Butlletí de la Societat Catalana de Matemàtiques
Vol. 30, núm. 1, 2015. Pàg. 5–56. DOI: 10.2436/20.2002.01.58

El llegat de Galton, Pearson, Fréchet i d'altres: com mesurar i interpretar l'associació estadística

Carles M. Cuadras

Resum: Presentem en tres parts els conceptes de correlació i d'associació estadística, començant per la noció de correlació de Galton, millorada per Pearson. Utilitzem com a il·lustració les dades clàssiques de Galton i Pearson sobre heretabilitat de pares i fills respecte a l'estatura. La segona part explica com s'han d'estudiar les mateixes dades des d'una perspectiva multivariant (anàlisi de correlació canònica i de correspondències). Utilitzem també dades de Fisher. Mostrem com podem associar dades de tipus general mitjançant distàncies. La tercera part la dediquem a les distribucions bivariants. Presentem la teoria de funcions i valors propis per a dos nuclis, que s'aplica al desenvolupament diagonal d'una distribució bivariant, incloent-hi els desenvolupaments continus en termes d'integrals. Proposem una família de còpules canòniques, que permet generar distribucions bivariants.

Paraules clau: dades de Galton i Pearson, correlació intraclàssica, correlació canònica, anàlisi de correspondències, associació basada en distàncies, operadors integrals, distribucions amb marginals donades, funcions canòniques.

Classificació MSC2010: 62H20, 60E05.

Part I: Estadística clàssica

1 Introduccció

Des de sempre hem volgut esbrinar la causa d'un fenomen. Per què s'ha produït un incendi? Què ha provocat una guerra? Quina és la principal causa del càncer de pulmó? Per què un cantant té éxit? En algunes parts de la física i la mecànica, el binomi causa-efecte és ben conegut. Sovint les relacions entre variables observables i mesurables són deterministes i es poden expressar amb una relació funcional $y = f(x)$, que lliga una variable dependent y amb una d'independent x. Aquesta relació és molt important i ben estudiada en matemàtiques i altres ciències exactes. Però en biologia, economia, sociologia, meteorologia, etc., a ningú se li acudiria imaginar que pugui existir una relació

tan exacta entre dues variables. Per exemple, la pressió arterial d'una persona té relació amb el pes, però la relació és «imprecisa».

Com podem relacionar dues variables? En principi podem suposar $y = f(x) + e$, on $f(x)$ representa la part determinista i e la part aleatòria de cada observació y. Altrament dit,

$$\text{observació} = \text{model} + \text{error},$$

entenent per «error» la desviació del model. Aleshores és natural que ens plantegem com cal mesurar aquesta relació mitjançant un coeficient entre 0 i 1, de manera que el valor 0 signifiqui manca de relació i el valor 1 es pugui interpretar com que hi ha relació perfecta, matemàticament parlant.[1]

Aleshores es planteja la determinació de la funció f i la naturalesa de la desviació e, sovint interpretada com una variable aleatòria amb distribució normal.

2 Galton i el coeficient de correlació

L'any 1886 Francis Galton [37], contemporani de Mendel i cosí de Darwin, es va plantejar la tasca de mesurar objetivament la relació entre dues variables observades sobre una mateixa població, a fi de donar una prova científica de la teoria de l'evolució. Aleshores va tenir la idea genial de proposar el coeficient de correlació r, que va il·lustrar amb les dades de $n = 205$ famílies. Va anotar les variables següents:

$$\begin{aligned} X_1 &= \text{alçada del pare}, & Y_1 &= \text{alçada del fill},\\ X_2 &= \text{alçada de la mare}, & Y_2 &= \text{alçada de la filla}. \end{aligned}$$

D'entrada Galton va trobar que els pares i els fills eren, en general, més alts que les mares i les filles, és a dir, hi havia dades de dues poblacions. Encertadament va intuir que el coeficient de correlació hauria de ser una mesura d'associació vàlida per a una sola població. De fet, quan barregem dues poblacions, es poden presentar paradoxes (vegeu la secció 7). Per tant, Galton va decidir introduir una correcció consistent a augmentar l'alçada de les dones. Concretament, va considerar les variables

$$X = (X_1 + 1.08X_2)/2, \quad Y = Y_1 \text{ o } Y = 1.08Y_2,$$

on, per a cada X, la variable Y pot prendre tants valors com fills i filles tenia el matrimoni, obtenint $n = 934$ parelles d'observacions. La variable X rep el nom de *mid-parent* i encara avui dia s'utilitza en antropologia.

1 Gairebé mai sabrem si el model representat per la funció f és el correcte. Un model és una simplificació de la realitat. M. Chasles va dir: «La geometria és l'art de raonar bé sobre figures falses».

Amb la correcció de multiplicar per 1.08 l'alçada de cada dona, Galton va tractar les dades com si tingués només una població [41]. La taula 1 il·lustra algunes dades. La família 18 té 3 fills i cap filla. La 102 té 3 fills i 3 filles i la 198 té 4 fills i 1 filla.

Família	Pare	Mare	Gènere	Fill/Filla
18	78	73	1	66.5-64.5-64
102	69	66	1	70-68.5-68
102	69	66	2	65-63-62.5
198	65.5	60	1	68-68-67-67
198	65.5	60	2	62

TAULA 1: Exemple de dades sobre alçades (en polzades) de pares i fills obtingudes per Galton el 1886.

El coeficient de correlació és el quocient $r = S_{xy}/(S_x S_y)$, on S_{xy} és la covariància i S_x, S_y són les desviacions típiques de les variables X, Y. El valor absolut d'aquest coeficient és invariant per transformacions lineals de les variables i satisfà $|r| \leq 1$. Va ser proposat el 1895 per K. Pearson, amb la qual cosa millorà la invenció de Galton.[2]

Galton va obtenir $r = 0.50$, que indicava una certa relació entre pares i fills respecte a l'alçada. Aleshores va representar les dades en un diagrama (figura 1), va dibuixar a ull una recta d'ajust i va afirmar que l'alçada d'un fill era aproximadament $2/3$ de l'alçada dels pares (variable mid-parents).

La recta $y = a + bx$ que millor s'ajusta a les dades, en el sentit dels mínims quadrats, on

$$a \text{ i } b \text{ són tals que } \sum_{i=1}^{n} (y_i - a - bx_i)^2 \text{ és mínima,}$$

és la recta de regressió, que és

$$y = \overline{y} + b(x - \overline{x}), \quad \text{on} \quad b = S_{xy}/S_x^2, \tag{1}$$

essent $\overline{x}$ i $\overline{y}$ les mitjanes de les variables i b el coeficient de regressió o pendent de la recta de regressió. Aquesta recta descriu el model abans esmentat, és a dir, la funció $f(x)$ és lineal. A la secció següent parlarem de l'origen del terme *regressió*.

2 El descobriment del coeficient de correlació és descrit per Galton (que tenia una bona visió matemàtica) a *Memories of My Life*. El coeficient r apareix implícitament en treballs de Gauss i Bravais sobre la distribució normal bivariant. K. Pearson, a qui hem d'atribuir la formulació actual, va proposar (com també ho proposaria Hilbert) una teoria de la relativitat. La seva definició de *coeficient de correlació* es basa en l'estructura d'espai vectorial de les variables aleatòries. Tot i la importància del coeficient de correlació, l'estadístic Sigmund Schott (1928) va afirmar: «És molt lamentable que la complicada fonamentació matemàtica del càlcul de la correlació i la seva prolixa resolució aritmètica excloguin per sempre la possibilitat d'aplicar en algun cas concret aquest notable procediment».

	Galton	$n = 934$	Pearson	$n = 1078$
	Mid-parents	Fills	Pare	Fill
Mitjana	69.20	69.23	67.20	68.16
Desviació típica	1.80	2.58	2.72	2.74
Correlació/regressió	$r = 0.50$	$b = 0.71$	$r = 0.51$	$b = 0.52$

TAULA 2: Mitjanes, desviacions típiques i coeficent de correlació i regressió que resulten de les dades originals de Galton i de Pearson-Lee. Dades en polzades; 67.2 polzades són 170.7 cm., 68.16 polzades són 173 cm.

EXEMPLE. La taula 2 conté alguns resultats estadístics per a les dades de Galton ($n = 934$) i de Pearson-Lee ($n = 1078$), publicades l'any 1903 [53] i que comentarem a la secció 8.

3 La regressió a la mitjania

El coeficient de regressió $b = 0.71$ (taula 2) va donar peu a Galton per afirmar que hi havia una «regressió a la mitjania». Altrament dit, en predir y donat x, s'esperava que un pare alt, l'alçada del qual era superior a la mitjana, tindria fills també alts, però no tant com el pare. És a dir, el fill (o filla) estaria més a prop de la mitjana que el seu progenitor. De manera similar, un pare baix tindria tendència a tenir fills baixos però no tant com el pare.

Aquest notable fet, que dóna nom al terme *model de regressió* (tot i que *regressió* és epistemològicament incorrecte), sembla paradoxal. Però no ho és en sentit biològic. Si els pares alts tinguessin fills més alts, a la generació següent els fills encara ho serien més, i al cap de moltes generacions hi hauria gegants. De la mateixa manera, si els pares baixos tinguessin fills més baixos, a la llarga hi hauria nans. Tampoc hi ha paradoxa estadística, atès que si ajuntem tots els fills, la mitjana es recupera i si hi ha regressió dins d'una família, no n'hi ha a la població.

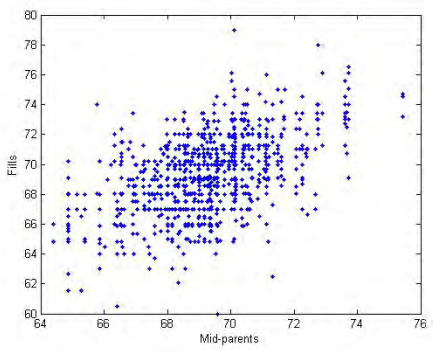

FIGURA 1: Diagrama que representa l'alçada combinada de pare i mare sobre la dels fills i filles, per a 934 observacions obtingudes per Galton.

4 Correlació múltiple

El coeficient de correlació en la forma que ara el coneixem va ser elaborat per K. Pearson. Podem generalitzar la correlació simple r, vàlida per a dues variables, definint la correlació múltiple R. Donada una variable resposta Y i p variables explicatives $X_1, \ldots, X_p$, el coeficient R es defineix com la correlació simple entre Y i $\widehat{Y}$, essent $\widehat{Y} = b_0 + b_1X_1 + \cdots + b_pX_p$ la combinació lineal que millor s'ajusta a Y en el sentit que els coeficients b_i verifiquen

$$E(Y - \widehat{Y})^2 = E(Y - b_0 - b_1X_1 - \cdots - b_pX_p)^2 \text{ és mínima.}$$

Es calcula mitjançant

$$R^2 = \mathbf{r}'\mathbf{R}^{-1}\mathbf{r},$$

on $\mathbf{R}$ és la matriu $p \times p$ de correlacions entre les variables X, i $\mathbf{r}$ és el vector amb les correlacions de la variable Y amb cadascuna de les X.

EXEMPLE. Per a les dades de Galton, si Y és l'alçada del pare (o de la mare), X_1 i X_2 són les alçades de fill i filla, obtenim:

$$\mathbf{r}_{\text{pare}} = \begin{bmatrix} 0.5457 \\ 0.5067 \end{bmatrix}, \quad \mathbf{r}_{\text{mare}} = \begin{bmatrix} 0.3650 \\ 0.3808 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1 & 0.5419 \\ 0.5419 & 1 \end{bmatrix}.$$

Els coeficients de correlació múltiple del pare sobre fill/filla i de la mare sobre fill/filla, són:

$$R_{\text{pare}} = 0.6067, \quad R_{\text{mare}} = 0.4249.$$

Curiosament, respecte a l'alçada, hi ha més influència del pare.

5 Un sofisma sobre la correlació

Plantegem una paradoxa aparent o sofisma sobre la correlació. Suposem que X i Y són dues variables correlacionades sobre la mateixa població, amb variàncies σ_x^2 i σ_y^2, respectivament. Denotem per σ_{xy} la covariància i per ρ_{xy} el coeficient de correlació. En tots els casos suposem que els paràmetres són poblacionals.

Considerem ara una mostra aleatòria simple de X de mida n. És a dir, considerem $X_1, \ldots, X_n$ variables independents, igualment distribuïdes amb la mateixa distribució que X.

La variància de la mitjana mostral $\overline{X}_n = (X_1 + \cdots + X_n)/n$ és σ_x^2/n i la covariància entre $\overline{X}_n$ i Y és

$$\operatorname{cov}(\overline{X}_n, Y) = \frac{1}{n}\operatorname{cov}(X_1 + \cdots + X_n, Y) = \frac{1}{n}n\sigma_{xy} = \sigma_{xy}.$$

Per tant, atès que $\rho_{xy} = \sigma_{xy}/(\sigma_x\sigma_y)$, el coeficient de correlació entre $\overline{X}_n$ i Y és

$$\operatorname{cor}(\overline{X}_n, Y) = \frac{\sigma_{xy}}{(\sigma_x/\sqrt{n})\sigma_y} = \sqrt{n}\rho_{xy}.$$

Ens trobem que si n és prou gran, podríem tenir $\sqrt{n}\rho_{xy} > 1$. Per exemple, si $\rho_{xy} = 0.5$ i $n > 4$ aleshores $\sqrt{n}\rho_{xy} > 1$. Això contradiu la propietat que el coeficient de correlació mai supera el valor 1. Tornarem a comentar aquesta aparent irregularitat més endavant.[3]

6 El tot pot ser més gran que la suma de les parts

Si les variables X estan incorrelacionades dues a dues, és a dir, la matriu de correlacions és $\mathbf{R} = \mathbf{I}$ (identitat), aleshores $R^2 = r_1^2 + \cdots + r_p^2$, essent $r_1, \ldots, r_p$ les correlacions simples de Y amb cadascuna de les variables X. En general les X estan correlacionades i hom esperaria que $R^2 < r_1^2 + \cdots + r_p^2$, atès que no hi ha redundància en la suma de quadrats però sí en R^2. No obstant això, hi ha situacions reals on

$$R^2 > r_1^2 + \cdots + r_p^2. \tag{2}$$

El cas $p = 2$ va ser estudiat per Hamilton i Routledge [40, 56]. En el cas general, Cuadras [9] va demostrar que la desigualtat (2), que podem escriure com a $R^2 = \mathbf{r}'\mathbf{R}^{-1}\mathbf{r} > \mathbf{r}'\mathbf{r}$, és equivalent a

$$\sum_{i=1}^{p} r_{Z_i}^2 (1 - \lambda_i) > 0,$$

essent r_{Z_i}, $i = 1, \ldots, p$, les correlacions simples entre la variable resposta Y i les components principals $Z_1, \ldots, Z_p$, i $\lambda_1, \ldots, \lambda_p$, els valors propis de $\mathbf{R}$. Recordem que les components principals Z_i són les combinacions lineals de les variables X amb variància màxima condicionada a que siguin incorrelacionades. Aquestes variàncies són precisament els valors propis de $\mathbf{R}$.

Atès que els primers valors propis són més grans que 1 i els últims, més petits, tenim que (2) es verifica si Y té una correlació alta amb les components principals de menor variància (figura 2). Per tant, és erroni creure que variables correlacionades són redundants.

Molt relacionada amb (2) és la possibilitat d'augmentar les correlacions simples però que, sorprenentment, disminueixi la correlació múltiple. Considerem la matriu de correlacions $\mathbf{R}$ de quatre variables X i dos vectors $\mathbf{r}_1$, $\mathbf{r}_2$ que contenen les correlacions amb les X de dues variables dependents Y_1, Y_2.

$$\mathbf{R} = \begin{bmatrix} 1 & 0.31 & 0.40 & 0.52 \\ 0.31 & 1 & 0.52 & 0.40 \\ 0.40 & 0.52 & 1 & 0.31 \\ 0.2 & 0.40 & 0.31 & \end{bmatrix}, \quad \mathbf{r}_1 = \begin{bmatrix} 0.60 \\ 0.50 \\ 0.40 \\ 0.30 \end{bmatrix}, \quad \mathbf{r}_2 = \begin{bmatrix} 0.59 \\ 0.49 \\ 0.10 \\ 0.10 \end{bmatrix}.$$

3 Alguns estadístics destacats van necessitar hores, fins i tot dies, a trobar una explicació a aquest desconcertant sofisma sobre la correlació.

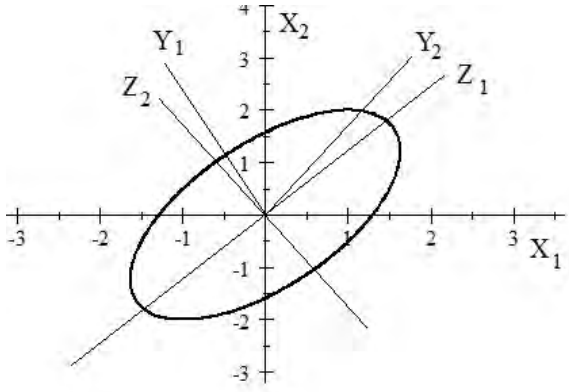

FIGURA 2: La variable dependent Y_1 segueix la direcció de la segona component principal Z_2, perpendicular a la direcció principal de les dades, representada per la primera component principal Z_1.

Com que les correlacions simples són més grans per a Y_1, s'esperaria que també ho sigui la correlació múltiple. No obstant això, es verifica

$$R_1^2 = \mathbf{r}_1' \mathbf{R}^{-1} \mathbf{r}_1 = 0.483 < R_2^2 = \mathbf{r}_2' \mathbf{R}^{-1} \mathbf{r}_2 = 0.688.$$

L'explicació d'aquesta anomalia es la següent: $\mathbf{r}_1$ és un vector amb direcció (mesurada per l'angle) pròxima als últims vectors propis de $\mathbf{R}$, mentre que $\mathbf{r}_2$ segueix una direcció més semblant als primers vectors propis. Altrament dit, Y_1 estaria més influïda per les dues últimes components principals, mentre que Y_2 estaria més relacionada amb les dues primeres. Vegeu la figura 2, amb només dues variables X, però igualment il·lustrativa. La conseqüència de tot això és que el costum de descartar variables X poc correlacionades amb Y podria ser inadequat. Per a més detalls, vegeu [10, 59].

7 Fal·làcia ecològica

Examinem la figura 3, que representa el risc de càncer respecte al consum de calories per a diferents països. Considerant només les mitjanes de cada país (cercle negre), veiem que el risc augmenta amb el consum de calories. Però examinant cada país per separat (cercles blancs agrupats), veiem que el risc disminueix si augmenta l'alimentació. La correlació és, de fet, negativa, però si agrupem poblacions, apareix una falsa correlació positiva. L'anomenada *fal·làcia ecològica* es deu al fet d'ajuntar poblacions diferents i aplicar la correlació, ignorant que aquest coeficient només es pot utilitzar a cada població per separat.

EXEMPLES. Per a una malaltia hereditària, es va detectar una certa correlació positiva entre l'edat X d'aparició en el pare i l'edat Y d'aparició en el fill [42].

En realitat X i Y són independents, i l'edat d'aparició en el fill no té res a veure amb la del pare. La falsa correlació era deguda a que hi ha dos gens A i B, que es presenten amb probabilitats p_A i p_B i que causen la mateixa malaltia. És a dir, les edats observades provenen de dues poblacions diferents. Si X i Y (igualment distribuïdes condicionades a cada gen) tenen mitjanes $E(X|A) = E(Y|A) = \mu_A$ i $E(X|B) = E(Y|B) = \mu_B$ i anàlogament variàncies σ_A^2 i σ_B^2, depenent del gen causant A o B, les esperances matemàtiques de les variables observades X i Y, els seus quadrats i el producte XY són:

$$\begin{aligned} E(X) = E(Y) &= p_A\mu_A + p_B\mu_B, \\ E(X^2) = E(Y^2) &= p_A(\mu_A^2 + \sigma_A^2) + p_B(\mu_B^2 + \sigma_B^2), \\ E(XY) &= p_A\mu_A^2 + p_B\mu_B^2. \end{aligned}$$

Operant, obtenim les variàncies:

$$\text{var}(X) = \text{var}(Y) = p_A\sigma_A^2 + p_B\sigma_B^2 + p_Ap_B(\mu_A - \mu_B)^2.$$

De manera similar, la covariància és $\text{cov}(X, Y) = p_Ap_B(\mu_A - \mu_B)^2$. El coeficient de correlació és, doncs,

$$\rho_{X\mathcal{Y}} = \left[1 + \frac{p_A\sigma_A^2 + p_B\sigma_B^2}{p_Ap_B(\mu_A - \mu_B)^2}\right]^{-1}.$$

Concretament, si $p_A = p_B = 1/2$, $\sigma_A^2 = \sigma_B^2 = \sigma^2$, i la diferència $|\mu_A - \mu_B|$ entre les dues mitjanes val 2σ, aleshores obtenim $\rho_{X\mathcal{Y}} = 0.5$, tot i que les variables X i Y són independents a cada població.

Un altre exemple apareix en el polèmic llibre *The Bell Curve*, de Berrstein i Murray, publicat el 1994. En aquest tractat s'afirma que els blancs són més intel·ligents que els negres. En conseqüència, argumenten els autors, els blancs també seran superiors en una altra habilitat mesurada per una variable H, que es pugui quantificar i que estigui correlacionada positivament amb la intel·ligència I. Per tant, a fi d'escollir entre un blanc i un negre per encarregar-se d'una feina (pilotar una aeronau, per exemple), que es faria millor si l'habilitat mesurada mitjançant H és alta, si ambdós candidats tenen el mateix coeficient d'intel·ligència, aleshores el valor de H en el blanc serà més alt que en el negre. Per tant, seria convenient proposar el candidat blanc. L'argumentació, fruit de barrejar dues poblacions, és falsa, com va denunciar Kaplan [46]. És més, per a una mateixa I, el negre tindria un valor H més alt que el blanc, com podem apreciar a la figura 4, que conté les dues rectes de regressió, i per tant el negre estaria més capacitat per a la feina. De manera més precisa, suposem la mateixa habilitat mitjana de H en blancs i negres, $E(H|B) = E(H|N) = 100$, però una certa superioritat en intel·ligència: $E(I|B) = 100$, $E(I|N) = 90$. Si les desviacions típiques són iguals a 15 i el coeficient de correlació entre H i I és 0.6, aleshores les prediccions (utilitzant les rectes de regressió) indicarien la superioritat del negre sobre el blanc:

$$\begin{aligned} &\text{Blanc} \quad H = 100 + 0.6 \times (105 - 100) = 103, \\ &\text{Negre} \quad H = 100 + 0.6 \times (105 - 90) = 109. \end{aligned}$$

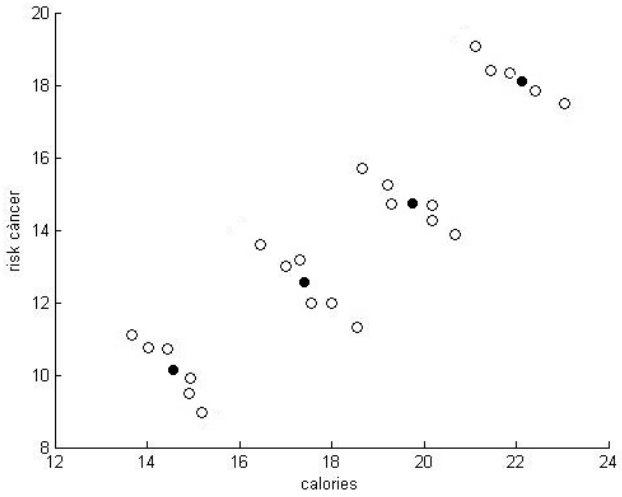

FIGURA 3: Fal·làcia ecològica. La correlació considerant només les mitjanes de les poblacions (cercles negres) és positiva. Però la correlació dintre de cada població (cercles blancs agrupats) és negativa.

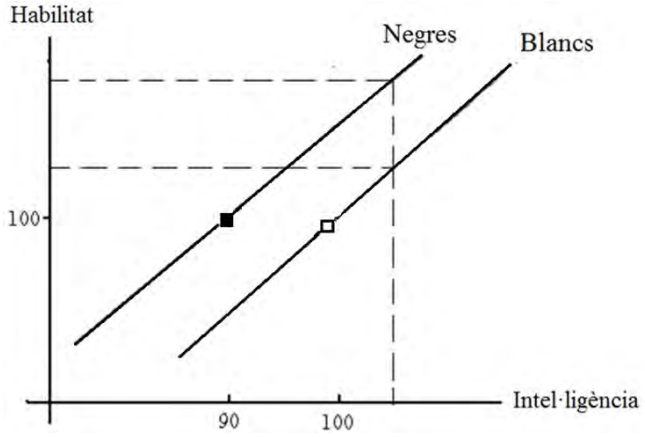

FIGURA 4: Suposant que la intel·ligència sigui superior en els blancs, si un negre té el mateix coeficient d'intel·ligència que un blanc, aleshores el superaria en una altra habilitat correlacionada amb la intel·ligència. Equivocadament, a *The Bell Curve* s'afirma el contrari.

8 Més sobre la correlació

8.1 Raó de correlació

Quan K. Pearson va definir el coeficient de correlació en la seva forma actual, de seguida deuria veure que les dades de Galton no responien perfectament a la definició. En efecte, per a cada valor de X (variable mid-parent), tenim 1, 2 o més valors de Y, perquè una parella pot tenir més d'un fill. A més, si bé podem suposar que les alçades dels pares són independents, les dels fills estan correlacionades dintre de cada família. De fet, el 1903, K. Pearson [53] va efectuar els càlculs amb noves dades i va considerar $n = 1078$ famílies amb només un o dos fills. Va obtenir un coeficient de correlació $r = 0.51$ entre pare i fill varó (vegeu la taula 2).

Podem estudiar les dades de Galton considerant un model lineal

$$y_{ij} = \mu + \beta_i x_i + e_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i, \tag{3}$$

on y_{ij} és l'alçada del fill j de la família i, μ és una mitjana general, β_i és el coeficient de regressió per a la família i corresponent a un mid-parent x_i. El terme e_{ij} és la desviació del model, i reflecteix el fet que els fills d'una mateixa família no tenen la mateixa estatura. Sigui $\hat{\mu}_H$ l'estimació per mínims quadrats suposant que totes les β_i són 0. Aquesta estimació és la mitjana general de les y_{ij}. Denotem per $R_H^2 = \sum_{i,j}(y_{ij} - \hat{\mu}_H)^2$ la suma de quadrats residual. Siguin $\hat{\mu}$, $\hat{\beta}_i$, $i = 1, \ldots, k$, les estimacions sense imposar restriccions i sigui $R_0^2 = \sum_{i,j}(y_{ij} - \hat{\mu} + \hat{\beta}_i x_i)^2$ la suma de quadrats residual. Tenim que $R_0^2 \leq R_H^2$, atès que R_H^2 és un mínim restringit. Aleshores

$$r^2 = 1 - \frac{R_0^2}{R_H^2} \tag{4}$$

és un coeficient que indica el grau de relació lineal entre X i Y. Veiem fàcilment que si el model és perfecte (no hi ha desviació e_{ij}), aleshores $r^2 = 1$. Per a les dades de Galton, r coincideix amb el coeficient de correlació, i per tant $r = 0.50$.

Plantegem ara un model general

$$y_{ij} = g(x_i) + e_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

on g és una funció possiblement no lineal. Si suposem $g = 0$, l'estimació de μ i la suma de quadrats residual R_H^2 són les mateixes d'abans. L'estimació no restringida de $g(x_i)$ és la mitjana corresponent al grup i. Aplicant la mateixa mesura d'ajust, obtenim

$$\hat{\eta}^2 = 1 - \sum_{i=1}^{k} \frac{n_i}{n} \frac{S_i^2}{S_Y^2},$$

on $n = n_1 + \cdots + n_k$, S_i^2 és la variància per a la família i, S_Y^2 és la variància global. Es compleix $\hat{\eta}^2 \geq r^2$. Si no hi ha desviació del model no lineal, és a dir, $S_i^2 = 0$, aleshores $\hat{\eta}^2 = 1$. En les aplicacions cal comparar $\hat{\eta}$ amb r.

Considerem la coneguda descomposició de la variabilitat en l'anàlisi de la variància, ANOVA (vegeu [57]),

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2=\sum_{i=1}^{k}n_i(\overline{y}_{i\cdot}-\overline{y})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\cdot})^2, \tag{5}$$

essent $\overline{y}$ la mitjana general, $\overline{y}_{i\cdot}$ la mitjana del grup i. Escriurem $Q_T = Q_E + Q_D$, on les quantitats Q_T, Q_E, Q_D són les sumes de quadrats totals, entre grups i dintre de grups. Aleshores també podem expressar la raó de correlació per

$$\hat{\eta}^2 = 1-\frac{Q_D}{Q_T}=\frac{Q_E}{Q_T}.$$

EXEMPLE. Per a les dades de Galton obtenim $\hat{\eta}=0.58$, un indici clar de relació no lineal, ja que $\hat{\eta}=0.58$ és significativament més gran que $r=0.50$. Possiblement sigui degut a les repeticions dels valors dels pares que tenen més d'un fill. Però Pearson va comentar que era degut a una manera amateur d'obtenir les dades [58]. De fet, Pearson va treballar amb 1078 famílies i unes dades més ben observades i millor ajustades als conceptes de *correlació* i *regressió*. Va obtenir $\hat{\eta}=0.52$, fet que indica una bona relació lineal, ja que la diferència entre $\hat{\eta}=0.52$ i $r=0.51$ no és significativa.

Si dividim per n els termes de (5) obtenim

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2=\sum_{i=1}^{k}\frac{n_i}{n}(\overline{y}_{i\cdot}-\overline{y})^2+\sum_{i=1}^{k}\frac{n_i}{n}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\cdot})^2, \tag{6}$$

que seria una versió estadística de la descomposició de la variància d'una variable aleatòria. És a dir, donat un parell de variables (X,Y), la descomposició (6) és la versió ANOVA de

$$\operatorname{var}(Y)=\operatorname{var}[E[Y|X]]+E[\operatorname{var}[Y|X]], \tag{7}$$

on $Y|X$ representa la variable Y condicionada a X. La corba de regressió de la mitjana és la millor corba que ajusta Y en funció de X, és a dir, $y=E[Y|X=x]$. El grau de concentració de les observacions (x,y) de (X,Y) al llarg d'aquesta corba és

$$\eta^2=1-\frac{E[\operatorname{var}[Y|X]]}{\operatorname{var}(Y)}.$$

8.2 Correlació intraclàssica

Quina és la correlació entre germans per a una característica física com l'alçada? La resposta no s'obté mitjançant el coeficient de correlació ordinari, atès que no podem disposar dels valors dels germans en parelles ordenades (x,y).

Suposem el model

$$y_{ij}=\mu+A_i+e_{ij},\quad i=1,\dots,k,\ j=1,\dots,n_i,$$

on A_i és una mena de variable aleatòria tal que $E(A_i) = 0$ i $\text{var}(A_i) = \sigma_A^2$ per a tota i. Suposem les condicions usuals d'independència completa entre les A_i, e_{ij} i també que $\text{var}(e_{ij}) = \sigma^2$.

Es defineix la correlació intraclàssica ρ_I com la que hi ha entre dues observacions y_{ij} i $y_{ij'}$ dintre del mateix grup (o família). Es demostra fàcilment [57] que $\rho_I = \sigma_A^2/(\sigma_A^2 + \sigma^2)$.

Siguin Q_A i Q_R les sumes de quadrats entre grups i dintre de grups (quantitats que abans hem indicat per Q_E i Q_D). Dividint pels graus de llibertat corresponents, considerem les mitjanes

$$\overline{Q}_A = \frac{Q_A}{k-1}, \quad \overline{Q}_R = \frac{Q_R}{n-k},$$

essent $n = n_1 + \cdots + n_k$. Segons [33], una estimació de ρ_I és

$$\hat{\rho}_I = \frac{\overline{Q}_A - \overline{Q}_R}{\overline{Q}_A + (n_0 - 1)\overline{Q}_R}, \tag{8}$$

on $n_0 = \left(n - \sum_{i=1}^k n_i^2/n\right)/(k-1)$.

Relacionem ara $\hat{\rho}_I$ amb la raó de correlació $\hat{\eta}^2$. Amb la present notació, tenint en compte que $Q_A/Q_R = \hat{\eta}^2/(1-\hat{\eta}^2)$, és fàcil veure que la correlació intraclàssica és

$$\hat{\rho}_I = \frac{(n-1)\hat{\eta}^2 - (k-1)}{(n - kn_0 + n_0 - 1)\hat{\eta}^2 + (n_0 - 1)(k-1)}.$$

A diferència de $\hat{\eta}$, la correlació intraclàssica $\hat{\rho}_I$ és un coeficient estadístic del qual no hi ha versió probabilística.

EXEMPLE. Per a les dades (no balancejades amb $n_i > 1$) de Galton, aplicant (8), obtenim $n_0 = 5.23$ i $\hat{\rho}_I = 0.38$, que indica una certa correlació intraclàssica entre les alçades de germans i germanes.

8.3 La perspectiva bayesiana

És oportú comentar aquí la condició fonamental de la metodologia ANOVA: la independència estocàstica i la igualtat de variàncies del terme d'error. Així, en el model ANOVA d'un sol factor,

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

suposem la distribució normal $N_{n_i}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$ per a cada successió e_{ij}, $j = 1, \ldots, n_i$. Aquesta hipòtesi és discordant amb l'enfocament típicament bayesià, que suposa les mostres amb distribució conjunta «intercanviable». Altrament dit, la distribució de cadascuna de les n_i observacions y_{ij} és (condicionalment a cada grup) la mateixa, però la distribució conjunta és simètrica.

Per alleugerir la notació, denotem n_i per n i els n_i errors per $x_1, \ldots, x_n$. Aleshores:

$$(x_1, \ldots, x_n) \sim (x_{j_1}, \ldots, x_{j_n}),$$

on $(j_1, \ldots, j_n)$ és una permutació de $(1, \ldots, n)$. El símbol $\sim$ significa «mateixa distribució que».

Segons el teorema de Bruno de Finetti, la densitat conjunta pren la forma [5]

$$p(x_1, \ldots, x_n) = \int_{\mathbf{I}_\alpha} \prod_{j=1}^{n} f(x_j|\alpha)\pi(\alpha)\, d\alpha,$$

on $f(x_j|\alpha)$ és una mena de model estadístic on el paràmetre α prové de l'observació d'una variable latent amb suport l'interval $\mathbf{I}_\alpha$ i densitat de probabilitat $\pi(\alpha)$. La distribució dels valors $x_1, \ldots, x_n$ són independents si els condicionem a un valor fix de α. Suposant que la variància és constant σ^2, indicant la mitjana condicionada per $\mu(\alpha) = \int_{\mathbb{R}} x f(x|\alpha)\, dx$, la mitjana global d'un error x és $a = \int_{\mathbf{I}_\alpha} \mu(\alpha)\pi(a)\, d\alpha$. A més, si posem $A = \int_{\mathbf{I}_\alpha} \mu(\alpha)^2\pi(\alpha)\, d\alpha$, el valor esperat del producte de dos errors x i x' és precisament

$$\int_{\mathbf{I}_\alpha} \int_{\mathbb{R}^2} xx' f(x|\alpha) f(x'|\alpha)\pi(\alpha)\, d\alpha\, dx\, dx' = A.$$

Com que la variància és $\sigma^2 + A - a^2$, fàcilment es demostra que la correlació entre dos errors és

$$\rho = \frac{A - a^2}{\sigma^2 + A - a^2}.$$

Si suposem α fix amb probabilitat 1 (enfocament clàssic), obtenim $A = a^2$ i $\rho = 0$.

Aplicant aquest enfocament bayesià, atès que no coneixem $\pi(\alpha)$, un model raonable seria suposar que $(x_1, \ldots, x_n)$ segueix la distribució normal simètrica $N_n(\mathbf{0}, \Sigma)$, i la matriu de covariàncies és

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

Aquesta distribució és intercanviable i pressuposa (dintre d'un grup) una mateixa correlació ρ per a cada parella d'observacions (o errors) diferents. En general, un bayesià postularia $\rho \neq 0$ i un clàssic consideraria la restricció $\rho = 0$, que en el cas de normalitat implicaria independència completa entre les observacions. És correcte el model restringit? Afortunadament per a l'estadística clàssica, Wilks [60] va demostrar que el test F de Fisher-Snedecor és també vàlid amb dades normals equicorrelacionades, és a dir, en el cas que sigui $\rho \neq 0$.

Ara és el moment de resoldre el sofisme de la secció 5, on la correlació entre $\overline{X}_n$ i Y semblava que podia ser més gran que 1. El parany consistia a

prendre $X_1, \ldots, X_n$ independents. L'acció d'obtenir una mostra aleatòria simple, tan usual en estadística, és incorrecta si X està correlacionada amb Y. Hem d'admetre, doncs, que $X_1, \ldots, X_n$ no són independents.

Dit això, hi ha maneres diferents de veure que $|\text{cor}(\overline{X}_n, Y)| \leq 1$. Per exemple, si suposem el model d'equicorrelació, que implica que les n variables X són intercanviables i no pas independents, aleshores $\text{cov}(X_i, X_j) = \sigma_x^2 \rho$, $i \neq j$. Per tant,

$$\text{var}(\overline{X}_n) = [n\sigma_x^2 + n(n-1)\sigma_x^2 \rho]/n^2,$$

i el coeficient de correlació és

$$\text{cor}(\overline{X}_n, Y) = \frac{\sqrt{n}\rho_{xy}}{\sqrt{1 + (n-1)\rho}},$$

que mai pot superar el valor 1.

Part II: Estadística multivariant

Els mètodes de les dues seccions següents es poden trobar amb detall a [8, 15].

9 Anàlisi de correlació canònica

La correlació múltiple entre Y i $X_1, \ldots, X_p$ és la màxima correlació entre Y i una combinació lineal de $X_1, \ldots, X_p$. La generalització, proposada per Hotelling,[4] considera dos vectors aleatoris $\mathbf{X} = (X_1, \ldots, X_p)$ i $\mathbf{Y} = (Y_1, \ldots, Y_q)$, i determina les combinacions lineals [44]

$$U = \mathbf{Xa} = a_1 X_1 + \cdots + a_p X_p, \quad V = \mathbf{Yb} = b_1 Y_1 + \cdots + b_q Y_q,$$

de tal manera que la correlació entre U i V sigui màxima, on $\mathbf{a} = (a_1, \ldots, a_p)'$ i $\mathbf{b} = (b_1, \ldots, b_q)'$ són dos vectors.

Indiquem per $\mathbf{S}_{11}$ i $\mathbf{S}_{22}$ les matrius de covariàncies mostrals de $\mathbf{X}$ i $\mathbf{Y}$, respectivament (el cas poblacional té un tractament molt semblant). Sigui $\mathbf{S}_{12}$ la matriu $p \times q$ amb les covariàncies de les variables $\mathbf{X}$ amb les variables $\mathbf{Y}$. Tenim aleshores la supermatriu $\mathbf{S} = (\mathbf{S}_{ij})$,

	$\mathbf{X}$	$\mathbf{Y}$
$\mathbf{X}$	$\mathbf{S}_{11}$	$\mathbf{S}_{12}$
$\mathbf{Y}$	$\mathbf{S}_{21}$	$\mathbf{S}_{22}$

on $\mathbf{S}_{21} = \mathbf{S}_{12}'$. Si $\mathbf{S}_{11} = (s_{ij})$, aleshores $\text{var}(U) = \sum_{i,j=1}^{p} a_i a_j s_{ij} = \mathbf{a}'\mathbf{S}_{11}\mathbf{a}$, i anàlogament $\text{var}(V) = \mathbf{b}'\mathbf{S}_{22}\mathbf{b}$ i $\text{cov}(U, V) = \mathbf{a}'\mathbf{S}_{12}\mathbf{b}$.

4 H. Hotelling va introduir l'anàlisi de correlació canònica el 1936, en un intent de relacionar aptituds físiques i mentals. El mètode tindria una gran influència en anàlisi factorial, anàlisi de correspondències, anàlisi discriminant, anàlisi canònica de poblacions i altres mètodes multivariants.

Podem suposar que $\text{var}(U) = \text{var}(V) = 1$. Aleshores el problema equival a:

$$\text{maximitzar} \quad \mathbf{a}'\mathbf{S}_{12}\mathbf{b} \quad \text{restringit a} \quad \mathbf{a}'\mathbf{S}_{11}\mathbf{a} = 1, \quad \mathbf{b}'\mathbf{S}_{22}\mathbf{b} = 1.$$

La solució no és única i els vectors de coeficients $\mathbf{a}$ i $\mathbf{b}$ que compleixen aquesta condició són els vectors canònics. La màxima correlació entre U i V, combinacions lineals de $\mathbf{X}$ i $\mathbf{Y}$, rep el nom de primera correlació canònica r_1. Clarament, $r_1 = 0$ si $\mathbf{X}$ és independent de $\mathbf{Y}$, i $r_1 = 1$ si hi ha una relació lineal entre els dos vectors. Per tant, r_1 té un paper destacat en la mesura de l'associació entre $\mathbf{X}$ i $\mathbf{Y}$.

Si $U_1 = \mathbf{X}\mathbf{a}_1$, $V_1 = \mathbf{Y}\mathbf{b}_1$ és la primera parella de variables canòniques, definim $U_2 = \mathbf{X}\mathbf{a}_2$, $V_2 = \mathbf{Y}\mathbf{b}_2$ com la parella de variables (també combinacions lineals de $\mathbf{X}$ i $\mathbf{Y}$) incorrelacionades amb U_1 i V_1. Aleshores $r_2 = \text{cor}(U_2, V_2)$ és la segona correlació canònica. Anàlogament obtenim la tercera i següents variables i correlacions canòniques, que satisfan les condicions d'optimització esmentades a dalt.

Podem formular una expressió conjunta per als vectors canònics $\mathbf{a}_i$ i $\mathbf{b}_i$ utilitzant la descomposició singular d'una matriu. Suposant que $p \geq q$, sigui $m = \min\{p, q\} = q$ i considerem la matriu $p \times q$

$$\mathbf{Q} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}. \tag{9}$$

Calculem la descomposició singular $\mathbf{Q} = \mathbf{U}\mathbf{D}_s\mathbf{V}'$, on $\mathbf{U}$ és una matriu $p \times q$ amb columnes ortonormals, $\mathbf{V}$ és una matriu $q \times q$ ortogonal i $\mathbf{D}_s$ és una matriu diagonal amb els valors singulars de $\mathbf{Q}$. Altrament dit, $\mathbf{U}'\mathbf{U} = \mathbf{I}_q, \mathbf{V}'\mathbf{V} = \mathbf{V}'\mathbf{V} = \mathbf{I}_q, \mathbf{D}_s = \text{diag}(s_1, \ldots, s_m)$. Aleshores els vectors canònics i les correlacions canòniques són

$$\mathbf{a}_i = \mathbf{S}_{11}^{-1/2}\mathbf{u}_i, \quad \mathbf{b}_i = \mathbf{S}_{22}^{-1/2}\mathbf{v}_i, \quad r_i = s_i.$$

Formalment: $U_i = \mathbf{X}\mathbf{a}_i$, $V_i = \mathbf{Y}\mathbf{b}_i$, $r_i = s_i = \text{cor}(U_i, V_i)$, $i = 1, \ldots, m$, és a dir, les correlacions canòniques són els valors singulars de $\mathbf{Q}$.

Cada parella de variables canòniques té correlació màxima, i manté la condició d'ortogonalitat, és a dir,

$$\text{cor}(U_i, V_i) = r_i, \quad \text{cor}(U_i, U_j) = \text{cor}(V_i, V_j) = \text{cor}(U_i, V_j) = 0 \quad \text{si } i \neq j.$$

A més es compleix la relació recursiva següent entre els vectors canònics:

$$\mathbf{a}_i = s_i^{-1}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b}_i, \quad \mathbf{b}_i = s_i^{-1}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{a}_i.$$

EXEMPLE. Per a les dades de Galton, podem considerar les quatre variables

$$\begin{aligned} X_1 &= \text{alçada del pare}, & Y_1 &= \text{alçada del fill}, \\ X_2 &= \text{alçada de la mare}, & Y_2 &= \text{alçada de la filla}. \end{aligned}$$

Volem relacionar (Y_1, Y_2) amb (X_1, X_2). Atès que necessitem quaternes completes, hem seleccionat famílies amb almenys un fill i una filla, prenent la mitjana

quan hi ha més d'un descendent. Obtenim $n = 150$ i les matrius de covariàncies:

$$\mathbf{S}_{11} = \begin{pmatrix} 6.9654 & 0.4516 \\ 0.4516 & 5.3950 \end{pmatrix}, \quad \mathbf{S}_{12} = \begin{pmatrix} 3.1759 & 2.7013 \\ 1.8697 & 1.7857 \end{pmatrix},$$

$$\mathbf{S}_{21} = \begin{pmatrix} 3.1759 & 1.8697 \\ 2.7013 & 1.7857 \end{pmatrix}, \quad \mathbf{S}_{22} = \begin{pmatrix} 4.8624 & 2.4138 \\ 2.4138 & 4.0804 \end{pmatrix}.$$

Els dos vectors canònics i els valors singulars són:

$$\mathbf{a}_1 = (0.3050; 0.2312)', \quad \mathbf{b}_1 = (0.2742; 0.2642)', \quad s_1 = 0.71,$$
$$\mathbf{a}_2 = (0.2265; -0.3646)', \quad \mathbf{b}_2 = (0.4648; -0.5265)', \quad s_2 = 0.038.$$

La primera correlació canònica és $r_1 = 0.71$ i les primeres variables canòniques són:

$$U_1 = 0.3050X_1 + 0.2312X_2, \quad V_1 = 0.2742Y_1 + 0.2642Y_2.$$

La correlació ha augmentat de 0.50 a 0.71. De fet $r_1 = 0.71$ és una mesura global d'associació entre les estatures de pares i fills.

Resulta interessant constatar que el càlcul de la correlació entre $U_1 = 0.3050X_1 + 0.2312X_2$ i la variable mid-parent $(X_1 + 1.08X_2)/2$ (però considerant el conjunt complet de les 934 dades inicials) dóna $r = 0.995$, un valor molt elevat. Veiem que la variable mid-parent emprada per Galton a fi de reduir diferències entre pares i mares és gairebé equivalent a la primera variable canònica.

Definim ara una mesura global d'associació entre els vectors aleatoris $\mathbf{X} = (X_1, \ldots, X_p)$ i $\mathbf{Y} = (Y_1, \ldots, Y_q)$. En situacions pràctiques interessa trobar una mesura d'associació entre dues matrius de dades $n \times p$ i $n \times q$. Una bona mesura es basa en la lambda de Wilks, que es fa servir per decidir si una hipòtesi és certa o hauria de ser rebutjada.

Sigui $f(x, \theta)$ un model estadístic que consisteix en una densitat de probabilitat indexada per un paràmetre $\theta \in \Theta$. Donada una mostra aleatòria simple de mida n la funció de versemblança és

$$L(x_1, \ldots, x_n; \theta) = f(x_1, \theta) \times \cdots \times f(x_n, \theta).$$

L'estimació màxim-versemblant de θ (proposada per R. A. Fisher) és el valor $\hat{\theta}$ que maximitza $L(x_1, \ldots, x_n; \theta)$. Plantegem una hipòtesi nul·la H_0 sobre el paràmetre θ, que es pot formular com una restricció sobre el conjunt Θ. Escriurem $H_0 : \theta \in \Theta_0 \subset \Theta$. Sigui $\hat{\theta}_0$ l'estimació màxim-versemblant de θ dins Θ_0. Tenim aleshores dues funcions de versemblança. La raó de versemblança és el quocient

$$\lambda = L(x_1, \ldots, x_n; \hat{\theta}_0)/L(x_1, \ldots, x_n; \hat{\theta}).$$

Atès que el numerador és un màxim restringit, clarament $0 \leq \lambda \leq 1$. Un valor de λ pròxim a 1 indicaria que les dues versemblances són semblants i aniria a

favor de H_0. Sota un model regular i suposant H_0 certa, l'estadístic $\chi^2 = -2\ln\lambda$ convergeix en llei a la distribució khi quadrat si $n \to \infty$, propietat que permet decidir sobre l'acceptació o el rebuig de H_0. Sovint, en anàlisi multivariant, la raó de versemblança està relacionada amb la lambda de Wilks, que es defineix com el quocient de determinants $\Lambda = \det(\mathbf{A})/\det(\mathbf{A}+\mathbf{B})$, on $\mathbf{A}$ i $\mathbf{B}$ són matrius de Wishart estocàsticament independents.

En el cas de dos vectors $\mathbf{X}$ i $\mathbf{Y}$ normals multivariants, la independència estocàstica entre ambdós vectors es formula plantejant la hipòtesi $H_0 : \Sigma_{12} = \mathbf{0}$, on Σ_{12} és la matriu $p \times q$ que conté les covariàncies poblacionals entre les variables de cada vector. Es demostra que la raó de versemblança λ està relacionada amb

$$\Lambda = |\mathbf{S}|/(|\mathbf{S}_{11}|\,|\mathbf{S}_{22}|) = (1-r_1^2)\times\cdots\times(1-r_m^2),$$

on $\mathbf{S}$ és la supermatriu de covariàncies $(p+q)\times(p+q)$. Sota H_0, aquesta Λ segueix la distribució lambda de Wilks i verifica $\Lambda = \lambda^{2/n}$. Com que Λ pròxima a 1 afavoreix la hipòtesi nul·la d'independència entre $\mathbf{X}$ i $\mathbf{Y}$, resulta raonable definir la mesura d'associació global següent:

$$A_W = 1 - \Lambda = 1 - \prod_{i=1}^{m}(1-r_i^2), \tag{10}$$

on $m = \min\{p,q\}$. Aleshores A_W val 0 en cas d'independència estocàstica (suposant normalitat) i val 1 si hi ha una relació lineal entre $\mathbf{X}$ i $\mathbf{Y}$. Per a les dades de Galton obtenim $A_W = 0.5060$.

10 Anàlisi de correspondències

Podem també estudiar l'associació entre les alçades de pares i fills definint intervals de classe per a les variables i aplicant anàlisi de correspondències,[5] mètode multivariant que permet visualitzar les files i columnes d'una taula de contingència $\mathbf{N} = [n_{ij}]$, d'ordre $I \times J$. Aquest mètode, amb la contribució prèvia de R. A. Fisher i d'altres, va esdevenir popular amb l'obra de J.-P. Benzécri.[6] Vegeu [4, 39].

Denotem per $\mathbf{P}$ la matriu de correspondències i per $\mathbf{r}$, $\mathbf{c}$ els totals marginals de les seves files i columnes:

$$\mathbf{P} = \tfrac{1}{n}\mathbf{N}, \quad \mathbf{r} = \mathbf{P}\mathbf{1}_J, \quad \mathbf{c} = \mathbf{P}'\mathbf{1}_I,$$

5 L'anàlisi de correspondències ha estat descoberta per diversos autors amb plantejaments diferents. H. O. Hirschfeld (1935) i R. A. Fisher (1940) van ser els pioners. Després K. Maung (1941) va relacionar els resultats de Fisher (donar valors a variables categòriques que maximitzen la correlació) amb l'anàlisi de correlació canònica de Hotelling. J.-P. Benzécri li donaria una interpretació geomètrica, continguda en els seus dos llibres sobre *L'analyse des données*, publicats el 1973. El mètode es difondria a partir de les obres de L. Lebart i d'altres, i en especial de M. Greenacre, publicades independentment el 1984.

6 Convidat per J. Torrens-Ibern, J. Robert, deixeble de Benzécri, va donar el 1974 a Barcelona un curs sobre anàlisi de correspondències.

essent $n = \sum n_{ij}$ i $\mathbf{1}_I$, $\mathbf{1}_J$ els vectors amb uns de dimensions I, J, respectivament. Podem interpretar $\mathbf{N}$ com el resum de les relacions entre dues matrius $\mathbf{X}$ i $\mathbf{Y}$ de dades binàries que relacionen dos conjunts de variables categòriques (possiblement ordinals) amb I i J categories. Si considerem les matrius diagonals $\mathbf{D}_r = \text{diag}(\mathbf{r})$ i $\mathbf{D}_c = \text{diag}(\mathbf{c})$, que contenen les freqüències relatives marginals per files i per columnes, es verifica

$$\mathbf{X}'\mathbf{X} = n\mathbf{D}_r, \quad \mathbf{Y}'\mathbf{Y} = n\mathbf{D}_c, \quad \mathbf{X}'\mathbf{Y} = n\mathbf{P} = \mathbf{N}.$$

Recordem que, donada una mostra $\mathbf{x} = (x_1, \ldots, x_n)'$, podem expressar la variància per $s^2 = n^{-1}\mathbf{x}'\mathbf{x} - (n^{-1}\mathbf{1}'\mathbf{x})^2$, on $\mathbf{1}$ és el vector columna amb n uns. D'una manera similar, una matriu de covariàncies és $n^{-1}\mathbf{X}'\mathbf{X} - \overline{\mathbf{x}}\,\overline{\mathbf{x}}'$, on $\overline{\mathbf{x}} = n^{-1}\mathbf{X}'\mathbf{1}$ és el vector de mitjanes. En el nostre cas el vector (columna) de mitjanes de les I variables fila és $\mathbf{r} = n^{-1}\mathbf{X}'\mathbf{1}$. Anàlogament per a les J columnes. Per tant, les matrius de covariàncies per a les variables files, columnes, i entre files i columnes són

$$\mathbf{S}_{11} = \mathbf{D}_r - \mathbf{r}\mathbf{r}', \quad \mathbf{S}_{22} = \mathbf{D}_c - \mathbf{c}\mathbf{c}', \quad \mathbf{S}_{12} = \mathbf{P} - \mathbf{r}\mathbf{c}'.$$

Si ara volem donar valors a les categories de les variables, de manera que les correlacions siguin màximes, aplicant (9) i calculant la descomposició singular, obtenim

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2} = \mathbf{U}\mathbf{D}_s\mathbf{V}', \tag{11}$$

on $\mathbf{D}_s$ és matriu diagonal amb els valors singulars, que coincideixen amb les correlacions canòniques. Els vectors canònics són

$$\mathbf{a}_i = \mathbf{D}_r^{-1/2}\mathbf{u}_i, \quad \mathbf{b}_i = \mathbf{D}_c^{-1/2}\mathbf{v}_i, \quad i = 1, \ldots, K = \min\{I, J\},$$

que matricialment podem expressar per

$$\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_s, \quad \mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_s.$$

La multiplicació a la dreta per $\mathbf{D}_s$ no altera les correlacions. Es compleixen les relacions següents entre les coordenades de files i de columnes

$$\mathbf{A} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{B}\mathbf{D}_s^{-1}, \quad \mathbf{B} = \mathbf{D}_c^{-1}\mathbf{P}'\mathbf{A}\mathbf{D}_s^{-1}. \tag{12}$$

Considerem ara els perfils de les files $(p_{i1}/r_i, \ldots, p_{iJ}/r_i)$, que podem interpretar com les «probabilitats condicionades» de les J columnes a una fila i determinada. Definim una distància (al quadrat) entre els perfils de dues files i, i',

$$\delta^2_{ii'} = \sum_{j=1}^{J} \frac{(p_{ij}/r_i - p_{i'j}/r_{i'})^2}{c_j}. \tag{13}$$

Si dues files tenen el mateix perfil aleshores $\delta^2_{ii'} = 0$. Aquesta distància es coneix amb el nom de *khi quadrat*.

Es defineix la *inèrcia* o *variabilitat geomètrica* com la mitjana ponderada de les distàncies khi quadrat entre files:

$$V_\delta = \tfrac{1}{2} \sum_{i=1}^{I} \sum_{i'=1}^{I} r_i \delta_{ii'}^2 r_{i'}. \tag{14}$$

V_δ és una mesura de dispersió de les dades, versió multivariant de la variància. Es demostra que $V_\delta = \sum_{k=1}^{K} s_k^2 = \chi^2/n$, on s_i, $i = 1, \ldots, K = \min\{I, J\}$, són els valors singulars de (11) i χ^2 és l'estadístic khi quadrat per contrastar la hipòtesi d'independència entre files i columnes de $\mathbf{N}$, és a dir, $H_0 : p_{ij} = p_{i\cdot} \times p_{\cdot j}$, on p_{ij} és la probabilitat conjunta i $p_{i\cdot}$, $p_{\cdot j}$ són les probabilitats marginals.

A més, les distàncies euclidianes (al quadrat) entre les files de la matriu $\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_s$ coincideixen amb les distàncies khi quadrat entre files. Les dues primeres coordenades (a_{i1}, a_{i2}) proporcionen una representació òptima en dimensió dos, en el sentit que la seva variabilitat geomètrica, considerant només dues coordenades, és màxima i val $V_\delta(2) = \sum_{k=1}^{2} s_k^2$, que en general aporta una proporció alta de la variabilitat total $V_\delta = \sum_{k=1}^{K} s_k^2$.

Anàlogament podem definir el perfil de les columnes i una distància khi quadrat entre columnes. Aleshores la matriu $\mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_s$ conté les coordenades euclidianes per a una representació de les columnes.

D'altra banda, les relacions (12), recíproques entre $\mathbf{A}$ i $\mathbf{B}$, demostren que les coordenades (a_{il}) de les files són les mitjanes, ponderades pels perfils de les files, de les coordenades (b_{jl}) de les columnes. Anàlogament per a les coordenades de les columnes. Per exemple, cada primera coordenada de les files verifica

$$a_{i1} = \frac{1}{s_1}\left(b_{11}\frac{p_{i1}}{r_i} + b_{21}\frac{p_{i2}}{r_i} + \cdots + b_{J1}\frac{p_{iJ}}{r_i}\right), \quad i = 1, \ldots, I,$$

i anàlogament les columnes. Però aquesta mitjana baricèntrica no és perfecta (per raons geomètriques, els punts $\mathbf{a}_i$ no poden ser mitjanes dels $\mathbf{b}_j$ i recíprocament). Hi ha, per tant, un factor dilatador $1/s_1 > 1$. Una conseqüència important d'aquesta relació és que podem representar conjuntament, en un mateix gràfic, les files i columnes d'una taula de contingència $I \times J$.

EXEMPLES. Agrupem les dades de Galton i formem intervals de classe per a les alçades de mid-parents i fills. Per a les alçades dels pares es consideren els cinc intervals:

$$\text{fins a } 66,\ (66, 68],\ (68, 70],\ (70, 72],\ \text{més de } 72,$$

l'últim indicat a la figura 5 per «∘74». En el cas dels fills hi ha més variabilitat i es consideren els set intervals:

$$\text{fins a } 64,\ (64, 66],\ (66, 68],\ (68, 70],\ (70, 72],\ (72, 74],\ \text{més de } 74,$$

l'últim indicat en el gràfic per «□76». Creuant aquests intervals obtenim una taula de contingència que representem per anàlisi de correspondències simples.

Veiem clarament l'associació, ja que les alçades dels pares són pròximes a les dels fills, en el sentit, per exemple, que les alçades entre 68 i 70 dels fills són mitjanes ponderades de les alçades dels pares, i el valor que predomina és precisament la «probabilitat condicionada» a l'interval $(68, 70]$ dels pares. La variabilitat geomètrica del gràfic (dimensió 2) representa el 96 % de la variabilitat total (dimensió 4). El primer valor singular o primera correlació canònica és 0.49, valor molt semblant a la correlació $r = 0.50$ que s'obté amb les dades originals.

La taula 3 (esquerra) és una taula 4×4 que conté dades de freqüències obtingudes per Galton sobre $n = 1000$ individus, classificats segons el color dels ulls (1 blau, 2 verd o gris, 3 gris fosc, 4 castany fosc) de pares i fills. Convé tenir en compte que ara les variables són categòriques i no podem calcular el coeficient de correlació clàssic. La figura 6 (esquerra) és el resultat de l'anàlisi de correspondències simples. L'associació entre pares i fills respecte al color dels ulls és notable. La variabilitat geomètrica és el 86.7 % de la variabilitat total. La primera correlació canònica és $r_1 = 0.39$.

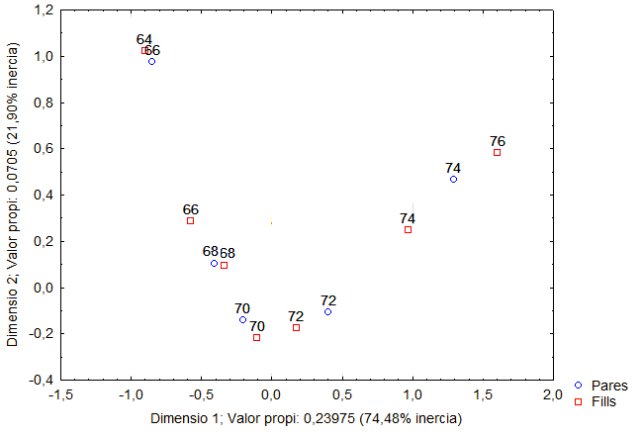

FIGURA 5: Representació per anàlisi de correspondències de les files (mid-parents) i columnes (fills), d'una taula de contingència 5×7 resultat de distribuir en intervals de longitud 2 les dades de Galton, que relacionen les estatures de pares i fills. El valor 70 representa l'interval de classe de 68 a 70.

10.1 Associació en taules d'ordre superior

La taula 3 (dreta) és una taula $2 \times 2 \times 2$ que conté dades de freqüències obtingudes per Galton sobre $n = 5008$ individus, classificats segons el color (clar o fosc) dels ulls, tenint en compte el color dels ulls dels pares i els avis. Galton va considerar 78 famílies amb molts fills, per tant hi ha relació (en el sentit

de correlació intraclàssica) entre germans i germanes. Ara tenim tres variables categòriques, que es poden representar aplicant anàlisi de correspondències ordinari a la taula de Burt, taula simètrica que conté les freqüències combinant les variables. Aquest tipus de representació es coneix per anàlisi de correspondències múltiples [15, 39]. La figura 6 (dreta) representa les freqüències de fills, pares i avis d'acord amb el color dels ulls. La separació entre color C i color F, que al seu torn estan agrupats, mostra que hi ha associació entre color i parentesc.

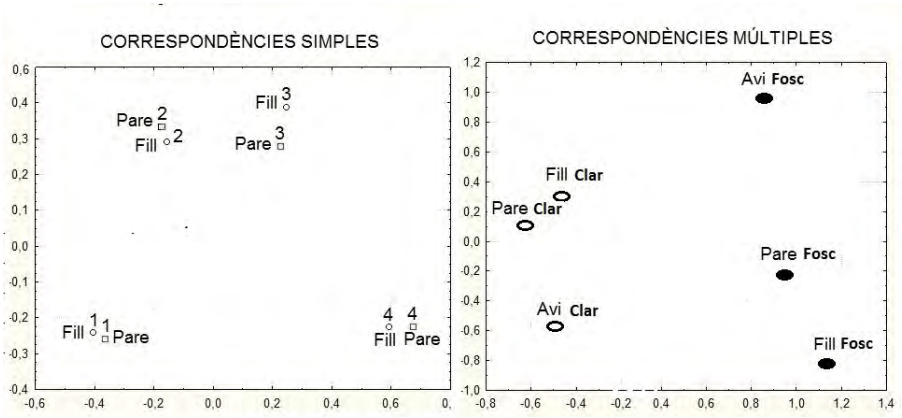

FIGURA 6: Anàlisi de correspondències simples (esquerra) i de correspondències múltiples (dreta), del colors dels ulls. Classificacions i freqüències obtingudes per Galton de 1000 individus (esquerra) i de 5008 individus (dreta).

		Pare			
	Color	1	2	3	4
Fill	1	194	70	41	30
	2	83	124	41	36
	3	25	34	55	23
	4	56	36	43	109

	Avi			
	Clar		Fosc	
Pare	C	F	C	F
Fill C	1928	552	596	508
Fill F	303	395	225	501

TAULA 3: Classificació i freqüències obtingudes per Galton de 1000 individus (esquerra) i 5008 individus (dreta), segons el color dels ulls, considerant quatre colors (esquerra) i color clar o fosc (dreta), i amb el registre també del color dels ulls dels seus pares i avis.

Intentem ara mesurar l'associació estadística entre fills, pares i avis respecte del color dels ulls. En primer lloc hem d'especificar un model. Siguin p_{ijk} les probabilitats, on $i, j, k \in \{1, 2\}$ són els índexs per a fill, pare i avi. Per exemple,

p_{111} és la probabilitat que un fill tingui els ulls clars, així com el seu pare i el seu avi. L'estimació és $\hat{p}_{111} = 1928/5008$. Posem $p_{ij\cdot} = p_{ij1} + p_{ij2}$ i anàlogament $p_{i\cdot k}$, $p_{\cdot jk}$, $p_{i\cdot\cdot}$, etc. Especifiquem ara un model. En teoria, el més correcte seria suposar que hi ha vuit probabilitats p_{ijk} que sumen 1. Però estadísticament no podem decidir la seva validesa perquè hi ha tants paràmetres lliures com observacions. Les estimacions dels p_{ijk} donarien freqüències esperades $n\hat{p}_{ijk}$ que coincidirien amb les observades (l'estadístic khi quadrat seria $\chi^2 = 0$). Hem d'afinar una mica més. Suposarem que el color dels ulls dels fills és independent entre pares i fills, entre avis i fills i entre avis i pares. En termes de probabilitats:

$$p_{ij\cdot} = p_{i\cdot\cdot} \times p_{\cdot j\cdot}, \quad p_{i\cdot k} = p_{i\cdot\cdot} \times p_{\cdot\cdot k}, \quad p_{\cdot jk} = p_{\cdot j\cdot} \times p_{\cdot\cdot k}.$$

Per a les dades de la taula 3 (dreta), el test per a aquesta hipòtesi aplicant la raó de versemblança dóna una khi quadrat de $\chi^2 = -2\ln\lambda = 16.8$ amb un grau de llibertat, bastant significativa i, per tant, rebutgem que hi ha independència entre fills i pares, etc. Les dades tampoc s'ajusten a un model genètic combinant dos al·lels. De fet, el model correcte no el coneixem (vegeu la nota al peu de la pàgina 6). Acceptem, doncs, els paràmetres $p_{ij\cdot}$, $p_{ij\cdot}$, $p_{\cdot jk}$ com a probabilitats genèriques, que es desvien del model de treball format pels productes $p_{i\cdot\cdot} \times p_{\cdot j\cdot}$, etc.

Considerem ara el model d'independència completa: el color dels ulls de fills, pares i avis és independent. En termes de probabilitats:

$$p_{ijk} = p_{i\cdot\cdot} \times p_{\cdot j\cdot} \times p_{\cdot\cdot k}.$$

Aquest model és una restricció de l'anterior. Atès que la khi quadrat és $\chi^2 = -2\ln\lambda = 868.3$ amb dos graus de llibertat, molt significativa, el model s'hauria de rebutjar. També podem afirmar que el model anterior ($\chi^2 = 16.8$) s'ajusta molt millor a les dades observades. Com cal mesurar aquest ajust, interpretat en termes d'associació?

Acceptem el valor khi quadrat dividit pels graus de llibertat com una distància entre freqüències observades i esperades, és a dir, una mesura de la desviació del model corresponent. Aplicant (4) obtenim

$$\theta = 1 - \frac{16.8}{868.3/4} = 0.9226.$$

El grau d'associació global (relatiu als dos models) que hi ha entre les tres generacions respecte al color (clar o fosc) dels ulls és alt.

El cas general de taules de contingència $J_1 \times \cdots \times J_q$ és semblant. La representació de les variables categòriques la podem fer aplicant l'anàlisi de correspondències múltiples. Aleshores cal establir dos models, un de general, l'altre més restringit, la desviació del qual expressi la dependència global. Aplicant el mateix procediment, obtindrem una mesura de dependència relativa. Usualment els càlculs es fan aplicant models log-lineals [15], atès que prenent logaritmes, els productes esdevenen sumes.

10.2 Alternatives a l'anàlisi de correspondències

Quan les freqüències no formen una taula $I \times J$ sino que provenen de I poblacions multinomials independents, aleshores és més adequada *la distància de Hellinger* entre perfils [23], seguint el plantejament degut a Rao [54]:

$$\delta^2_{ii'} = \sum_{j=1}^{J} \left(\sqrt{p_{ij}/r_i} - \sqrt{p_{i'j}/r_{i'}} \right)^2 .$$

Hi ha altres alternatives que responen al nom d'anàlisi no simètric, d'anàlisi log-ràtio, etc. Totes es poden unificar parametritzant la distància, que depèn de dos paràmetres $0 \leq \alpha, \beta \leq 1$, per

$$\delta^2_{ii'} = \sum_{j=1}^{J} \left[\left(\frac{p_{ij}}{r_i c_j} \right)^{\alpha} - \left(\frac{p_{i'j}}{r_{i'} c_j} \right)^{\alpha} \right]^2 c_j^{2\beta},$$

vegeu [19]. Aleshores les coordenades de les files i de les columnes s'obtenen a partir de la descomposició singular:

$$\mathbf{D}_r^{1/2} \mathbf{H}_c \left\{ \frac{1}{\alpha} [(\mathbf{D}_r^{-1} \mathbf{P} \mathbf{D}_c^{-1})^{(\alpha)} - \mathbf{1}\mathbf{1}'] \right\} \mathbf{D}_c^{\beta} = \mathbf{U} \mathbf{D}_s \mathbf{V}', \tag{15}$$

on $\mathbf{M}^{(\alpha)} = [m_{ij}^{\alpha}]$, $\mathbf{H}_c = \mathbf{I} - \mathbf{1}\mathbf{r}'$. La taula 4 resumeix els diferents mètodes, amb els acrònims CA (anàlisi de correspondències), HD (anàlisi amb distància Hellinger [54]), LR (anàlisi log-ràtio [1]), NSCA (anàlisi de correspondències no simètriques). En general, les coordenades de les files són $\mathbf{A} = \mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{D}_s$ i les de les columnes, $\mathbf{B} = \mathbf{D}_c^{\beta - 1} \mathbf{V} \mathbf{D}_s$. La matriu centradora $\mathbf{H}_c = \mathbf{I} - \mathbf{1}\mathbf{r}'$, que premultiplica a (15), no apareix a (11) perquè no afecta les solucions CA i NSCA. La representació en dimensió 2 és una forma geomètrica d'explicar una part de la inèrcia $V_\delta = \sum s_i^2$, on $s_1, s_2, \ldots$ són els valors singulars.

En el mètode CA, la inèrcia és el coeficient de contingència de Pearson i està lligat a la khi quadrat per a taules de contingència. Si considerem HD, aleshores la inèrcia té a veure amb el coeficient d'afinitat de Matusita, que mesura el grau de concordància entre dues densitats. La inèrcia en el mètode NSCA està relacionada amb el coeficient de Goodman-Kruskal, que mesura la predictibilitat de les columnes donades les files. Per tant, NSCA és adequat si les columnes fan el paper de variables resposta i les files, de variables predictores. Finalment, LA és útil per a dades composicionals [1] (els valors de cada fila són positius i sumen una quantitat fixa).

Hi ha encara més mètodes [38], alguns basats en freqüències acumulades [12]. Però el mètode CA és el més racional i el que millor respecta les propietats de les probabilitats. Vegeu [22] per a una perspectiva general.

Coordenades files, columnes SVD, inèrcia			$\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_s$ $\quad \mathbf{B} = \mathbf{D}_c^{\beta-1}\mathbf{V}\mathbf{D_s}$ $\mathbf{Q} = \mathbf{U}\mathbf{D_s}\mathbf{V}'$ $\quad V_\delta = \mathrm{tra}(\mathbf{Q}'\mathbf{Q})$
Mètode	α	β	$\mathbf{Q}$
CA (Benzécri, Greenacre)	1	1/2	$\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}')\mathbf{D}_c^{1/2}$
HD (Hellinger, Rao)	1/2	1/2	$\mathbf{D}_r^{1/2}\mathbf{H}_c(\mathbf{D}_r^{-1/2}\mathbf{P}^{(1/2)}\mathbf{D}_c^{-1/2} - \mathbf{1}\mathbf{1}')\mathbf{D}_c^{1/2}$
LR (Aitchison, Greenacre)	0	1/2	$\mathbf{D}_r^{1/2}\mathbf{H}_c \ln(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1})\mathbf{D}_c^{1/2}$
NSCA (Lauro, D'Ambra)	1	1	$\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}')\mathbf{D}_c$

TAULA 4: Quatre mètodes per representar la relació entre dos conjunts de variables categòriques.

11 Associació estadística amb dades generals

Fem una ullada a la taula 5, que combina un dels *Reports de la recerca a Catalunya (1996–2002)*, publicat per l'IEC, i una taula extreta de [7]. La diagonal conté el nombre d'articles de matemàtiques i estadística publicats per cada país sense col·laboració amb els altres països de la llista. Sobre la diagonal hi ha el nombre d'articles publicats amb autors d'altres països. La part per sota de la diagonal indica la relació comercial (1 si és important, 0 si no ho és). Per exemple, Espanya (incloent-hi Catalunya) ha publicat 10 560 articles, dels quals 8597 tenen autors espanyols i cap dels altres països de la llista, 692 amb coautors d'EUA, 473 amb França, etc.; i Espanya ha mantingut amb els EUA una relació comercial important (valor 1).

Hi ha associació estadística entre col·laboració científica i relació comercial? Vegem una manera general de mesurar-la.

	EUA	Esp.	Fran.	R. U.	Ital.	Alem.	Can.	Japó	Xina	Russ.
EUA	63 446	692	2281	2507	1642	2812	2739	1039	1773	893
Espanya	1	8597	473	347	352	278	163	69	104	177
França	1	1	17 155	532	916	884	496	269	167	606
Regne Unit	1	1	1	12 585	490	810	480	213	339	365
Itàlia	0	1	1	0	13 197	677	290	169	120	512
Alemanya	1	1	1	1	1	16 588	499	350	408	984
Canadà	1	0	0	1	0	0	7927	228	601	204
Japó	1	0	0	0	0	0	1	20 001	371	193
Xina	1	0	0	0	0	0	1	1	39 140	64
Rússia	0	0	0	0	0	0	0	0	1	18 213

TAULA 5: Nombre d'articles publicats amb autors d'un sol país (diagonal) i amb autors d'altres països (sobre la diagonal). Relació comercial significativa entre països (sota la diagonal).

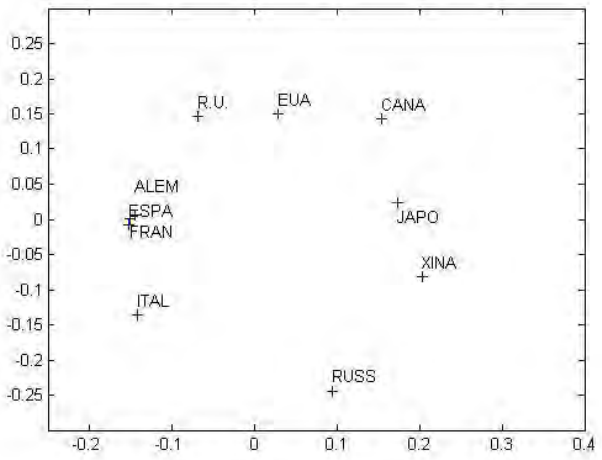

FIGURA 7: Representació RMDS de deu països tenint en compte la col·laboració científica i la relació comercial.

Sigui $\Omega = \{\omega_1, \ldots, \omega_n\}$ un conjunt finit amb n objectes o individus. Suposem que per algun procediment podem definir una mesura de dissimilaritat o distància $\delta\colon \Omega\times\Omega \to \mathbb{R}_+$ entre cada parella d'objectes: $\delta_{ij} = \delta(\omega_i, \omega_j)$. Suposem que δ és simètrica i no negativa: $\delta_{ij} = \delta_{ji} \geq \delta_{ii} = 0$. Obtenim aleshores una matriu $n \times n$ de distàncies $\Delta_x = (\delta_{ij})$.

Suposarem que aquesta matriu de distàncies és euclidiana. És a dir, existeix una configuració de punts $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, amb coordenades $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$, de tal manera que $\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$. Tenim, doncs, que les coordenades dels elements de Ω formen una matriu $n \times p$, que denotarem per $\mathbf{X} = [x_{ij}]$, de tal manera que la distància euclidiana entre cada parella de files i, j coincideix amb la distància inicial δ_{ij}.

A fi de saber si Δ_x és euclidiana, sigui $\mathbf{I}_n$ la matriu identitat i $\mathbf{1}_n$ el vector d'uns. És ben conegut que una manera d'obtenir $\mathbf{X}$ (que no és única) a partir de Δ_x, consisteix a trobar primer les matrius $\mathbf{A}_x = -\frac{1}{2}\Delta_x^{(2)}$ i $\mathbf{G}_x = \mathbf{H}_c\mathbf{A}_x\mathbf{H}_c$, on $\Delta_x^{(2)} = (\delta_{ij}^2)$ i $\mathbf{H}_c = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n'$ és la matriu de centrat. Seguidament calculem la descomposició espectral $\mathbf{G}_x = \mathbf{U}\Lambda_x^2\mathbf{U}'$, que proporciona la matriu de coordenades $\mathbf{X} = \mathbf{U}\Lambda_x$. Aleshores Δ_x és una matriu de distàncies euclidianes si i només si $\mathbf{G}_x$ és una matriu semidefinida positiva (els valors propis són no negatius).

Suposant que els valors propis continguts en la matriu diagonal Λ_x^2 estan disposats en ordre decreixent, les matrius $\mathbf{X}$ i $\mathbf{U}$ contenen les coordenades principals i estàndard, respectivament, dels n individus en relació amb les distàncies Δ_x. Aquest procediment, conegut per anàlisi de coordenades principals [8, 15], permet fer una representació gràfica dels n individus en dimensió reduïda (usualment 2) prenent les primeres coordenades principals, és a dir, les

dues primeres columnes de $\mathbf{X}$. El resultat és semblant a l'anàlisi de correspondències, que de fet seria una anàlisi de coordenades principals ponderat. Hem de dir, però, que aquí volem relacionar aquestes coordenades amb un segon conjunt de dades observades sobre el mateix conjunt Ω.

Considerem, doncs, un altre conjunt de dades amb observacions sobre els mateixos n individus, i que, mitjançant una distància o dissimilaritat apropiada, podem obtenir una segona matriu de distàncies Δ_y, de la qual calculem $\mathbf{A}_y$ i a continuació $\mathbf{G}_y = \mathbf{V}\Lambda_y^2\mathbf{V}'$, com hem fet abans. Si els valors propis a la diagonal de Λ_y^2 estan ordenats, les coordenades principals de Ω respecte a Δ_y són $\mathbf{Y} = \mathbf{V}\Lambda_y$. Les coordenades estàndard són les n files de $\mathbf{V}$.

Amb aquestes coordenades, l'associació entre els dos conjunts de dades es pot plantejar quantificant l'associació entre les matrius $\mathbf{X}(n \times p)$ i $\mathbf{Y}(n \times q)$. Tenint en compte que són matrius centrades (la mitjana de cada columna és 0), plantegem ara un model matricial que generalitzi (1). Seguint el plantejament iniciat a [14], el model és $\mathbf{Y} = \mathbf{XB} + \Xi$, essent $\mathbf{B}$ una matriu $p \times q$ de paràmetres i Ξ un matriu $n \times q$ de desviacions aleatòries.

L'estimació de $\mathbf{B}$ pel criteri dels mínims quadrats és $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ i la matriu de predicció és $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \mathbf{PY}$, on $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ és la matriu de projecció. Òbviament, no hi ha cap relació entre $\mathbf{X}$ i $\mathbf{Y}$ si $\hat{\mathbf{B}} = 0$, mentre que la relació és perfecta si $\hat{\mathbf{Y}} = \mathbf{Y}$.

En el test de la hipòtesi $H_0 : \mathbf{B} = \mathbf{0}$, sigui $\mathbf{E} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})$ la «matriu error». De $\mathbf{P}^2 = \mathbf{P}$ obtenim $\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$. Denotem per $\mathbf{H} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ la «matriu hipòtesi». Clarament $\mathbf{Y}'\mathbf{Y} = \mathbf{H} + \mathbf{E}$. Atès que $\mathbf{Y}$ és una matriu centrada, tenim que $\mathbf{Y}'\mathbf{Y}$ és proporcional a una matriu de covariàncies. Tenim, doncs, la versió matricial de la descomposició de la variabilitat en dues parts: la deguda a la hipòtesi i la deguda a la desviació de la hipòtesi (error).

Apliquem ara el criteri de la raó de versemblança, en aquest cas equivalent a la lambda de Wilks, $W = \det(\mathbf{E}) / \det(\mathbf{E} + \mathbf{H})$. Valors de W propers a 0 indiquen que la hipòtesi H_0 hauria de ser rebutjada i, per tant, $\mathbf{B} \neq \mathbf{0}$.

En el nostre context de matrius de coordenades principals, tenim $\mathbf{Y} = \mathbf{V}\Lambda_y$ així com $\hat{\mathbf{Y}} = \mathbf{PY} = \mathbf{UU}'\mathbf{V}\Lambda_y$. Per tant, $\mathbf{E} = \Lambda_y(\mathbf{I} - \mathbf{V}'\mathbf{UU}'\mathbf{V})\Lambda_y$ la qual cosa implica $\mathbf{E} + \mathbf{H} = \mathbf{Y}'\mathbf{Y} = \Lambda_y\mathbf{V}'\mathbf{V}\Lambda_y = \Lambda_y^2$. Obtenim $W = \det(\mathbf{I} - \mathbf{V}'\mathbf{UU}'\mathbf{V})$.

D'altra banda, d'acord amb (9), les correlacions canòniques r_i entre les columnes de $\mathbf{X}$ i les de $\mathbf{Y}$ satisfan l'equació $\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{v}_i = r_i^2\mathbf{Y}'\mathbf{Y}\mathbf{v}_i$, és a dir, $\mathbf{H}\mathbf{v}_i = r_i^2(\mathbf{E} + \mathbf{H})\mathbf{v}_i$, on $\mathbf{v}_i$ és el corresponent vector propi. Això implica $\mathbf{E}\mathbf{v}_i = (1 - r_i^2)(\mathbf{E} + \mathbf{H})\mathbf{v}_i$. Per tant, podem expressar W en termes de correlacions canòniques: $W = (1 - r_1^2) \times \cdots \times (1 - r_m^2)$, on $m = \min(p, q)$. Trobem la mateixa mesura d'associació $A_W = 1 - W$ que hem obtingut abans, vegeu (10), però ara per a dades generals. A_W val 0 si no hi ha relació entre els dos conjunts de dades, i val 1 si hi ha una relació lineal en el sentit que les matrius de distàncies defineixen espais equivalents.

EXEMPLE. Considerem la taula 5 i la matriu de similaritat $\mathbf{S} = (s_{ij})$, on $s_{ij} = 1$ si hi ha relació comercial significativa, $s_{ij} = 0$ en cas contrari. La relació comercial interna de cada país és molt intensa, per tant convindrem que $s_{ii} = 3$.

Considerem també la matriu $\mathbf{T} = (t_{ij})$, on per a dos països i i j definim $t_{ij} = n_{ij}/\min\{n_i, n_j\}$, essent n_{ij} el nombre d'articles publicats en col·laboració. Els nombres n_i, n_j són el total d'articles publicats pels països i i j. En el cas d'Espanya i EUA tenim $s_{12} = 1$, $t_{12} = 692/8597 = 0.0805$. Transformant les similaritats en distàncies (al quadrat)

$$\delta^2_{ij}(x) = 2(1 - s_{ij}), \quad \delta^2_{ij}(y) = 2(1 - t_{ij}),$$

tenim que $\mathbf{G}_x = \mathbf{H}_c\mathbf{S}\mathbf{H}_c = \mathbf{U}\Lambda_x^2\mathbf{U}'$ i $\mathbf{G}_y = \mathbf{H}_c\mathbf{T}\mathbf{H}_c = \mathbf{V}\Lambda_y^2\mathbf{V}'$. Considerem les coordenades principals $\mathbf{U}\Lambda_x$ i $\mathbf{V}\Lambda_y$ i prenem les dues primeres columnes de $\mathbf{U}$ i de $\mathbf{V}$. Denotem aquestes dues matrius de coordenades estàndard per $\mathbf{U}_2$ i $\mathbf{V}_2$. Fent una anàlisi de correlació canònica, aplicant (9) i atès que $\mathbf{S}_{11}$ i $\mathbf{S}_{22}$ són matrius identitat, obtenim $\mathbf{Q} = \mathbf{U}_2'\mathbf{V}_2$, que té dos valors singulars o correlacions canòniques $r_1 = 0.9182$, $r_2 = 0.6437$. Per tant, $A_W = 0.9081$, que indica una bona associació entre col·laboració científica i comercial.

Podem triar altres coeficients amb propietats semblants, com ara el producte $\prod r_i^2$, que en aquest cas donaria un valor bastant baix [14]. Però per diverses raons, A_W és millor, com es demostra en el tractament i la comparació d'imatges hiperespectrals [31].

La figura 7 és el resultat d'aplicar la *related metric scaling* [27], tècnica que permet representar els països tenint en compte les dues matrius de similaritats considerades. Consisteix a diagonalizar la matriu

$$\mathbf{G} = \mathbf{G}_x + \mathbf{G}_y - \frac{1}{2}(\mathbf{G}_x^{1/2}\mathbf{G}_y^{1/2} + \mathbf{G}_y^{1/2}\mathbf{G}_x^{1/2}),$$

i calcular les coordenades principals conjuntes. Podem veure que $\mathbf{G} = \mathbf{G}_x + \mathbf{G}_y$ en el cas d'ortogonalitat de les configuracions i que $\mathbf{G} = \mathbf{G}_x = \mathbf{G}_y$ en el cas que les matrius de distàncies Δ_x, Δ_y coincideixin. La representació es pot fer també aplicant una tècnica coneguda per *multiple factor analysis* [35], que diagonalitza $\mathbf{G}_x + \mathbf{G}_y$ i, per tant, ignora la redundància entre les dues matrius, perquè prendre la suma implícitament suposa ortogonalitat, és a dir, $\mathbf{U}'\mathbf{V} = \mathbf{0}$. Considerant $\mathbf{G}_x^{1/2}\mathbf{G}_y^{1/2} + \mathbf{G}_y^{1/2}\mathbf{G}_x^{1/2}$ es tenen en compte les correlacions entre coordenades, contingudes en els productes $\mathbf{U}'\mathbf{V}$ i $\mathbf{V}'\mathbf{U}$, i es gaudeix d'algunes de les propietats de la distància de Mahalanobis (vegeu la secció 13.2).

12 MANOVA basat en distàncies

Suposem que tenim $k \geq 2$ conjunts de dades provinents de les poblacions $\Omega_1, \ldots, \Omega_k$, obtinguts observant p variables quantitatives. En el model MANOVA d'una via, sigui $\mathbf{T} = \mathbf{B} + \mathbf{W}$ la descomposició de la matriu de productes creuats «total» en suma de «entre grups» i «dintre grups», vegeu (5), partint de n_i observacions provinents de Ω_i. A fi de contrastar la hipòtesi: $H_0 : \Omega_1 = \cdots = \Omega_k$, en cas de normalitat multivariant i homogeneïtat de covariàncies, tenim que $\mathbf{W}$ i $\mathbf{B}$ fan el paper de $\mathbf{E}$ i $\mathbf{H}$, i l'estadístic clàssic per decidir sobre H_0 és la lambda de Wilks:

$$W = \det(\mathbf{W})/\det(\mathbf{B} + \mathbf{W}).$$

Per tractar amb dades generals, suposem que mitjançant una funció distància δ entre observacions obtenim les matrius d'intradistàncies $\Delta_{11}, \ldots, \Delta_{kk}$, i les matrius d'interdistàncies $\Delta_{12}, \ldots, \Delta_{k-1k}$. La matriu de distàncies global és Δ. Calculant coordenades principals (no centrades) per a cada matriu de distàncies, i generalitzant la igualtat $(1/n)\sum_i (x_i - \overline{x})^2 = (1/2n^2)\sum_i \sum_j (x_i - x_j)^2$, obtenim les matrius $p \times p$ següents:

$$\mathbb{T} = \sum_{g,h=1}^{k} \sum_{i,i'=1}^{n_g,n_h} (\mathbf{x}_{gi} - \mathbf{x}_{hi'})(\mathbf{x}_{gi} - \mathbf{x}_{hi'})',$$

$$\mathbb{B} = \sum_{g,h=1}^{k} n_g n_h (\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_h)(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_h)',$$

$$\mathbb{W}_j = \sum_{i,i'=1}^{n_j} (\mathbf{x}_{ji} - \mathbf{x}_{ji'})(\mathbf{x}_{ji} - \mathbf{x}_{ji'})',$$

que verifiquen $\mathbb{T} = \mathbb{B} + n\sum_{g=1}^{k} n_g^{-1}\mathbb{W}_g$. Per tant, tenim que $\mathrm{tr}(\mathbb{T}) = \mathrm{tr}(\mathbb{B}) + n\sum_{g=1}^{k} n_g^{-1}\,\mathrm{tr}(\mathbb{W}_g)$.

Calculant les mitjanes de les distàncies (al quadrat), aplicant (14) amb pes $1/n$, podem obtenir la variabilitat geomètrica per a Δ i per a la Δ_{ii} de cada població Ω_i separadament. D'aquesta manera, per a dades generals i treballant només amb distàncies, és possible descompondre la variabilitat geomètrica

$$V_\delta \text{ (total)} = V_\delta \text{ (entre)} + \sum_{i=1}^{k} \frac{n_i}{n} V_\delta \text{ (dintre } i\text{)}, \tag{16}$$

que seria una generalització de (6). Aleshores un test per decidir si hi ha diferències significatives entre les k poblacions, podria estar basat en el quocient $\gamma = V_\delta$ (entre)$/V_\delta$ (total); vegeu [14].

Recentment, M. Anderson i d'altres, amb la tècnica anomenada PERMANOVA, han continuat el mateix enfocament amb dades depenent de dos o més factors (incloent-hi interaccions), tot i que la justificació teòrica és encara insuficient.

Hi ha una versió probabilística de (16), relacionada amb (7). Considerem un vector aleatori $\mathbf{X}$ que té per densitat la mixtura $f(\mathbf{x}) = w_1 f_1(\mathbf{x}) + \cdots + w_k f_k(\mathbf{x})$, on totes les densitats tenen el mateix suport $\mathbb{S}$. Suposem que existeix una representació $\psi\colon \mathbb{S} \to \mathbb{E}$, on $\mathbb{E}$ és un espai euclidià (o de Hilbert separable). Aleshores la variabilitat geomètrica de $\mathbf{X}$, respecte a una distància δ, definida com a [30]

$$V_\delta(\mathbf{X}) = \frac{1}{2}\int_{S\times S} \delta^2(\mathbf{x},\mathbf{y}) f(\mathbf{x}) f(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \tag{17}$$

verifica

$$V_\delta(\mathbf{X}) = V(\mu_1, \ldots, \mu_k) + \sum_{i=1}^{k} w_i V_i,$$

essent $V(\mu_1,\ldots,\mu_k) = \frac{1}{2}\sum_{i,j=1}^{k} w_i\delta^2(\mu_i,\mu_j)w_j = \sum_{i=1}^{k} w_i\delta^2(\mu_i,\mu)$, on $\mu_i = E_i[\psi(\mathbf{X})]$, $\delta^2(\mu_i,\mu_j) = ||\mu_i-\mu_j||^2$, $\mu = w_1\mu_1+\cdots+w_k\mu_k$. La variabilitat geomètrica V_i és la de (17) però canviant f per f_i; vegeu [21].

13 Més conceptes multivariants

13.1 Correlació intraclàssica multivariant

Considerem de nou la correlació intraclàssica (8). De fet, la definició univariant de correlació entre dues observacions y_{ij} i $y_{ij'}$ és pràctica però convencional. Més aviat hem d'acceptar que $\widehat{\rho}_I$ és una mesura d'homogeneïtat entre les mostres d'un mateix grup o família. Des d'aquesta perspectiva, resulta raonable definir una mesura de correlació intraclàssica multivariant entre les observacions d'una mateixa subpoblació quan disposem de k subpoblacions.

En el model MANOVA esmentat abans, sigui $\mathbf{T} = \mathbf{B}+\mathbf{W}$, i considerem $n_0 = \left(n-\sum_{i=1}^{k} n_i^2/n\right)/(k-1)$. La mesura proposada, que generalitza (8), és

$$\widehat{\Phi}_I = \frac{\operatorname{tra}(\mathbf{B})/(k-1) - \operatorname{tra}(\mathbf{W})/(n-k)}{\operatorname{tra}(\mathbf{B})/(k-1) + (n_0-1)\operatorname{tra}(\mathbf{W})/(n-k)}. \tag{18}$$

Més generalment, relacionant (16) amb les traces de $\mathbf{B}$ i $\mathbf{W}$, una definició de *correlació intraclàssica* per a dades generals, treballant amb una distància δ entre observacions, seria:

$$\widehat{\Phi}_I = \frac{V_\delta\ (\text{entre})/(k-1) - \sum_{i=1}^{k}(\frac{n_i}{n})V_\delta\ (\text{dintre } i)/(n-k)}{V_\delta\ (\text{entre})/(k-1) + (n_0-1)\sum_{i=1}^{k}(\frac{n_i}{n})V_\delta\ (\text{dintre } i)/(n-k)}.$$

EXEMPLE. Per a les dades de flors iris, formades per les espècies Iris setosa, Iris versicolor i Iris virginica, emprades per R. A. Fisher per il·lustrar l'anàlisi discriminant [15], tenim $k = 3$, $n_1 = n_2 = n_3 = 50$. Per tant, $n_0 = 50$. Aplicant (18), obtenim $\widehat{\Phi}_I = 0.907$. Hi ha, per tant, una bona homogeneïtat entre les flors dintre de cada espècie.

13.2 Fal·làcia ecològica multivariant

La fal·làcia ecològica amb dues variables (figura 3) es pot interpretar en el sentit que les dades globals segueixen una direcció, gairebé perpendicular a la direcció de les dades de cada subpoblació.

A fi de detectar la fal·làcia amb dades relativament complicades, suposem dues poblacions amb vectors de mitjanes $\mu_i = (\mu_{i1},\ldots,\mu_{ip})'$, $i = 1,2$, i matriu de covariàncies (comuna) $\Sigma = [\sigma_{ij}]$. La distància de Mahalanobis entre les dues poblacions es defineix per $(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$. Aquesta distància (al quadrat) és invariant per transformacions lineals de les variables, té en compte les correlacions i apareix de manera natural en molts models multivariants.

Si suposem que les variables estan incorrelacionades, Σ seria diagonal, i la distància de Mahalanobis esdevindria

$$(\mu_1-\mu_2)'[\mathrm{diag}(\Sigma)]^{-1}(\mu_1-\mu_2) = \sum_{j=1}^{p}(\mu_{1j}-\mu_{2j})/\sigma_j^2,$$

que es coneix per *distància de Pearson*.

Aleshores, tenint en compte que $R^2 = \mathbf{r}'\mathbf{R}^{-1}\mathbf{r}$ i $\mathbf{r}'\mathbf{r}$ són formalment similars a una distància de Mahalanobis i de Pearson, respectivament, entre r i $\mathbf{0}$, (2) suggereix la desigualtat

$$(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) > (\mu_1-\mu_2)'[\mathrm{diag}(\Sigma)]^{-1}(\mu_1-\mu_2). \tag{19}$$

Cuadras i Fortiana [28] proven que (19) es presenta si la direcció del segment que uneix els punts μ_1 i μ_2 és essencialment ortogonal a la direcció principal (comuna) de les dades de cada subpoblació, que és donada per les primeres components principals (variables combinació lineal amb màxima variabilitat). Per tant, la fal·làcia ecològica estaria relacionada amb la desigualtat (19).

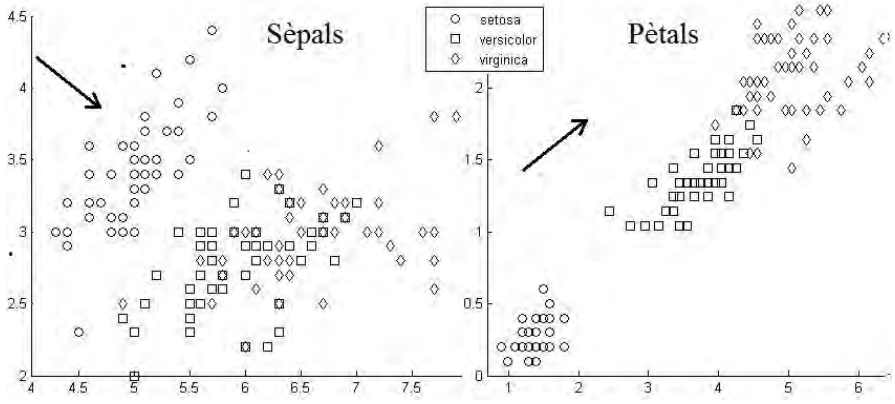

FIGURA 8: Representació de tres espècies de flors del gènere *Iris*, segons longitud i amplada de sèpals i de pètals. Per als sèpals, la distància de Mahalanobis predomina sobre la de Pearson, i els centres de les poblacions segueixen una direcció gairebé ortogonal a la principal de les dades de cada espècie. Per als pètals, la distància de Pearson predomina sobre la de Mahalanobis i els centres de les poblacions segueixen la mateixa direcció que la principal de les dades.

EXEMPLE. La figura 8 il·lustra la fal·làcia amb les dades sobre flors de les espècies Iris setosa, versicolor i virginica, considerades a la secció anterior. La figura 8 (esquerra) representa les tres espècies segons longitud i amplada de sèpals de la flor. La taula 6 (esquerra) conté les distàncies de Mahalanobis (a sobre de la diagonal), que predominen sobre les de Pearson (a sota de la diagonal).

Aleshores la direcció dels centres de les poblacions és gairebe ortogonal a la direccio principal de les dades, i, per tant, hi ha fal·làcia ecològica. A més, les correlacions ajuntant les cent cinquanta observacions són negatives, però les correlacions dintre de cada espècie són positives. També a la figura 8 (dreta), utilitzant ara longitud i amplada dels pètals, observem el contrari, i per tant no hi ha fal·làcia. Les distàncies de Mahalanobis, taula 6 (centre), són més petites que les de Pearson i la direcció dels centres de les espècies és gairebé la mateixa que la de les dades de cada espècie. Això concorda amb les correlacions, que resulten totes positives considerant les cent cinquanta observacions, i també calculant-les per a cada espècie per separat.

No resulta tan fàcil visualitzar la fal·làcia considerant les quatre variables. Aleshores és quan resulta útil la desigualtat (19). Aquesta es compleix comparant *setosa* amb *versicolor* i amb *virginica*. D'acord amb la taula 6 (dreta), la suma de les distàncies de Mahalanobis i de Pearson, considerant longituds i amplades de sèpals i pètals, dóna 286.7 i 277.0, respectivament. Mahalanobis predomina sobre Pearson i, per tant, hi ha fal·làcia ecològica en sentit multivariant.

	Sèpals			Pètals			Sèpals i pètals		
Iris	*set.*	*vers.*	*virg.*	*set.*	*vers.*	*virg.*	*set.*	*vers.*	*virg.*
setosa	–	14.9	21.6	–	48.2	112	–	89.8	179
versicolor	7.0	–	1.61	70.1	–	14.0	77.1	–	17.2
virginica	11.2	1.96	–	166	20.7	–	177	22.7	–

TAULA 6: Distàncies de Mahalanobis (sobre la diagonal) i Pearson (sota la diagonal), entre tres espècies de flors del gènere *Iris*, considerant sèpals, pètals i les quatre mesures juntes.

13.3 Coeficients d'asimetria i curtosi

Siguin $\mathbf{x}_1, \ldots, \mathbf{x}_n$ les observacions de p variables sobre una població. Denotem per $\overline{\mathbf{x}}$ el vector de mitjanes i per $\mathbf{S}$ la matriu $p \times p$ de covariàncies. Considerant els productes escalars de Mahalanobis $d_{ij} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(x_j - \overline{\mathbf{x}})$, són mesures multivariants d'asimetria i curtosi (proposades per K. V. Mardia) les quantitats

$$b_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^3, \qquad b_2 = \frac{1}{n}\sum_{i=1}^{n} d_{ii}^2.$$

En el cas de distribució normal multivariant $N_p(\mu, \Sigma)$, els valors poblacionals són $\beta_1 = 0$ i $\beta_2 = p(p+2)$.

En la mateixa línia de la secció 12, suposem dades generals (possiblement mixtes), descrites per una matriu $n \times n$ de distàncies Δ. Calculem la corresponent matriu de productes interns $\mathbf{G} = \mathbf{U}\Lambda^2\mathbf{U}'$, que proporciona la matriu de coordenades principals $\mathbf{X} = \mathbf{U}\Lambda$. Aleshores, posant $(a_{ij})^{(m)} = (a_{ij}^m)$, la generalització, en termes de distàncies, de les mesures d'asimetria i curtosi és:

$$\hat{\beta}_1 = n\mathbf{1}_n'[(\mathbf{G}\mathbf{G}^-)^{(3)}]\mathbf{1}_n, \quad \hat{\beta}_2 = n\,\mathrm{tra}[(\mathbf{G}\mathbf{G}^-)^{(2)}],$$

on $\mathbf{G}^- = \mathbf{U}\Lambda^{-2}\mathbf{U}'$ és la inversa generalitzada de $\mathbf{G}$. Els valors teòrics de referència serien $\beta_1 = 0$ i $\beta_2 = m(m+2)$, on m és el nombre de coordenades principals, és a dir, el nombre de columnes de $\mathbf{X}$.

EXEMPLES. Amb les dades de Galton s'obtenen $b_1 = 0.04$ i $b_2 = 9.0$, que són bastant semblants als valors teòrics $\beta_1 = 0$ i $\beta_2 = 8$ de la normal bivariant. Per a les cinquanta flors de l'espècie *Iris virginica*, s'obté $b_1 = 3.15$ i $b_2 = 24.3$, essent els valors teòrics $\beta_1 = 0$ i $\beta_2 = 24$ en cas de multinormalitat. Les dades tenen asimetria i una curtosi molt propera a la normal. Per a la taula 5 (part superior), si considerem les dues primeres coordenades principals calculades sobre la matriu $\mathbf{T}$ (secció 11), obtenim $\widehat{\beta}_1 = 2.2$ i $\widehat{\beta}_2 = 5.9$, que es desvien una mica de $\beta_1 = 0$ i $\beta_2 = 8$.

Part III: Distribucions bivariants amb marginals donades

14 Classes de Fréchet-Hoeffding

Quan K. Pearson va estudiar les dades de Galton sobre estatures de pares i fills, així com les seves pròpies dades (amb A. Lee) [53], va tenir molta cura d'ajustar-les a una distribució normal bivariant. Aquesta distribució va ser un model matemàtic que resumia correctament les dades observades. Per tant, una altra manera d'abordar l'associació estadística consisteix a utilitzar la teoria de distribucions bivariants amb marginals donades, que està relacionada amb la teoria de còpules.

Suposem que $H(x,y) = P[X \le x, Y \le y]$ és la funció de distribució conjunta del vector aleatori (X,Y), definit sobre un espai de probabilitats $(\Omega, \mathcal{A}, P)$. Les distribucions marginals són F i G essent

$$F(x) = P[X \le x] = H(x,\infty), \quad G(y) = P[Y \le y] = H(\infty, y).$$

Considerarem que $H \in \mathcal{F}(F,G)$, essent $\mathcal{F}(F,G)$ la classe de funcions de distribució bivariants amb marginals F i G. Un exemple de distribució bivariant $H_0 \in \mathcal{F}(F,G)$ és $H_0(x,y) = F(x)G(y)$, que correspon al cas d'independència estocàstica.

Fréchet [36] va introduir les funcions H_- i H_+ següents:

$$H_-(x,y) = \max\{F(x) + G(y) - 1, 0\}, \quad H_+(x,y) = \min\{F(x), G(y)\},$$

i va demostrar[7] la desigualtat

$$H_-(x,y) \le H(x,y) \le H_+(x,y).$$

7 Quan el 1977, a la biblioteca del Seminari Matemàtic de Barcelona, vaig consultar el número corresponent a 1951 de la revista on Fréchet va publicar el seu fonamental article, els plecs estaven sense tallar. El document no havia estat consultat abans. Per cert, és una coincidència que els únics funcionaris per oposició del Seminari Matemàtic —l'eminent analista F. Sunyer-Balaguer (1967) i C. M. Cuadras (1974–1979)— eren de Figueres (Girona).

Aleshores les funcions $H_-, H_+ \in \mathcal{F}(F,G)$, anomenades *cotes inferior* i *superior de Fréchet-Hoeffding*, verifiquen:

$$H = H_- \Leftrightarrow F(X) = 1 - G(Y) \quad \text{(q.s)},$$
$$H = H_+ \Leftrightarrow F(X) = G(Y) \qquad \text{(q.s)}.$$

Això significa que si la distribució és H_-, hi ha dependència funcional negativa (Y és funció decreixent de X). Si la distribució és H_+, hi ha dependència funcional positiva (Y és funció creixent de X).

EXEMPLE. Suposem que dues variables X i Y mesuren el mateix fenomen però seguint procediments diferents i que hi ha una relació no lineal entre X i Y. Si disposem de n parelles (x_i, y_i), aleshores podem aplicar algun model de regressió. Però si només disposem de la distribució de cada variable per separat, la relació $F(x) = G(y)$ pot resoldre el problema. Com a il·lustració, suposem que $F(x) = (x-1)^2$ si $1 \le x \le 2$, $G(y) = [1 - e^{-(y-2)}/(1 - e^{-2})$ si $2 \le y \le 4$. De $y = G^{-1}(F(x))$ obtenim $y = 2 - \ln[1 - (1 - e^{-2})(x-1)^2]$. Tanmateix en una situació pràctica haurem de treballar amb les funcions de distribució empíriques $F_n(x)$ i $G_n(y)$, com en el cas de la mesura del diàmetre d'un bacteri, seguint dos procediments diferents (analitzador de partícules i citometria). A [45], seguint aquest procediment, els autors obtenen polinomis de segon grau per a diverses classes de bacteris. Per exemple, per a *E. coli* obtenen $y = 8.74x^2 - 6.71x + 2.08$, on y (que va de 0.7×10^{-3} fins a 1.3×10^{-3}) és el diàmetre en micres i x (que va de 0.35×10^{-3} fins a 0.65×10^{-3}) és una mesura de flux citomètric que proporciona un analitzador de partícules.

14.1 Correlacions de Hoeffding

Hoeffding [43] va demostrar que la covariància, si existeix, es pot calcular en termes de les funcions acumulatives H, F i G, aplicant

$$\operatorname{cov}(X,Y) = \int_{R^2} [H(x,y) - F(x)G(y)]\, dx\, dy. \tag{20}$$

Aleshores, si denotem per ρ el coeficient de correlació per a la distribució H, es compleix la desigualtat

$$\rho_- \le \rho \le \rho_+,$$

essent ρ_- i ρ_+ els coeficients de correlació corresponents a les cotes H_- i H_+.

EXEMPLE. Si desconeixem la distribució conjunta de X i Y (com a l'exemple anterior), les correlacions màxima i mínima que podem obtenir, fixades les marginals, són ρ_+ i ρ_-. Per exemple, si X és uniforme sobre $(0,1)$, $F(x) = x$, $0 < x < 1$, i Y és exponencial, $G(y) = 1 - e^{-y}$, $y > 0$, aleshores $\rho_+ = \sqrt{3}/2$ i $\rho_- = -\sqrt{3}/2$.

14.2 Desenvolupaments diagonals

D'acord amb (11), en un model bivariant discret, on el rang de les variables és numerable, podem expressar la densitat de probabilitat conjunta p_{ij} com una suma

$$p_{ij} = r_i c_j \left(1 + \sum_{n\geq 1} s_n a_{in} b_{in}\right),$$

on $\{s_n\}$ són valors singulars, que fan el paper de correlacions canòniques. L'anàlisi de correspondències seria un cas finit del desenvolupament diagonal que presentem tot seguit.

Suposem que la derivada de Radon-Nikodym $dH(x,y)/dF(x)\,dG(y)$ existeix i que la mesura dH és absolutament contínua respecte a $dF \times dG$. És a dir, podem expressar H integrant respecte a $dF(x)\,dG(y)$. A més suposem que el coeficient de contingència de Pearson Φ^2, definit com

$$\Phi^2 + 1 = \int_a^b \int_c^d [dH(x,y)]^2/[dF(x)\,dG(y)],$$

és finit. Aleshores podem obtenir el desenvolupament:

$$dH(x,y) = dF(x)\,dG(y)\left[1 + \sum_{n\geq 1} \rho_n a_n(x) b_n(y)\right]. \tag{21}$$

Si les densitats conjunta i marginals (respecte de la mesura de Lebesgue) són $h(x,y)$, $f(x)$ i $g(y)$, llavors podem escriure el desenvolupament (21) com

$$h(x,y) = f(x)g(y)\left[1 + \sum_{n\geq 1} \rho_n a_n(x) b_n(y)\right]. \tag{22}$$

Aleshores $\{a_n(x)\}$, $\{b_n(y)\}$, anomenades funcions canòniques, són conjunts de funcions ortonormals, que són complets respecte de f i g. D'altra banda

$$E[a_n(X)] = E[b_n(Y)] = 0, \quad E[a_n(X)b_m(Y)] = \rho_n \delta_{mn},$$

essent ρ_n l'enèsima correlació canònica entre X i Y. Les $a_n(X)$, $b_m(Y)$ són les variables canòniques. El desenvolupament diagonal (22) va ser introduït per Lancaster [47].

Disposant les correlacions canòniques en ordre descendent, aleshores ρ_1 és la primera correlació canònica i és la màxima correlació entre una funció de X i una funció de Y

$$\rho_1 = \sup \operatorname{cor}(\alpha(X), \beta(Y)).$$

Es verifica $0 \leq \rho_1 \leq 1$ i es pot provar que $\rho_1 = 0$ si i només si X i Y són estocàsticament independents. A més Rényi [55] va demostrar que ρ_1 és l'única mesura de dependència que verifica set postulats proposats per ell mateix. Per trobar ρ_1, vegeu [52]. Si tenim n mostres (x_i, y_i), un procediment pràctic per estimar ρ_1 consisteix a aplicar anàlisi de correlació canònica (secció 9) a potències de x_i i de y_i, és a dir, relacionar $(x_i, x_i^2, \ldots, x_i^m)$ amb $(y_i, y_i^2, \ldots, y_i^m)$.

EXEMPLE. Un element de $\mathcal{F}(F,G)$, que apareix sovint a les aplicacions i en la construcció de models bivariants, és la distribució FGM (deguda a Farlie, Gumbel i Morgenstern), que depèn del paràmetre θ,

$$H_\theta(x,y) = F(x)G(y)\{1+\theta[1-F(x)][1-G(y)]\}, \quad -1 \le \theta \le 1.$$

El coeficient de correlació és $\rho = \theta/3$. Aquest coeficient verifica $-1/3 \le \rho \le 1/3$ i és tal que $|\theta/3|$ és la primera correlació canònica. De fet, només n'hi ha una, i la FGM seria el cas més senzill de desenvolupament diagonal.

14.3 Còpules

Sigui (U,V) un vector aleatori amb valors sobre $\mathbf{I}^2$, essent $\mathbf{I} = [0,1]$. Una còpula és una funció de distribució $C(u,v)$ amb marginals uniformes $(0,1)$. Per tant, $C(u,v)$ verifica

$$C(0,v) = C(u,0) = 0, \quad C(u,1) = u, \quad C(1,v) = v.$$

El cas d'independència entre U i V correspon a la còpula $C_0(u,v) = uv$. Les cotes de Fréchet-Hoeffding són

$$C_-(u,v) = \max\{u+v-1,0\}, \quad C_+(u,v) = \min\{u,v\}.$$

Per a tota còpula C es verifica

$$C_-(u,v) \le C(u,v) \le C_+(u,v), \quad u,v \in \mathbf{I}. \tag{23}$$

Les còpules són importants pel teorema de Sklar, que diu que a tota distribució H li podem associar una còpula C_H tal que $H(x,y) = C_H(F(x),G(y))$. Per tant, n'hi ha prou a estudiar còpules per generar distribucions bivariants. Per exemple, la família de còpules

$$C_\theta(u,v) = uv[1+\theta(1-u)(1-v)], \quad -1 \le \theta \le 1, \tag{24}$$

genera la distribució FGM. Només hem de substituir u i v per $F(x)$ i $G(y)$.

Les còpules, també anomenades *funcions de dependència* o *representacions uniformes*, capturen l'associació d'un parell (U,V) i en general, aplicant el teorema de Sklar, descriuen l'associació de qualsevol parell de variables aleatòries (X,Y).

Per mesurar l'associació entre X i Y amb distribució conjunta H i còpula C, hom utilitza els coeficients de correlació ρ_S de Spearman i τ de Kendall, que es defineixen per

$$\rho_S = 12\int_0^1\int_0^1 C(u,v)\,du\,dv - 3,$$

$$\tau = 4\int_0^1\int_0^1 C(u,v)\,dC(u,v) - 1.$$

Tot i que ρ_S i τ estan definits en termes de còpules, tenen sentit per a qualsevol distribució conjunta $H(x,y)$. Ambdues mesures d'associació verifiquen:

1. Són invariants respecte de transformacions estrictament creixents de X i Y.
2. $\tau = \rho_s = 0$ si $H = H_0$ (independència estocàstica).
3. $\tau = \rho_s = -1$ si $H = H_-$ (cota inferior de Fréchet-Hoeffding).
4. $\tau = \rho_s = +1$ si $H = H_+$ (cota superior de Fréchet-Hoeffding).

Numèricament ρ_s i τ són semblants. S'han estudiat molt les relacions entre ρ_S i τ, la més coneguda de les quals és

$$-1 \leq 3\tau - 2\rho_S \leq 1.$$

Generalment es fa servir τ si es pot calcular fàcilment amb la còpula triada (cas de les anomenades còpules arquimedianes). Es fa servir ρ_S si el seu càlcul resulta més assequible que el de τ.

EXEMPLES. Per a la distribució FGM, amb còpula (24), s'obté $\rho_S = \theta/3$ i $\tau = 2\theta/9$. Es pot donar el cas que ρ_S valgui 0 però les variables siguin dependents. Per exemple:

$$C(u,v) = uv + (2u-1)u(1-u)(2v-1)v(1-v)/2,$$

no és la còpula independència, però $\rho_S = 0$. A més a més, en aquest exemple també és $\tau = 0$.

Un altre exemple ens el dóna la distribució de Plackett, que s'aplica com a alternativa a la normal bivariant. Aquesta distribució apareix en el context de l'estudi de la dependència en taules de contingència 2×2 i es defineix com a $H(x,y)$ amb marginals $F(x)$ i $G(y)$, de tal manera que

$$\theta = \frac{H(x,y)[1 - F(x) - G(y) + H(x,y)]}{[F(x) - H(x,y)][(G(y) - H(x,y))]}$$

és constant per a tot x, y. El paràmetre $\theta > 0$ mesura la depèndència. Hi ha independència estocàstica si $\theta = 1$. Suposem ara que les distribucions marginals són de Cauchy, distribució que no té moments (la mitjana i la variància no existeixen). Aleshores no podem calcular el coeficient de correlació lineal de Pearson, però, emprant la còpula corresponent, podem calcular la rho de Spearman, que per a la distribució de Plackett és

$$\rho_\theta = \frac{\theta+1}{\theta-1} - \frac{2\theta}{(\theta-1)^2}\log(\theta).$$

Fins aquí hem resumit els aspectes més bàsics de les distribucions amb marginals donades i les seves còpules. Per a un estudi ampli d'aquest tema, vegeu [51].

Si la modelització de les dades bivariants la podem descriure mitjançant una còpula, tenim un camí idoni per esbrinar el tipus d'associació entre les variables. Però hi ha moltes famílies de còpules: simètriques, arquimedianes, max-estables,

singulars, etc. Recentment es tendeix a identificar primer l'estructura d'una còpula partint de les dades [48] i després triar la còpula escaient.

Les seccions següents contenen resultats que resumeixen tres articles [13, 17, 20] sobre còpules.

15 La covariància entre dues funcions

Sigui (X,Y) un vector aleatori amb funció de distribució H i marginals F i G. Vegem primer dues generalitzacions de (20) i del desenvolupament (21). D'una banda tenim que

$$\operatorname{cov}(\alpha(X),\beta(Y)) = \int_{R^2} [H(x,y) - F(x)G(y)]\, d\alpha(x)\, d\beta(y),$$

suposant que les funcions α i β verifiquen les condicions d'existència de la variància [3, 11]. D'altra banda tenim que

$$H(x,y) - F(x)G(y) = \sum_{n\geq 1} \rho_n \int_a^b M_1(x,s)\, da_n(s) \int_c^d M_2(t,y)\, db_n(t), \quad (25)$$

on $M_1(x,y) = \min\{F(x),F(y)\} - F(x)F(y)$, $M_2(x,y) = \min\{G(x),G(y)\} - G(x)\,G(y)$; vegeu [12].

Convé observar que (25) és un desenvolupament de la funció de distribució en termes de correlacions canòniques i cotes de Fréchet. De fet, estem interessats a trobar les correlacions i les funcions canòniques en el cas de còpules. Altrament dit, hem de trobar funcions ϕ_1 i ϕ_2 tals que

$$\rho = \sup \operatorname{cor}(\phi_1(U),\phi_2(V)). \quad (26)$$

Per a aquest fi ens serà útil el resultat següent.

TEOREMA 1. *Siguin C, C^* dues còpules corresponents a dos vectors aleatoris (U,V) i (U^*,V^*). Si $\alpha(u)$ i $\beta(v)$ són funcions de variació afitada definides sobre $\mathbf{I} = [0,1]$, aleshores*

$$\operatorname{cov}(\alpha(U),\beta(V)) - \operatorname{cov}(\alpha(U^*),\beta(V^*)) = \int_{\mathbf{I}}\int_{\mathbf{I}} [C(u,v) - C^*(u,v)]\, d\alpha(u)\, d\beta(v).$$

En particular, la covariància entre $\alpha(U)$ i $\beta(V)$ és

$$\operatorname{cov}(\alpha(U),\beta(V)) = \int_{\mathbf{I}}\int_{\mathbf{I}} [C(u,v) - uv]\, d\alpha(u)\, d\beta(v), \quad (27)$$

i la variància de $\alpha(U)$ és

$$\operatorname{var}[\alpha(U)] = \int_{\mathbf{I}}\int_{\mathbf{I}} [\min\{u,v\} - uv]\, d\alpha(u)\, d\alpha(v). \quad (28)$$

A partir d'ara suposarem que (U,V) és intercanviable. Altrament dit, la còpula $C(u,v)$ és una funció simètrica. Aleshores podem suposar $\phi_1 = \phi_2$ a (26). El problema es redueix a trobar ρ i ϕ de tal manera que

$$\rho = \sup \operatorname{cov}(\phi(U),\phi(V)) / \operatorname{var}[\phi(U)].$$

16 Operadors integrals

Considerem una còpula C, la cota superior $\min\{u, v\}$, que denotem per M, i la còpula d'independència uv, que anomenem Π. Tenim que $C \prec M$ en el mateix sentit que a (23). Aleshores és clar que a les integrals de (27) i (28) hi intervenen els nuclis simètrics $K = C - \Pi$ i $L = M - \Pi$, que verifiquen $K \prec L$.

Definim a continuació els operadors integrals i els productes escalars relacionats amb una còpula C. Advertim que fem servir el mateix símbol K per al nucli i per a l'operador integral associat. Suposarem que totes les funcions pertanyen al conjunt $\mathbb{B}$ de funcions de variació afitada a l'interval $\mathbf{I} = [0, 1]$. Sabem que $\mathbb{B} \subset L^2(\mathbf{I})$.

DEFINICIÓ 2. L'operador integral K sobre $\mathbb{B}$ és

$$K\varphi(u) = \int_{\mathbf{I}} K(u, v)\, d\varphi(v).$$

De manera similar definim $L\varphi$ i $T\varphi$ essent $T = aK + bL$.

K i L proporcionen la covariància i la variància, respectivament. Els operadors integrals també serveixen per resoldre equacions diferencials.

EXEMPLE. Considerem l'equació $y'' + f(x) = 0$, $y(0) = y(1) = 0$, on $f(x)$ és una funció contínua a l'interval $(0, 1)$. Es pot provar que la solució és $y(x) = \int_{\mathbf{I}} L(x, s)\, dF(s)$, essent $F(x)$ una primitiva de $f(x)$ i $L(s, t) = \min\{s, t\} - st$.

DEFINICIÓ 3. El producte escalar generalitzat entre les funcions integrables ϕ, φ de $\mathbb{B}$ és

$$(\phi, K\varphi) = \int_{\mathbf{I}^2} K(u, v)\, d\phi(u)\, d\varphi(v).$$

De manera similar, definim $(\phi, L\varphi)$, $(\phi, T\varphi)$ i $(K\phi, L\varphi)$.

DEFINICIÓ 4. L'operador K és L-compacte si per a tota successió L-afitada $\{\phi_n\}$, en el sentit que $(\phi_n, L\phi_n)$ és finit, la successió $\{K\phi_n\}$ conté una subsuccessió convergent $\{K\phi_{n(i)}\}$. És a dir, $K\phi_{n(i)} \to \phi^*$, en el sentit que existeix ϕ^* tal que $||K\phi_{n(i)} - \phi^*||_L \to 0$ quan $i \to \infty$.

Són ben conegudes les propietats dels operadors a l'espai $L^2(\mathbf{I})$ quan K és general i L és l'operador identitat. Però aquí L, que permet trobar la variància, no és la identitat.

Les propietats següents es poden demostrar aplicant (27) i (28) i adaptant propietats conegudes en el cas que L sigui l'operador identitat.

1. Els operadors integrals K i L són lineals i satisfan

   $$(\phi, L\phi) \geq \max\{(\phi, K\phi), 0\} \geq 0.$$

   Per tant, L és definit positiu i el producte escalar $(\phi, L\varphi)$ defineix la norma $\|\phi\|_L = [(\phi, L\phi)]^{1/2}$.
2. K és de Hilbert-Schmidt, *i. e.*, $\int_{\mathbf{I}^2} K(u, v)^2\, du\, dv < \infty$.
3. L'operador integral K és L-compacte.

4. Si K és lineal i L-compacte, aleshores $T = aK + bL$ és també compacte per a qualsevol parella de constants $a, b \in \mathbb{R}$.

Clarament, si canviem $K = C - \Pi$ per $L = M - \Pi$, les propietats anteriors segueixen essent vàlides, atès que L és també la diferència de dues còpules.

17 Anàlisi generalitzada de funcions pròpies

Vegem ara alguns resultats sobre funcions pròpies i valors propis relacionats amb els operadors K i L.

DEFINICIÓ 5. Una funció pròpia generalitzada de K respecte de L és la parella (ϕ, λ), on ϕ és una funció i λ és un nombre real, tals que

$$\int_{\mathbf{I}} K(u,v)\, d\phi(v) = \lambda \int_{\mathbf{I}} L(u,v)\, d\phi(v), \tag{29}$$

uniformement en $u \in \mathbf{I}$. Ho expressem per $K\phi = \lambda L\phi$.

EXEMPLE. Sigui la còpula $F_\lambda = \lambda M + (1-\lambda)\Pi$. Aleshores $K = F_\lambda - \Pi = \lambda(M - \Pi)$. Per tant, $K\phi = \lambda L\phi$ on $L(s,t) = \min\{s,t\} - st$ i ϕ és qualsevol funció integrable sobre l'interval $(0,1)$.

Si (ϕ, λ) és una parella pròpia (funció i valor propi) de K respecte de L, amb $\lambda \neq 0$, tenim que

$$\frac{(\phi, K\phi)}{(\phi, L\phi)} = \lambda. \tag{30}$$

No és fàcil provar el contrari: si (ϕ, λ) satisfà (30) uniformement en $u \in \mathbf{I}$, aleshores és una parella pròpia que satisfà (29).

Presentem a continuació algunes propietats de les parelles pròpies. Aquí $\|K\|_L$, $\|K\|$ i $\|T\|$ són normes dels operadors K i T.

TEOREMA 6. *Sigui* $\|K\|_L = \sup|(\xi, K\xi)/(\xi, L\xi)|$, $\|K\| = \sup|(\xi, K\xi)/(\xi,\xi)|$ *i* $\|T\| = \sup|(\xi, T\xi)/(\xi,\xi)|$, *essent* $\xi \neq \mathbf{0}$, *on* $T = aK + bL$, $a, b \in \mathbb{R}$. *Si* α *i* β *són dues funcions de* $\mathbb{B}$, *aleshores:*

1. *Si* K *és definit positiu,*

$$(\alpha, K\beta) \leq \min\{\|K\|_L \|\alpha\|_L \|\beta\|_L, \|K\| \|\alpha\| \|\beta\|\}. \tag{31}$$

2. *En general,*

$$|(\alpha, T\beta)| \leq \|T\| \|\alpha\| \|\beta\|. \tag{32}$$

Amb l'ajut d'aquest teorema es demostra que existeix almenys un valor propi, que en el nostre cas significa que existeix la màxima correlació canònica. El teorema següent és ben conegut quan L és l'operador identitat. En el nostre

context de dos operadors generats per còpules, la demostració canvia bastant i no és fàcil.[8]

TEOREMA 7. *K té almenys un valor propi respecte de L.*

Com en el cas clàssic (on L és la identitat), la demostració consisteix a provar que el valor propi, que sempre existeix, és

$$\lambda_1 = \sup_{\xi \neq 0} (\xi, K\xi)/(\xi, L\xi). \tag{33}$$

Però aquí L no és la identitat i la demostració necessita incorporar algunes innovacions [17].

Ara que sabem que hi ha valors propis, mencionem algunes propietats.

1. Si λ és un valor propi, aleshores $\lambda \leq \lambda_1 \leq 1$, on λ_1 es defineix a (33).
2. Si ψ i ψ^* són funcions pròpies amb valors propis $\lambda \neq \lambda^*$, aleshores $(\psi, K\psi^*) = (\psi, L\psi^*) = 0$.
3. Sigui $\{\psi_n, \lambda_n\}$ un conjunt de parelles pròpies tals que $(\psi_n, L\psi_n) > 0$. Si $\{\psi_n, \lambda_n\}$ és infinit numerable, aleshores $\lim_{n\to\infty} \lambda_n = 0$.
4. Sigui $\{\psi_n, \lambda_n\}$ un conjunt de parelles pròpies tals que $(\psi_n, L\psi_n) > 0$. Definim
$$\Phi_n(u) = \int_{\mathbf{I}} L(u,s)\, d\psi_n(s). \tag{34}$$
Es verifica $\Phi_n(0) = \Phi_n(1) = 0$, $\Phi_n' = -\psi_n$ i la funció contínua Φ_n té diversos zeros a $\mathbf{I}$.

Recordem ara el teorema de Mercer [2]. Si $A(s,t)$ és el nucli d'un operador integral continu, simètric i definit no negatiu, i $\{\psi_n, \lambda_n\}$ és el conjunt numerable de parelles pròpies, on els λ_n són positius i ψ_n és un sistema ortonormal, aleshores

$$A(s,t) = \sum_{n\geq 1} \lambda_n \psi_n(s)\psi_n(t), \tag{35}$$

on la sèrie convergeix absolutament i uniformement en s, t.

Si A és l'operador amb nucli $L(u,v) = \min\{u,v\} - uv$, i considerem les funcions $h_n(x) = \int_0^x \psi_n(s)\, ds$, podem obtenir una successió ortogonal de variables aleatòries, que permet desenvolupar U (i en general qualsevol variable X) en sèrie del tipus $X = \sum_{n\geq 1} d_n h_n(X)$, on la convergència és en mitjana quadràtica. Vegeu [16, 24, 26].

EXEMPLE. Suposem U amb distribució uniforme sobre (0,1). Aleshores $\min\{s,t\} - st = \sum_{n\geq 1} \lambda_n \psi_n(s)\psi_n(t)$, essent $\psi_n(x) = \sqrt{2}\sin(n\pi x)$, $\lambda_n = 1/(n\pi)^2$, $h_n(x) = [\sqrt{2}/(n\pi)][1 - \cos(n\pi x)]$.

En el cas de dos operadors integrals K i L, el desenvolupament en sèrie (35) presenta algunes modificacions.

8 Donades dues matrius simètriques **A** i **B**, la segona definida positiva, no és immediat provar que hi ha almenys un valor propi de **A** respecte de **B** (un λ tal que $\mathbf{Av} = \lambda\mathbf{Bv}$), sense fer ús de matrius inverses ni determinants, conceptes que no podem aplicar (en general) treballant amb operadors integrals.

TEOREMA 8. *Sigui $\{\psi_n, \lambda_n\}$ el conjunt numerable de parelles pròpies de K respecte de L, tals que $(\psi_n, L\psi_n) = 1$. Suposem que K i L són dos operadors integrals continus, simètrics, essent L definit positiu. Aleshores és vàlid el desenvolupament*

$$K = \sum_{n\geq 1} \lambda_n L\psi_n \otimes \psi_n L, \tag{36}$$

en el sentit que $\left[K(u,v) - \sum_{i=1}^{n} \lambda_i \int_{\mathbf{I}} L(u,s)\, d\psi_i(s) \int_{\mathbf{I}} L(t,v)\, d\psi_i(t)^2\right]^2 \to 0$ si $n \to \infty$, uniformement en $u, v \in \mathbf{I}$. A més, si $\{\lambda_n\}$ és infinit numerable, es compleix que $\lim_{n\to\infty} \lambda_n = 0$.

Ressaltem que els valors propis poden ser negatius i que per a certes còpules, el desenvolupament (36) podria ser una integral en lloc d'una sèrie.

Si fem ús de la funció (34), veiem fàcilment que

$$K(u,v) = \sum_{n\geq 1} \lambda_n \Phi_n(u)\Phi_n(v). \tag{37}$$

EXEMPLE. Considerem la còpula

$$C(u,v) = uv + u(1-u)v(1-v) + (2u-1)u(1-u)(2v-1)v(1-v)/2.$$

Si $K(u,v) = C(u,v) - uv$ i $L(u,v) = \min\{u,v\} - uv$, els valors propis de K respecte a L són $\lambda_1 = 1/3$ i $\lambda_2 = 1/10$. Es verifica

$$K = \tfrac{1}{3} L\psi_1 \otimes \psi_1 L + \tfrac{1}{10} L\psi_2 \otimes \psi_2 L,$$

essent $\psi_1(u) = \sqrt{3}(1-2u)$ i $\psi_2(u) = \sqrt{5}(6u^2 - 6u + 1)$ les corresponents funcions pròpies.

18 Anàlisi canònica

Apliquem ara aquest esquema a les còpules. El resultat següent relaciona les parelles pròpies amb les funcions i correlacions canòniques.

TEOREMA 9. *Sigui (U,V) un vector aleatori que té per distribució la còpula C. Suposem que $\phi\colon \mathbf{I} \to \mathbb{R}$ és una funció tal que $\operatorname{var}[\phi(U)]$ és finita. Considerem els operadors integrals $K = C - \Pi$ i $L = M - \Pi$. Es verifica:*

1. *El coeficient de correlació entre $\phi(U)$ i $\phi(V)$ és*

$$\operatorname{cor}(\phi(U), \phi(V)) = \frac{(\phi, K\phi)}{(\phi, L\phi)}, \tag{38}$$

   essent $(\phi, K\phi) = \operatorname{cov}(\phi(U), \phi(V))$, $(\phi, L\phi) = \operatorname{var}[\phi(U)]$.
2. *Si ϕ és una funció pròpia de K respecte a L, aleshores ϕ és una funció canònica i $(\phi, K\phi)/(\phi, L\phi) = \rho$ és la corresponent correlació canònica.*

Per tant, a partir de (36), podem expressar la còpula C per

$$C = \Pi + \sum_{n\geq 1} \rho_n L\phi_n \otimes \phi_n L, \tag{39}$$

que seria una variant de (25).

Tenint en compte propietats geomètriques basades en la distància khi quadrat (13), vegeu [16, 29], podem definir la dimensió d'una còpula com la cardinalitat del conjunt $\{\lambda_n\}$ de valors propis de K respecte a L.

Tenim aleshores:

1. Còpula de dimensió 0: és la còpula independència $\Pi(u,v) = uv$.
2. Còpula de dimensió 1: un exemple és la còpula FGM, definida a (24).
3. Còpula de dimensió 2: un exemple és

$$C_2(u,v) = uv + \theta_1 u(1-u)v(1-v) + \theta_2(2u-1)u(1-u)(2v-1)v(1-v), \tag{40}$$

   on (θ_1, θ_2) pertany a una regió continguda a $[-2,2]\times[-1,2]$.
4. Còpula de dimensió infinita numerable: un exemple és la còpula AMH (Ali, Mikhail i Haq):

$$AMH_\theta(u,v) = uv/[1-\theta(1-u)(1-v)], \quad -1 \leq \theta \leq 1. \tag{41}$$

5. Còpula de dimensió contínua, és a dir, infinita no numerable: un exemple és la còpula de Cuadras-Augé:

$$CA_\theta(u,v) = (uv)^{1-\theta}(\min\{u,v\})^\theta, \quad 0 \leq \theta \leq 1. \tag{42}$$

Aquesta còpula presenta una novetat important en la teoria de les funcions pròpies, tradicionalment basada en el supòsit que els valors propis formen un conjunt numerable. Sobre això en parlarem a la secció següent.

18.1 Ajust d'una còpula per una altra de coneguda

Un problema de les còpules (i dels models en general), és que no sempre sabem quina és la que podem ajustar a les dades observades. Suposem que tenim dues còpules: C_M, la còpula que fem servir com un model, i C_T, la còpula «veritable», la que veritablement segueixen les nostres dades bivariants.

Suposem que el desenvolupament canònic

$$dC_M(u,v) = du\,dv + \sum_n \rho_n A_n(u)\,du B_n(v)\,dv$$

de C_M existeix i és coneguda, essent A_n i B_n funcions canòniques unitàries (mitjana 0 i variància 1). Ens interessa aproximar C_T per una combinació lineal finita de correlacions i funcions canòniques. Dit amb més precisió, volem obtenir l'aproximació (vegeu [25])

$$\frac{dC_M(u,v)}{du\,dv} \simeq 1 + \sum_{i=1}^{k} c_i A_i(u) B_i(v),$$

on $c_1, \ldots, c_k$ són coeficients reals que minimitzen l'expressió

$$\int_{\mathbf{I}}\int_{\mathbf{I}} \left(\frac{dC_M(u,v) - du\,dv}{du\,dv} - \sum_{i=1}^{k} c_i A_i(u) B_i(v) \right)^2 du\,dv. \tag{43}$$

TEOREMA 10. *Suposem* $(U,V) \sim C_T$. *Els coeficients que minimitzen* (43) *són* $c_i = \rho_i$, *essent*

$$\rho_i = \operatorname{cor}(A_i(U), B_i(V)), \quad i = 1, \ldots, k.$$

L'avantatge d'aquest procediment és clar: coneixem les funcions canòniques A_i i B_i i, per tant, podrem calcular (mitjançant estimació estadística) les correlacions entre les parelles $A_i(U)$ i $B_i(V)$ respecte de la veritable còpula C_T, que és la que segueixen les nostres dades, però que en general desconeixem.

θ	$\rho_1 = \rho_S(C_2)$	ρ_2	a	$\rho_S(AMH)$	$\tau(AMH)$	$\tau(C_2)$
−1	−0.2711	0.0217	0.0055	−0.2710	−0.1817	−0.1815
−0.5	−0.1489	0.0080	0.0017	−0.1489	−0.0995	−0.0995
0.5	0.1924	0.0223	0.0032	0.1924	0.1288	0.1286
1	0.4783	0.2323	0.0261	0.4784	0.3333	0.3335

TAULA 7: Aproximació de la còpula AMH (model cert però desconegut) per una còpula coneguda que té dimensió dos. L'ajust és força bo i els coeficients de correlació rho i tau són gairebé els mateixos.

EXEMPLE. Una aproximació en dimensió 1 consisteix a utilitzar la còpula (24). La FGM és un model molt utilitzat en estadística, tot i que dóna un ajust pobre. Una aproximació bastant millor seria la còpula (40) (o una versió amb dimensió superior). Suposem, per exemple, que la còpula AMH, definida a (41), és la C_T, i que l'aproximem per la còpula (40), que fa el paper de model C_M. Calculant

$$a = \max_{u,v \in I} |C_T(u,v) - C_M(u,v)|,$$

obtenim valors petits, és a dir, un bon ajust, depenent del paràmetre θ de la còpula correcta AMH. A més, les correlacions de Spearman i Kendall donen pràcticament el mateix. La taula 7 mostra els valors dels coeficients de l'aproximació, expressats en termes de correlacions canòniques, on ρ_1 coincideix amb la rho de Spearman per a la còpula C_2.

Ara bé, si reduïm la dimensió d'una còpula eliminant termes del desenvolupament diagonal (36), o de (37), hem de tenir en compte que no sempre obtenim una nova còpula. De manera similar, en anàlisi de correspondències, quan representem les files i columnes en dimensió dos, el gràfic podria reflectir una taula amb valors negatius.

19 Dimensió contínua

Considerem el cas de parelles pròpies (ϕ, λ) on la funció ϕ té norma 0. Això significa, en el nostre context, tractar amb variables aleatòries no constants però amb variància 0. Però aquesta suposició fa trontollar la teoria dels valors propis. Abans de seguir, comentem algunes diferències amb la teoria ordinària dels operadors aplicats a funcions que no són constants.

1. Les funcions pròpies corresponents a un mateix valor propi generen un subespai de dimensió finita. Si la norma és 0 haurem de considerar valors propis amb multiplicitat infinita.
2. De $K\phi = \lambda\phi$, deduïm que les funcions pròpies ordinàries són contínues a $\mathbf{I} = [0, 1]$. Ara tenim $K\varphi = \lambda L\phi$ i podem trobar funcions contínues a trossos amb discontinuïtats de salt.
3. Les funcions es normalitzen segons $(\phi, L\phi) = 1$. Ara hem de considerar el cas $(\phi, L\phi) = 0$. Aleshores suposarem que existeix una família de funcions $\{\phi_\varepsilon\}$ de tal manera que $(\phi_\varepsilon, L\phi_\varepsilon) > 0$ i a més

$$\lim_{\varepsilon \to 0} \phi_\varepsilon = \phi, \quad \lim_{\varepsilon \to 0} \frac{K\phi_\varepsilon}{L\phi_\varepsilon} = \lambda > 0,$$

   en el sentit que $K\phi_\varepsilon(u)/L\phi_\varepsilon(u) \to \lambda$ uniformement en $u \in \mathbf{I}$.
4. El conjunt $\mathbb{S}_\lambda$ de valors propis és finit o infinit numerable. En aquest nou escenari, els valors propis de K respecte a L poden constituir un conjunt $\mathbb{S}_\lambda$ no numerable. Per exemple, $\mathbb{S}_\lambda$ podria ser un interval que té la potència del continu.

EXEMPLE. Definim $\phi_\varepsilon(u) = 1$ si $u \in (1 - \varepsilon, 1]$, $\phi_\varepsilon(u) = 0$ en altre cas, on $\varepsilon > 0$ és arbitràriament petit. Siguin ϕ l'indicador de $\{u = 1\}$ i $L = M - \Pi$, $K = M^{1/2}\Pi^{1/2} - \Pi$. Es pot provar (vegeu [12, 16]) que $(\phi_\varepsilon, L\phi_\varepsilon) = \varepsilon(1 - \varepsilon)$, $(\phi_\varepsilon, K\phi_\varepsilon) = (1 - \varepsilon)^{3/2} - (1 - \varepsilon)^2$, per tant, $\lim_{\varepsilon \to 0} K\phi_\varepsilon / L\phi_\varepsilon = 1/2$. Tenim que $(\phi, 1/2)$ és una parella pròpia de K respecte a L de tal manera que $(\phi, K\phi) = (\phi, L\phi) = 0$.

Ressaltem que les funcions pròpies de norma 0 poden coexistir amb les de norma positiva.

20 Còpules amb part singular

Una família general que conté còpules amb dimensió finita, numerable o contínua, és

$$D_\theta(u, v) = \min\{u, v\} G_\theta(\max\{u, v\}),$$

on θ és un paràmetre i G_θ és una funció de la qual parlarem tot seguit. Un membre d'aquesta família és la CA_θ, definida a (42), que és la mitjana geomètrica de les còpules M i Π. Tenim que $CA_0 = \Pi$ i $CA_1 = M$. Així doncs θ és una mesura de dependència.

Aquesta família, introduïda per Cuadras i Augé [18], ha estat estudiada i aplicada a [32, 49, 51] i a d'altres treballs. Té una part singular, ja que: $P[U = V] = \theta/(2-\theta)$. De fet, si $\delta_{\mathbf{A}}$ representa la funció indicadora de l'esdeveniment $\mathbb{A}$, la densitat és

$$f_\theta(u,v) = (1-\theta)\max\{u,v\}^{-\theta}\delta_{\{u\neq v\}} + \theta u^{1-\theta}\delta_{[u=v\}}, \tag{44}$$

respecte de la mesura $\mu = \lambda^2 + \lambda^1$, essent λ^2 la mesura de Lebesgue sobre $\mathbb{R}^2$ i λ^1 la mateixa mesura concentrada a $\{u = v\}$. La rho de Spearman i la tau de Kendall de CA_θ són: $\rho_S = 3\theta/(4-\theta)$ i $\tau = \theta/(2-\theta)$.

EXEMPLE. La família CA_θ s'ha utilitzat per definir un índex d'estabilitat financera, anomenat «Cuadras-Augé index»; vegeu [6]. Comentem, però, una situació més senzilla: la distribució dels sous de les parelles d'una població. Siguin (X,Y) els sous de l'home i la dona que són parella. Els dos sous poden ser independents, però hi ha la possibilitat que siguin iguals (cas freqüent en professors i funcionaris). Suposem que les funcions de distribució són de Pareto: $F(x) = 1-(m_0/x)^\alpha$ si $x \geq m_0$, $G(y) = 1-(m_0/y)^\beta$ si $y \geq m_0$, on m_0 representa el sou mínim i α i β són paràmetres tals que $2 < \alpha < \beta$. Observem que la mitjana de X és més gran que la de Y, i que si el paràmetre β tendeix a infinit, el sou mitjà tendeix al mínim m_0. Podem prendre CA_θ com a còpula relacionada amb la distribució conjunta $H(x,y)$ de (X,Y), on els sous mínim i màxim correspondrien als valors 0 i 1, respectivament. Però la densitat (44) de (U,V) indica que els sous iguals tenen més probabilitat a mesura que augmenten. Considerem, doncs, $(1-U, 1-V)$. Es fàcil veure que la còpula és

$$C_S(u,v) = u + v - 1 + CA_\theta(1-u, 1-v).$$

Per a C_S també els sous mínim i màxim corresponen als valors 0 i 1. La densitat adopta l'expressió

$$g_\theta(u,v) = (1-\theta)\max\{1-u,1-v\}^{-\theta}\delta_{\{u\neq v\}} + \theta(1-u)^{1-\theta}\delta_{[u=v\}}.$$

Ara els sous alts tenen menys probabilitat de coincidir.

Aplicant el teorema de Sklar (secció 14.3), la funció de distribució conjunta, possible model per a (X,Y), és $H(x,y) = C_S(F(x),G(y))$, és a dir, $H(x,y) = 0$ si $x, y < m$, i

$$H(x,y) = 1-(m_0/x)^\alpha - (m_0/y)^\beta + CA_\theta((m_0/x)^\alpha,(m_0/y)^\beta) \text{ si } x,y \geq m_0.$$

Continuant amb la teoria, a fi de trobar els valors i les funcions pròpies per als nuclis $K = CA_\theta - \Pi$ i $L = M - \Pi$, definim per a $0 \leq y \leq 1$ i $\varepsilon > 0$ arbitràriament petit, la funció

$$\mathcal{H}_{y,\varepsilon}(x) = \begin{cases} 0 & \text{si } x < y, \\ 1 & \text{si } y \leq x < y+\varepsilon, \\ \varepsilon/(y+\varepsilon) & \text{si } x \geq y+\varepsilon, \end{cases}$$

que té per límit la funció indicadora $\phi_\gamma(x)$,

$$\lim_{\varepsilon\to 0}\mathcal{H}_{\gamma,\varepsilon}(x)=\phi_\gamma(x)=\begin{cases}1 & \text{si } x=\gamma,\\ 0 & \text{si } x\neq\gamma.\end{cases}$$

Presentem ara la descomposició en valors i funcions pròpies de K respecte a L. El resultat següent es pot provar aplicant teoria de funcions generalitzades. Aquí fem l'anàlisi sobre els nuclis $K_{\theta_i}=M^{\theta_i}\Pi^{1-\theta_i}-\Pi$, $i=1,2$. Observem que $L=M-\Pi=K_1$.

TEOREMA 11. *Si $0\le\theta_1<\theta_2\le 1$, el conjunt de parelles pròpies de K_{θ_1} respecte a K_{θ_2} està format per $(\phi_\gamma,\lambda_\gamma)$ on*

$$\phi_\gamma=\mathcal{H}_{\gamma,0},\quad \lambda_\gamma=(\theta_1/\theta_2)\gamma^{\theta_2-\theta_1},\quad 0\le\gamma\le 1.$$

Tenint ara en compte que les parelles pròpies són funcions i correlacions canòniques, obtenim el resultat següent.

TEOREMA 12. *El conjunt de funcions i correlacions canòniques de la família CA_θ és $(\phi_\gamma,\theta\gamma^{1-\theta})$, $0\le\gamma\le 1$.*

Per tant, el conjunt de correlacions canòniques és un interval que té la potència del continu. Les variables canòniques tenen, en el límit, variància 0, però les correlacions canòniques són positives. Una conseqüència del teorema anterior és el resultat següent.

COROL·LARI 13. *La màxima correlació per a la família CA_θ és el paràmetre θ, essent $\mathcal{H}_1$ la funció canònica, on $\mathcal{H}_1$ és la distribució de Heaviside.*

La prova que θ és la màxima correlació apareix primer a [12]. Recordem que la màxima correlació, introduïda per H. Gebelein el 1941, és l'única mesura de dependència que compleix els postulats de Rényi [55]. Per tant, θ és una bona mesura de dependència per a aquesta família.

Considerem a continuació la sèrie (36), construïda suposant que les funcions pròpies tenen norma 1. Però ara la norma és 0 i la sèrie esdevé una integral. La successió de correlacions canòniques és, en aquest context, una funció, en general contínua.

DEFINICIÓ 14. Una funció de correlació canònica és una funció integrable $f_\theta\colon \mathbf{I}\to\mathbf{I}$, que satisfà $\int_{\mathbf{I}} f_\theta(\rho)\,d\rho\le 1$.

TEOREMA 15. *La còpula Cuadras-Augé es pot expressar en termes del desenvolupament diagonal continu*

$$CA_\theta(u,v)=uv+\int_{\max\{u,v\}}^1\theta\rho^{1-\theta}(u/\rho)(v/\rho)\,d\rho.$$

En general, la família $D_\theta(u,v)=\min\{u,v\}G_\theta(\max\{u,v\})$ admet el desenvolupament integral

$$D_\theta(u,v)=uv+\int_{\max\{u,v\}}^1 f_\theta(u)(u/\rho)(v/\rho)\,d\rho, \tag{45}$$

on $f_\theta(\rho)=G_\theta(\rho)-\rho G'_\theta(\rho)$ és una funció de correlació canònica.

21 Còpules canòniques

El terme $\max\{u,v\}$ que apareix a (45) verifica $\max\{u,v\} = uv/\min\{u,v\}$. Doncs bé, si C és una còpula, en general

$$uv + \int_{uv/C(u,v)}^{1} f_\theta(\rho)(u/\rho)(v/\rho)\,d\rho$$

també ho és. De fet Π/C es podria substituir per qualsevol quocient de dues còpules.

Encara amb més generalitat, podem obtenir la família (46), de la qual (47) seria un cas particular.

TEOREMA 16. *Sigui f_θ una funció de correlació canònica indexada per $\theta \in \mathbf{I}$. Definim les funcions $F_\theta(\rho) = \int[f_\theta(\rho)/\rho^2]\,d\rho$ i $G_\theta(\rho) = -\rho F_\theta(\rho) + \rho F_\theta(1) + \rho$. Per a quocients adequats Q de dues còpules, la descomposició integral* (45) *permet generar la família de còpules canòniques*

$$C_\theta = \Pi \times G_\theta(Q)/Q. \tag{46}$$

En particular, si $Q = \Pi/M$, obtenim la família

$$D_\theta(u,v) = \min\{u,v\}G_\theta(\max\{u,v\}), \tag{47}$$

on $G_\theta\colon \mathbf{I} \to \mathbf{I}$ és una funció creixent contínua de tal manera que $G_\theta(\rho)/\rho$ és decreixent dins $(0,1]$.

f_θ	G_θ	Q	Família
0	ρ	C/Π	Independència
1	1	Π/C	Qualsevol còpula
θ	$\theta + \rho\overline{\theta}$	Π/M	Fréchet
$\theta\rho^2$	$-\theta\rho^2 + (1+\theta)\rho$	Π/AMH_1	Farlie-Gumbel-Morgenstern
$\theta\rho^2/[\overline{\theta} + \rho\theta]^2$	$-\rho/[\overline{\theta} + \rho\theta]$	FGM_{-1}/Π	Ali-Mikhail-Haq
$\theta\rho^{1-\theta}$	$\rho^{1-\theta}$	Π/M	Cuadras-Augé
$f_\theta(\rho)$	$G_\theta(\rho)$	Π/M	Durante
$\theta\rho$	$-\theta\rho\log(\rho) + \rho$	Π/W	Nova família (48)

TAULA 8: Algunes còpules generades per funcions canòniques i el desenvolupament diagonal en versió integral.

La taula 8 descriu algunes construccions, on $\overline{\theta} = 1 - \theta$. Les famílies Farlie-Gumbel-Morgenstern (FGM), Ali-Mikhail-Haq (AMH) han estat descrites abans, vegeu (24), (41). La família de Fréchet és la mitjana ponderada de la cota superior i la independència:

$$F_\theta(u,v) = \theta\min\{u,v\} + (1-\theta)uv.$$

D'altra banda, D_θ de (47) va ser definida per Cuadras i Augé [18] i estudiada per Durante [34]. És interessant observar que si $Q = \Pi/M$, la densitat (44) adopta la forma simètrica

$$f_\theta(u,v) = (1-\theta)(1/Q)^\theta \delta_{\{u\neq v\}} + \theta(Q)^{1-\theta}\delta_{\{u=v\}}.$$

Finalment, amb $f_\theta(\rho) = \theta\rho$ i $Q = \Pi/W$, on W és la còpula $W(u,v) = [u^{-2}+v^{-2}-1]^{-1/2}$, obtenim la nova família de còpules

$$NC(\mathrm{u},v) = uv[1-\theta\log(uv/W(u,v))]. \tag{48}$$

EXEMPLE. Sovint volem descriure un fet real que ha estat observat des de diverses perspectives. Com a il·lustració, sigui θ_1 la probabilitat d'un esdeveniment $\mathbb{A}$, com ara l'eficàcia d'un tractament mèdic. Realitzem n_1 proves independents i sigui $L_1(x_1,\theta_1) = \binom{n}{x_1}\theta_1^{x_1}(1-\theta_1)^{n_1-x_1}$ la funció de versemblança, on x_1 és la freqüència observada de $\mathbb{A}$. L'estimador màxim versemblant de θ_1 és la freqüència relativa x_1/n_1.

Suposem ara que θ_1 és un paràmetre aleatori amb distribució uniforme $(0,1)$. La versemblança, sota aquesta perspectiva bayesiana, és

$$L(x_1) = \int_0^1 \binom{n}{x_1}\theta_1^{x_1}(1-\theta_1)^{n_1-x_1}\,d\theta_1.$$

Considerem dos experiments independents sota els quals obtenim les freqüències relatives x_1/n_1 i x_2/n_2 per a estimar θ_1, θ_2. Suposem que θ_1 i θ_2 són descripcions d'un metaparàmetre θ, també aleatori, que ens donaria la «veritable» probabilitat de $\mathbb{A}$ en condicions ideals. (Imaginem que un tractament mèdic s'ha estudiat en dos centres diferents.) Suposem que la distribució conjunta de (θ_1,θ) és una còpula (FGM, Plackett o una còpula canònica), i anàlogament la de (θ_2,θ). La versemblança de θ és

$$L(x_1,x_2|\theta) = \prod_{i=1}^{2}\int_0^1 \binom{n}{x_i}\theta_i^{x_i}(1-\theta_i)^{n_i-x_i}p(\theta_i|\theta)\,d\theta_i,$$

on $p(\theta_i|\theta)$ és la densitat de θ_i condicionada a θ i que es calcula a partir de la còpula. La distribució *a priori* de θ és uniforme a $(0,1)$. Aplicant el teorema de Bayes, la densitat *a posteriori* del metaparàmetre és

$$p(\theta|x_1,x_2) = \frac{L(x_1,x_2|\theta)}{\int_0^1 L(x_1,x_2|\theta)\,d\theta}.$$

Un cop tenim $p(\theta|x_1,x_2)$ podem prendre la moda (o la mitjana) com una estimació de la probabilitat θ; vegeu [50].

Resumint, la teoria de còpules, i en particular les còpules canòniques, defineixen una classe àmplia de models probabilístics que, correctament utilitzats, descriuen i sintetitzen l'associació estadística d'un conjunt de dades bivariants.

Referències

[1] Aitchison, J.; Greenacre, M. «Biplots of compositional data». *J. Roy. Statist. Soc. Ser. C*, 51 (4) (2002), 375–392.

[2] Ash, R. B. *Information Theory*. Nova York: Dover Publications, Inc., 1990. Reedició corregida de l'original de 1965.

[3] Beare, B. K. «A generalization of Hoeffding's lemma, and a new class of covariance inequalities». *Statist. Probab. Lett.*, 79 (5) (2009), 637–642.

[4] Benzécri, J.-P. *L'analyse des données. II. L'analyse des correspondances*. 2a ed. París: Dunod, 1976.

[5] Bernardo, J. M. «Una introducció a l'estadística bayesiana». *Butlletí de la Societat Catalana de Matemàtiques*, 17 (1) (2002), 7–64.

[6] Cherubini, U.; Mulinacci, S.; Gobbi, F.; Romagnoli, S. *Dynamic copula methods in finance*. Nova York: Wiley, 2011.

[7] Cox, T. F.; Cox, M. A. A. *Multidimensional Scaling*. Londres: Chapman & Hall, 1994. (Monographs on Statistics and Applied Probability; 59)

[8] Cuadras, C. M. *Métodos de Análisis Multivariante*. Barcelona: PPU, 1991.

[9] Cuadras, C. M. «Interpreting an inequality in multiple regression». *Amer. Statist.*, 47 (4) (1993), 256–258.

[10] Cuadras, C. M. «Increasing the correlations with the response variable may not increase the coefficient of determination: a PCA interpretation». A: *New Trends in Probability and Statistics*. Vol. 3. Utrecht: VSP, 1995, 75–83.

[11] Cuadras, C. M. «On the covariance between functions». *J. Multivariate Anal.*, 81 (1) (2002), 19–27.

[12] Cuadras, C. M. «Correspondence analysis and diagonal expansions in terms of distribution functions». *J. Statist. Plann. Inference*, 103 (1–2) (2002), 137–150.

[13] Cuadras, C. M. «Continuous canonical correlation analysis». *Research Letters in the Information and Mathematical Sciences*, 8 (2005), 97–103.

[14] Cuadras, C. M. «Distance-based association and multi-sample tests for general multivariate data». A: *Advances in Mathematical and Statistical Modeling*. Boston: Birkhäuser, 2008, 61–71. (Stat. Ind. Technol.)

[15] Cuadras, C. M. *Nuevos Métodos de Análisis Multivariante*. Barcelona: CMC Editions, 2014.

[16] Cuadras, C. M. «Nonlinear principal and canonical directions from continuous extensions of multidimensional scaling». *Open J. Stat.*, 4 (2) (2014), 132–149.

[17] Cuadras, C. M. «Contributions to the diagonal expansion of a bivariate copula with continuous extensions». *J. Multivariate Anal.*, 139 (2015), 28–44.

[18] Cuadras, C. M.; Augé, J. «A continuous general multivariate distribution and its properties». *Comm. Statist. A—Theory Methods*, 10 (4) (1981), 339–353.

[19] Cuadras, C. M.; Cuadras, D. «A parametric approach to correspondence analysis». *Linear Algebra Appl.*, 417 (1) (2006), 64–74.

[20] Cuadras, C. M.; Cuadras, D. «Eigenanalysis on a bivariate covariance kernel». *J. Multivariate Anal.*, 99 (10) (2008), 2497–2507.

[21] Cuadras, C. M.; Cuadras, D. «Partitioning the geometric variability in multivariate analysis and contingency tables». A: *Classification and Multivariate Analysis for Complex Data Structures*. Heidelberg: Springer, 2011, 237–244. (Stud. Classification Data Anal. Knowledge Organ.)

[22] Cuadras, C. M.; Cuadras, D. «A unified approach for the multivariate analysis of contingency tables». *Open J. Stat.*, 5 (2015), 223–232.

[23] Cuadras, C. M.; Cuadras, D.; Greenacre, M. J. «A comparison of different methods for representing categorical data». *Comm. Statist. Simulation Comput.*, 35 (2) (2006), 447–459.

[24] Cuadras, C. M.; Cuadras, D.; Lahlou, Y. «Principal directions of the general Pareto distribution with applications». *J. Statist. Plann. Inference*, 136 (8) (2006), 2572–2583.

[25] Cuadras, C. M.; Díaz, W. «Another generalization of the bivariate FGM distribution with two-dimensional extensions». *Acta Comment. Univ. Tartu. Math.*, 16 (1) (2012), 3–12.

[26] Cuadras, C. M.; Fortiana, J. «A continuous metric scaling solution for a random variable». *J. Multivariate Anal.*, 52 (1) (1995), 1–14.

[27] Cuadras, C. M.; Fortiana, J. «Visualizing categorical data with related metric scaling». A: Blasius, J.; Greenacre, M. (ed.). *Visualization of Categorical Data*. Nova York: Academic Press, 1998, 365–376.

[28] Cuadras, C. M.; Fortiana, J. «The importance of geometry in multivariate analysis and some applications». A: *Statistics for the 21st Century*. Nova York: Dekker, 2000, 93–108. (Statist. Textbooks Monogr.; 161)

[29] Cuadras, C. M.; Fortiana, J.; Greenacre, M. «Continuous extensions of matrix formulations in correspondence analysis, with applications to the FGM family of distributions». A: Heijmans, R. D. H.; Pollock, D. S. G.; Satorra, A. (ed.). *Innovations in Multivariate Statistical Analysis*. Dordrecht: Kluwer Ac. Publ., 2000, 101–116.

[30] Cuadras, C. M.; Fortiana, J.; Oliva, F. «The proximity of an individual to a population with applications in discriminant analysis». *J. Classification*, 14 (1) (1997), 117–136.

[31] Cuadras, C. M.; Valero, S.; Cuadras, D.; Salembier, P.; Chanussot, J. «Distance-based measures of association with applications in relating hyperspectral images». *Comm. Statist. Theory Methods*, 41 (13–14) (2012), 2342–2355.

[32] Dobrowolski, E.; Kumar, P. «Some properties of the Marshall-Olkin and generalized Cuadras-Augé families of copulas». *Aust. J. Math. Anal. Appl.*, 11 (1) (2014), 13 p.

[33] Donner, A.; Wells, G. «A comparison of confidence interval methods for the intraclass correlation coefficient». *Biometrics*, 42 (2) (1986), 401–412.

[34] Durante, F. «A new family of symmetric bivariate copulas». *C. R. Math. Acad. Sci. Paris*, 344 (3) (2007), 195–198.

[35] Escofier, B.; Pagès J. *Analyses factorielles simples et multiples*. París: Dunod, 1990.

[36] Fréchet, M. «Sur les tableaux de corrélation dont les marges sont données». *Ann. Univ. Lyon. Sect. A. (3)*, 14 (1951), 53–77.

[37] Galton, F. «Regression towards mediocrity in hereditary stature». *J. of the Anthropological Institute*, 15 (1886), 246–263.

[38] Goodman, L. A. «Correspondence analysis, association analysis, and generalized nonindependence analysis of contingency tables: Saturated and un saturated models, and appropriate graphical displays». A: Cuadras, C. M.; Rao, C. R. (ed.). *Multivariate Analysis: Future Directions 2*. Amsterdam: Elsevier, 1993, 265–294.

[39] Greenacre, M. J. *Theory and Applications of Correspondence Analysis*. Londres: Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], 1984.

[40] Hamilton, D. «Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$. Correlated variables are not always redundant». *Amer. Statist.*, 41 (2) (1987), 129–132.

[41] Hanley, J. A. «"Transmuting" women into men: Galton's family data on human stature». *Amer. Statist.*, 58 (3) (2004), 237–243.

[42] Harris, H.; Smith, C. A. B. «The sib-sib age of onset correlation among individuals suffering from a hereditary syndrome produced by more than one gene». *Ann. Eugenics*, 14 (1949), 309–318.

[43] Hoeffding, W. «Masstabinvariante Korrelationtheorie». *Schriften Math. Inst. Univ. Berlin*, 5 (1940), 181–233.

[44] Hotelling, H. «Relations between two sets of variates». *Biometrika*, 28 (3–4) (1936), 321–377.

[45] Julià, O.; Comas, J.; Vives-Rego, J. «Second-order functions are the simplest correlations between flow cytometric light scatter and bacterial diameter». *Journal of Microbiological Methods*, 40 (1) (2000), 57–61.

[46] Kaplan, J. «A statistical error in The Bell Curve». *Chance*, 10 (1997), 20–21.

[47] Lancaster, H. O. «The structure of bivariate distributions». *Ann. Math. Statist.*, 29 (1958), 719–736.

[48] Li, B.; Genton, M. G. «Nonparametric identification of copula structures». *J. Amer. Statist. Assoc.*, 108 (502) (2013), 666–675.

[49] Mai, J.-F.; Scherer, M. «Efficiently sampling exchangeable Cuadras-Augé copulas in high dimensions». *Inform. Sci.*, 179 (17) (2009), 2872–2877.

[50] Moreno, E.; Vázquez-Polo, F. J.; Negrín, M. A. «Objective Bayesian meta-analysis for sparse discrete data». *Stat. Med.*, 33 (21) (2014), 3676–3692.

[51] Nelsen, R. B. *An Introduction to Copulas*. 2a ed. Nova York: Springer, 2006. (Springer Series in Statistics)

[52] Papadatos, N.; Xifara, T. «A simple method for obtaining the maximal correlation coefficient and related characterizations». *J. Multivariate Anal.*, 118 (2013), 102–114.

[53] Pearson, K.; Lee, A. «On the laws of inheritance in man: I. Inheritance of physical characters». *Biometrika*, 2 (1903), 357–462.

[54] Rao, C. R. «A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance». *Qüestiió (2)*, 19 (1–3) (1995), 23–63.

[55] Rényi, A. «On measures of dependence». *Acta Math. Acad. Sci. Hungar.*, 10 (1959), 441–451.

[56] Routledge, R. D. «When stepwise regression fails: correlated variables some of which are redundant». *Internat. J. Math. Ed. Sci. Tech.*, 21 (3) (1990), 403–410.

[57] Scheffé, H. *The Analysis of Variance*. Nova York: John Wiley & Sons; Londres: Chapman & Hall, 1959.

[58] Wachsmuth, A.; Wilkinson, L.; Dallal, G. E. «Galton's bend: a previously undiscovered nonlinearity in Galton's family stature regression data». *Amer. Statist.*, 57 (3) (2003), 190–192.

[59] Waller, N. G. «The geometry of enhancement in multiple regression». *Psychometrika*, 76 (4) (2011), 634–649.

[60] Wilks, S. S. «Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution». *Ann. Math. Statistics*, 17 (1946), 257–281.

Departament d'Estadística
Facultat de Biologia
Universitat de Barcelona
Av. Diagonal, 643
08028 Barcelona
ccuadras@ub.edu