[\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1} b^{-1}(\text{ilr}(\mathbf{x}^*) - (\text{ilr}(\mathbf{a}) + b\boldsymbol{\xi}))\right], \end{aligned}$$

i per tant, es conclou que $\mathbf{x}^* \sim \mathcal{SN}_{\mathcal{S}}^{D}(\text{ilr}(\mathbf{a}) + b\boldsymbol{\xi}, b^2\boldsymbol{\Upsilon}, \boldsymbol{\varrho})$. □

Propietat 4.12 Sigui $\mathbf{x} \sim \mathcal{SN}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$ i $\mathbf{a} \in \mathcal{S}^D$ una composició constant. Aleshores es compleix la igualtat $f^*_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x}) = f^*_{\mathbf{x}}(\mathbf{x})$, on $f^*_{\mathbf{a}\oplus\mathbf{x}}$ i $f^*_{\mathbf{x}}$ representen les funcions de densitat de les composicions aleatòries $\mathbf{x}$ i $\mathbf{a} \oplus \mathbf{x}$ respectivament.

Demostració. Per la propietat anterior sabem que $\mathbf{a} \oplus \mathbf{x} \sim \mathcal{SN}_{\mathcal{S}}^{D}(\text{ilr}(\mathbf{a}) + \boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$ i per tant, a partir de les respectives funcions de densitat tenim que

$$\begin{aligned} f^*_{\mathbf{a}\oplus\mathbf{x}}(\mathbf{a} \oplus \mathbf{x}) = {} & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \\ & \times \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{a} \oplus \mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi})\right)' \boldsymbol{\Upsilon}^{-1}\left(\text{ilr}(\mathbf{a} \oplus \mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi})\right)\right] \\ & \times \Phi\left[\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\text{ilr}(\mathbf{a} \oplus \mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi}))\right] \\ = {} & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \\ & \times \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{a}) + \text{ilr}(\mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi})\right)' \boldsymbol{\Upsilon}^{-1}\left(\text{ilr}(\mathbf{a}) + \text{ilr}(\mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi})\right)\right] \\ & \times \Phi\left[\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\text{ilr}(\mathbf{a}) + \text{ilr}(\mathbf{x}) - (\text{ilr}(\mathbf{a}) + \boldsymbol{\xi}))\right] \\ = {} & 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1}\left(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi}\right)\right] \\ & \times \Phi\left[\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\text{ilr}(\mathbf{x}) - \boldsymbol{\xi})\right] = f^*_{\mathbf{x}}(\mathbf{x}). \end{aligned}$$

tal i com s'indicava a (4.9). □

Propietat 4.13 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{SN}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$. Sigui $\mathbf{x}_P = \mathbf{P}\mathbf{x}$ la composició $\mathbf{x}$ amb les components reordenades per la matriu permutació $\mathbf{P}$. Llavors $\mathbf{x}_P$ es distribueix segons un model $\mathcal{SN}_{\mathcal{S}}^{D}(\boldsymbol{\xi}_P, \boldsymbol{\Upsilon}_P, \boldsymbol{\varrho}_P)$ amb

$$\boldsymbol{\xi}_P = \mathbf{U}'\mathbf{P}\mathbf{U}\boldsymbol{\xi}, \qquad \boldsymbol{\Upsilon}_P = (\mathbf{U}'\mathbf{P}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}'\mathbf{P}\mathbf{U})', \qquad \boldsymbol{\varrho}_P = \frac{\boldsymbol{\upsilon}_P\boldsymbol{\Upsilon}_P^{-1}\mathbf{B}'\boldsymbol{\varrho}}{\sqrt{1+\boldsymbol{\varrho}'(\boldsymbol{\upsilon}^{-1}\boldsymbol{\Upsilon}\boldsymbol{\upsilon}^{-1}-\mathbf{B}\boldsymbol{\Upsilon}_P^{-1}\mathbf{B}')\boldsymbol{\varrho}}},$$

on $\mathbf{U}$ és la matriu d'ordre $D \times (D-1)$ les columnes de la qual són les coordenades clr dels vectors d'una base ortonormal de $\mathcal{S}^D$, $\mathbf{B} = \boldsymbol{\upsilon}^{-1}\boldsymbol{\Upsilon}(\mathbf{U}'\mathbf{P}'\mathbf{U})$, i $\boldsymbol{\upsilon}$ i $\boldsymbol{\upsilon}_P$ són matrius diagonals iguals a l'arrel quadrada de la diagonal de $\boldsymbol{\Upsilon}$ i $\boldsymbol{\Upsilon}_P$, respectivament.

Demostració. En la propietat 4.7 hem vist que $\operatorname{ilr}(\mathbf{x}_P) = (\mathbf{U}'\mathbf{P}\mathbf{U})\operatorname{ilr}(\mathbf{x})$. Aplicant el teorema del canvi de variable o la propietat 1.15 de la distribució normal asimètrica, obtenim que la densitat de les coordenades de la composició $\mathbf{x}_P$ és

$$f^*_{\mathbf{x}_P}(\mathbf{x}_P) = 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon}_P \mid^{-1/2} \exp\left[-\frac{1}{2}\left(\operatorname{ilr}(\mathbf{x}_P)-\boldsymbol{\xi}_P\right)'\boldsymbol{\Upsilon}_P^{-1}\left(\operatorname{ilr}(\mathbf{x}_P)-\boldsymbol{\xi}_P\right)\right]$$
$$\times\, \Phi\left[\boldsymbol{\varrho}_P'\boldsymbol{\upsilon}_P^{-1}(\operatorname{ilr}(\mathbf{x})-\boldsymbol{\xi}_P)\right],$$

i per tant, $\mathbf{x}_P \sim \mathcal{SN}_{\mathcal{S}}^{D}(\boldsymbol{\xi}_P, \boldsymbol{\Upsilon}_P, \boldsymbol{\varrho}_P)$. □

Propietat 4.14 Sigui $\mathbf{x}$ una composició aleatòria amb D parts que es distribueix segons un model $\mathcal{SN}_{\mathcal{S}}^{D}(\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$. Sigui $\mathbf{s} = \mathcal{C}(\mathbf{S}\mathbf{x})$ una subcomposició amb parts seleccionades mitjançant la matriu $\mathbf{S}$ d'ordre $C \times D$. Llavors la subcomposició $\mathbf{s}$ es distribueix segons una llei $\mathcal{SN}_{\mathcal{S}}^{C}(\boldsymbol{\xi}_S, \boldsymbol{\Upsilon}_S, \boldsymbol{\varrho}_S)$ amb

$$\boldsymbol{\xi}_S = \mathbf{U}^{*\prime}\mathbf{S}\mathbf{U}\boldsymbol{\xi}, \qquad \boldsymbol{\Upsilon}_S = (\mathbf{U}^{*\prime}\mathbf{S}\mathbf{U})\boldsymbol{\Upsilon}(\mathbf{U}^{*\prime}\mathbf{S}\mathbf{U})', \qquad \boldsymbol{\varrho}_S = \frac{\boldsymbol{\upsilon}_S\boldsymbol{\Upsilon}_S^{-1}\mathbf{B}'\boldsymbol{\varrho}}{\sqrt{1+\boldsymbol{\varrho}'(\boldsymbol{\upsilon}^{-1}\boldsymbol{\Upsilon}\boldsymbol{\upsilon}^{-1}-\mathbf{B}\boldsymbol{\Upsilon}_S^{-1}\mathbf{B}')\boldsymbol{\varrho}}},$$

on $\mathbf{U}$ és la matriu d'ordre $D \times (D-1)$ les columnes de la qual són les coordenades clr d'una base ortonormal de $\mathcal{S}^D$, $\mathbf{U}^*$ és la matriu d'odre $C \times (C-1)$ les columnes de la qual són les coordenades clr d'una base ortonormal de $\mathcal{S}^C$, $\mathbf{B} = \boldsymbol{\upsilon}^{-1}\boldsymbol{\Upsilon}(\mathbf{U}'\mathbf{S}'\mathbf{U}^*)$, i $\boldsymbol{\upsilon}$ i $\boldsymbol{\upsilon}_S$ són matrius diagonals iguals a l'arrel quadrada de la diagonal de $\boldsymbol{\Upsilon}$ i $\boldsymbol{\Upsilon}_S$, respectivament.

Demostració. En la propietat 4.8 hem vist que $\operatorname{ilr}(\mathbf{s}) = \left(\mathbf{U}^{*\prime}\mathbf{S}\mathbf{U}\right)\operatorname{ilr}(\mathbf{x})$. Donada la densitat de les coordenades $\operatorname{ilr}(\mathbf{x})$, obtenim la densitat de les coordenades $\operatorname{ilr}(\mathbf{s})$ aplicant el teorema del canvi de variable o la propietat 1.15 de la distribució normal asimètrica, i per tant es

conclou que la subcomposició $\mathbf{s}$ es distribueix segons un model normal asimètric a $\mathcal{S}^C$ amb paràmetres $\boldsymbol{\xi}_S, \boldsymbol{\Upsilon}_S$ i $\boldsymbol{\varrho}_S$. □

Veiem doncs que la família normal asimètrica a $\mathcal{S}^D$ és tancada per les operacions pertorbació, potència, permutació i pel pas a subcomposicions. Igual que passava amb el model alsn, no es possible descriure la distribució de qualsevol amalgama $\mathbf{x}_A$ en termes de la distribució de la composició $\mathbf{x}$. La raó d'aquesta dificultat és la impossibilitat d'obtenir una relació matricial entre els coeficients de $\mathbf{x}_A$ i de $\mathbf{x}$ en una base ortonormal. A diferència del model alsn, hem vist que la funció de densitat compleix la propietat (4.9).

Propietat 4.15 Sigui $\mathbf{x}$ una composició aleatòria amb distribució $\mathcal{SN}_{\mathcal{S}}^D(\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$. Llavors $\mathrm{E}[\mathbf{x}] = (\beta_1 \otimes \mathbf{e}_1) \oplus (\beta_2 \otimes \mathbf{e}_2) \oplus \ldots \oplus (\beta_{D-1} \otimes \mathbf{e}_{D-1})$, amb $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ base ortonormal de $\mathcal{S}^D$ i $\boldsymbol{\beta} = \boldsymbol{\xi} + \boldsymbol{\upsilon}\boldsymbol{\delta}\sqrt{2/\pi}$, on $\boldsymbol{\delta}$ és l'expressió alternativa del paràmetre de forma relacionat amb $\boldsymbol{\varrho}$ segons (1.14), i $\boldsymbol{\upsilon}$ és una matriu diagonal igual a l'arrel quadrada de la diagonal de $\boldsymbol{\Upsilon}$.
Demostració. L'esperança és un element de l'espai suport. Per tant, a partir de les coordenades de $\mathbf{x}$ respecte de la base ortonormal $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ i la densitat (4.14) obtindrem les coordenades del vector esperança en la mateixa base ortonormal. Per les propietats de la distribució normal asimètrica a l'espai real sabem que $\mathrm{E}[\mathrm{ilr}(\mathbf{x})] = \boldsymbol{\xi} + \boldsymbol{\upsilon}\boldsymbol{\delta}\sqrt{2/\pi}$, vector que denotem com $\boldsymbol{\beta}$. Obtenim $\mathrm{E}[\mathbf{x}]$ amb la combinació lineal $(\beta_1 \otimes \mathbf{e}_1) \oplus (\beta_2 \otimes \mathbf{e}_2) \oplus \ldots \oplus (\beta_{D-1} \otimes \mathbf{e}_{D-1})$.□

Per la igualtat (2.27) del capítol 2, sabem que les coordenades en la base ortonormal de $\mathrm{cen}[\mathbf{x}]$ són $\mathrm{E}[\mathrm{ilr}(\mathbf{x})]$. Sabem també que les coordenades d'un vector respecte d'una base són úniques. Així doncs obtenim que $\mathrm{cen}[\mathbf{x}] = \mathrm{E}[\mathbf{x}] = (\beta_1 \otimes \mathbf{e}_1) \oplus (\beta_2 \otimes \mathbf{e}_2) \oplus \ldots \oplus (\beta_{D-1} \otimes \mathbf{e}_{D-1})$. Aquesta és una diferència essencial entre la llei normal asimètrica a $\mathcal{S}^D$ i la llei normal asimètrica logística additiva. Recordem que pels models definits segons la perspectiva MOVE, no obteníem en cap cas la igualtat entre les composicions $\mathrm{cen}[\mathbf{x}]$ i $\mathrm{E}[\mathbf{x}]$.

Propietat 4.16 Sigui $\mathbf{x} \sim \mathcal{SN}_{\mathcal{S}}^D(\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho})$. Llavors una mesura de dispersió al voltant de l'esperança és $\mathrm{Mvar}[\mathbf{x}] = \mathrm{traça}\left(\boldsymbol{\Upsilon} - (2/\pi)\boldsymbol{\upsilon}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\upsilon}\right)$, on $\boldsymbol{\upsilon}$ és una matriu diagonal igual a l'arrel quadrada de la diagonal de $\boldsymbol{\Upsilon}$, i $\boldsymbol{\delta}$ és l'expressió alternativa del paràmetre de forma, relacionat amb $\boldsymbol{\varrho}$ segons l'expressió (1.14).
Demostració. La variància mètrica es defineix com $\mathrm{Mvar}[\mathbf{x}] = \mathrm{E}[d_a^2(\mathbf{x}, \mathrm{cen}[\mathbf{x}])]$. Donat que $\mathrm{cen}[\mathbf{x}] = \mathrm{E}[\mathbf{x}]$, la variància mètrica serà una mesura de dispersió al voltant de l'esperança

$\text{Mvar}[\mathbf{x}] = \text{E}[d_a^2(\mathbf{x}, \text{E}[\mathbf{x}])]$. Sabem que la distància d_a entre dos elements és igual a la distància euclidiana d_{eu} entre les coordenades dels elements respecte d'una base ortonormal. Per tant podem escriure $\text{Mvar}[\mathbf{x}] = \text{E}[d_{eu}^2(\text{ilr}(\mathbf{x}), \text{E}[\text{ilr}(\mathbf{x})])]$, valor que coincideix amb la traça de la matriu de covariàncies de les coordenades $\text{ilr}(\mathbf{x})$. Utilitzant que la densitat de les coordenades és normal asimètrica amb paràmetres $\boldsymbol{\xi}$, $\boldsymbol{\Upsilon}$ i $\boldsymbol{\varrho}$, obtenim que $\text{Mvar}[\mathbf{x}] = \text{traça}(\boldsymbol{\Upsilon} - (2/\pi)\boldsymbol{\upsilon}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\upsilon})$.□

4.4.2 Aspectes d'inferència estadística

Per calcular una estimació puntual dels paràmetres $\boldsymbol{\xi}$, $\boldsymbol{\Upsilon}$ i $\boldsymbol{\varrho}$ a partir dels valors d'una mostra, aplicarem el mètode de màxima versemblança a les coordenades de la mostra respecte d'una base ortonormal. Recordem que no existeix una expressió analítica dels estimadors en funció de la mostra sinó que cal utilitzar procediments numèrics per calcular el màxim de la funció de versemblança i obtenir el valor de $\hat{\boldsymbol{\xi}}$, $\widehat{\boldsymbol{\Upsilon}}$ i $\hat{\boldsymbol{\varrho}}$ (vegeu apartat 1.3.3).

Les estimacions puntuals dels paràmetres ens permetran calcular una estimació per a l'esperança i la variància mètrica de la composició aleatòria $\mathbf{x}$:

$$\widehat{\text{E}[\mathbf{x}]} = (\hat{\beta}_1 \otimes \mathbf{e}_1) \oplus (\hat{\beta}_2 \otimes \mathbf{e}_2) \oplus \cdots \oplus (\hat{\beta}_{D-1} \otimes \mathbf{e}_{D-1}),$$

$$\widehat{\text{Mvar}[\mathbf{X}]} = \text{traça}\left(\widehat{\boldsymbol{\Upsilon}} - \frac{2}{\pi}\widehat{\boldsymbol{\upsilon}}\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}'\widehat{\boldsymbol{\upsilon}}\right),$$

on $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\xi}} + \hat{\boldsymbol{\upsilon}}\sqrt{\frac{2}{\pi}}\hat{\boldsymbol{\delta}}$ i $\hat{\boldsymbol{\upsilon}}$ és una matriu diagonal igual a l'arrel quadrada de la diagonal de $\widehat{\boldsymbol{\Upsilon}}$.

El model normal a $\mathcal{S}^D$ és un cas particular del model normal asimètric a $\mathcal{S}^D$ ja que correspon al cas $\boldsymbol{\varrho} = \mathbf{0}$. Donada una mostra aleatòria, podem ajustar els dos models i comparar-los amb un test de raó de versemblança. Tan sols cal contrastar la hipòtesi nul·la $\text{H}_0 : \boldsymbol{\varrho} = \mathbf{0}$ vers la hipòtesi contrària utilitzant les coordenades de la mostra respecte de la base ortonormal i aplicant el test descrit a la secció 1.3.3

Un aspecte interessant utilitzat sovint en la modelització de vectors aleatoris, és la validació del model amb una prova de bondat d'ajust. Per validar una distribució normal asimètrica a $\mathcal{S}^D$ haurem d'aplicar un test multivariant de normalitat asimètrica a les coordenades ilr de la mostra. Reservem el següent capítol per presentar el desenvolupament i l'aplicació d'aquest tipus de contrastos.

4.4.3 Altres parametritzacions

Hem definit la llei normal asimètrica a $\mathcal{S}^D$ utilitzant el vector de coordenades de la composició aleatòria respecte d'una base ortonormal. No obstant això, obtenim la mateixa llei de probabilitat si utilitzem el vector de coordenades respecte de la base B. Per veure amb detall aquesta propietat, utilitzarem la notació $\mathbf{v}$ i $\mathbf{y}$ per les coordenades de $\mathbf{x}$ en la base ortonormal i la base B respectivament.

En termes de les coordenades $\mathbf{v}$, la probabilitat d'un esdeveniment A de $\mathcal{S}^D$ és

$$P(A) = \int_{A^*} 2(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Upsilon} \mid^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{v}-\boldsymbol{\xi})' \boldsymbol{\Upsilon}^{-1}(\mathbf{v}-\boldsymbol{\xi})\right] \Phi\left[\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}(\mathbf{v}-\boldsymbol{\xi})\right] d\mathbf{v},$$

on A^* representa les coordenades de A respecte de la mateixa base ortonormal. Apliquem la transformació $\mathbf{v} = (\mathbf{FU})^{-1}\mathbf{y}$, on $\mathbf{FU}$ és la matriu del canvi a la base B, i obtenim

$$\begin{aligned} P(A) &= \int_{\mathbf{FU}A^*} \frac{2(2\pi)^{-(D-1)/2}}{\mid \boldsymbol{\Upsilon} \mid^{1/2}} \exp[\mathbf{M}] \, \Phi\left[\boldsymbol{\varrho}'\boldsymbol{\upsilon}^{-1}((\mathbf{FU})^{-1}\mathbf{y}-\boldsymbol{\xi})\right] \frac{1}{|\mathbf{FU}|} d\mathbf{y}, \\ &= \int_{\mathbf{FU}A^*} \frac{2(2\pi)^{-(D-1)/2}}{\mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right] \Phi\left[\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right] d\mathbf{y} \end{aligned}$$

amb

$$\mathbf{M} = -\frac{1}{2}\left((\mathbf{FU})^{-1}\mathbf{y}-\boldsymbol{\xi}\right)' \boldsymbol{\Upsilon}^{-1}\left((\mathbf{FU})^{-1}\mathbf{y}-\boldsymbol{\xi}\right),$$

$\boldsymbol{\mu} = \mathbf{FU}\boldsymbol{\xi}$, $\boldsymbol{\Sigma} = (\mathbf{FU})\boldsymbol{\Upsilon}(\mathbf{FU})'$, $\boldsymbol{\alpha} = \boldsymbol{\omega}((\mathbf{FU})^{-1})'\boldsymbol{\upsilon}^{-1}\boldsymbol{\varrho}$ i $\boldsymbol{\omega}$ és una matriu diagonal amb l'arrel quadrada de la diagonal de $\boldsymbol{\Sigma}$. Observem que les relacions entre els paràmetres $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}$ i $\boldsymbol{\xi}, \boldsymbol{\Upsilon}, \boldsymbol{\varrho}$ coincideixen amb les relacions (3.4) que obteníem a l'apartat 3.3.4. i que el conjunt $\mathbf{FU}A^*$ representa les coordenades de l'esdeveniment A en la base B.

Podem doncs definir un model normal asimètric a $\mathcal{S}^D$ en termes de les coordenades alr de la composició aleatòria. Amb el canvi que hem aplicat, hem obtingut la funció de densitat de les coordenades $\mathbf{y}$. És evident que les probabilitats es conserven, però, recordem que la base B no és ortonormal i per tant, la distància euclidiana entre coordenades no és igual a la distància d'Aitchison entre les respectives composicions. Així doncs, serà incorrecte utilitzar la densitat de les coordenades en la base B en procediments on hi intervinguin distàncies o productes escalars com, per exemple, en el càlcul de la variància mètrica Mvar[$\mathbf{x}$].

Podríem també treballar amb les coordenades respecte del sistema de generadors B^*. No obstant això, caldria una densitat normal asimètrica singular o degenerada. Actualment, no

disposem encara d'aquest model i per tant, no és immediat donar l'expressió de la funció de densitat del vector de coordenades respecte del sistema de referència B^*.

Capítol 5

Proves de bondat d'ajust

En la modelització de vectors aleatoris utilitzem una determinada distribució per representar una població. Sovint escollim la distribució en base a una mostra representativa d'aquesta. Una vegada seleccionat el model, podem procedir a valorar el seu ajust a la mostra mitjançant un contrast d'hipòtesi. Aquests tipus de contrasts s'anomenen tests o proves de bondat d'ajust. Es parteix de la hipòtesi nul·la H_0 : ≪la mostra aleatòria $y_1, y_2, \ldots, y_n$ prové d'una població amb funció de distribució F(y) ≫. Seguidament, es valora la versemblança de la hipòtesi nul·la a partir d'un estadístic (funció de la mostra) la distribució del qual és coneguda suposant certa H_0. Finalment, i a partir d'una regla de decisió basada en el valor de l'estadístic, es decideix si tenim motius suficients per rebutjar la hipòtesi H_0.

En els capítols 3 i 4 hem descrit i introduït diverses famílies de distribucions per modelitzar dades composicionals. El següent pas natural és validar el model a partir d'un test de bondat d'ajust. Malauradament, l'aplicació d'aquestes tècniques és bastant limitada, entre altres coses perquè els programes informàtics habituals disposen d'un nombre relativament baix d'estadístics, sobretot en el cas de distribucions multivariants. Les possibilitats es redueixen dràsticament si el model a contrastar és la distribució normal asimètrica logística additiva o la distribució normal asimètrica a $\mathcal{S}^D$.

En aquest capítol del treball d'investigació ens proposem donar les pautes per aplicar els test sobre els models definits segons les metodologies MOVE i STAY. L'estudi es centra en contrastos univariants basats en la funció de distribució empírica, abreviada amb les sigles EDF (de l'anglès *Empirical Distribution Function*). Per aquesta raó, iniciem el capítol amb

un apartat introductori dedicat a la notació, definicions i procediment general d'una prova de bondat d'ajust d'aquest tipus. El segon apartat conté un estudi original amb una adaptació d'aquesta tècnica al cas particular de la distribució normal asimètrica a la recta real en el que analitzem les dificultats principals, calculem els percentils més rellevants de cada estadístic a partir de tècniques de Monte Carlo i estudiem la seva potència davant diverses distribucions alternatives. Acabem aquest apartat amb una anàlisi breu d'una tècnica de bondat d'ajust similar que ha estat desenvolupada recentment per Dalla-Valle (2001). Dediquem el tercer apartat a estudiar les proves de bondat d'ajust per diverses famílies de distribucions sobre el símplex. Recordem en primer lloc els treballs d'Aitchison (1986) i d'Aitchison et al. (2003) on es detallen els passos a seguir per validar el model normal logístic additiu. Combinant la proposta d'Aitchison i els tests univariants per a la distribució normal asimètrica, proposem una metodologia per validar el model normal asimètric logístic additiu. Un contrast de normalitat i de normalitat asimètrica multivariants seran suficients per validar els models normal a $\mathcal{S}^D$ i normal asimètric a $\mathcal{S}^D$. Tan sols cal aplicar-los a les components de la mostra en una base ortonormal. Finalment fem referència al treball d'Egozcue et al. (2001) on es defineix un nou estadístic basat en la distància d'Aitchison que podria aplicar-se per validar els models definits sobre el símplex.

5.1 Proves univariants basades en la funció de distribució empírica

A grans trets podem dir que els tests de bondat d'ajust basats en la funció de distribució empírica mesuren la diferència entre dues funcions de distribució: la funció de distribució teòrica, suposada en la hipòtesi nul·la, i la funció de distribució empírica, calculada a partir de la mostra. Per mesurar aquesta diferència podem utilitzar diversos estadístics. Si el valor de l'estadístic és $\ll$massa gran$\gg$ direm que les dues distribucions són significativament diferents i per tant haurem de rebutjar la hipòtesi nul·la.

A continuació, introduïm la notació i la definició dels elements que apareixen en un test d'aquest tipus. Seguidament, detallem el procediment a seguir fins a decidir si rebutgem o no la hipòtesi nul·la.

5.1.1 Notació i definicions

Stephens i D'Agostino (1986) defineixen la funció de distribució empírica com:

Definició 5.1 Sigui $y_1, y_2, \ldots, y_n$ una mostra de mida n. Sigui $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ la mateixa mostra ordenada. La *funció de distribució empírica* (EDF) $\mathrm{F}_n(y)$ avaluada al punt y es defineix com

$$\mathrm{F}_n(y) = \frac{\text{nombre d'observacions} \leq y}{n} \qquad (-\infty < y < +\infty),$$

o de manera més precisa com:

$$\mathrm{F}_n(y) = \begin{cases} 0, & y < y_{(1)}; \\ i/n, & y_{(i)} \leq y < y_{(i+1)}; \\ 1, & y_{(n)} < y. \end{cases}$$

□

Aquesta funció mesura la proporció d'individus de la mostra amb un valor més petit o igual a y i per tant, és una funció esglaonada.

Utilitzarem la notació $\mathrm{F}(\cdot)$ per indicar la funció de distribució teòrica. Així doncs, $\mathrm{F}(y)$ serà la probabilitat d'obtenir una observació més petita o igual a y.

Els estadístics basats en la funció de distribució empírica mesuren la diferència entre les funcions F_n i F. Coneixem diferents estadístics que Stephens i D'Agostino (1986) divideixen en dues famílies.

1. Família Cramér-von Mises

   Els estadístics d'aquesta família mesuren la diferència entre les dues funcions de distribució utilitzant l'expressió

   $$n \int_{-\infty}^{\infty} (\mathrm{F}_n(y) - \mathrm{F}(y))^2 \psi(\mathrm{F}(y)) d\mathrm{F}(y),$$

   on $\psi(\mathrm{F}(y))$ és una funció pes que pondera la diferència $(\mathrm{F}_n(y) - \mathrm{F}(y))^2$ allà on volem que el test sigui més sensible.

   Si volem que totes les diferències tinguin el mateix pes prendrem $\psi(\mathrm{F}(y)) = 1$. En aquest cas l'estadístic s'anomena *Cramér-von Mises* i es denota W^2.

Si volem donar més pes a les diferències $(\mathrm{F}_n(y) - \mathrm{F}(y))^2$ de les cues de la distribució prendrem $\psi(\mathrm{F}(y)) = (\mathrm{F}(y)(1 - \mathrm{F}(y)))^{-1}$. Fixem-nos que el valor de $\psi(\mathrm{F}(y))$ augmenta a mesura que $\mathrm{F}(y)$ s'apropa a 0 o 1. En aquest cas el test tindrà una bona potència quan la mostra presenti dades anòmales, però sacrificarem potència quan $\mathrm{F}_n(y)$ i $\mathrm{F}(y)$ siguin diferents a prop de la mediana (Anderson i Darling, 1954). Aquest estadístic es coneix amb el nom de *Anderson-Darling* i es denota A^2.

És habitual també una modificació de l'estadístic W^2, anomenat estadístic de *Watson* (U^2) i definit com

$$U^2 = n \int_{-\infty}^{\infty} \left(\mathrm{F}_n(y) - \mathrm{F}(y) - \int_{-\infty}^{\infty} (\mathrm{F}_n(x) - \mathrm{F}(x)) d\mathrm{F}(x) \right)^2 d\mathrm{F}(y).$$

Aquest estadístic fou introduït per Watson (1961) per tal de treballar amb dades circulars, és a dir, per treballar amb dades l'espai mostral de les quals és una circumferència. Tot i això, es possible utilitzar aquest estadístic quan l'espai mostral és la recta real.

2. Família Kolmogorov-Smirnov

En aquesta família trobem els estadístics D^+ i D^- que mesuren respectivament la diferència ≪vertical≫ més gran quan $\mathrm{F}_n(y)$ és superior a $\mathrm{F}(y)$, i la diferència ≪vertical≫ més gran quan $\mathrm{F}_n(y)$ és inferior a $\mathrm{F}(y)$, és a dir:

$$D^+ = \sup_y (\mathrm{F}_n(y) - \mathrm{F}(y)),$$

$$D^- = \sup_y (\mathrm{F}(y) - \mathrm{F}_n(y)).$$

L'estadístic més conegut i utilitzat d'aquesta família és l'estadístic de *Kolmogorov-Smirnov* (D), definit com

$$D = \sup_y \mid \mathrm{F}_n(y) - \mathrm{F}(y) \mid = \max(D^+, D^-)$$

Un altre estadístic estretament relacionat és l'estadístic de *Kuiper* (V), definit com

$$V = D^+ + D^-$$

Aquest estadístic, al igual que U^2, fou introduït per treballar amb dades circulars però es pot utilitzar també amb dades de l'espai real.

Sigui y una variable aleatòria amb funció de distribució $\mathrm{F}(\cdot)$. La imatge per F del domini de y, dóna lloc als valors d'una variable p uniforme a l'interval $[0,1]$ i per tant, la seva funció de distribució és $\mathrm{F}^*(p) = p$, $0 \leq p \leq 1$.

Donada una mostra $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ de la variable y, podem calcular la mostra $p_{(1)} < p_{(2)} < \cdots < p_{(n)}$ de la variable p aplicant la transformació $p_{(i)} = \mathrm{F}(y_{(i)})$, per a $i = 1, 2, \ldots, n$. Sigui F_n^* la funció de distribució empírica dels valors $p_{(i)}$. Podem realitzar el contrast de bondat d'ajust mesurant les diferències entre les funcions $\mathrm{F}^*(p)$ i $\mathrm{F}_n^*(p)$ ja que es compleix l'igualtat

$$\mathrm{F}_n(y) - \mathrm{F}(y) = \mathrm{F}_n^*(p) - \mathrm{F}^*(p).$$

Conseqüentment, els estadístics calculats amb la mostra $p_{(i)}$ tindran els mateixos valors que els calculats amb la mostra $y_{(i)}$. Aquesta transformació permet obtenir unes fórmules de càlcul dels estadístics més senzilles i que detallem tot seguit:

$$\begin{aligned}
W^2 &= \sum_{i=1}^{n} \left(p_{(i)} - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}; \\
A^2 &= -n - \frac{1}{n} \sum_{i=1}^{n} ((2i-1) \ln p_{(i)} + (2n+1-2i) \ln(1-p_{(i)})); \\
U^2 &= W^2 - n \left(\bar{p} - \frac{1}{2} \right)^2; \\
D^+ &= \max_i \left(\frac{i}{n} - p_{(i)} \right); \\
D^- &= \max_i \left(p_{(i)} - \frac{i-1}{n} \right); \\
D &= \max(D^+, D^-), \\
V &= D^+ + D^-.
\end{aligned} \tag{5.1}$$

on $\bar{p}$ és la mitjana aritmètica de la mostra $p_{(i)}$.

És possible conèixer, sota la suposició de H_0, la distribució de cada un dels estadístics. Cal però tenir en compte que aquesta distribució varia amb la mida de la mostra i amb el valor dels paràmetres de la distribució suposada a H_0.

5.1.2 Procediment general

L'objectiu d'una prova de bondat d'ajust és valorar la versemblança de la hipòtesi nul·la $H_0 : \ll$ La mostra $y_1, y_2, \ldots, y_n$ prové d'una població amb funció de distribució $\mathrm{F}(y; \boldsymbol{\theta})$, on

F és una funció contínua i $\boldsymbol{\theta}$ és el vector de paràmetres≫.

A la pràctica ens caldrà distingir dos casos, depenent de si coneixem o no el valor dels paràmetres de la distribució. Quan el vector de paràmetres $\boldsymbol{\theta}$ sigui conegut ho anomenem Cas 0, i quan $\boldsymbol{\theta}$ sigui totalment desconegut ho anomenem Cas 1. És possible conèixer només algunes de les components del vector de paràmetres $\boldsymbol{\theta}$, però aquesta és una situació poc usual a la pràctica i per tant, no la considerarem. Stephens (1974) fa un estudi detallat de tots els casos quan $\mathrm{F}(\cdot)$ és una distribució normal univariant.

En el Cas 0 la transformació $p_i = \mathrm{F}(y_i; \boldsymbol{\theta})$ dóna una mostra que, sota H_0, es distribueix uniformement a l'interval $[0, 1]$. En aquest cas, es pot comprovar que la distribució dels estadístics depèn únicament de la mida (n) de la mostra. Aquestes distribucions no responen a cap dels models coneguts però podem trobar-les numèricament mitjançant mètodes de Monte Carlo. Així doncs, per a cada valor de n disposem d'unes taules (Stephens, 1970) que recullen els principals percentils de cada estadístic. Per als estadístics de la família de Cràmer-von Mises, és possible trobar també els percentils de la seva distribució asimptòtica. Per eliminar la dependència de la mida de la mostra i evitar així l'ús de taules molt extenses, Stephens (1970) modifica els estadístics. D'aquesta manera tenim una única distribució per a cada estadístic transformat. A Stephens i D'Agostino (1986, pàg. 105, taula 4.2) trobem una taula amb les modificacions i els principals percentils dels estadístics transformats.

Resumim a continuació, i de forma esquemàtica, el procediment a seguir per aplicar un test de bondat d'ajust a una mostra en el Cas 0:

a. Ordenar la mostra: $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$.

b. Calcular els valors $p_{(i)}$ utilitzant l'expressió $p_{(i)} = \mathrm{F}(y_{(i)}; \boldsymbol{\theta})$, on $\mathrm{F}(\cdot)$ és la distribució suposada en H_0.

c. Calcular l'estadístic d'interès segons les fórmules 5.1.

d. Modificar el valor de l'estadístic utilitzant la fórmula apropiada de l'esmentada taula (Stephens i D'Agostino, 1986, pàg. 105). Anomenarem T a aquest valor.

e. Escollir el nivell de significació α i buscar el percentil utilitzant l'esmentada taula. Anomenarem T^* a aquest percentil.

f. Comparar els valors T i T^*. Si T és superior a T^*, rebutgem la hipòtesi H_0 a un nivell de significació α.

En el Cas 1, donat que no coneixem el vector de paràmetres $\boldsymbol{\theta}$, ens caldrà estimar-lo a partir de la mostra. A continuació podem calcular els valors $p_{(i)}$ utilitzant la transformació $p_{(i)} = \mathrm{F}(y_{(i)}; \hat{\boldsymbol{\theta}})$, on $\hat{\boldsymbol{\theta}}$ representa l'estimació de $\boldsymbol{\theta}$. Stephens i D'Agostino (1986) adverteixen que, en aquest cas, la distribució de la mostra $p_{(i)}$ sota H_0, no serà uniforme a l'interval $[0, 1]$. Adverteixen, a més, que la distribució dels estadístics serà molt diferent a la distribució que obteníem en el Cas 0 ja que ara depèn de la distribució contrastada, de la mida de la mostra, del veritable valor dels paràmetres desconeguts així com del mètode d'estimació escollit. Per tant, si utilitzem la taula esmentada anteriorment (Stephens i D'Agostino, 1986, pàg. 105, taula 4.2) per contrastar la hipòtesi nul·la cometrem un greu error.

Stephens i D'Agostino (1986) observen que quan les components desconegudes del vector $\boldsymbol{\theta}$ són paràmetres de localització o d'escala, si aquests s'estimen amb un mètode apropiat, llavors la distribució dels estadístics no depèn del veritable valor d'aquests paràmetres. No succeeix el mateix amb els paràmetres de forma.

Així doncs, en el Cas 1, no podem donar unes taules amb els percentils de cada estadístic ja que seran diferents en cada situació. Caldrà que cada usuari, una vegada decideixi la distribució a contrastar i la mida de la mostra, calculi la distribució dels estadístics per a diferents valors dels paràmetres desconeguts (no de localització ni d'escala). En aquest cas també és possible trobar la distribució asimptòtica dels estadístics de la família de Cramér-von Mises. Trobem a la literatura nombrosos treballs d'aquest tipus. Per exemple Stephens i D'Agostino (1986) estudien la distribució dels estadístics quan el model de contrast és exponencial, normal, Weibull, Gamma i Cauchy, entre d'altres. Puig i Stephens (1998, 2000, 2001) fan un estudi complet per a la distribució de Laplace i la distribució hiperbòlica.

5.2 Proves per a la distribució $\mathcal{SN}^1(\mu, \sigma, \lambda)$

En aquesta secció desenvolupem tests de bondat d'ajust per a la distribució normal asimètrica utilitzant la metodologia descrita a l'apartat anterior. Gupta i Tuhao (2001) utilitzen l'estadístic de Kolmogorov-Smirnov per validar l'ajust d'un model normal asimètric a unes dades però es limiten a considerar el Cas 0. El nostre objectiu és trobar la distribució de cada un

dels estadístics de contrast en el Cas 1, és a dir, en el cas on tots els paràmetres del model normal asimètric són desconeguts.

La densitat normal asimètrica univariant té tres paràmetres que hem anomenat μ, σ i λ. Els paràmetres μ i σ són de localització i d'escala, però λ és un paràmetre de forma. Per aquesta raó, ens caldrà trobar la distribució dels estadístics per a diferents valors de λ així com per a diferents valors de la mida de la mostra. Iniciem aquest apartat amb el càlcul, mitjançant mètodes de Monte Carlo, d'unes taules que resumiran la distribució dels estadístics. Seguidament, indiquem de forma esquemàtica el procediment a seguir per contrastar un model normal asimètric a partir d'una mostra. Acabem l'apartat amb un estudi de la potència dels tests utilitzant diverses distribucions alternatives.

5.2.1 Taules

Habitualment es calcula la distribució dels estadístics EDF utilitzant mètodes de Monte Carlo. L'estratègia consisteix en trobar una mostra de mida gran (10000 individus) de cada un dels estadístics i descriure la distribució a partir dels seus percentils més importants.

En el nostre cas i donat que la distribució depèn de la mida de la mostra n i del paràmetre λ, caldrà repetir tot el procés per a cada possible combinació de valors n i λ.

Així doncs, fixats n i λ, es generen 10000 mostres, $y_1, y_2, \ldots, y_n$, de mida n, d'una distribució $\mathcal{SN}(\mu = 0, \sigma = 1, \lambda)$. Fixem els paràmetres $\mu = 0$ i $\sigma = 1$ perquè el seu valor no afecta al resultat final. Per a cada mostra generada es calculen els estimadors dels paràmetres, $\widehat{\mu}$, $\widehat{\sigma}$ i $\widehat{\lambda}$ utilitzant el mètode de màxima versemblança. En aquest procés d'estimació, hem escollit la parametrització μ, σ i λ en comptes de la parametrització centrada amb μ^*, σ^* i γ_1 (vegeu discussió a la secció 1.3.3). La raó principal d'escollir aquesta parametrització és perquè no fem cap tipus d'inferència posterior amb els resultats. Amb els valors de les estimacions i a partir de la mostra simulada prèviament ordenada, calculem les probabilitats acumulades $p_{(i)} = \mathrm{F}(y_{(i)}; \widehat{\mu}, \widehat{\sigma}, \widehat{\lambda})$, per a $i = 1, 2, \ldots, n$, on $\mathrm{F}(\cdot)$ representa la funció de distribució de la normal asimètrica univariant. Tal i com hem indicat en la secció 1.3.1, aquesta funció de distribució és igual a la diferència entre la funció de distribució d'una normal estàndard, $\mathcal{N}(0, 1)$, i la funció d'Owen, T, multiplicada per 2. Així doncs podem calcular $p_{(i)}$ utilitzant

l'expressió

$$p_{(i)} = \Phi\left(\frac{y_{(i)} - \widehat{\mu}}{\widehat{\sigma}}\right) - 2\mathrm{T}\left(\frac{y_{(i)} - \widehat{\mu}}{\widehat{\sigma}}, \widehat{\lambda}\right).$$

Per avaluar la funció d'Owen, apliquem els algorismes AS 76 (Young i Minder, 1974) i AS R55 (Youn-Min, 1985). L'algorisme AS 76 conté la funció d'Owen original mentre que l'algorisme AS R55 incorpora una correcció. Amb aquestes subrutines calculem les probabilitats $p_{(i)}$ amb 6 decimals exactes.

A continuació utilitzem les fórmules (5.1) per calcular el valor de cada un dels estadístics. D'aquesta manera, obtenim una mostra de 10000 valors per a cada estadístic. Seguidament ordenem la mostra i calculem els percentils $P_q\%$, utilitzant l'expressió

$$P_q = x_{\left(\frac{qN}{100}+0.5\right)}$$

on ara $x_{(i)}$ denota l'i-èsim element de la mostra de cada estadístic, N és igual a 10000 i escollim $q = 50, 75, 85, 95, 97.5$ i 99. Cal recordar que si el valor $\frac{qN}{100} + 0.5$ no és un nombre enter, es calcula la mitjana aritmètica dels dos valors $x_{(i)}$ d'ordre més proper a $\frac{qN}{100} + 0.5$. Repetirem tot aquest procés per a cada combinació de valors dels paràmetres n i λ.

El rang de variació del paràmetre λ d'una distribució normal asimètrica és tota la recta real. Hem comprovat però, que la distribució dels estadístics A^2, W^2, U^2, D i V no depèn del signe de λ. És a dir, s'obté la mateixa distribució amb mostres provinents d'una $\mathcal{SN}(\mu, \sigma, \lambda)$ i d'una $\mathcal{SN}(\mu, \sigma, -\lambda)$. Per aquesta raó, tan sols considerem valors de λ a l'interval $[0, +\infty)$. Per altra banda, sabem que per a valors de λ superiors a 20 les distribucions normals asimètriques resultants són pràcticament indistingibles entre elles. Així doncs, restringim λ a l'interval $[0, 20]$. Dins d'aquest interval decidim prendre 8 valors, concretament $\lambda = 0, 1, 2, 3, 5, 7, 10$ i 20. Inicialment fem increments de mida 1 però en augmentar el valor de λ podem també augmentar aquest increment.

Pel que fa a la mida de la mostra n, hem decidit no prendre valors inferiors a 50 per evitar problemes en el càlcul dels estimadors de màxima versemblança (vegeu secció 1.3.3). A mesura que anem augmentant la mida de la mostra, la distribució dels estadístics tendeix cap a la seva distribució asimptòtica. Hem calculat la distribució per a diferents valors de n grans i hem pogut comprovar empíricament que quan $n > 500$ la distribució dels estadístics és pràcticament igual a la distribució que s'obté pel valor $n = 500$. Així doncs hem escollit valors de n dins de l'interval $[50, 500]$, concretament $n = 50, 100, 150, 200, 300$ i 500. Inicialment ens

cal fer increments de valor 50 però posteriorment podem fer increments de mida més gran.

De la taula 5.1 fins a la taula 5.5 detallem el valor dels percentils finals. Notem que a les taules corresponents als estadístics de Kolmogorov-Smirnov i de Kuiper, es dóna el valor dels percentils de $\sqrt{n}D$ i $\sqrt{n}V$ en comptes dels percentils dels estadístics originals D i V. Aquesta és una pràctica comú en aquest tipus de taules, introduïda per Lilliefords (1967), ja que quan n tendeix a ∞ els estadístics D i V tendeixen a una distribució nul·la, en canvi $\sqrt{n}D$ i $\sqrt{n}V$ s'estabilitzen. Així doncs, si volem aplicar el contrast utilitzant els estadístics D i V, haurem de multiplicar el seu valor per $\sqrt{n}$ abans de comparar amb els percentils de les taules 5.3 i 5.4.

Taula 5.1: *Percentils de l'estadístic* A^2.

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
0	50	0.2738	0.3696	0.4334	0.4859	0.5846	0.6925	0.8423
	100	0.2768	0.3756	0.4380	0.4879	0.5762	0.6691	0.7894
	150	0.2806	0.3750	0.4420	0.4948	0.5815	0.6647	0.7838
	200	0.2798	0.3789	0.4485	0.5041	0.5980	0.6845	0.7943
	300	0.2844	0.3803	0.4480	0.5022	0.5916	0.6801	0.7993
	500	0.2881	0.3843	0.4526	0.5065	0.5956	0.6829	0.8095
1	50	0.2724	0.3692	0.4367	0.4874	0.5796	0.6916	0.8220
	100	0.2781	0.3754	0.4407	0.4904	0.5865	0.6672	0.7951
	150	0.2803	0.3725	0.4380	0.4881	0.5700	0.6542	0.7868
	200	0.2799	0.3754	0.4420	0.4890	0.5750	0.6648	0.7740
	300	0.2803	0.3777	0.4429	0.4946	0.5776	0.6696	0.7891
	500	0.2849	0.3804	0.4473	0.5025	0.5922	0.6946	0.7998
2	50	0.2768	0.3793	0.4536	0.5124	0.6188	0.7264	0.8689
	100	0.2777	0.3704	0.4395	0.4928	0.5844	0.6888	0.8289
	150	0.2758	0.3737	0.4420	0.4923	0.5927	0.6899	0.8235
	200	0.2751	0.3679	0.4363	0.4879	0.5792	0.6781	0.7787
	300	0.2754	0.3708	0.4348	0.4878	0.5762	0.6581	0.7771
	500	0.2782	0.3727	0.4413	0.4948	0.5722	0.6551	0.7863
3	50	0.2844	0.3939	0.4735	0.5414	0.6534	0.7679	0.9086
	100	0.2849	0.3896	0.4678	0.5291	0.6410	0.7608	0.9253
	150	0.2826	0.3861	0.4620	0.5246	0.6350	0.7503	0.9128
	200	0.2845	0.3847	0.4576	0.5170	0.6186	0.7317	0.9039
	300	0.2864	0.3876	0.4594	0.5215	0.6189	0.7131	0.8658
	500	0.2848	0.3829	0.4526	0.5067	0.6078	0.7035	0.8450

(*continua a la pàgina següent)*

Taula 5.1 *(continuació)*

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
5	50	0.3116	0.4429	0.5378	0.6128	0.7472	0.8838	1.0594
	100	0.3133	0.4425	0.5378	0.6201	0.7607	0.9162	1.1303
	150	0.3080	0.4361	0.5252	0.6019	0.7569	0.9102	1.1142
	200	0.3075	0.4318	0.5217	0.5991	0.7406	0.8774	1.1178
	300	0.3062	0.4330	0.5301	0.6092	0.7338	0.8853	1.0988
	500	0.3137	0.4394	0.5327	0.6052	0.7382	0.8719	1.0391
7	50	0.3327	0.4727	0.5743	0.6605	0.8067	0.9781	1.2175
	100	0.3345	0.4813	0.5954	0.6827	0.8575	1.0304	1.2643
	150	0.3309	0.4772	0.5904	0.6814	0.8507	1.0321	1.2437
	200	0.3300	0.4754	0.5817	0.6720	0.8523	1.0186	1.2830
	300	0.3302	0.4785	0.5922	0.6853	0.8510	1.0240	1.2498
	500	0.3328	0.4822	0.5874	0.6798	0.8464	1.0106	1.2160
10	50	0.3565	0.5187	0.6372	0.7316	0.9079	1.1064	1.4119
	100	0.3584	0.5214	0.6482	0.7591	0.9325	1.1135	1.3743
	150	0.3620	0.5286	0.6598	0.7705	0.9672	1.1465	1.4166
	200	0.3579	0.5293	0.6593	0.7663	0.9593	1.1527	1.4249
	300	0.3555	0.5234	0.6528	0.7574	0.9410	1.1460	1.4119
	500	0.3618	0.5247	0.6482	0.7489	0.9190	1.1356	1.4197
20	50	0.3932	0.5832	0.7259	0.8416	1.0723	1.3744	1.7777
	100	0.3996	0.6023	0.7486	0.8644	1.0716	1.2946	1.6351
	150	0.3985	0.6027	0.7465	0.8718	1.0896	1.3395	1.6461
	200	0.4057	0.6065	0.7575	0.8756	1.1098	1.3506	1.6820
	300	0.4064	0.6160	0.7703	0.8972	1.1366	1.3815	1.7149
	500	0.4098	0.6185	0.7723	0.8993	1.1470	1.3851	1.7439

Taula 5.2: *Percentils de l'estadístic W^2.*

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
0	50	0.0411	0.0581	0.0703	0.0803	0.1003	0.1222	0.1579
	100	0.0415	0.0588	0.0706	0.0800	0.0959	0.1156	0.1369
	150	0.0418	0.0588	0.0715	0.0807	0.0973	0.1127	0.1352
	200	0.0416	0.0597	0.0722	0.0817	0.0992	0.1171	0.1350
	300	0.0423	0.0598	0.0728	0.0822	0.0990	0.1158	0.1369
	500	0.0432	0.0601	0.0727	0.0834	0.1003	0.1157	0.1397
1	50	0.0408	0.0583	0.0709	0.0813	0.0996	0.1241	0.1567
	100	0.0413	0.0587	0.0708	0.0809	0.0977	0.1155	0.1405
	150	0.0418	0.0591	0.0707	0.0797	0.0949	0.1104	0.1360
	200	0.0417	0.0587	0.0708	0.0801	0.0961	0.1129	0.1337
	300	0.0418	0.0591	0.0706	0.0805	0.0955	0.1143	0.1373
	500	0.0425	0.0599	0.0722	0.0828	0.0993	0.1169	0.1384
2	50	0.0419	0.0604	0.0747	0.0861	0.1101	0.1319	0.1660
	100	0.0417	0.0587	0.0714	0.0811	0.1002	0.1181	0.1479
	150	0.0412	0.0588	0.0716	0.0815	0.0986	0.1168	0.1482
	200	0.0406	0.0576	0.0698	0.0791	0.0977	0.1143	0.1378
	300	0.0406	0.0577	0.0702	0.0795	0.0964	0.1148	0.1351
	500	0.0410	0.0578	0.0710	0.0817	0.0964	0.1121	0.1341
3	50	0.0435	0.0645	0.0809	0.0951	0.1178	0.1437	0.1751
	100	0.0435	0.0627	0.0777	0.0906	0.1136	0.1412	0.1810
	150	0.0431	0.0624	0.0774	0.0900	0.1106	0.1364	0.1693
	200	0.0431	0.0615	0.0756	0.0871	0.1081	0.1302	0.1646
	300	0.0432	0.0621	0.0763	0.0879	0.1084	0.1273	0.1575
	500	0.0432	0.0613	0.0745	0.0853	0.1043	0.1247	0.1526

(*continua a la pàgina següent)*

Taula 5.2 *(continuació)*

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
5	50	0.0497	0.0761	0.0955	0.1123	0.1414	0.1685	0.2105
	100	0.0503	0.0755	0.0958	0.1128	0.1454	0.1805	0.2255
	150	0.0489	0.0745	0.0935	0.1101	0.1432	0.1754	0.2218
	200	0.0491	0.0739	0.0931	0.1083	0.1385	0.1708	0.2267
	300	0.0487	0.0744	0.0939	0.1106	0.1379	0.1707	0.2172
	500	0.0500	0.0756	0.0944	0.1092	0.1378	0.1684	0.2031
7	50	0.0544	0.0826	0.1044	0.1216	0.1527	0.1866	0.2359
	100	0.0546	0.0856	0.1083	0.1275	0.1659	0.2017	0.2569
	150	0.0544	0.0851	0.1082	0.1286	0.1654	0.2051	0.2534
	200	0.0542	0.0845	0.1068	0.1273	0.1652	0.2022	0.2574
	300	0.0545	0.0846	0.1089	0.1288	0.1653	0.2013	0.2544
	500	0.0547	0.0860	0.1086	0.1276	0.1640	0.2026	0.2462
10	50	0.0591	0.0911	0.1152	0.1362	0.1723	0.2073	0.2758
	100	0.0597	0.0933	0.1204	0.1423	0.1828	0.2224	0.2764
	150	0.0603	0.0949	0.1233	0.1466	0.1907	0.2303	0.2858
	200	0.0604	0.0957	0.1230	0.1459	0.1879	0.2295	0.2856
	300	0.0598	0.0949	0.1219	0.1445	0.1832	0.2310	0.2890
	500	0.0608	0.0956	0.1216	0.1442	0.1820	0.2247	0.2833
20	50	0.0645	0.1014	0.1298	0.1519	0.1995	0.2524	0.3367
	100	0.0669	0.1068	0.1374	0.1610	0.2071	0.2520	0.3200
	150	0.0676	0.1075	0.1392	0.1642	0.2111	0.2645	0.3288
	200	0.0686	0.1096	0.1402	0.1672	0.2137	0.2671	0.3362
	300	0.0694	0.1115	0.1440	0.1708	0.2219	0.2728	0.3441
	500	0.0692	0.1114	0.1430	0.1724	0.2247	0.2761	0.3545

Taula 5.3: *Percentils de l'estadístic U^2.*

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
0	50	0.0405	0.0569	0.0675	0.0770	0.0928	0.1100	0.1303
	100	0.0412	0.0582	0.0699	0.0789	0.0951	0.1131	0.1327
	150	0.0415	0.0584	0.0710	0.0803	0.0968	0.1117	0.1335
	200	0.0413	0.0592	0.0717	0.0810	0.0985	0.1161	0.1345
	300	0.0420	0.0593	0.0721	0.0814	0.0980	0.1152	0.1363
	500	0.0428	0.0595	0.0723	0.0828	0.0996	0.1148	0.1387
1	50	0.0401	0.0568	0.0683	0.0781	0.0931	0.1074	0.1336
	100	0.0410	0.0582	0.0700	0.0797	0.0964	0.1137	0.1346
	150	0.0415	0.0587	0.0701	0.0791	0.0938	0.1090	0.1347
	200	0.0413	0.0583	0.0703	0.0795	0.0955	0.1123	0.1330
	300	0.0415	0.0586	0.0700	0.0799	0.0946	0.1137	0.1364
	500	0.0422	0.0593	0.0716	0.0822	0.0985	0.1162	0.1371
2	50	0.0407	0.0576	0.0703	0.0802	0.0972	0.1156	0.1357
	100	0.0408	0.0571	0.0690	0.0779	0.0954	0.1111	0.1337
	150	0.0405	0.0573	0.0695	0.0791	0.0953	0.1127	0.1366
	200	0.0399	0.0565	0.0680	0.0771	0.0944	0.1111	0.1304
	300	0.0399	0.0566	0.0686	0.0779	0.0937	0.1114	0.1311
	500	0.0405	0.0570	0.0698	0.0800	0.0944	0.1094	0.1302
3	50	0.0410	0.0593	0.0716	0.0822	0.1001	0.1161	0.1401
	100	0.0413	0.0586	0.0710	0.0814	0.0991	0.1185	0.1452
	150	0.0412	0.0584	0.0716	0.0810	0.0998	0.1179	0.1442
	200	0.0412	0.0580	0.0699	0.0799	0.0979	0.1156	0.1416
	300	0.0414	0.0585	0.0707	0.0809	0.0986	0.1144	0.1388
	500	0.0413	0.0579	0.0696	0.0790	0.0955	0.1114	0.1383

(*continua a la pàgina següent)*

Taula 5.3 *(continuació)*

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
5	50	0.0446	0.0644	0.0792	0.0905	0.1096	0.1273	0.1571
	100	0.0448	0.0649	0.0794	0.0918	0.1126	0.1340	0.1614
	150	0.0441	0.0639	0.0785	0.0901	0.1111	0.1329	0.1589
	200	0.0438	0.0636	0.0782	0.0897	0.1095	0.1297	0.1602
	300	0.0436	0.0639	0.0786	0.0913	0.1121	0.1337	0.1656
	500	0.0447	0.0651	0.0792	0.0908	0.1103	0.1327	0.1575
7	50	0.0468	0.0673	0.0814	0.0935	0.1142	0.1336	0.1632
	100	0.0470	0.0695	0.0854	0.0982	0.1201	0.1448	0.1716
	150	0.0464	0.0698	0.0852	0.0981	0.1211	0.1457	0.1768
	200	0.0468	0.0692	0.0844	0.0979	0.1213	0.1439	0.1709
	300	0.0469	0.0693	0.0855	0.0984	0.1205	0.1438	0.1764
	500	0.0469	0.0699	0.0859	0.0982	0.1210	0.1456	0.1726
10	50	0.0489	0.0714	0.0877	0.1007	0.1222	0.1473	0.1791
	100	0.0498	0.0732	0.0905	0.1053	0.1287	0.1503	0.1828
	150	0.0498	0.0739	0.0919	0.1070	0.1310	0.1559	0.1859
	200	0.0497	0.0743	0.0916	0.1064	0.1305	0.1557	0.1911
	300	0.0497	0.0734	0.0914	0.1056	0.1307	0.1554	0.1874
	500	0.0499	0.0739	0.0912	0.1049	0.1286	0.1542	0.1882
20	50	0.0528	0.0775	0.0949	0.1085	0.1362	0.1599	0.1942
	100	0.0535	0.0795	0.0991	0.1138	0.1390	0.1632	0.1986
	150	0.0537	0.0792	0.0984	0.1145	0.1407	0.1644	0.2008
	200	0.0546	0.0807	0.1004	0.1150	0.1399	0.1672	0.2085
	300	0.0546	0.0815	0.1013	0.1167	0.1426	0.1723	0.2123
	500	0.0545	0.0816	0.1019	0.1172	0.1452	0.1751	0.2133

Taula 5.4: *Percentils de l'estadístic $\sqrt{n}D$.*

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
0	50	0.5450	0.6346	0.6911	0.7333	0.8048	0.8707	0.9579
	100	0.5531	0.6422	0.6995	0.7389	0.8004	0.8605	0.9300
	150	0.5576	0.6481	0.7057	0.7439	0.7974	0.8506	0.9099
	200	0.5600	0.6529	0.7074	0.7481	0.8096	0.8560	0.9178
	300	0.5655	0.6578	0.7126	0.7498	0.8142	0.8674	0.9317
	500	0.5711	0.6617	0.7173	0.7556	0.8195	0.8781	0.9385
1	50	0.5438	0.6353	0.6945	0.7358	0.8108	0.8737	0.9512
	100	0.5523	0.6443	0.6973	0.7363	0.7997	0.8611	0.9332
	150	0.5597	0.6489	0.7021	0.7393	0.7956	0.8547	0.9194
	200	0.5573	0.6446	0.7010	0.7418	0.8057	0.8618	0.9310
	300	0.5615	0.6516	0.7068	0.7444	0.8020	0.8577	0.9260
	500	0.5698	0.6612	0.7176	0.7556	0.8148	0.8723	0.9423
2	50	0.5533	0.6491	0.7128	0.7619	0.8342	0.9029	0.9909
	100	0.5536	0.6485	0.7054	0.7469	0.8127	0.8802	0.9614
	150	0.5560	0.6478	0.7055	0.7486	0.8170	0.8813	0.9525
	200	0.5566	0.6487	0.7000	0.7413	0.8079	0.8682	0.9446
	300	0.5581	0.6497	0.7044	0.7450	0.8063	0.8667	0.9341
	500	0.5615	0.6534	0.7097	0.7493	0.8104	0.8667	0.9322
3	50	0.5649	0.6701	0.7387	0.7877	0.8669	0.9361	1.0356
	100	0.5673	0.6686	0.7347	0.7875	0.8682	0.9417	1.0442
	150	0.5712	0.6750	0.7384	0.7882	0.8685	0.9397	1.0187
	200	0.5705	0.6712	0.7327	0.7792	0.8603	0.9405	1.0290
	300	0.5750	0.6762	0.7432	0.7879	0.8586	0.9237	1.0067
	500	0.5747	0.6733	0.7370	0.7815	0.8557	0.9244	1.0040

(*continua a la pàgina següent)*

Taula 5.4 *(continuació)*

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
5	50	0.5973	0.7169	0.7936	0.8484	0.9294	1.0108	1.1022
	100	0.6047	0.7302	0.8043	0.8610	0.9553	1.0393	1.1383
	150	0.6051	0.7271	0.8027	0.8566	0.9440	1.0405	1.1297
	200	0.6073	0.7254	0.8013	0.8553	0.9437	1.0290	1.1366
	300	0.6086	0.7275	0.8059	0.8619	0.9528	1.0365	1.1324
	500	0.6172	0.7377	0.8072	0.8642	0.9465	1.0297	1.1245
7	50	0.6174	0.7410	0.8162	0.8690	0.9567	1.0361	1.1453
	100	0.6324	0.7599	0.8383	0.8984	0.9884	1.0714	1.1749
	150	0.6314	0.7602	0.8452	0.9036	1.0012	1.0780	1.1979
	200	0.6339	0.7671	0.8456	0.9051	1.0068	1.0872	1.1909
	300	0.6367	0.7690	0.8487	0.9098	1.0022	1.0921	1.2015
	500	0.6409	0.7721	0.8524	0.9128	1.0029	1.0865	1.1907
10	50	0.6371	0.7653	0.8453	0.9018	0.9953	1.0755	1.2034
	100	0.6485	0.7829	0.8708	0.9333	1.0279	1.1102	1.2076
	150	0.6547	0.7939	0.8827	0.9404	1.0419	1.1285	1.2249
	200	0.6580	0.7944	0.8826	0.9474	1.0411	1.1278	1.2315
	300	0.6604	0.7978	0.8801	0.9407	1.0389	1.1276	1.2431
	500	0.6632	0.8005	0.8870	0.9479	1.0364	1.1323	1.2363
20	50	0.6575	0.7952	0.8762	0.9359	1.0455	1.1414	1.2671
	100	0.6767	0.8183	0.9068	0.9699	1.0692	1.1557	1.2532
	150	0.6803	0.8250	0.9083	0.9715	1.0818	1.1687	1.2729
	200	0.6853	0.8300	0.9211	0.9847	1.0807	1.1790	1.2838
	300	0.6892	0.8357	0.9254	0.9959	1.0986	1.1963	1.3061
	500	0.6939	0.8423	0.9338	0.9953	1.0986	1.1999	1.3182

Taula 5.5: *Percentils de l'estadístic* $\sqrt{n}V$.

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
0	50	0.7104	0.8157	0.8772	0.9224	0.9844	1.0526	1.1322
	100	1.0046	1.1536	1.2405	1.3044	1.3922	1.4886	1.6012
	150	1.0109	1.1604	1.2534	1.3150	1.4152	1.4995	1.5989
	200	1.0175	1.1693	1.2574	1.3258	1.4267	1.5133	1.6116
	300	1.0285	1.1771	1.2641	1.3274	1.4244	1.5202	1.6139
	500	1.0337	1.1813	1.2712	1.3358	1.4400	1.5243	1.6288
1	50	0.9846	1.1288	1.2127	1.2731	1.3644	1.4562	1.5580
	100	1.0047	1.1572	1.2400	1.3041	1.3968	1.4884	1.6011
	150	1.1395	1.1625	1.2492	1.3093	1.4067	1.4916	1.5955
	200	1.0146	1.1589	1.2527	1.3139	1.4192	1.5129	1.6186
	300	1.0207	1.1700	1.2571	1.3205	1.4180	1.5026	1.6139
	500	1.0346	1.1842	1.2743	1.3363	1.4333	1.5317	1.6276
2	50	0.9863	1.1324	1.2234	1.2867	1.3833	1.4733	1.5773
	100	0.9991	1.1460	1.2302	1.2914	1.3845	1.4728	1.5764
	150	1.0048	1.1511	1.2410	1.3045	1.4044	1.4875	1.5864
	200	1.0048	1.1524	1.2397	1.3070	1.4033	1.4796	1.5746
	300	1.0103	1.1549	1.2459	1.3072	1.4073	1.4908	1.6115
	500	1.0201	1.1650	1.2567	1.3242	1.4186	1.5015	1.5959
3	50	0.9891	1.1376	1.2250	1.2852	1.3786	1.4646	1.5718
	100	1.0023	1.1481	1.2376	1.3024	1.4033	1.4929	1.6007
	150	1.0055	1.1544	1.2450	1.3096	1.4104	1.4924	1.6056
	200	1.0100	1.1554	1.2439	1.3059	1.4029	1.4940	1.6149
	300	1.0176	1.1660	1.2518	1.3132	1.4132	1.4993	1.6016
	500	1.0185	1.1636	1.2468	1.3103	1.4145	1.4955	1.6100

(*continua a la pàgina següent*)

Taula 5.5 *(continuació)*

		nivell de significació α						
λ	**n**	**0.5**	**0.25**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**
5	50	1.0136	1.1680	1.2587	1.3208	1.4238	1.5215	1.6397
	100	1.0276	1.1798	1.2742	1.3432	1.4429	1.5405	1.6477
	150	1.0284	1.1822	1.2718	1.3393	1.4408	1.5347	1.6489
	200	1.0290	1.1817	1.2763	1.3410	1.4460	1.5333	1.6405
	300	1.0292	1.1858	1.2809	1.3465	1.4542	1.5483	1.6558
	500	1.0427	1.1994	1.2869	1.3522	1.4501	1.5366	1.6607
7	50	1.0274	1.1798	1.2731	1.3407	1.4395	1.5313	1.6382
	100	1.0434	1.2031	1.2988	1.3649	1.4731	1.5632	1.6793
	150	1.0435	1.2091	1.3063	1.3754	1.5178	1.6063	1.7379
	200	1.0488	1.2117	1.3057	1.3738	1.5302	1.6268	1.7447
	300	1.0564	1.2176	1.3129	1.3803	1.5313	1.6309	1.7572
	500	1.0588	1.2256	1.3197	1.3855	1.5257	1.6350	1.7383
10	50	0.0429	1.2043	1.3004	1.3678	1.4740	1.5697	1.7039
	100	1.0614	1.2265	1.3307	1.4012	1.5096	1.6002	1.7163
	150	1.0650	1.2356	1.3369	1.4136	1.5178	1.6063	1.7379
	200	1.0714	1.2394	1.3404	1.4105	1.5302	1.6268	1.7447
	300	1.0737	1.2424	1.3444	1.4187	1.5313	1.6309	1.7572
	500	1.0802	1.2448	1.3490	1.4177	1.5257	1.6350	1.7383
20	50	1.0715	1.2366	1.3418	1.4122	1.5271	1.6326	1.7391
	100	1.0890	1.2629	1.3699	1.4404	1.5569	1.6527	1.7740
	150	1.0982	1.2720	1.3749	1.4484	1.5658	1.6686	1.7873
	200	1.1063	1.2773	1.3848	1.4564	1.5667	1.6729	1.8074
	300	1.1066	1.2850	1.3900	1.4603	1.5748	1.6901	1.8244
	500	1.1160	1.2938	1.4016	1.4785	1.5916	1.6918	1.8296

5.2.2 Procediment per fer un test

Amb les taules que hem construït per a cada un dels estadístics, podem procedir a aplicar un test de bondat d'ajust a una mostra. En aquest cas la hipòtesi nul·la és:
H_0 : $\ll$ La mostra $y_1, y_2, \ldots, y_n$ prové d'una població amb distribució normal asimètrica univariant amb paràmetres desconeguts$\gg$.

Resumim a continuació, i de forma esquemàtica, el procediment a seguir per aplicar aquest test de bondat d'ajust:

a. Ordenar la mostra: $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$.

b. Trobar els estimadors $\widehat{\mu}, \widehat{\sigma}$ i $\widehat{\lambda}$ dels paràmetres μ, σ i λ amb el mètode de màxima versemblança.

c. Calcular els valors $p_{(i)}$ $(i = 1, 2, \ldots, n)$ utilitzant la transformació $p_{(i)} = \mathrm{F}(y_{(i)}; \widehat{\mu}, \widehat{\sigma}, \widehat{\lambda})$ on $\mathrm{F}(\cdot)$ és la funció de distribució d'una normal asimètrica univariant.

d. Calcular l'estadístic d'interès (el denotem T) a partir de les fórmules (5.1).

e. Escollir el nivell de significació, α, i buscar el percentil de l'estadístic d'interès, que denotarem per T^*, amb la taula apropiada (taules 5.1-5.5). Trobarem aquest percentil a la línia que correspon als valors $\widehat{\lambda}$ i n. Si algun d'aquests valors, $\widehat{\lambda}$ o n, no apareix directament a la taula, suggerim aplicar interpolació lineal entre els dos valors més propers. Si cap dels dos valors, $\widehat{\lambda}$ i n, apareixen a la taula, suggerim aplicar interpolació lineal bivariant. En el cas que $n > 500$ o $\widehat{\lambda} > 20$ suggerim utilitzar els valors corresponents a $n = 500$ i $\widehat{\lambda} = 20$. Finalment, en cas que $\widehat{\lambda} < 0$ caldrà utilitzar els percentils que corresponen al valor $|\widehat{\lambda}|$.

f. Comparar els valors T i T^*. Si T és superior a T^* rebutgem la hipòtesi nul·la a un nivell de significació α.

5.2.3 Càlcul del veritable nivell de significació

Quan utilitzem aquest tipus de contrast no coneixem el valor real del paràmetre λ i l'aproximem per la seva estimació $\widehat{\lambda}$. A més, a la pràctica, no trobem a les taules la línia corresponent

als valors exactes de $\widehat{\lambda}$ i n, i conseqüentment apliquem interpolació lineal o interpolació bivariant. El fet de treballar amb l'estimació $\widehat{\lambda}$ i posteriorment aplicar interpolació provoca una variació en el nivell de significació. En aquesta secció estudiarem la diferència que existeix entre el nivell de significació real i el nivell de significació α que escollim.

Per fer aquest estudi simulem 10000 mostres de mida $n = 100$ i 350 d'una variable aleatòria amb distribució $\mathcal{SN}(0, 1, \lambda)$. Escollim $\lambda = 0, 1, 2, 3, 5, 7, 10$ i 20. Apliquem el test de bondat d'ajust a cada una de les mostres seguint el procediment descrit a l'apartat anterior. Escollim un nivell de significació del 5%, és a dir, $\alpha = 0.05$. Aquest nivell de significació ens indica la probabilitat d'error de tipus 1, és a dir, la probabilitat de rebutjar H_0 en cas que sigui certa. Així doncs, hauríem de rebutjar la hipòtesi nul·la aproximadament en uns 500 casos. Les taules 5.5 i 5.6 contenen el nombre de mostres que no passen el test quan $n = 100$ i $n = 375$ respectivament.

En el cas $n = 100$ hem d'aplicar tan sols interpolació lineal quan el valor $\widehat{\lambda}$ no apareix a les taules 5.1-5.5. Observant els valors de la taula 5.6 trobem que el veritable nivell de significació es mou entre el 3.61% i el 5.76%, és a dir, tenim un nivell de significació bastant semblant al 5% escollit.

En el cas $n = 375$ ens cal aplicar sempre interpolació ja que no apareix aquesta mida mostral a les taules. La interpolació serà del tipus bivariant quan el valor de l'estimador $\widehat{\lambda}$ no coincideixi amb cap dels valors λ de les taules. Els valors de la taula 5.7 tornen a constatar que el veritable nivell de significació es mou en valors propers al 5%. En particular obtenim un nivell de significació que es mou entre els valors 4.5% i 5.63%. Cal notar que aquests valors són millors que els obtinguts en el cas $n = 100$. Això pot semblar paradoxal ja que en aquest cas utilitzem interpolació doble mentre que en el cas $n = 100$ utilitzem tan sols interpolació lineal simple. L'explicació d'aquest fenomen és senzilla. En primer lloc sabem que amb valors de n grans l'estimació del paràmetre λ és més precisa. Sabem també que la distribució dels estadístics és diferent si variem el valor de n. Però aquesta, a mesura que el valor de n augmenta, tendeix a la seva distribució asimptòtica. Per tant, amb un valor de n gran, els percentils que utilitzarem en la interpolació seran molt similars i el corresponent valor interpolat serà més aproximat al real.

Hem repetit aquest estudi simulant 10000 mostres de mida $n > 500$ d'una distribució $\mathcal{SN}(0, 1, \lambda)$, per a $\lambda = 0, 1, 2, 3, 5, 7, 10$ i 20. Hem aplicat a cada mostra el test de bondat

d'ajust utilitzant les taules corresponents a $n = 500$ amb un nivell de significació del 5%. El veritable nivell de significació ha resultat molt similar a l'escollit 5%, concretament varia entre els valors 4.5% i 5.5%. Així doncs, podem utilitzar les taules anteriors quan $n > 500$ sense cap problema. No recomanem aplicar aquests tests de bondat d'ajust quan la mostra tingui menys de 50 individus per evitar els problemes derivats de l'estimació dels paràmetres del model normal asimètric.

Taula 5.6: *Nombre de mostres rebutjades entre 10000. Mostres de mida 100 i generades amb un model* $\mathcal{SN}(0, 1, \lambda)$.

	λ							
	0	**1**	**2**	**3**	**5**	**7**	**10**	**20**
A^2	471	445	409	385	403	396	423	576
W^2	447	432	389	361	399	394	444	557
U^2	475	462	437	421	435	443	492	560
D	437	434	393	397	428	455	471	542
V	529	513	459	453	443	449	477	517

Taula 5.7: *Nombre de mostres rebutjades entre 10000. Mostres de mida 375 i generades amb un model* $\mathcal{SN}(0, 1, \lambda)$.

	λ							
	0	**1**	**2**	**3**	**5**	**7**	**10**	**20**
A^2	563	539	498	464	484	470	471	526
W^2	533	530	470	450	487	485	468	523
U^2	538	540	487	458	497	508	495	508
D	522	531	447	457	507	511	480	486
V	503	547	486	451	471	486	493	505

5.2.4 Estudi de potència

En aquesta secció estudiarem la potència de cada estadístic utilitzant diverses distribucions alternatives. Green i Hegazy (1976) adverteixen de la dificultat en descriure el comportament

dels estadístics. La seva potència depèn de la distribució alternativa escollida així com de la mida de la mostra.

Pel que fa a distribucions alternatives, escollim famílies asimètriques com per exemple la lognormal $\Lambda(\mu, \sigma)$, l'exponencial $\text{Exp}(\lambda)$ i la χ^2 amb γ graus de llibertat. Escollim també famílies amb paràmetre de forma com per exemple la Gamma $\text{Ga}(\alpha, \beta)$ i la Weibull $\text{W}(\alpha, \beta)$. Considerem també diverses mides mostrals, en concret prenem els valors de n utilitzats per construir les taules, és a dir, $n = 50, 100, 150, 200, 300$ i 500.

Per a cada valor de n i per a cada distribució alternativa, simulem 10000 mostres. Seguidament apliquem el test de bondat d'ajust a cada una amb un nivell de significació del 5%. Per calcular la potència, tan sols cal comptar el nombre de vegades en què rebutgem la hipòtesi nul·la. A la taula 5.8 mostrem els resultats. La primera columna de la taula especifica l'estadístic utilitzat per fer el contrast. La segona columna conté la mida de la mostra i les altres contenen la proporció de mostres rebutjades en cada cas. En la primera fila de la taula especifiquem la distribució alternativa utilitzada.

Observant els resultats de la taula 5.8 podem concloure que l'estadístic d'Anderson-Darling, A^2, presenta una potència superior quan la distribució alternativa és lognormal, exponencial, Weibull o χ^2, però la potència dels l'estadístics U^2 i V és més gran quan la distribució alternativa és una Gamma. En general, podem afirmar que la família Cramér-von Mises té una potència millor que la família de Kolmogorov-Smirnov. Concretament, els estadístics d'Anderson-Darling (A^2) i Cramér-von Mises (W^2) presenten una potència més gran que l'estadístic de Kolmogorov-Smirnov (D). Els estadístics de Watson (U^2) i Kuiper (V), tenen un comportament bastant similar però U^2 té una potència lleugerament superior.

Una altra característica que s'observa a la taula 5.8 és que la potència dels estadístics augmenta amb la mida de la mostra.

Si analitzem amb detall el comportament dels estadístics per a cada una de les distribucions alternatives, observarem que la seva potència és molt gran quan la distribució és $\Lambda(0, 1)$, $\text{W}(1, 2)$, $\text{Exp}(1)$ i χ^2 amb 2 i 4 graus de llibertat. Així doncs, podem afirmar que els tests aconsegueixen discriminar perfectament la distribució normal asimètrica de les distribucions anteriors. La probabilitat de rebutjar la hipòtesi nul·la és considerablement alta en aquests casos, sobretot quan la distribució alternativa és $\Lambda(0, 1)$, $\text{Exp}(1)$ i χ^2_2. Per altra banda, observem que el test no aconsegueix discriminar entre una distribució normal asimètrica i una

distribució Gamma. En els casos Ga$(4, 1)$ i Ga$(5, 1)$ la proporció de mostres rebutjades és molt baixa, en alguns casos és menor que el nivell de significació. Els estadístics tampoc presenten una bona potència quan la distribució alternativa és $\Lambda(0, 0.3)$.

5.2.5 Prova de bondat d'ajust independent del paràmetre de forma

Dalla-Valle (2001) ha desenvolupat recentment un test de bondat d'ajust per a la distribució normal asimètrica utilitzant l'estadístic d'Anderson-Darling (A^2). Tal i com hem indicat en els apartats anteriors, la distribució de l'estadístic A^2 depèn de la mida de la mostra i del veritable valor del paràmetre de forma. L'objectiu del treball de Dalla-Valle és trobar una transformació de l'estadístic per eliminar la influència que hi exerceix el paràmetre de forma.

L'estudi comença amb la simulació de 12000 mostres de mida n d'una normal asimètrica $\mathcal{SN}(0, 1, \lambda)$. Dalla-Valle pren 9 valors per al paràmetre de forma, $\lambda = 0, 1, 2, 3, 4, 5, 7, 10, 20$, i 5 valors per a la mida de la mostra, $n = 50, 100, 150, 250, 500$. Amb les mostres simulades es calculen els estimadors de màxima versemblança. En el procés de maximització s'utilitza la parametrització centrada i en els casos on el coeficient d'asimetria arriba al seu valor màxim s'utilitza la modificació de l'algorisme del màxim versemblant. Recordem que aquesta modificació consistia en reiniciar el procés iteratiu de maximització i parar-lo quan la funció de logversemblança arribés a un valor no significativament inferior al màxim (vegeu secció 1.3.3). Amb els estimadors trobats es procedeix a calcular l'estadístic d'Anderson-Darling. Una vegada acabada aquesta primera fase s'obté per a cada n i cada λ, una mostra de 12000 valors de l'estadístic A^2. D'aquesta mostra es calcula el percentil 95%, que denotarem $P_{0.95,n,\lambda}$.

Tal i com hem indicat a l'inici d'aquest apartat, es pretén eliminar la influència del paràmetre λ sobre l'estadístic de Anderson-Darling. El procediment que s'aplica consisteix en, fixat un valor del paràmetre n, prendre els 9 percentils $P_{0.95,n,\lambda}$ ($\lambda = 0, 1, 2, 3, 4, 5, 7, 10, 20$) i aplicar-los una transformació de manera que els 9 valors transformats siguin pràcticament iguals. Per buscar la transformació adient, Dalla-Valle utilitza la següent estratègia: en primer lloc i amb cada mostra de 9 percentils (una mostra per a cada valor de n) es valora la bondat d'ajust d'un model lineal expressat com una suma de dues parts, una d'elles depenent del paràmetre λ i l'altra totalment independent de λ. En el seu treball, Dalla-Valle (2001)

Taula 5.8: *Potència dels estadístics amb un nivell de significació del 5%. Nombre de mostres rebutjades entre 10000.*

		Alternativa							
Test	n	$\Lambda(0,1)$	$\Lambda(0,0.3)$	$\mathrm{W}(1,2)$	$\mathrm{Ga}(4,1)$	$\mathrm{Ga}(5,1)$	$\mathrm{Exp}(1)$	χ^2_2	χ^2_4
A^2	50	9396	478	1216	397	322	7601	7595	1190
	100	9979	646	1557	544	485	9428	9467	1572
	150	9999	845	1986	682	548	9899	9914	2003
	200	9999	941	2476	723	666	9982	9978	2388
	300	10000	1256	3468	942	792	10000	10000	3430
	500	10000	1793	5284	1329	1052	10000	10000	5296
W^2	50	9384	426	1240	409	299	7331	7298	1185
	100	9978	548	1495	529	439	9298	9298	1505
	150	9999	722	1878	621	486	9867	9876	1894
	200	9999	795	2325	636	613	9975	9973	2192
	300	10000	994	3151	802	659	10000	9999	3146
	500	10000	1421	4772	1082	917	10000	10000	4786
U^2	50	8648	525	924	469	403	5392	5284	890
	100	9879	643	1295	590	534	7953	7981	1305
	150	9995	777	1736	726	581	9269	9292	1793
	200	9999	835	2319	773	695	9770	9755	2249
	300	10000	1035	3404	963	782	9975	9975	3340
	500	10000	1517	5169	1486	1096	9999	10000	5180
D	50	9090	469	1098	436	375	6521	6439	1057
	100	9948	533	1265	547	475	8765	8808	1311
	150	9998	695	1579	613	518	9685	9721	1626
	200	9999	779	2021	630	584	9907	9914	1922
	300	10000	923	2689	752	618	9996	9995	2674
	500	10000	1237	4100	975	803	10000	10000	4160
V	50	8443	522	813	508	441	5052	4994	799
	100	9855	630	1090	589	530	7720	7783	1123
	150	9988	775	1496	705	604	9211	9216	1551
	200	9999	840	1952	777	670	9749	9744	1919
	300	10000	1001	2961	974	786	9977	9973	2925
	500	10000	1342	4506	1513	1059	9999	10000	4548

examina diversos models però comprova que el més adequat és

$$\hat{P}_{0.95,n,\lambda} = a_n + b_n(1-\delta^2)^{k_n} + \epsilon \tag{5.2}$$

on a_n i b_n són els paràmetres del model lineal estimats a partir de la mostra, ϵ és una variable aleatòria amb $\mathrm{E}(\epsilon) = 0$ i $\mathrm{var}(\epsilon) = \sigma^2$, δ és el paràmetre relacionat amb λ a partir de les fórmules (1.8), i k_n és una constant escollida dins un ampli rang de valors. Seguidament es calcula $\hat{P}_{0.95,n,\lambda}$ ($\lambda = 0, 1, 2, 3, 4, 5, 7, 10, 20$), el valor estimat del percentil 95% de l'estadístic A^2 a partir del model lineal ajustat. Per reduir la variabilitat que indueix el paràmetre λ en els valors reals $P_{0.95,n,\lambda}$ Dalla-Valle proposa aplicar la següent transformació:

$$\mathrm{h}(P_{0.95,n,\lambda}) = \frac{P_{0.95,n,\lambda}}{\hat{P}_{0.95,n,\lambda}} \qquad \lambda = 0, 1, 2, 3, 4, 5, 7, 10, 20. \tag{5.3}$$

Anomenem a $\mathrm{h}(P_{0.95,n,\lambda})$ l'estadístic Anderson-Darling transformat. Certament els 9 valors de l'estadístic transformat són molt similars. A continuació, cal establir un criteri per seleccionar només un dels nou valors $\mathrm{h}(P_{0.95,n,\lambda})$ com a representant del percentil 95% de l'estadístic Anderson-Darling transformat. Es contemplen diversos criteris de selecció com ara, el valor màxim, el valor mínim o la mitjana mostral dels nou valors transformats, però basant-se en un estudi comparatiu dels diferents criteris, Dalla-Valle proposa escollir el mínim. Així doncs, obtenim com a resultat final un percentil 95% per a l'estadístic Anderson-Darling transformat.

Seguidament es repeteix aquest procés per a cada valor de n. En la taula 5.9 reproduïm el valor del percentil 95% de l'estadístic transformat així com els paràmetres del model lineal (5.2) que cal utilitzar en la transformació.

Taula 5.9: *Coeficients de la transformació de l'estadístic A^2 i el seu percentil 95%.*

n	a_n	b_n	k_n	**percentil 95%**
50	0.4450	0.1246	-1/3	0.9586
100	0.4732	0.0782	-0.4	0.9367
150	0.4913	0.0603	-0.45	0.9572
250	0.5146	0.0429	-0.5	0.9392
500	0.5199	0.0471	-0.45	0.9401

Per contrastar la hipòtesi nul·la que una mostra $y_1, y_2, \ldots, y_n$ prové d'un model normal asimètric amb aquest test, caldrà seguir els passos que detallem a continuació:

a. Ordenar la mostra: $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$.

b. Trobar els estimadors $\widehat{\mu}, \widehat{\sigma}$ i $\widehat{\lambda}$ dels paràmetres μ, σ i λ amb el mètode de màxima versemblança i utilitzant la parametrització centrada.

c. Calcular els valors $p_{(i)}$ utilitzant la transformació $p_{(i)} = \mathrm{F}(y_{(i)}; \widehat{\mu}, \widehat{\sigma}, \widehat{\lambda})$ per a $i = 1, \ldots, n$, on $\mathrm{F}(\cdot)$ és la funció de distribució d'una normal asimètrica univariant.

d. Calcular l'estadístic d'Anderson-Darling a partir de les fórmules (5.1). Anomenarem T a aquest valor.

e. Aplicar a T la transformació (5.3) utilitzant els coeficients a_n, b_n i k_n de la taula 5.9. Si el valor de n no apareix a la taula, aplicar interpolació lineal entre els dos valors més propers. Si el valor de n és superior a 500, utilitzar el valor corresponent a $n = 500$.

f. Comparar el valor de l'estadístic transformat amb el percentil 95% que apareix a la taula 5.9. Si el primer és més gran que el segon, rebutgem H_0 amb un nivell de significació 0.05.

Dalla-Valle (2001) completa el seu treball calculant la potència del test amb diverses distribucions alternatives, concretament utilitza la distribució lognormal amb $\mu = 0$ i $\sigma = 1$ i les distribucions $\mathrm{Ga}(4, 1)$ i $\mathrm{Ga}(5, 1)$. Per a cada distribució genera 10000 mostres de mida 100, aplica el test i calcula la proporció de mostres rebutjades. Els resultats són similars als obtinguts en el nostre estudi de potència; en el cas de la lognormal, la proporció de mostres rebutjades és de l'ordre del 75%. Així doncs, es pot considerar que el test aconsegueix discriminar perfectament la distribució normal asimètrica de la distribució lognormal. No podem dir el mateix per a la distribució Gamma ja que la proporció de mostres rebutjades és bastant inferior, concretament de l'ordre del 5%.

5.3 Proves per a la distribució $\mathcal{L}^D(\mu, \mathbf{\Sigma})$

Per validar l'ajust del model normal logístic additiu a un conjunt de dades composicionals $\mathbf{X}$, és suficient aplicar un test de normalitat multivariant a les dades alr transformades,

$\mathbf{Y} = \text{alr}(\mathbf{X})$. Sabem que existeixen una gran quantitat i una gran varietat de contrastos de normalitat multivariant. No obstant això, Aitchison (1986, secció 7.3) utilitza els contrastos de bondat d'ajust basats en la funció de distribució empírica i proposa valorar la normalitat multivariant en tres fases:

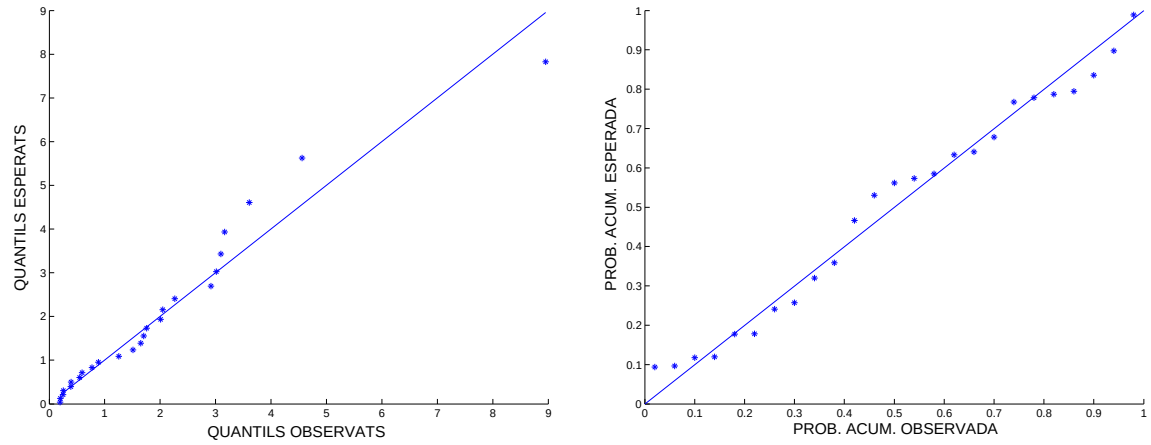
1. Proves de normalitat de les marginals. Valorar la normalitat de cada marginal aplicant un test univariant basat en els estadístics d'Anderson-Darling, Cramér-von Mises i Watson.

2. Proves bivariants dels angles. Aplicar un test a cada parell de variables basat en el següent resultat: si $(\mathrm{u}_1, \mathrm{u}_2) \sim \mathcal{N}^2(\mathbf{0}, \mathbf{I}_2)$ llavors l'angle entre el vector d'origen $(0, 0)$ i extrem $(\mathrm{u}_1, \mathrm{u}_2)$ i l'eix u_1, es distribueix uniformement a l'interval $[0, 2\pi]$. Amb les dades mostrals calculem els angles i utilitzant els mateixos estadístics valorem la bondat d'ajust dels angles a la distribució uniforme.

3. Prova multivariant dels radis. Per a cada observació o fila de la matriu $\mathbf{Y}$, denotada per $\mathbf{y}_i$ $(i = 1, 2, \ldots, n)$, es calculen les distàncies o radis $d_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}})' \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}})$ on $\hat{\boldsymbol{\mu}}$ i $\widehat{\mathbf{\Sigma}}$ són el vector de mitjanes mostrals i la matriu de covariàncies mostrals corregides de les dades alr transformades. Sota la hipòtesis de normalitat multivariant de les dades, aquestes distàncies es distribueixen aproximadament segons una llei χ^2 amb $D-1$ graus de llibertat. Utilitzant els mateixos estadístics valorem la bondat d'ajust dels radis a la distribució χ^2.

Aitchison (1986) dóna les taules amb les modificacions que cal aplicar als estadístics mostrals i els percentils dels estadístics Anderson-Darling, Cramér-von Mises i Watson necessaris per dur a terme els tres tipus de contrastos. Suggereix també complementar l'anàlisi amb un mètode gràfic, com per exemple un diagrama Q-Q o un diagrama P-P.

En el capítol 3 hem intuït que un model aln ajustava adequadament la subcomposició de les tres primeres components (A,B i C) de la base de dades Hongite. Si utilitzem la tercera component com a denominador en la transformació alr i apliquen els contrastos descrits, es conclou, en tots els casos, acceptar la hipòtesi nul·la i per tant acceptar el model normal logístic additiu. Arribem a la mateixa conclusió si observem el diagrama Q-Q o el diagrama P-P de les distàncies d_i (figura 5.1).



(a) (b)

Figura 5.1: *Diagrames Q-Q*(a) *i diagrama P-P* (b) *de les distàncies de Mahalanobis*

Veiem doncs, que validar l'ajust d'un model normal logístic additiu a unes dades composicionals és equivalent a validar el model normal a les dades alr transformades. Barceló-Vidal (1996) aplica la mateixa tècnica per validar els models basats en les transformacions Box-Cox però adverteix d'una dificultat: els estadístics utilitzats en els punts 1 i 2 depenen, en general, de la component utilitzada com a divisor en la transformació logquocient additiva. En la propietat 3.5 hem vist que la distribució normal logística additiva és tancada per la permutació de les seves components, és a dir, podem utilitzar qualsevol altre part com a comú denominador en la transformació alr i el resultat és també una distribució normal logística additiva. Malgrat tot, si apliquem aquests tests de bondat d'ajust podem arribar a conclusions diferents. Amb les dades de l'exemple anterior, calculem els estadístics sobre cada marginal després d'aplicar les transformacions $\text{alr}_1, \text{alr}_2$ i alr_3 i observem com efectivament obtenim resultats diferents. A la taula 5.10 recollim el valor de l'estadístic de Cramér-von Mises convenientment transformat segons indica Aitchison (1986, secció 7.3). Podem observar clarament la dependència del denominador.

Taula 5.10: *Estadístic Cramér-von Mises de cada marginal.*

Transformació	$\text{alr}_1 = \ln(\mathbf{x}_{-1}/x_1)$	$\text{alr}_2 = \ln(\mathbf{x}_{-2}/x_2)$	$\text{alr}_3 = \ln(\mathbf{x}_{-3}/x_3)$
marginal 1	0.2034	0.2034	0.3150
marginal 2	0.3150	0.2531	0.2531

Tot i tenir resultats diferents, en aquest cas arribem sempre a la conclusió que el model normal logístic additiu és adequat. Els problemes sorgiran quan amb un denominador acceptem el model i amb un altre el rebutgem. Cal tenir en compte que aquesta dificultat desapareix si treballem amb el contrast dels radis.

Donada $\mathbf{x} \sim \mathcal{L}^D(\boldsymbol{\mu}, \mathbf{\Sigma})$, sabem per definició que $\text{alr}(\mathbf{x}) \sim \mathcal{N}^{D-1}(\boldsymbol{\mu}, \mathbf{\Sigma})$. Hem vist que existeix una relació matricial entre els vectors $\text{alr}(\mathbf{x}), \text{clr}(\mathbf{x})$ i $\text{ilr}(\mathbf{x})$. Així doncs, i per les propietats de la distribució normal multivariant, sabem que

$$\text{alr}(\mathbf{x}) \sim \mathcal{N}^{D-1}(\boldsymbol{\mu}, \mathbf{\Sigma}) \Leftrightarrow \text{ilr}(\mathbf{x}) \sim \mathcal{N}^{D-1}(\boldsymbol{\xi}, \mathbf{\Upsilon}) \Leftrightarrow \text{clr}(\mathbf{x}) \sim \mathcal{N}^{D}(\boldsymbol{\lambda}, \mathbf{\Gamma}).$$

Validar la normalitat multivariant de les dades alr transformades és equivalent a validar la normalitat de les dades clr o ilr transformades. No aconsellem treballar amb la transformació logquocient centrada ja que dóna lloc a una distribució degenerada. Per valorar l'ajust d'un model normal a la mostra ilr transformada, aplicarem les tres mateixes fases anteriors. En aplicar la transformació ilr a les dades, tenim la llibertat d'escollir la base ortonormal. Així doncs, la dificultat de la dependència del denominador que presentaven les dades alr transformades, es tradueix en aquest cas, a una dependència de la base ortonormal escollida en la transformació. Com a exemple, prenem la subcomposició de les tres primeres components de la base de dades Hongite i apliquem-li la transformació ilr utilitzant tres bases ortonormals diferents. Denotem per $\text{ilr}_{B_1}, \text{ilr}_{B_2}$ i ilr_{B_3} les transformacions logquocients isomètriques amb les bases

$$\begin{aligned}
B_1 &= \{\mathcal{C}(e^{\sqrt{1/2}}, e^{-\sqrt{1/2}}, 1), \mathcal{C}(e^{\sqrt{1/6}}, e^{\sqrt{1/6}}, e^{-\sqrt{2/3}})\} \\
B_2 &= \{\mathcal{C}(e^{\sqrt{2/3}}, e^{-\sqrt{1/6}}, e^{-\sqrt{1/6}}), \mathcal{C}(1, e^{\sqrt{1/2}}, e^{-\sqrt{1/2}})\} \\
B_3 &= \{\mathcal{C}(e^{-\sqrt{2/7}}, e^{\sqrt{9/14}}, e^{-\sqrt{1/14}}), \mathcal{C}(e^{\sqrt{8/21}}, e^{\sqrt{1/42}}, e^{-\sqrt{25/42}})\}
\end{aligned}$$

Observem que la base B_2 s'obté reordenant les composicions de la base B_1. Contràriament, la base B_3 és totalment diferent. A la taula 5.11, hem recollit el valor de l'estadístic de Anderson-Darling convenientment transformat segons indica Aitchison (1986, secció 7.3). Podem comprovar com els valors dels estadístics EDF en els contrastos de normalitat marginal depenen de la base ortonormal.

Tot i les diferències entre els valors obtinguts, es conclou en tots els casos acceptar la normalitat de cada marginal.

Taula 5.11: *Estadístic Anderson-Darling de cada marginal.*

Transformació	ilr_{B_1}	ilr_{B_2}	ilr_{B_3}
marginal 1	0.2034	0.2894	0.2201
marginal 2	0.2748	0.2531	0.2974

Les marginals i els angles bivariants de les dades transformades, retenen una part de la variabilitat total del conjunt de dades multivariant. A més, aquestes variabilitats parcials canvien si apliquem al vector una transformació lineal. Per aquesta raó, podem obtenir conclusions diferents en aplicar els contrastos. Si les nostres dades tenen un perfil marcadament normal, com per exemple el conjunt Hongite, podem aplicar transformacions lineals i tot i obtenir valors diferents dels estadístics, gairebé en cap cas tindrem motius per rebutjar la hipòtesis nul·la de normalitat. No passarà el mateix si les dades tenen un perfil diferent al d'un model normal. Pensem per exemple en un conjunt de dades amb un perfil semblant al de la distribució normal asimètrica. Si cada distribució marginal presenta una clara asimetria, rebutjarem la normalitat. No obstant això, sabem de l'existència d'una transformació lineal que provoca que una component absorbeixi tota la asimetria i per tant, és possible que totes les marginals, excepte una, passin el test. Si a més, la asimetria d'aquesta marginal té un valor moderat, és possible acceptar la hipòtesi de normalitat. Entre aquests dos casos hi ha un ampli ventall de possibilitats de manera que, quan la normalitat no sigui massa clara, podem acceptar-la o rebutjar-la depenent de la transformació lineal que apliquem a les dades. És per aquesta raó que observem la dependència del denominador o la dependència de la base ortonormal ja que un canvi de denominador o de base en les transformacions alr o ilr correspon a una transformació lineal.

Per donar una solució definitiva al problema, s'ha desenvolupat recentment una metodologia equivalent que evita la dificultat de la dependència del denominador o de la base (Aitchison et al., 2003). Aquest treball es basa en la descomposició en valors singulars de les dades composicionals i la caracterització de la variabilitat composicional en termes de les operacions $\oplus$ i $\otimes$ de $\mathcal{S}^D$. Sigui $\mathbf{X}$ la mostra de mida n d'una composició de D parts i sigui $\hat{\mathbf{g}}$ l'estimació del centre de la composició aleatòria. Mitjançant la descomposició en valors

singulars, sabem que qualsevol matriu $\mathbf{X}$ es pot descomposar de la forma

$$\mathbf{x}_j = \hat{\mathbf{g}} \oplus (u_{j,1} p_1 \otimes \mathbf{b}_1) \oplus \cdots \oplus (u_{j,D-1} p_{D-1} \otimes \mathbf{b}_{D-1}) \qquad (j = 1, 2, \ldots, n), \tag{5.4}$$

on $\mathbf{x}_j$ indica la fila j de la matriu $\mathbf{X}$, $p_1 > p_2 > \cdots > p_{D-1}$ són valors singulars positius, $\mathbf{b}_i$ $(i = 1, 2, \ldots, D-1)$ són composicions, i els valors $u_{j,i}$ són les components específiques de cada fila. Observem que si una composició aleatòria $\mathbf{x}$ descomposa de la forma

$$\mathbf{x} = \text{cen}[\mathbf{x}] \oplus (\mathrm{u}_1 \pi_1 \otimes \boldsymbol{\beta}_1) \oplus \cdots \oplus (\mathrm{u}_{D-1} \pi_{D-1} \otimes \boldsymbol{\beta}_{D-1}), \tag{5.5}$$

llavors la distribució del vector $\mathbf{u} = (\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_{D-1})'$ caracteritza la distribució de $\mathbf{x}$. Per exemple, si $\mathbf{u} \sim \mathcal{N}^{D-1}(\mathbf{0}, \mathbf{I})$, llavors la distribució de $\mathbf{x}$ és normal logística additiva. Així doncs podem validar la distribució aln amb un contrast de normalitat multivariant sobre el vector $\mathbf{u}$.

En el fons, la descomposició en valors singulars és equivalent a buscar les coordenades en una base ortonormal específica. No obstant això, ens proporciona dos grans avantatges. En primer lloc, obtenim que les components del vector $\mathbf{u}$ són independents i per tant, la normalitat de les marginals és condició necessària i suficient per a la normalitat conjunta. En segon lloc, sabem que les components de $\mathbf{u}$ estan ordenades de major a menor variabilitat i que les r primeres retenen una proporció

$$q_r = \frac{\sum_{i=1}^{r} p_i^2}{\sum_{i=1}^{D-1} p_i^2}$$

de la variabilitat total de la composició $\mathbf{x}$. Aquesta propietat serà important des del punt de vista de les proves de bondat d'ajust.

A la pràctica, abans de validar el model normal logístic additiu, caldrà calcular els elements que intervenen en l'expressió (5.4). La descomposició en valors singulars de la matriu $\mathbf{Z} = (\mathbf{I}_n - (1/n)\mathbf{J}_n) \ln \mathbf{X} (\mathbf{I}_D - (1/D)\mathbf{J}_D)$ és igual a $\mathbf{Z} = \mathbf{V}\mathbf{P}\mathbf{W}'$, amb $\mathbf{P} = \text{diag}(p_1, p_2, \ldots, p_{D-1})$, i $\mathbf{V}$ i $\mathbf{W}$ matrius amb columnes ortonormals i de suma 0. Tal i com s'indica a Aitchison et al. (2003) la matriu $\text{clr}^{-1}(\mathbf{W})$ té per files les composicions $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{D-1}$. En aquest cas la notació $\text{clr}^{-1}(\mathbf{W})$ indica que s'aplica la transformació clr inversa a cada fila de $\mathbf{W}$. La matriu $\mathbf{U} = \sqrt{n}\mathbf{V}$ conté la mostra del vector $\mathbf{u}$ la distribució del qual és, sota la hipòtesis nul·la, normal multivariant. Per validar aquesta hipòtesis es proposa utilitzar els contrastos de normalitat univariant, el contrast bivariant dels angles i el contrast multivariant dels radis seguint el següent esquema:

1. Aplicar el contrast de normalitat a la primera columna de **U**.

2. Aplicar el contrast de normalitat a la segona columna de **U**, el contrast dels angles i el contrast dels radis a les dues primeres columnes de **U**.

3. Aplicar el contrast de normalitat a la tercera columna, el contrast dels angles entre la primera i la tercera columna, el contrast dels angles entre la segona i la tercera columna i el contrast dels radis a les tres primeres columnes de **U**.

Podem utilitzar les taules que apareixen a Aitchison (1986, secció 7.3) per obtenir els percentils dels estadístics transformats.

Teòricament, caldria continuar amb l'esquema anterior fins a l'última columna de la matriu **U**. A la pràctica no sempre serà necessari, ja que és possible que les primeres columnes retinguin un gran percentatge de la variabilitat total. A Aitchison et al. (2003) s'aplica aquesta metodologia a certes bases de dades, entre elles, el conjunt Hongite amb 5 parts (Aitchison, 1986). Les dues primeres columnes de la matriu **U** que en resulta, retenen una proporció del 99.6% de la variabilitat total. Donat que aquestes dues columnes passen tots els tests, es conclou que el conjunt de dades es pot ajustar amb un model normal logístic additiu sense aplicar cap mena de contrast a les altres tres columnes. Òbviament, la pregunta important és si el valor 99.6% es pot considerar suficient, però la resposta dependrà totalment dels objectius concrets de la investigació que es du a terme.

Abans hem indicat que la normalitat de cada component del vector **u** és condició necessària i suficient per garantir la normalitat logística additiva de la composició **x**. És important remarcar que hem arribat a la matriu **U** amb un procés d'estimació dels paràmetres desconeguts. El fet que les columnes de **U** tinguin mitjana 0 i estiguin incorrelacionades, prové d'un procés de centratge on hi intervé l'estimació de cen[**x**] i d'un procés d'ortogonalització on hi intervé l'estimació de la covariància. Per aquesta raó, a Aitchison et al. (2003) es recomana aplicar, a banda de les proves marginals, les proves de bondat d'ajust dels angles i dels radis com una mesura extra de control.

5.4 Proves per a la distribució $\mathcal{LS}^D(\mu, \mathbf{\Sigma}, \alpha)$

Donada $\mathbf{x} \sim \mathcal{LS}^D(\boldsymbol{\mu}, \mathbf{\Sigma}, \boldsymbol{\alpha})$, sabem per definició que $\mathbf{y} = \text{alr}(\mathbf{x}) \sim \mathcal{SN}^{D-1}(\boldsymbol{\mu}, \mathbf{\Sigma}, \boldsymbol{\alpha})$. Per tant, validar l'ajust del model normal asimètric logístic additiu a unes dades $\mathbf{X}$, és equivalent a validar el model normal asimètric a la mostra alr transformada, $\mathbf{Y} = \text{alr}(\mathbf{X})$.

A la secció 5.2. hem desenvolupat tan sols proves de bondat d'ajust per a la distribució normal asimètrica univariant. No obstant això, sabem que cada component d'un vector normal asimètric té una distribució normal asimètrica univariant. Per tant, un primer pas raonable és aplicar, a cada component del vector $\mathbf{y}$, els tests de bondat d'ajust desenvolupats a la secció 5.2. Si tan sols una d'aquestes components no passa el test, podrem rebutjar la hipòtesi nul·la. En cas contrari, no tenim una condició suficient per assegurar la distribució del vector i caldrà contrastar la suposició amb un test multivariant.

Azzalini i Capitanio (1999) indiquen que, sota hipòtesis de normalitat asimètrica, la distància de Mahalanobis $d = (\mathbf{y} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ es distribueix segons una χ^2 amb $D-1$ graus de llibertat (aquest resultat es demostra a partir de la propietat 1.16). Si substituïm $\boldsymbol{\mu}$ i $\mathbf{\Sigma}$ de l'expressió pel valor de les estimacions màxim versemblants $\hat{\boldsymbol{\mu}}$ i $\hat{\mathbf{\Sigma}}$, podem calcular una distància per a cada observació $\mathbf{y}_i$,

$$d_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}})'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}) \qquad (i = 1, 2, \ldots, n). \tag{5.6}$$

Sobre la mostra de distàncies, aplicarem una prova de bondat d'ajust per validar la distribució χ^2_{D-1}, la mateixa que aplicàvem per al model aln. Azzalini i Capitanio (1999) suggereixen utilitzar també un mètode gràfic com, per exemple, un Q-Q plot o un P-P plot.

En el capítol 3 hem utilitzat la distribució alsn per modelitzar certs conjunts de dades composicionals, però tan sols hem validat l'ajust des d'un punt de vista intuïtiu. A continuació anem a aplicar els contrastos descrits, és a dir, una prova de normalitat asimètrica a cada marginal i una prova multivariant als radis.

El model normal asimètric logístic additiu semblava proporcionar un bon ajust a la subcomposició $(Al_2O_3, SiO_2, Fe_2O_3)$ de la base de dades Halimba. Més concretament, hem vist que l'ajust amb el model alsn millorava l'ajust obtingut amb un model aln. Validarem en primer lloc, la normalitat asimètrica de cada marginal de la mostra alr transformada amb la component Al_2O_3 com a denominador. La taula 5.12 recull el valor dels estadístics de contrast mostrals calculats a partir de les fórmules (5.1).

Taula 5.12: *Estadístics de contrast (base de dades Halimba).*

Estadístic	$\ln(SiO_2/Al_2O_3)$	$\ln(Fe_2O_3/Al_2O_3)$
A^2	1.2275	1.4522
W^2	0.1649	0.2123
U^2	0.1310	0.2093
$\sqrt{n}D$	1.1591	1.1656
$\sqrt{n}V$	1.6651	2.1199

Fixat un nivell de significació $\alpha = 0.05$ i utilitzant les taules 5.1-5.5, podrem trobar el percentil de cada estadístic. En aquest cas $n = 332$ i el valor del paràmetre de forma és $\hat{\lambda} = -4.396$ per a la primera marginal i $\hat{\lambda} = -0.810$ per a la segona. En cada cas caldrà aplicar interpolació lineal doble per obtenir els percentils. A la taula 5.13 es recullen els percentils 95% dels estadístics.

Taula 5.13: *Percentil 95% dels EDF estadístics.*

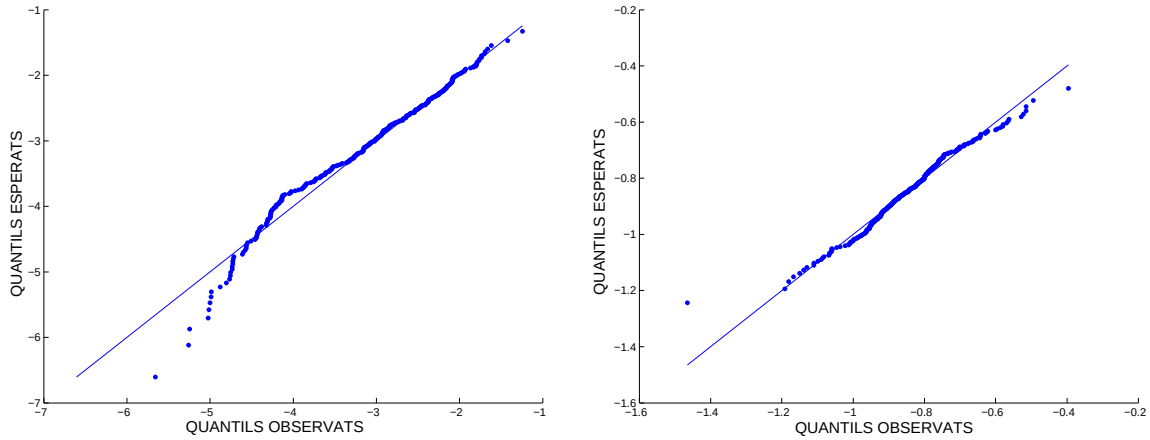
Estadístic	$\lambda = -4.396$ **i** $n = 332$	$\lambda = -0.810$ **i** $n = 332$
A^2	0.6990	0.5823
W^2	0.1288	0.0967
U^2	0.1077	0.0958
$\sqrt{n}D$	0.9235	0.8061
$\sqrt{n}V$	1.4414	1.4217

En tots els casos, el valor dels estadístics mostrals supera al valor dels percentils trobats a les taules. A la vista d'aquests resultats, cal rebutjar la hipòtesi de normalitat asimètrica per a les dues marginals. Hem completat l'estudi amb les tècniques de bondat d'ajust gràfiques. A la figura 5.2 hem representat els diagrames quantil-quantil del model normal asimètric univariant ajustat a cada marginal de la mostra alr transformada. Observant el gràfics, resulta evident rebutjar la hipòtesi de normalitat asimètrica per a la primera marginal però no obtenim uns resultats tan clars per a la segona.

Malauradament, tenim la dificultat de la dependència del denominador utilitzat en la

transformació alr. Per aquesta raó, hem repetit els contrastos anteriors utilitzant les components SiO_2 i Fe_2O_3 com a denominadors. Tot i així, la conclusió final és la mateixa en tots els casos: rebutjar la hipòtesi de normalitat asimètrica d'ambdues marginals. També arribem a la mateixa conclusió si apliquem els contrastos de normalitat asimètrica a les components de les dades ilr transformades.

No és necessari aplicar el test dels radis ja que, la normalitat asimètrica de les marginals és una condició necessària per garantir la normalitat asimètrica multivariant del vector. Així doncs, si bé el model alsn és millor que el model aln per ajustar el conjunt de dades, no proporciona un ajust suficientment raonable.



(a) *variable* $\ln(SiO_2/Al_2O_3)$ (b) *variable* $\ln(Fe_2O_3/Al_2O_3)$

Figura 5.2: *Diagrames quantil-quantil (Base de dades Halimba).*

En el capítol 3 també hem vist intuïtivament que el model alsn és adequat per modelitzar el conjunt de dades Convex (vegeu pàgina 93). A continuació validarem el model amb els contrastos de bondat d'ajust. Apliquem els test univariants descrits a la secció 5.2 a les dues marginals de la mostra alr transformada, utilitzant la tercera component com a denominador. El valor dels estadístics de contrast es recull a la taula 5.14.

Fixat un nivell de significació $\alpha = 0.05$ i utilitzant les taules 5.1-5.5, podrem trobar el percentil de cada estadístic. En aquest cas $n = 200$ i el valor del paràmetre de forma és $\hat{\lambda} = -0.861$ per a la primera marginal i $\hat{\lambda} = 3.027$ per a la segona. Aplicant interpolació lineal a les taules 5.1-5.5, obtenim els percentils que es recullen a la taula 5.15.

En tots els casos s'observa que el valor dels estadístics mostrals és inferior al valor dels

Taula 5.14: *Estadístics de contrast (base de dades Convex).*

Estadístic	$\ln(x_1/x_3)$	$\ln(x_2/x_3)$
A^2	0.1398	0.3244
W^2	0.0176	0.0594
U^2	0.0176	0.0548
$\sqrt{n}D$	0.4238	0.6305
$\sqrt{n}V$	0.7375	1.1648

Taula 5.15: *Percentil 95% dels EDF estadístics.*

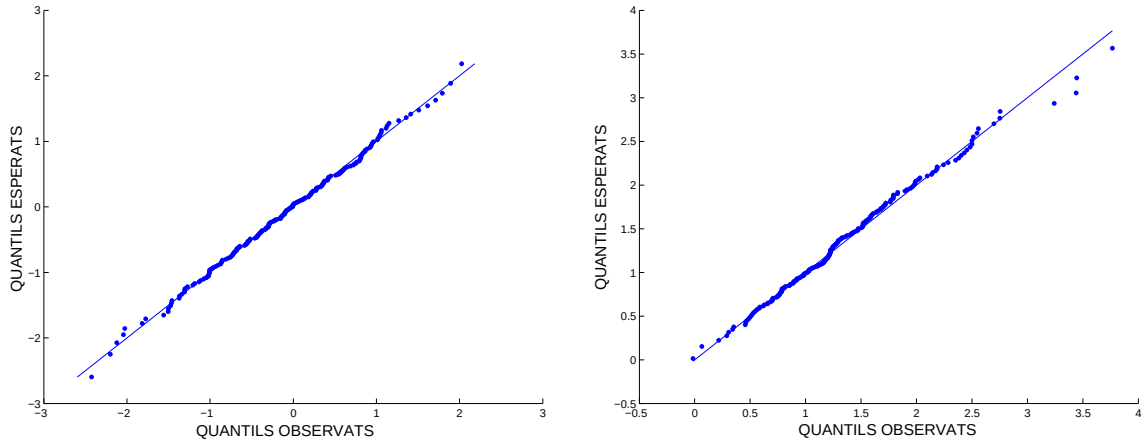
Estadístic	$\lambda = -0.861$ **i** $n = 200$	$\lambda = 3.027$ **i** $n = 200$
A^2	0.5782	0.6203
W^2	0.0965	0.1085
U^2	0.0959	0.0981
$\sqrt{n}D$	0.8062	0.8614
$\sqrt{n}V$	1.4202	1.4035

percentils trobats amb les taules 5.1-5.5. Així doncs no tenim motius suficients per rebutjar la hipòtesi de normalitat asimètrica per a les dues marginals. Hem completat l'estudi amb un diagrama quantil-quantil per a cada una de les marginals (figura 5.3). Podem observar que en els dos casos el núvol de punts s'ajusta raonablement a la línia central del gràfic i per tant es conclou acceptar el model normal asimètric univariant.

El següent pas raonable és aplicar el contrast de bondat d'ajust multivariant dels radis. Ens caldrà primer calcular les diferències

$$d_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}), \qquad (i = 1, 2, \ldots, 200) \tag{5.7}$$

on $\mathbf{y}_i$ indica la transformació logquocient additiva de la i-èsima observació. Amb els valors d_i calculem els estadístics EDF, els transformem segons indica Aitchison (1986, pàg 146, taula 7.3) o bé Stephens i D'Agostino (1986, pàg 105, taula 4.2) i obtenim $A^2 = 0.5263$, $W^2 = 0.0854$, $U^2 = 0.0608$, $D = 0.6584$ i $V = 1.0183$. En tots els casos el valor de l'estadístic és inferior al corresponent percentil 95% i, per tant, no tenim motius suficients per rebutjar



(a) *variable* $\ln(x_1/x_3)$ (b) *variable* $\ln(x_2/x_3)$

Figura 5.3: *Diagrames quantil-quantil (base de dades Convex).*

la distribució χ^2_2.

Seguint els suggeriments d'Azzalini i Capitanio (1999) realitzem el diagrama quantil-quantil de les dades d_i vers els valors teòrics suposant una distribució χ^2_2. Si inspeccionem visualment la figura 5.4 arribem a la mateixa conclusió: el model χ^2_2 ajusta adequadament la mostra de les distàncies de Mahalanobis. D'aquest estudi es conclou que el model normal asimètric logístic additiu ajusta satisfactòriament el conjunt de dades Convex.

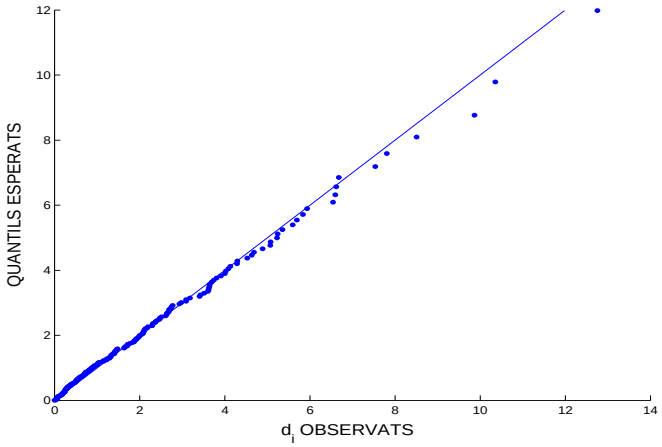


Figura 5.4: *Diagrama quantil-quantil (base de dades Convex).*

Donat que coneixem la dependència del denominador utilitzat en la transformació alr, hem repetit els contrastos de les marginals prenent les components 1 i 2 en el denominador.

Les conclusions a les que arribem són en tots els casos les mateixes.

L'altra possibilitat és aplicar la transformació ilr. Donada $\mathbf{x}$ composició aleatòria amb distribució alsn, la propietat de les transformacions lineals de la distribució normal asimètrica ens garanteix aquesta distribució per al vector ilr($\mathbf{x}$). Així doncs, és equivalent validar l'ajust del model normal asimètric a les dades ilr transformades, ilr($\mathbf{X}$). Amb la base de dades Convex arribem a la mateixa conclusió: no rebutjar el model normal asimètric logístic additiu.

Finalment, una altra possibilitat seria utilitzar les propietats 1.13 i 1.14 per desenvolupar uns tests de bondat d'ajust semblants als descrits a Aitchison et al. (2003). Deixem aquesta possibilitat com a tema obert per realitzar en futures investigacions.

5.5 Proves per a distribucions segons la metodologia STAY

En aquest treball de recerca hem definit dues distribucions segons la metodologia STAY: la distribució normal i la normal asimètrica a $\mathcal{S}^D$. El procediment per validar l'ajust d'aquests models a una mostra és senzill. Treballant amb les coordenades de la mostra respecte d'una base ortonormal, haurem de validar el model normal o el model asimètric. Aplicarem els contrastos de bondat d'ajust de normalitat o normalitat asimètrica a les marginals i el contrast dels radis a la mostra de les distàncies de Mahalanobis. En el cas del model normal a $\mathcal{S}^D$, podem aplicar també els contrastos bivariants dels radis.

Òbviament ens apareixerà la dificultat de la dependència de la base ortonormal. En el cas del model normal a $\mathcal{S}^D$, podem utilitzar la descomposició en valors singulars que es proposa a Aitchison et al. (2003).

5.6 La distància d'Aitchison d_a com a estadístic de bondat d'ajust

El popular test de bondat d'ajust χ^2 redueix una mostra de mida n de qualsevol variable a una mostra d'una distribució multinomial. El procediment és simple ja que tan sols cal una partició en k intervals del domini de la variable i un recompte del nombre d'observacions que obtenim en cada interval, $n_1, n_2, \ldots, n_k$. A partir de la distribució suposada en la hipòtesi nul·la, podem calcular la probabilitat de cada interval, que denotarem com $q_1, q_2, \ldots, q_k$. Així

doncs, es pot considerar que les freqüències observades n_i, $i = 1, 2, \ldots, k$, és una mostra d'una variable multinomial on la probabilitat de l'i-èsim interval és q_i, $i = 1, 2, \ldots, k$. Mitjançant un estadístic, per exemple l'estadístic χ^2, es procedeix a comparar les freqüències observades amb el model multinomial.

Egozcue et al. (2001) proposen un nou estadístic, basat en la distància d'Aichison entre composicions, per dur a terme aquesta comparació. Donat que els vectors $(q_1, q_2, \ldots, q_k)'$ i $(n_1/n, n_2/n, \ldots, n_k/n)'$ es poden considerar composicions, sembla adequat mesurar la seva diferència a partir de la distància d'Aitchison. Així doncs, es defineix un nou estadístic de bondat d'ajust com $D_A^2 = d_a^2((q_1, q_2, \ldots, q_k)', (n_1/n, n_2/n, \ldots, n_k/n)')$. Observem que l'estadístic serà sempre positiu i, si l'ajust és perfecte, obtenim el valor 0. Malgrat això, ens trobem amb una dificultat: la mesura esdevé infinita si alguna de les freqüències observades, n_i, és nul·la. Aquest inconvenient es pot solucionar fàcilment escollint una partició del domini de la variable que garanteixi que totes les freqüències observades siguin estrictament positives. En particular, Egozcue et al. (2001) proposen dos procediments diferents per definir els intervals. El primer consisteix en agrupar les dades en diversos grups i prendre com a partició del domini els punts mitjos entre cada grup. El segon és similar però abans de construir els grups apliquem la funció de distribució suposada en la hipòtesi nul·la a cada element de la mostra. Seguidament agrupem les dades transformades i calculem el punts mitjos entre cada grup. Per obtenir la partició del domini, apliquem als punts mitjos la funció de distribució inversa. Egozcue et al. (2001) completen el treball amb un estudi de potència comparatiu entre diversos estadístics, D_A^2, χ^2 i D entre altres, aplicant els diferents procediments per definir la partició.

L'estadístic D_A^2 proporciona un test de bondat d'ajust alternatiu per a la distribució normal asimètrica. Tot i que actualment s'ha aplicat només en el Cas 0, el cas on la distribució de la hipòtesi nul·la està totalment especificada, Egozcue et al. (2001) indiquen que es pot estendre fàcilment al cas de tenir algun paràmetre de la distribució desconegut. Existeix també la possibilitat de generalitzar aquests contrastos al cas multivariant. Deixem com a tema obert per a futures investigacions la generalització d'aquests contrastos per validar el model normal asimètric multivariant.

.

Epíleg

En aquest capítol presentem les conclusions relacionades amb els objectius marcats a l'inici d'aquesta tesi doctoral i exposem un llistat d'idees, problemes i altres aportacions que han anat sorgint a mesura que s'avançava en la investigació. És del tot evident que no hem tancat el tema d'investigació, més aviat el contrari, a mesura que hem anat aprofundint, han anat sorgint nous problemes i noves vies de desenvolupament. Som conscients que l'estudi amb profunditat de cada una de les qüestions obertes és, en realitat, una línia de recerca per estudiar en un futur.

Conclusions

En el desenvolupament d'aquesta tesi doctoral, hem realitzat una revisió de les distribucions més importants definides sobre el símplex utilitzant la metodologia MOVE. Dins d'aquest context, hem introduït el model normal asimètric logístic additiu com a generalització natural del model normal logístic additiu. A banda de les bones propietats algebraiques que presenta el model, hem comprovat que aporta la solució a una de les principals mancances de la distribució normal logística additiva, la manca d'ajust en conjunts de dades que presenten asimetria.

Per altra banda, hem estudiat el perfil de la distribució de l'amalgama de components d'una composició normal logística additiva. Hem observat que, en certs casos, la seva transformació logquocient es pot modelitzar adequadament amb una distribució normal però, en altres casos, el biaix que presenta fa necessari l'ajust amb una distribució normal asimètrica. Tot i que aquest model no sempre proporciona un bon ajust, les conclusions del nostre estudi han ampliat el ventall de possibles distribucions a utilitzar.

L'estructura d'espai euclidià ha permès definir models paramètrics utilitzant la metodologia STAY, és a dir definint directament els models sobre el símplex, sense necessitat de recórrer a les transformacions. Hem introduït el model normal a $\mathcal{S}^D$ i el model normal asimètric a $\mathcal{S}^D$ mitjançant la seva funció de densitat sobre els coeficients de la composició aleatòria respecte d'una base ortonomal. Aquestes densitats són les derivades de Radon-Nikodým de la probabilitat respecte la mesura de $\mathcal{S}^D$ coherent amb la seva estructura d'espai vectorial euclidià. Les lleis normal a $\mathcal{S}^D$ i normal asimètrica a $\mathcal{S}^D$ són equivalents, sobre $\mathcal{S}^D$, a les lleis normal logística additiva i normal asimètrica logística additiva respectivament. No obstant això, les seves propietats i els seus elements característics són diferents. En particular, el valor de l'esperança i la variància mètrica són coherents amb el centre i la variabilitat d'una composició aleatòria. En tots els càlculs i demostracions utilitzem les coordenades dels elements de $\mathcal{S}^D$ respecte d'una base ortonormal sobre les quals podem aplicar tot l'anàlisi estàndard real multivariant. Fins al moment, no coneixem cap altre treball de recerca en aquesta direcció, és a dir, on s'utilitzi l'estructura de l'espai per definir les lleis de probabilitat.

Per entendre en un cas senzill aquesta perspectiva, hem estudiat primer la recta real positiva, l'estructura de la qual permet aplicar l'anàlisi real estàndard als coeficients respecte d'una base unitària, es a dir als logaritmes de qualsevol element. Tenint en compte aquesta estructura i treballant amb els logaritmes, hem definit la llei normal a $\mathbb{R}^+$. La seva funció de densitat és la derivada de Radon-Nikodým respecte la mesura de $\mathbb{R}^+$ coherent amb l'estructura d'espai euclidià real, i la seva expressió coincideix amb la llei de freqüències de la distribució lognormal que va introduir McAlister al 1879. El model compleix les mateixes propietats que la distribució normal a $\mathbb{R}$. En particular, es tracta d'una densitat que compleix la igualtat (3.1) i que és simètrica respecte de la mitjana. Els seus intervals d'isodensitat estan centrats en la mitjana i la seva longitud és un múltiple de la desviació estàndard. La llei normal a $\mathbb{R}^+$ és, sobre $\mathbb{R}^+$, coincident amb la llei lognormal clàssica. No obstant això, la primera és més fàcil de manejar i les propietats del model normal a $\mathbb{R}$ es transfereixen directament prenent exponencials.

Finalment i per validar l'ajust del model normal asimètric a unes dades, hem desenvolupat unes proves de bondat d'ajust univariants utilitzant els estadístics basats en la funció de distribució empírica. Hem calculat la distribució de cada estadístic, que depèn del paràmetre de forma així com de la mida de la mostra, mitjançant tècniques de Monte Carlo i l'hem

resumida en un conjunt de taules. De l'estudi de potència realitzat es conclou que aquests contrastos presenten una bona potència davant la distribució lognormal, la distribució exponencial i la distribució χ^2, però aquesta disminueix quan la veritable distribució és una Gamma.

Línies d'investigació futures

Presentem a continuació una relació de qüestions obertes que han anat sorgint al llarg del nostre treball de recerca.

- Sabem que la distribució de Dirichlet i la distribució de Dirichlet escalada són la distribució resultant de la clausura de D components independents amb distribució Gamma. Hem vist que la seva funció de densitat i les seves propietats s'han desenvolupat considerant el símplex com un subconjunt de l'espai real. L'estudi d'aquesta classe de distribucions des d'una perspectiva STAY, obre una línia futura d'investigació. És possible trobar l'expressió de la seva funció de densitat de probabilitat sobre les coordenades d'una composició aleatòria respecte d'una base ortonormal, desenvolupar les seves propietats en relació a l'estructura d'espai vectorial del símplex, analitzar la viabilitat en els seus àmbits d'aplicació i realitzar un estudi comparatiu amb la distribució de Dirichlet clàssica.

- Hem comprovat empíricament que el model normal asimètric resulta adequat per modelitzar, en certs casos, el logaritme de la suma de variables lognormals. Des d'un punt de vista teòric, hem proposat l'aproximació de la distribució utilitzant un model normal asimètric amb els mateixos tres primers moments. Donada la impossibilitat d'obtenir una expressió analítica exacta d'aquests moments, hem optat pel seu càlcul mitjançant desenvolupaments de Taylor. Queda doncs pendent, l'estudi de la viabilitat del model normal asimètric així com la valoració de la qualitat de les aproximacions dels moments.

- La descomposició en valors singulars d'una composició aleatòria permet aplicar proves de bondat d'ajust de normalitat logística additiva independents del denominador escollit en la transformació logquocient, o contrastos de normalitat a $\mathcal{S}^D$ independents de la base ortonormal. Queda pendent d'estudi la generalització d'aquesta metodologia per als

models normal asimètric logístic additiu i normal asimètric a $\mathcal{S}^D$. En aquesta mateixa línia, queda també com una qüestió oberta la utilització de la distància d'Aitchison com a estadístic de bondat d'ajust aplicable a tots els models multivariants.

Referències

Ahrens, L. (1954). The lognormal distribution of the elements. *Geochimica et Cosmochimica Acta (5)*, pp. 49–73.

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *J. R. Statist. Soc. B* vol. *44*(no. 2), pp. 139–177.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* vol. *70*(no. 1), pp. 57–65.

Aitchison, J. (1985). A general class of distributions on the simplex. *J.R. Statist. Soc, B* vol. *47*(no. 1), pp. 136–146.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London (UK): Chapman and Hall.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology* vol. *24*(no. 4), pp. 365–379.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. See Pawlowsky-Glahn (1997), pp. 3–35.

Aitchison, J. (2002). Simplicial inference. In M. A. G. Viana and D. S. P. Richards (Eds.), *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics Series*, pp. 1–22. American Mathematical Society, Providence, Rhode Island (USA).

Aitchison, J. and J. Bacon-Shone (1999). Convex linear combinations of compositions. *Biometrika* vol. *86*(no. 2), pp. 351–364.

Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernandez, and V. Pawlowsky-Glahn (2000).

Logratio analysis and compositional distance. *Mathematical Geology* vol. *32*(no. 3), pp. 271–275.

Aitchison, J. and J. A. C. Brown (1957). *The Lognormal Distribution*. Cambridge (UK): Cambridge University Press.

Aitchison, J., G. Mateu-Figueras, and K. Ng (2003). Characterisation of distributional forms for compositional data and associated distributional tests. *Submitted to Mathematical Geology*.

Aitchison, J. and S. Shen (1980). Logistic-normal distributions: some properties and uses. *Biometrika* vol. *67*(no. 2), pp. 261–272.

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

Alabert, A. (1996). *Mesura i Probabilitat*. Bellaterra (CAT): Universitat Autònoma de Barcelona, Servei de Publicacions. Materials, no. 23.

Anderson, T. W. and D. A. Darling (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.* vol. *49*, pp. 765–769.

Ash, R. (1972). *Real Analysis and Probability*. New York (USA): Academic Press.

Azzalini, A. (1985). A class of distribution which includes the normal ones. *Scand. J. Statist.* vol. *12*, pp. 171–178.

Azzalini, A. and A. Capitanio (1999). Statistical applications of the multivariate skew-normal distribution. *J.R. Statist. Soc. B* vol. *61*(no. 3), pp. 579–602.

Azzalini, A. and A. Dalla-Valle (1996). The multivariate skew-normal distribution. *Biometrika* vol. *83*(no. 4), pp. 715–726.

Barakat, R. (1976). Sums of independent lognormally distributes random variables. *J. Opt. Soc. Am.* vol. *66*(no. 3), pp. 211–216.

Barceló-Vidal, C. (1996). *Mixturas de datos composicionales*. Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona (E).

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

cada, Universitat de Girona.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. See Ross (2001). CD-ROM.

Barndorff-Nielsen, O. E. and B. Jorgensen (1991). Some parametric models on the simplex. *Journal of multivariate Analysis* vol. *39*, pp. 106–116.

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation on species composition. *J. Amer. Statist. Assoc.* vol. *96*(no. 456), pp. 1205–1213.

Bruna, J. (1996). *Anàlisi Real*. Bellaterra (CAT): Universitat Autònoma de Barcelona, Servei de Publicacions. Materials, no. 26.

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

Buccianti, A., V. Pawlowsky-Glahn, C. Barceló-Vidal, and E. Jarauta-Bragulat (1999). Visualization and modeling of natural trends in ternary diagrams: a geochemical case study. See Lippard, Næss, i Sinding-Larsen (1999), pp. 139–144.

Castellet, M. and I. Llerena (1990). *Àlgebra lineal i geometria. 2a edició*. Bellaterra (CAT): Publicacions de la Universitat Autònoma de Barcelona.

Clark, I. and W. Harper (2000). *Practical Geostatistics 2000*. Columbus Ohio, (USA): Ecosse North America.

Connor, J. R. and J. E. Mosimann (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Amer. Statist. Assoc.* vol. *64*, pp. 194–206.

Crow, E. L. and K. Shimizu (1988). *Lognormal Distributions. Theory and Applications*. New York (USA): Marcel Dekker, Inc.

Dalla-Valle, A. (2001). A test for the hypothesis of skew normality in a population. *Submitted to Journal of Statistical Computation and Simulation*.

Darroch, J. N. and I. R. James (1974). F-independence and null correlations of bounded-sum, positive variables. *J.R. Statist. Soc. B* vol. *36*, pp. 467–483.

Daunis-i Estadella, J., J. Egozcue, and V. Pawlowsky-Glahn (2002). Least squares regression in the simplex. *Terra Nostra, ISBN 0946-8978; special issue: 8th Annual Conference of the International Association for Mathematical Geology* vol. *3*, pp. 411–416.

Davis, J. (1986). *Statistics and Data Analysis in Geology*. New York (USA): Wiley & Sons.

Doob, J. (1994). *Measure Theory*. New York (USA): Springer-Verlag.

Egozcue, J. J., E. Pardo-Igúzquiza, and V. Pawlowsky-Glahn (2001). Looking for powerful goodness of fit tests. See Ross (2001). CD-ROM.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelò-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Submitted to Mathematical Geology*.

Fenton, L. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems CS-8*, pp. 57–67.

Galton, F. (1879). The geometric mean in vital and social statistics. *Proceedings of the Royal Society of London* vol. *29*, pp. 365–367.

Green, J. R. and Y. A. S. Hegazy (1976). Powerful modified-EDF goodness-of-fit tests. *J. Amer. Statist. Assoc.* vol. *71*(no. 353), pp. 204–209.

Gupta, A. K. and C. Tuhao (2001). Goodness-of-fit tests for the skew-normal distribution. *Commun. Statist. Simula.* vol. *30*(no. 4), pp. 907–930.

Gupta, R. D. and D. S. P. Richards (1987). Multivariate Liouville distributions. *Journal of multivariate Analysis* vol. *23*, pp. 233–256.

Hamdan, M. A. (1971). The logarithm of the sum of two correlated log-normal variates. *J. Amer. Statist. Assoc.* vol. *66*(no. 333), pp. 105–106.

Ho, C. (1995). Calculating the mean and variance of power sums with two log-normal components. *IEEE Transactions on Vehicular Technology* vol. *44*(no. 4), pp. 756–762.

James, I. R. and J. E. Mosimann (1980). A new characterization of the Dirichlet distribution through neutrality. *Ann. Statist.* vol. *8*, pp. 183–189.

Koch, G. and R. Link (1980). *Statistical Analysis of Geological Data*. New York (USA): Dover Publications, Inc.

Kosniowski, C. (1989). *Topología Algebraica*. Barcelona, (CAT): Editorial Reverté.

Krige, D. (1981). *Lognormal-de Wijsian geostatistics for ore evaluation*. Johannesburg (South Africa): South African Institute of Mining and Metallurgy.

Lang, S. (1971). *Linear algebra. 2nd Ed.* Addison-Wesley.

Lilliefords, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Amer. Statist. Assoc.* vol. *62*, pp. 399–402.

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

Malliavin, P. (1995). *Integration and Probability*. New York (USA): Springer-Verlag.

Marlow, N. A. (1967). A normal limit theorem for power sums of independent random variables. *The Bell System Technical Journal* vol. *46*, pp. 2081–2089.

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

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1997). Different classifications of the Darss Sill data set based on mixture models for compositional data. See Pawlowsky-Glahn (1997), pp. 151–156.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998a). A critical approach to non-parametric classification of compositional data. In A. Rizzi, M. Vichi, and H.-H. Bock (Eds.), *Advances in Data Science and Classification (Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS'98), Università "La Sapienza", Rome, 21–24 July*, pp. 49–56. Springer-Verlag, Berlin (D).

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

Martín-Fernández, J. A., M. Bren, C. Barceló-Vidal, and V. Pawlowsky-Glahn (1999). A measure of difference for compositional data based on measures of divergence. See Lippard, Næss, i Sinding-Larsen (1999), pp. 211–216.

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

Mateu-Figueras, G., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Una aproximación a la distribución del logaritmo de la suma de variables lognormales. In *Actas del XXV Congreso Nacional de la Sociedad de Estadística e Investigación Operativa (SEIO)*, pp. 549–550. Servicio de Publicacións Universidade de Vigo. Vigo (E).

Mateu-Figueras, G. and V. Pawlowsky-Glahn (2003). Una alternativa a la distribución lognormal. See Saralegui i Ripoll (2003), pp. 1849–1858.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. A. Martín-Fernández (2002). Normal in $\mathbb{R}^+$ vs lognormal in $\mathbb{R}$. *Terra Nostra, ISBN 0946-8978; special issue: 8th Annual Conference of the International Association for Mathematical Geology* vol. *3*, pp. 305–310.

McAlister, D. (1879). The law of the geometric mean. *Proceedings of the Royal Society of London* vol. *29*, pp. 367–376.

Narayanan, A. (1991). Algorithm AS 266: maximum likelihood estimation of the parameters of the dirichlet distribution. *Appl. Statist.* vol. *40*, pp. 365–374.

Naus, J. I. (1969). The distribution of the logarithm of the sum of two log-normal variates. *J. Amer. Statist. Assoc.* vol. *64*, pp. 655–659.

Naus, J. I. (1973). Power sum distributions. *J. Amer. Statist. Assoc.* vol. *68*(no. 343), pp. 740–742.

Pawlowsky, V. (1986). *Räumliche Strukturanalyse und Schätzung ortsabhängiger Kompositionen.* Ph. D. thesis, Freie Universität Berlin, Berlin (D).

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

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *SERRA* vol. *15*(no. 5), pp. 384–398.

Pawlowsky-Glahn, V. and J. J. Egozcue (2002). BLU estimators and compositional data. *Mathematical Geology* vol. *34*(no. 3), pp. 259–274.

Pawlowsky-Glahn, V. and R. Olea (en impremta). *Geostatistical Analysis of Compositional Data*. New York (USA): Oxford University Press.

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* vol. *LX*, pp. 489–502.

Pirinen, P. (2000). Conditional outage probability evaluation in WCDMA at hight data rates. new Jersey (USA). IEEE 6th Int. Symp. on Spread-Spectrum Tech. & Appli.: Springer-Verlag, Berlin (D).

Puig, P. and M. A. Stephens (1998). Test of fit for the laplace distribution. *Preprint, centre de recerca matemàtica* (no. 396).

Puig, P. and M. A. Stephens (2000). Test of fit for the laplace distribution. *Technometrics* vol. *42*(no. 4), pp. 417–424.

Puig, P. and M. A. Stephens (2001). Goodness-of-fit for the hyperbolic distribution. *The Canadian Journal of Statistics* vol. *29*(no. 2), pp. 1–12.

Rayens, W. S. and C. Srinivasan (1994). Dependence properties of generalized Liouville distributions on the simplex. *J. Amer. Statis. Assoc.* vol. *89*(no. 428), pp. 1465–1470.

Rendu, J.-M. (1981). *An Introduction to Geostatistical Methods of Mineral Evaluation*. Johannesburg (South Africa): South African Institute of Mining and Metallurgy.

Ross, G. (Ed.) (2001). *Proceedings of IAMG'01 — The sixth annual conference of the International Association for Mathematical Geology*. CR-ROM.

Saralegui, J. and E. Ripoll (Eds.) (2003). *Actas del XXVII Congreso Nacional de la Sociedad de Estadística e Investigación Operativa (SEIO)*. Sociedad de Estadística e Investigación Operativa, Lleida (E).

Schleher, D. (1977). Generalized Gram-Charlier series with application to the sum of log-normal variates. *IEEE Transactions of Information Theory* vol. *Mar77*, pp. 275–280.

Schwartz, S. and Y. Yeh (1982). On the distribution function and moments of power sums

with log-normal components. *The Bell System Technical Journal* vol. *61*(no. 7), pp. 1441–1462.

Stephens, M. A. (1970). Use of the Kolmogorov-Smirnov Crámer-von Mises and related statistics without extensive tables. *J. R. Statist. Soc. B* vol. *32*(no. 1), pp. 115–122.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *J. Amer. Statist. Assoc.* vol. *69*(no. 347), pp. 730–737.

Stephens, M. A. and R. B. D'Agostino (1986). *Goodness-of-Fit Techniques*. New York (USA): Marcel Dekker.

Stuart, A. and J. Ort (1986). *Kendall's Advanced Theory of Statistics. Vol.I: Distribution Theory*. London (UK): Edward Arnold.

Thompson, R., J. Esson, and A. Duncan (1972). Major element chemical variation in the Eocene lavas of the Isle of Skye, Scotland. *J. Petrology* (no. 13), pp. 219–253.

Tolosana Delgado, R., V. Pawlowsky-Glahn, and G. Mateu Figueras (2003). Krigeado de variables positivas. un modelo alternativo. See Saralegui i Ripoll (2003), pp. 1387–1389.

von Eynatten, H., V. Pawlowsky-Glahn, and J. J. Egozcue (2002). Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams. *Mathematical Geology* vol. *34*(no. 3), pp. 249–257.

Watson, J. J. (1961). Goodness-of-fit test on a circle. *Biometrika* vol. *48*(no. 1), pp. 109–114.

Weir, A. (1974). *General integration and measure*. Cambridge (UK): Cambridge University press.

Wilkinson, R. I. (1934). Unpublished work.

Xambó, S. (1977). *Álgebra Lineal i Geometrías lineales*. Barcelona (CAT): Eunibar. 2 vol.

Xambó, S. (1997). *Geometria*. Barcelona (CAT): Edicions UPC. Politext, no. 60.

Youn-Min, C. (1985). Remark AS R55. *Applied Statistics* vol. *34*(no. 1), pp. 100–101.

Young, J. C. and C. E. Minder (1974). Algorithm AS 76: an integral useful in calculating non-central t and bivariate normal probabilities. *Applied Statistics* vol. *23*(no. 3), pp. 455–457.