El Condicionament en les Anàlisis Factorials Descriptives: L'Anàlisi Parcial Interna i Simultània

Ramon Nonell i Torrent

TESI DOCTORAL

dirigida pel Dr. Tomàs Aluja i Banet
Dep. Estadística i Inv. Operativa
Facultat d'Informàtica de Barcelona
Universitat Politècnica de Catalunya

Barcelona, 1992

El Condicionament en les Anàlisis Factorials Descriptives: L'Anàlisi Parcial Interna i Simultània

Ramon Nonell i Torrent

TESI DOCTORAL

dirigida pel Dr. Tomàs Aluja i Banet
Dep. Estadística i Inv. Operativa
Facultat d'Informàtica de Barcelona
Universitat Politècnica de Catalunya

Barcelona, 1992

a les persones amigues i amades

"en la destreza de las armas no hay ángulo recto, ni lo pudo haber, porque ángulo recto es caer una línea recta perpendicular sobre otra y en el hombre, estando inhiesto, no cae línea alguna que forme los ángulos obtusos ni agudos, de que se compone el recto".

Don Miguel Pérez de Miranda, en *Principio de los cinco sujetos de que se compone la filosofía y matemática de las armas y práctica especulativa*, destruint la base de l'*ángulo recto*, pilar incommovible de l'esgrima filosòfica.

Joan Perucho, a *Històries apòcrifes*.

Agraïments

En primer lloc, agraeixo la dedicació del Dr. Tomàs Aluja, i li agraeixo sobretot les bones idees generadores d'aquesta Tesi Doctoral.

Agraeixo també al Dr. Manuel Martí la confiança que m'ha demostrat i els consells que d'ell he rebut al llarg d'aquests anys.

També, a tots aquells de qui he obtingut, en un moment o altre, l'ajuda que necessitava per continuar. Sobretot dono les gràcies, per ordre d'aparició, a la Dra. Pilar Nivela, al Sr. Santi Thió, a la Sra. Roser Rius, a la Dra. Elena Fernández i a la Sra. Karina Gibert.

Aquesta Tesi s'insereix en els estudis descriptius propis de les anàlisis de grans matrius de dades en què es tracten variables (p variables) sobre individus (n individus), amb l'objectiu de trobar i caracteritzar les relacions i oposicions tant entre variables com entre individus mitjançant projeccions, generalment en plans, de les variables i del individus considerats com a punts dels respectius espais $\mathcal{R}^n$ i $\mathcal{R}^p$, aconseguint així les representacions que conserven més informació -o millor encara, les representacions que fan més palesa l'estructura informativa que les dades ens poden arribar a proporcionar.

Els objectius de la Tesi es configuren de la següent manera:

- Presentar l'Anàlisi Parcial Interna i Simultània com a aportació al millorament d'aquest estudis descriptius de dades.

  Aquesta anàlisi pretén una depuració de les relacions (bàsicament entre les variables) a través del control de la influència que les pròpies variables de l'anàlisi poden exercir-se unes damunt les altres. Se situa, doncs, en el camp dels estudis del condicionament. És una anàlisi que comparada amb les anàlisis habituals pot

ajudar-nos en una estructuració més fidedigna de les dades -i tal vegada, en una recerca acurada de les variables més adequades que caldria mantenir en eventuals estudis posteriors de les dades.

L'Anàlisi Parcial Interna i Simultània es presenta al capítol setè de la Tesi; es fa tant per a variables numèriques (Anàlisi en Components Principals Parcial Interna i Simultània) com per a categòriques (Anàlisi de Correspondències Múltiples Parcial Interna i Simultània). Es presenten i demostren els resultats, remarcant els punts de partença que han permès el desenvolupament d'aquesta anàlisi (l'estudi de les correlacions parcials en el cas de variables numèriques, l'Anàlisi Local[Alu 85] damunt d'un graf en el cas de les categòriques). En el capítol vuitè i últim es mostra una aplicació, que no pretén pas de ser exhaustiva, de l'Anàlisi Parcial Interna i Simultània sobre unes dades que allà es detallen.

- Presentar algunes anàlisis de diversos autors que ens proporcionin una visió de l'estat actual de les anàlisis descriptives de dades -de les direccions en què es desenvolupen aquestes anàlisis. És, és clar, una tria de metodologies, però, creiem, força completa i adequada de cara a situar en els seu lloc l'Anàlisi Parcial Interna i Simultània. En aquest sentit, tot el capítol sisè està dedicat a l'Anàlisi Local -anàlisi que tracta les relacions i oposicions existents en unes dades havent eliminat la influència que factors externs hi poden exercir. Les altres anàlisis es presenten al capítol cinquè.

  Òbviament, en un punt o altre havíem d'aturar-nos; no parlem de les anàlisis més recents, per exemple amb metodologia de Panel o d'Empelt[Mor 91], que tanmateix ens ajudarien en aquesta visió de la zona d'estudis descriptius de dades en què situem la Tesi.

- Presentar una formalització uniforme de les anàlisis descriptives de dades. És un punt important, també, d'aquesta Tesi.

  Presentar amb detall, i de manera formal i general, l'anàlisi d'una matriu de dades X en la recerca dels factors principals. Seguint A. Carlier[Car 90] es desenvolupa la que anomenem Anàlisi Factorial Descriptiva d'una matriu X_{nxp} amb mètrica M_{pxp} en l'espai dels idividus i amb mètrica de pesos per a aquests individus D_{nxn}. Es fa al primer capítol de la Tesi.

  En el segon capítol es demostra que l'Anàlisi Canònica Generalitzada no és sinó una Anàlisi Factorial Descriptiva, la qual cosa ens direcciona cap a la presentació uniforme de totes les anàlisis com a diferents casos particulars de l'Anàlisi Factorial Descriptiva: l'Anàlisi Canònica i l'Anàlisi en Components Principals, que així es presenten al capítol tercer; l'Anàlisi de Correspondències Múltiples que igualment es presenta com a cas particular de l'Anàlisi Factorial Descriptiva en el capítol quart, i l'Anàlisi de Correspondències Simples que, en el mateix capítol, es presenta com a cas particular de l'Anàlisi Canònica i per tant, com a una Anàlisi Factorial Descriptiva.

  Òbviament, les anàlisis-extensió que presentem en el capítol cinquè, com també l'Anàlisi Local i la pròpia Anàlisi Parcial Interna i Simultània, es formalitzen adequadament com a anàlisis factorials descriptives.

En definitiva, la Tesi té l'objectiu d'aportar l'Anàlisi Pracial Interna i Simultània a l'estudi del condicionament en les anàlisis descriptives de dades alhora que presenta aquestes anàlisis, les clàssiques i algunes de recents, d'una manera general i formal -a partir, és clar, dels estudis de diferents autors-, donant-ne una visió uniforme i rigorosa.

Índex

Capítol 1

Anàlisi Factorial Descriptiva

1.1 Elements

Considerem la matriu X de les n observacions de p variables sobre n individus:

$$X = (x_{ij})$$

amb x_{ij} l'observació de la j-èsima variable, $j \in J = \{1, \ldots, p\}$, sobre l'individu i-èsim, $i \in I = \{1, \ldots, n\}$.

Cada individu s'identifica amb un vector

$$x_i = (x_{i1}, \ldots, x_{ip})^t$$

de l'espai $\mathcal{R}^p$, i cada variable amb un vector

$$x^j = (x_{1j}, \ldots, x_{nj})^t$$

de l'espai $\mathcal{R}^n$. Aleshores, anomenem $\mathcal{R}^p$ espai dels individus i $\mathcal{R}^n$ espai de les variables.

Sigui M una mètrica euclídea (M és doncs una matriu simètrica definida i positiva) a l'espai $\mathcal{R}^p$, i D una mètrica diagonal a l'espai $\mathcal{R}^n$ que pondera els individus segons uns pesos p_i, $i \in I, p_i > 0, \sum_{i \in I} p_i = 1$.

Si les variables x^j estan centrades (és a dir: $\sum_{i \in I} p_i x_{ij} = 0, \forall j \in J$), la covariància entre dues d'elles s'obté:

$$cov(x^j, x^{j'}) = (x^j)^t D x^{j'}.$$

Aleshores, X^tDX és la matriu de covariàncies.

En aquest cas en què les variables estan centrades, fixem-nos a més que el D-angle $\theta_{jj'}$ entre x^j i $x^{j'}$ compleix:

$$cos\theta_{jj'} = \frac{x^{j^t}Dx^{j'}}{\sqrt{x^{j^t}Dx^j x^{j'^t}Dx^{j'}}},$$

expressió que no és sinó la correlació entre les variables x^j i $x^{j'}$ caracteritzades pels seus valors sobre els n individus.

La norma associada a la mètrica M la denotarem $\| \ \|_M$.

La relació que lliga els espais que intervenen en la nostra anàlisi és:

$$\begin{array}{ccc} \mathcal{R}^p & \xleftarrow{X^t} & \mathcal{R}^{n*} \\ \downarrow M & & \uparrow D \\ \mathcal{R}^{p*} & \xrightarrow{X} & \mathcal{R}^n \end{array}$$

entenent les mètriques clàssicament com a aplicacions de l'espai sobre el seu dual[Cai 76].

Si bé les mètriques podrien fer-se implícites en les dades si aquestes les transformés-sim convenientment[Car 90], mantenim explícitament les matrius M i D i parlem del triplet (X, D, M).

1.2 Mètrica en l'espai de les matrius

Denotem per $\mathcal{L}_{n,p}$ l'espai de les matrius amb n files i p columnes, i hi definim la norma Hilbert Schmidt segons M i D d'una matriu X com:

$$\| X \|_{DM} = [tr(X^tDXM)]^{1/2},$$

on tr és la funció traça.

En un context multivariant, $\| X \|^2_{DM}$ es considera una mesura de la variabilitat de les variables[Car 90]. Ho veiem més clarament amb les següents propietats:

- $\| X \|^2_{DM} = \sum_{i \in I} p_i \ \| x_i \|^2_M$ és la inèrcia respecte a l'origen del conjunt dels individus ponderats $\{(x_i, p_i), i \in I\}$.

- Si les variables x^j estan centrades i M és també una mètrica diagonal:

$$M = diag(m_j, j \in J),$$

aleshores:

$$\| X \|^2_{DM} = \sum_{j \in J} m_j var(x^j).$$

En aquest cas, clarament, doncs, la norma Hilbert Schmidt ens dóna la variabilitat total de les variables.

1.3 Operadors de projecció ortogonal

Indiquem P_W l'operador de projecció M-ortogonal sobre un subespai W de l'espai del individus $(\mathcal{R}^n, M)$. P_W és, doncs, una aplicació idempotent i M-simètrica[Cai 76]. $I - P_W$ és la projecció sobre $W^\perp$.

L'operador P_W aplicat a la matriu X ens dóna:

$$X = \begin{pmatrix} \vdots \\ x_i^t \\ \vdots \end{pmatrix} \longrightarrow XP_W^t = \begin{pmatrix} \vdots \\ (P_W x_i)^t \\ \vdots \end{pmatrix}$$

que transforma cada fila x_i^t de la matriu X en $(P_W x_i)^t$, i que per tant transforma la matriu X en la seva projecció sobre l'espai de les matrius $Z \in \mathcal{L}_{n,p}$ tals que les seves files z_i^t engendren subespais continguts en W.

Això ens permet de tenir:

$$\| X \|^2_{DM} = \| XP_W^t \|^2_{DM} + \| X - XP_W^t \|^2_{DM}.$$

1.4 Descomposició en valors singulars

El conegut Teorema de descomposició en valors singulars[Car 90], ens dóna la descomposició següent:

Si $X \in \mathcal{L}_{n,p}$ té $rang = r$, aleshores

$$X = V\Lambda U^t,$$

amb $V^tDV = U^tMU = I_{rxr}$, i essent U, V i Λ les següents matrius:

- U és la matriu pxr de columnes u_s, s=1,...,r, formant una base M-ortonormal de les files de la matriu X a l'espai dels individus $(\mathcal{R}^p, M)$.

- V és la matriu nxr de columnes v_s, s=1, ...,r, formant una base D-ortonormal de les columnes de la matriu X a l'espai de les variables $(\mathcal{R}^n, D)$.

- Λ és la matriu diagonal rxr dels valor singulars $\lambda_s(\lambda_s > 0, s = 1, \ldots, r)$ -que són les arrels quadrades dels valors propis de la matriu MX^tDX, o bé de la matriu $DXMX^t$, doncs els valors propis són els mateixos[1]-, ordenats en ordre decreixent.

Aquesta descomposició la denotem **d.v.s. del triplet** (X, M, D).[2]

Les propietats d'aquesta descomposició són prou conegudes, i només les recordem (veure [Car 90]):

- $X^tDXM = \sum_{s=1}^{r} \lambda_s^2 u_s u_s^t M$.

- $XMX^tD = \sum_{s=1}^{r} \lambda_s^2 v_s v_s^t D$.

  Notem que $v_s v_s^t D$ i $u_s u_s^t M$ són els operadors projecció D i M-ortogonals respectivament, sobre els espais $< v_s >$ i $< u_s >$ engendrats el primer per v_s, i l'altre per u_s[Car 90].

[1] M, per ser mètrica, sempre es pot descomposar en dues matrius idèntiques que escrivim $M^{1/2}$: $M = M^{1/2}M^{1/2}$. Les demostracions de molts dels resultats enunciats en aquests capítol usen d'aquesta descomposició. A més, aleshores podem dir que aquests valors propis també ho són de $M^{1/2}X^tDXM^{1/2}$ i de $D^{1/2}XMX^tD^{1/2}$.

[2] La descomposició en valors singulars ens assegura la unicitat dels vectors u_s i v_s, llevat de reflexions en el cas dels valors singulars de multiplicitat 1, i de "rotacions" en el cas de multiplicitat superior.

- $X = \sum_{s=1}^{r} \lambda_s v_s u_s^t$.

  Que és l'expressió de la matriu X en el sistema ortonormal (segons la mètrica Hilbert Schmidt) de matrius $(v_s u_s^t)_{s=1,\ldots,r}$. Aleshores:

- $\| X \|_{DM}^2 = \sum_{s=1}^{r} \lambda_s^2$.

  Recordant la interpretació de la norma Hilbert Schmidt, resulta que la variabilitat total de les variables ve determinada pels valors singulars.

Finalment tenim les fórmules de transició o de dualitat entre les matrius V i U:

$$\begin{aligned} v_s &= \tfrac{1}{\lambda_s} X M u_s \\ u_s &= \tfrac{1}{\lambda_s} X^t D v_s, \end{aligned}$$

que de fet són l'aplicació, ponderada per λ_s, de l'esquema

$$\begin{array}{ccc} \mathcal{R}^p & \xleftarrow{X^t} & \mathcal{R}^{n*} \\ \downarrow M & & \uparrow D \\ \mathcal{R}^{p*} & \xrightarrow{X} & \mathcal{R}^n \end{array}$$

i que ens permeten d'obtenir els vectors u_s a partir dels vectors v_s, i a l'inrevés. En forma matricial:

$$\begin{aligned} V &= XMU\Lambda^{-1} \\ U &= X^t DV\Lambda^{-1}. \end{aligned}$$

1.5 Anàlisi Factorial Descriptiva: Aproximació per projecció

Siguin $< V_k >$ i $< U_k >$ els subespais engendrats per les k primeres columnes de V i U respectivament, i Λ_k la matriu diagonal dels k primers valors singulars, $k \leq r$.

Indiquem Q_k la projecció D-ortogonal sobre $< V_k >$ i P_k la projecció M-ortogonal sobre $< U_k >$.

Les matrius formades per les k primeres columnes de V i U les indiquem, simplement, V_k i U_k respectivament.

Aleshores:

$$Q_k = V_k V_k^t D$$
$$P_k = U_k U_k^t M.$$

I resulta[Car 90]:

$$V_k \Lambda_k U_k^t = X P_k^t = Q_k X.$$

Doncs bé, si Q_T i P_W són els operadors projecció sobre T i W, subespais qualsevols respectivament dels espais $\mathcal{R}^n$ i $\mathcal{R}^p$ de dimensió k, es compleix:

$$\| Q_T X \|^2_{DM} = tr(Q_T X M X^t D) \leq \| Q_k X \|^2_{DM} = \sum_{s=1}^{k} \lambda_s^2$$
$$\| X P_W^t \|^2_{DM} = tr(P_W X^t D X M) \leq \| X P_k^t \|^2_{DM} = \sum_{s=1}^{k} \lambda_s^2.$$

Anem a interpretar aquests resultats essencials:

- $\| X P_W^t \|^2_{DM} = \sum_{i \in I} p_i \| P_W x_i \|^2_M$, i per tant el subespai $< U_k >$ és òptim en el sentit que maximitza la inèrcia de les projeccions sobre subespais de dimensió k del núvol de punts $(x_i, p_i), i \in I$.

  Fixem-nos que si W està engendrat per un sol vector y de l'espai $\mathcal{R}^p$ dels individus, aleshores:

$$\| X P_W^t \|^2_{DM} = \sum_{i \in I} p_i \| y y^t M x_i \|^2_M,$$

  i aplicant la definició de norma i les propietats dels operadors projecció:

$$\| X P_W^t \|^2_{DM} = \sum_{i \in I} p_i y^t M x_i x_i^t M y = y^t M X^t D X M y.$$

  Per tant, podem dir que hem anat a la recerca del vector $y \in \mathcal{R}^p$ que maximitza l'expressió:

$$y^t M X^t D X M y \quad \textit{sota la restricció} \quad y^t M y = 1. \tag{1.1}$$

  Aquest vector és, doncs, la primera columna de la matriu U.

  Utilitzant multiplicadors de Lagrange es veu que la solució al problema (1.1) és el vector y_1 tal que My_1 és vector propi de la matriu MX^tDX associat al seu valor propi més gran h_1, valor màxim de l'expressió $y_1^t M X^t D X M y_1 = h_1$.

Efectivament, la primera columna de la matriu U, u_1, és tal que Mu_1 és vector propi de la matriu MX^tDX associat al seu valor propi més gran, que és λ_1^2, essent λ_1 el primer valor singular de la descomposició en valors singulars de la matriu X.

Fixem-nos que si les variables x^j estan centrades i $M = I$, estem buscant els valors propis i els vectors propis de la usual matriu de covariàncies X^tDX.

Podríem anar buscant els vectors que successivament anirien fent màxima l'expressió y^tMX^tDXMy sota la restricció $y^tMy = 1$ i que serien tals que My fossin vectors propis associats als successius (en ordre decreixent) valors propis més grans de la matriu MX^tDX.

Coincidirien, aquests vectors, amb les successives columnes de la matriu U (llevat, és clar, de reflexions i en el seu cas de rotacions degudes a multiplicitats superiors a 1 dels diferents valors propis).

La presentació de l'Anàlisi Factorial Descriptiva com, primerament, la recerca d'aquell vector y tal que maximitza y^tMX^tDXMy sota la restricció $y^tMy = 1$, és la més clàssica, però ens ha semblat interessant d'introduir la norma de Hilbert Schmidt en l'espai de les matrius i de presentar l'Anàlisi com la maximització de la norma Hilbert Schmidt per a les projeccions de la matriu X.

- Si les variables x^j estan centrades tenim la igualtat:

$$\| X - XP_W^t \|_{DM}^2 = \frac{1}{2} \sum_{i \in I} \sum_{i' \in I} p_i p_{i'} (\| x_i - x_{i'} \|_M^2 - \| P_W x_i - P_W x_{i'} \|_M^2)$$

que pot interpretar-se com una mesura de la distorsió dels individus al projectar-se sobre W.

Aleshores, és clar que $< U_k >$ és el subespai de dimensió K que minimitza la distorsió dels individus en ser projectats.

- Encara, si M també és diagonal, amb les x^j centrades, obtenim:

$$\| Q_T X \|^2_{DM} = \sum_{j \in J} m_j var(Q_T x^j),$$

i per tant $< V_k >$ és òptim en el sentit d'explicar la major part de variabilitat de les variables.

En definitiva, tot el tractament realitzat sobre el triplet (X, M, D), ens permet d'aproximar tant els individus x_i com les variables x^j mitjançant les projeccions als espais $< U_k >$ i $< V_k >$ respectivament - espais que són els millors en el sentit dels resultats abans esmentats.

Recordem que $< U_k >$ està engendrat pels vectors $(u_{s1}, \ldots, u_{sp}), s = 1, \ldots, k$, i $< V_k >$ pels vectors $(v_{s1}, \ldots, v_{sn}), s = 1, \ldots, k$, i que les projeccions tenen les expressions:

$$P_k x_i = \sum_{s=1}^{k} \lambda_s v_{si} u_s$$
$$Q_k x^j = \sum_{s=1}^{k} \lambda_s u_{sj} v_s.$$

Els eixos u_s i v_s els anomenem **eixos o factors principals**, els primers de l'espai dels individus i els segons de l'espai de les variables, i, òbviament, $\lambda_s v_{si}$ i $\lambda_s u_{sj}$, coordenades de les projeccions de l'individu i i de la variable j sobre els factors principals, les anomenem **coordenades principals.**

Fixem-nos que tenim d'altres expressions de les coordenades de les projeccions:

- $\lambda_s v_{si} = u_s^t M x_i$.

- $\lambda_s u_{sj} = v_s^t D x^j$.

Tota l'Anàlisi fins aquí obtinguda l'anomenem **Anàlisi Factorial Descriptiva (AFD) del triplet (X, M, D).**

L'Anàlisi Factorial Descriptiva del triplet (X, M, D), doncs, ens porta a trobar, a través de la descomposició en valors singulars els "millors subespais", tant en l'espai $(\mathcal{R}^p, M)$ dels individus, com en l'espai $(\mathcal{R}^n, D)$ de les variables, en el sentit de conservar

millor la informació i fer més palesa l'estructura de les dades en ser projectades sobre subespais de dimensió reduïda.

Capítol 2

Anàlisi Canònica Generalitzada

Tractem primerament, l'Anàlisi Canònica Generalitzada (ACG) com una Anàlisi Factorial Descriptiva aplicada a un cert triplet, i en segon lloc veiem la seva presentació usual geomètrica[Vol 85] com a recerca dels "millors" representants (factors comuns) d'un conjunt de vectors-variable.

2.1 Elements

Considerem la matriu X de dimensió nxp dels valors que sobre n individus prenen p variables que estan agrupades en m paquets. Denotem $X_k, k = 1, \ldots, m$, els diferents paquets.

Cada paquet X_k conté els valors de p_k variables. Aleshores, $\sum_{k=1}^{m} p_k = p$. Escrivim X_{jk} la variable j-èsima del paquet X_k, $j = 1, \ldots, p_k$. Identifiquem la variable X_{jk}, és clar, amb els seus valors sobre els n individus: $x_{ijk}, i = 1, \ldots, n$ (columna de la matriu X).

Així doncs, la matriu X és la juxtaposició de les m matrius X_k:

$$X = \begin{pmatrix} & X_1 & & & X_m & \\ & \vdots & & & \vdots & \\ \cdots & x_{ik1} & \cdots & \cdots & x_{ikm} & \cdots \\ & \vdots & & & \vdots & \\ & \underbrace{\qquad\qquad}_{p_1 \text{ variables}} & & & \underbrace{\qquad\qquad}_{p_m \text{ variables}} & \end{pmatrix}$$

Suposem que els p_k vectors que representen les p_k variables del paquet X_k són vectors linealment independents de l'espai $\mathcal{R}^n$, amb la qual cosa hi definim m subespais vectorials engendrats per les variables de cada paquet i de dimensions p_k:

$$E_k = < X_{1k}, \ldots, X_{p_k k} > .$$

El rang de cada matriu X_k, és, doncs, p_k.

La matriu X ens dóna els valors de totes les variables sobre els individus. Els individus s'identifiquen amb els vectors de l'espai $\mathcal{R}^p$:

$$(x_{i11}, \ldots, x_{ip_1 1}, \ldots, x_{i1m}, \ldots, x_{ip_m m})^t.$$

Tenim, doncs, n vectors-individu x_i.

Considerem en $\mathcal{R}^p$ la mètrica M definida:

$$M = \begin{pmatrix} X_1^t X_1 & & & \\ & \ddots & & \vec{0} \\ & & X_k^t X_k & \\ \vec{0} & & & \ddots \\ & & & X_m^t X_m \end{pmatrix}$$

Es demostra que M^{-1} també és una matriu bloc-diagonal de terme k-èsim $(X_k^t X_k)^{-1}$.

2.2 L'*ACG* com a una Anàlisi Factorial Descriptiva

Doncs bé, l'Anàlisi Canònica Generalitzada consisteix en l'Anàlisi Factorial Descriptiva del triplet (X, M^{-1}, I).

Des del punt de vista més clàssic de l'AFD, es tracta de buscar com a factors principals del núvol d' individus definit per les files de la matriu X, els vectors $y_s \in \mathcal{R}^p$ tals que ens donen els més grans valors de l'expressió:

$$y_s^t M^{-1} X^t X M^{-1} y_s,$$

sota la restricció $y_s^t M^{-1} y_s = 1$.

Sabem que els vectors y_s que solucionen el problema són aquells tals que $M^{-1}y_s$ són vectors propis associats als valors propis, h_s, més grans de la matriu $M^{-1}X^tX$:

$$\begin{aligned} M^{-1}y_s &= z_s \\ M^{-1}X^tXz_s &= h_s z_s. \end{aligned}$$

2.3 Interpretacions de l'ACG

Anem a interpretar l'Anàlisi Factorial Descriptiva del triplet (X, M^{-1}, I) -la qual cosa acaba donant la visió més usual i geomètrica de l'ACG.

Tenim els factors principals y_s de l'Anàlisi del triplet (X, M^{-1}, I). Per tant, podem considerar les projeccions dels individus en l'espai $(\mathcal{R}^p, M^{-1})$ sobre els subespais engendrats pels k primers factors principals, $y_1, \ldots, y_k, k \leq rang X$.

Les interpretacions són les de l'AFD presentat al capítol anterior, és clar. Podem, doncs, representar els individus amb les seves coordenades principals.

L'individu i té per coordenada principal sobre l'eix y_s:

$$F_s(i) = y_s^t M^{-1} x_i.$$

$(F_s(i))^2$ és la M^{-1}-norma al quadrat de la projecció sobre y_s de l'individu i.

I la projecció de l'individu i sobre el subespai engendrat pels k primers factors principals és:

$$P_k x_i = \sum_{s=1}^{k} (y_s^t M^{-1} x_i) y_s = \sum_{s=1}^{k} F_s(i) y_s.$$

La norma Hilbert Schmidt al quadrat de la matriu de les projeccions és:

$$\| XP_k^t \|^2_{DM^{-1}} = \sum_{i=1}^{n} \| P_k x_i \|^2_{M^{-1}} = \sum_{i=1}^{n} \sum_{s=1}^{k} (F_s(i))^2.$$

Considerem ara el vector u_1 de l'espai de les variables $\mathcal{R}^n$:

$$u_1 = \frac{1}{\sqrt{h_1}} X z_1 = \frac{1}{\sqrt{h_1}} X M^{-1} y_1.$$

Fixem-nos que u_1 s'obté simplement d'aplicar la fórmula de transició descrita al capítol anterior. Per tant, és el factor principal de l'Anàlisi Factorial Descriptiva de l'espai $(\mathcal{R}^n, I)$ de les variables.

Anem a veure les propietats essencials d'aquest vector u_1[Car 90]. Els resultats provenen del fet que estem a l'Anàlisi Factorial Descriptiva de les variables definides per la matriu X a l'espai $(\mathcal{R}^n, I)$.

- Sigui Π_k l'operador projecció canònica sobre E_k a l'espai $\mathcal{R}^n$. Resulta:

$$\Pi_k = X_k (X_k^t X_k)^{-1} X_k^t.$$

- Sigui $\Pi = \sum_{k=1}^{m} \Pi_k$. Resulta:

$$\Pi = X M^{-1} X^t,$$

  i $rang\Pi = rang X$, per tant.

- Tenim que u_1 és el vector propi associat al més gran valor propi h_1 de la matriu $XM^{-1}X^t$, i per tant és solució del problema:

  Trobar el vector u tal que $u^t \Pi u$ sigui màxim sota la restricció $u^t u = 1$, essent $u_1^t \Pi u_1 = h_1$.

- A més, considerem les projeccions $Z_{1k} = \frac{\Pi_k u_1}{\|\Pi_k u_1\|_c}$, on $\| \ \|_c$ és la norma canònica. Aleshores, resulta que els vectors $Z_{1k} \in E_k$ són tals que, unitaris per la norma

canònica, fan màxima la suma:

$$\sum_{k=1}^{m} cos^2(u_1, Z_{1k}).$$

- Si considerem, per exemple, els dos primers factors principals de les variables:

$$u_1 = \frac{1}{\sqrt{h_1}} X z_1$$
$$u_2 = \frac{1}{\sqrt{h_2}} X z_2$$

segons és usual a l'AFD, podem projectar les variables X_{jk} sobre el subespai engendrat per u_1 i u_2:

$$Q_{<u_1,u_2>} X_{jk} = (X_{jk}^t u_1)u_1 + (X_{jk}^t u_2)u_2.$$

- I ara, a més, essent Z_{1k} la projecció unitària d'u_1 sobre E_k i Z_{2k} la d'u_2, cada paquet pot representar-se sobre aquest pla principal segons la combinació:

$$(Z_{1k}^t u_1)u_1 + (Z_{2k}^t u_2)u_2.$$

2.4 L'*ACG* des del punt de vista geomètric clàssic

Interpretem els vectors u_s i Z_{sk} com habitualment es fa a l'ACG.

L'Anàlisi va a la recerca dels vectors "representants" de cada paquet de variables "més a prop" dels factors principals de les variables definides per la matriu X. És a dir:

Busquem el vector $u \in \mathcal{R}^n$ i els vectors $Z_k \in E_k$ tals que $\| u \|_c = \| Z_k \|_c = 1$ i $\sum_{k=1}^{m} cos^2(u, Z_k)$ sigui màxim.

Es demostra[Vol 85] que si $(u, Z_1, \ldots, Z_k)$ és solució del problema plantejat, Z_k és forçosament projecció canònica unitària del vector u sobre E_k.

Com que

$$\sum_{k=1}^{m} cos^2(u, Z_k) = u^t \Pi u,$$

es tracta només de trobar el vector u, i seqüencialment els que van fent successivament màxima l'expressió $u^t \Pi u$ sota la restricció $u^t u = 1$, i de considerar les seves projeccions sobre els subespais E_k.

L'equivalència entre els dos punts de vista és clara, si bé, fixem-nos que l'ACG presentat com una AFD sobre (X, M^{-1}, I), afegeix a les projeccions de les variables sobre els eixos principals u_s, les projeccions dels individus sobre els eixos y_s.

Les fórmules de transició són, és clar:

$$u_s = \frac{1}{\sqrt{h_s}} X M^{-1} y_s$$
$$y_s = \frac{1}{\sqrt{h_s}} X^t u_s.$$

En els capítols següents presentem diferents anàlisis habituals com a casos particulars de l'ACG.

Capítol 3

Anàlisi Canònica i Anàlisi en Components Principals

Anem a desenvolupar l'Anàlisi Canònica (AC) i a veure que és un cas particular de l'ACG. En el següent capítol veiem com l'AC ens porta a l'Anàlisi de Correspondències Simples.

Igualment desenvolupem l'Anàlisi en Components Principals (ACP) formulant-lo també com a cas particular de l'ACG.

3.1 Anàlisi Canònica

3.1.1 Elements

Considerem una població d'n individus, i els valors sobre ells observats de p variables $X_1, \ldots, X_p$ i de q variables $Y_1, \ldots, Y_q$.

Identifiquem les variables amb els valors que prenen sobre els n individus:

$$X_j = (x_{1j}, \ldots, x_{nj})^t, j = 1, \ldots, p$$
$$Y_k = (y_{1k}, \ldots, y_{nk})^t, k = 1, \ldots, q.$$

Sigui X la matriu nxp formada pels valors:

$$(x_{ij}) \begin{matrix} j = 1, \ldots, p \\ i = 1, \ldots, n \end{matrix}$$

i sigui Y la matriu nxq formada pels valors:

$$(y_{ik}) \begin{matrix} k = 1, \ldots, q \\ i = 1, \ldots, n. \end{matrix}$$

Suposem que els p vectors X_j són vectors linealment independents de l'espai $\mathcal{R}^n$, i que igualment ho són els q vectors Y_k.

Podem representar les dades:

$$\begin{matrix} n & \updownarrow & (X & Y) \\ & & \leftrightarrow & \leftrightarrow \\ & & p & q \end{matrix}$$

E_X és l'espai engendrat pels p vectors X_j, i E_Y l'engendrat pels q vectors Y_k.

3.1.2 Plantejament de l'*AC*. Solució geomètrica clàssica i solució segons l'*ACG*

L'*AC* planteja el problema de trobar dos vectors, $Z \in E_X$ i $T \in E_Y$ tals que $cos(Z,T)$ sigui màxim. És a dir: trobar les dues direccions dels espais E_X i E_Y que estan més a prop.

Recordem a més, que si les variables X_j i Y_k són centrades, també ho són qualsevols $Z \in E_X$ i $T \in E_Y$, i que aleshores $cos(Z,T)$=coeficient de correlació $\rho(Z,T)$. És a dir, en aquest cas busquem els vectors representants dels dos grups de variables que estiguin més correlacionats.

Com que és obvi que Z i T els podem agafar unitaris per la norma canònica, el problema el formulem:

Trobar $Z \in E_X$ i $T \in E_Y$ tals que $\| Z \|_c = \| T \|_c = 1$ i que $cos(Z,T)$ sigui màxim.

Anem a veure[Vol 85] com estem davant d'un cas particular de l'ACG.

Sigui Π_X l'operador projecció ortogonal sobre E_X i Π_Y l'operador projecció ortogonal sobre E_Y.

Sigui $\Pi = \Pi_X + \Pi_Y$. A l'ACG volem trobar el vector $u_1 \in \mathcal{R}^n$ tal que $u_1^t \Pi u_1$ sigui màxim sota la restricció $u_1^t u_1 = 1$.

Vegem que les solucions del problema plantejat a l'AC no són més que les de l'ACG. És a dir: Z és la projecció unitària sobre E_X i T sobre E_Y del vector u_1.

Tenim u_1 que és vector propi de Π associat al seu més gran valor propi h_1. Si agafem les projeccions unitàries:

$$Z_X = \frac{\Pi_X u_1}{\| \Pi_X u_1 \|_c}$$

i:

$$Z_Y = \frac{\Pi_Y u_1}{\| \Pi_Y u_1 \|_c},$$

que són les solucions de l'ACG, es demostra que Z_X i Z_Y són, també, els vectors unitaris d'E_X i E_Y, respectivament, més propers; és a dir, formant angle mínim, amb $cos(Z_X, Z_Y) = h_1 - 1$. A més, u_1 és el vector unitari en la direcció de la seva bisectriu, Z_X és la projeció unitària de Z_Y sobre E_X, i recíprocament, Z_Y és la projecció unitària sobre E_Y de Z_X.[1]

Si successivament considerem les solucions de l'ACG, (u_s, Z_{sX}, Z_{sY}), es demostra [Vol 85] que (Z_{sX}, Z_{sY}) són els vectors successivament més propers entre ells -és a dir, formant angle mínim- tals que, si $m \neq l$:

$$Z_{lX} \perp Z_{mX}$$
$$Z_{lY} \perp Z_{mY}.$$

En definitiva, vist l'AC com a cas particular de l'ACG, l'AC és una AFD sobre el triplet (Δ, M^{-1}, I), essent Δ la matriu $nx(p+q)$ de juxtaposició de les matrius X i Y,

[1] Vegeu en Volle[Vol 85] la discusió sobre el cas $h_1 = 1$; és a dir, quan Z_X i Z_Y són ortogonals.

M^{-1} la mètrica en $\mathcal{R}^{p+q}$

$$M^{-1} = \begin{pmatrix} (X^tX)^{-1} & \vec{0} \\ \vec{0} & (Y^tY)^{-1} \end{pmatrix},$$

i I la matriu identitat en $\mathcal{R}^n$.

3.2 L'*ACP*: cas particular de l'*ACG*

3.2.1 Elements

Sigui una matriu A de dimensions nxp de manera que cada columna la formen les observacions sobre n individus d'una certa variable. Considerem la matriu de les variables centrades i reduïdes; és a dir:

$$X = (x_{ij}) \begin{matrix} j = 1, \ldots, p \\ i = 1, \ldots, n \end{matrix}$$

amb

$$x_{ij} = \frac{a_{ij} - \frac{\sum_{i=1}^{n} a_{ij}}{n}}{\sqrt{\sum_{i=1}^{n} \frac{(a_{ij} - \sum_{i=1}^{n} \frac{a_{ij}}{n})^2}{n}}}.$$

Al tractar els termes x_{ij} estem considerant les correlacions entre les variables de l'anàlisi; si tractéssim

$$y_{ij} = a_{ij} - \frac{\sum_{i=1}^{n} a_{ij}}{n},$$

estaríem considerant les covariàncies. Les dues anàlisis no són equivalents. La que presentem és més habitual, i és l'adequada en estudis amb variables de variablilitat diferent o, per exemple, variables mesurades en unitats diferents[Le2 85].

Representem les variables per les columnes x^j i els individus per les files trasposades x_i, com és ususal en els estudis que presentem.

La metodologia de l'*ACP* de la matriu X és ben coneguda. Ens remetem a diferents textos i principalment a Volle[Vol 85], Mardia[Mar 79], Lebart[Le2 85],

Cuadras[Cua 81], Saporta[Sap 78], i sobretot per a les interpretacions -és a dir, per a les lectures dels resultats obtinguts.

Considerem la matriu Z de valors

$$z_{ij} = \frac{1}{\sqrt{n}} x_{ij},$$

aconseguint així que les variables z^j (columnes de la matriu Z) representin les originals de la matriu A projectades sobre el subespai de dimensió $n-1$ ortogonal al vector de components totes igual a 1, i normalitzades (per la norma canònica).

Considerem

$$L = X^t X$$

i

$$\frac{1}{n} X X^t = Z Z^t.$$

Fixem-nos que la correlació entre dues variables z^j, $z^{j'}$, que és igual a la correlació entre x^j, $x^{j'}$, s' obté:

$$\rho_{jj'} = \frac{1}{n} l_{jj'},$$

essent $l_{jj'}$ el terme general de la matriu L.

La distància entre dues variables és:

$$\| z^j - z^{j'} \|^2 = 2(1 - \rho_{jj'}).$$

L'anàlisi de les variables s'ha de fer, doncs, en termes de correlacions. Són prou conegudes totes les interpretacions al respecte.

3.2.2 L'*ACP* com a cas particular de l'*ACG*. L'*ACP* com a una *AFD*

L'*ACP* és un cas paricular de l'*ACG* amb $p_k = 1, k = 1, \ldots, m = p$, de manera que en cada paquet k hi ha els valors sobre els n individus de la variable z^k, que és la que

engendra el subespai vectorial E_k. Resulta: $\Pi = ZZ^t$, ja que ara

$$M^{-1} = \begin{pmatrix} z^{1t}z^1 & & & \\ & \ddots & & \vec{0} \\ & & z^{kt}z^k & \\ \vec{0} & & & \\ & & & z^{pt}z^p \end{pmatrix}^{-1} = I_p.$$

Els vectors v que solucionen el problema de l'ACG de maximitzar $v^t\Pi v$ sota la restricció $v^tv = 1$, són els que fan màxima l'expressió

$$\sum_{k=1}^{p} cos^2(v, \phi_k),$$

essent ϕ_k la projecció canònica unitària de v sobre $< x^k >=< z^k >$. De fet, $\phi_k = z^k$, llevat del signe, és clar, i per tant, tenim que la coordenada al quadrat de la projecció de z^k sobre v és

$$(z^kv)^2 = cos^2(z^k, v) = cos^2(v, \phi_k),$$

i per tant

$$\sum_{k=1}^{p} cos^2(v, \phi_k) = \sum_{k=1}^{p} (z^kv)^2.$$

És a dir, estem buscant els vectors v màximament correlacionats amb les variables z^k, i la correlació al quadrat de z^k amb un vector v no és sinó la coordenada al quadrat de la seva projecció sobre v.

A l'apartat següent veurem com tots aquests resultats són els usuals de l'ACP clàssicament presentada, però de moment, dins del nostre esquema general, diem que l'ACP és l'AFD del triplet:[2]

$$(Z, I_p, I_n).$$

[2]Tal vegada, hem de tenir en compte que hi ha triplets l'anàlisi dels quals és equivalent a l'anàlisi aquí presentada:

- $(X = (x_{ij}), M = I_p, D_{nxn} = diag(1/n))$.
- $(Y = (y_{ij}), M = S^{-2}, D_{nxn} = diag(1/n))$, essent S^{-2} la matriu diagonal pxp de les inverses de les variàncies empíriques de les variables y^j columnes de la matriu Y.
- $(1/\sqrt{n}Y, M = S^{-2}, D = I_n)$.

3.2.3 Plantejament i solució clàssics del problema de l'ACP

Des del punt de vista clàssic, l'ACP busca els vectors (eixos o factors principals) de l'espai dels individus $\mathcal{R}^p$, unitaris per la norma canònica, de manera que es maximitzi la inèrcia de les projeccions ortogonals dels individus x_i sobre els subespais per ells engendrats. És a dir, volem buscar u unitaris que successivament vagin fent màxima l'expressió:

$$\sum_{i=1}^{n} \| (x_i^t u)u \|_c^2 = \sum_{i=1}^{n} (x_i^t u)^2 = u^t (\sum_{i=1}^{n} x_i x_i^t)u = u^t X^t X u = u^t L u.$$

Ja sabem que el primer factor u_1 és vector propi associat a h_1, el valor propi més gran de la matriu L. Successivament aniríem obtenint els altres factors principals de l'anàlisi dels individus.

La coordenada $F_s(i)$ de l'individu x_i sobre l'eix factorial u_s seria la component i-èsima del vector Xu_s.

Si ara considerem l'anàlisi de les variables z^j i volem trobar els vectors unitaris v de l'espai $\mathcal{R}^n$ que maximitzin

$$\sum_{j=1}^{p} \| (z^j v)v \|_c^2 = \sum_{j=1}^{p} (z^j v)^2 = v^t Z Z^t v,$$

les solucions són els vectors propis v_s associats als més grans valors propis de ZZ^t, que són $\frac{h_s}{n}$.

Resulta:

$$v_s = \frac{1}{\sqrt{h_s}} X u_s.$$

També:

$$u_s = \frac{1}{\sqrt{h_s}} X^t v_s.$$

I encara, considerant l'anàlisi de la matriu de covariàncies en lloc de la de correlacions, són equivalents les anàlisis dels triplets:

- $(Y, I_p, D_{nxn} = diag(1/n))$.
- $(1/\sqrt{n}Y, M = I_p, D = I_n)$.

Són, simplement, les habituals fórmules de transició o dualitat.

La coordenada $G_s(j)$ de la variable z^j sobre l'eix principal v_s és la j-èsima component del vector $Z^t v_s$.

Recordem les interpretacions dels eixos principals com a noves variables, combinació lineal de les inicials, incorrelacionades i de variància màxima: $(F_s(i))_{i=1,\ldots,n}$ compleix:

- $F_s(i) = \sum_{j=1}^{p} x_{ij} u_{sj}$, de manera que u_{sj} és l'escalar que multiplica la variable x^j.
- $\sum_{i=1}^{n} \frac{F_s(i)}{n} = 0$.
- $\sum_{i=1}^{n} \frac{F_s(i)^2}{n} = \frac{h_s}{n}$.
- $\sum_{i=1}^{n} F_s(i) F_{s'}(i) = 0, s \neq s'$.

Entès així el factor s-èsim, la coordenada $G_s(j)$ de la variable z^j no és sinó el coeficient de correlació entre la pròpia variable z^j i l'eix u_s identificat amb $(F_s(i))_{i=1,\ldots,n}$.

La lectura, doncs, de l'anàlisi de les variables en termes de correlacions, és clara.

Evidentment, els factors principals de l'anàlisi de les variables, és a dir, els vectors v_s associats als valors propis t_s de ZZ^t, són els vectors que solucionen l'ACG presentat a l'apartat anterior, i per tant els factors u_s, entesos com a vectors de la forma $(F_s(i))_{i=1,\ldots,n}$, és a dir, identificats amb $\sqrt{h_s} v_s$, són els vectors màximament correlacionats amb les variables z^j.

Recordem que $t_s = \frac{h_s}{n}$, essent h_s els valors propis de la matriu $X^t X$. La matriu X i la matriu Z només es diferencien en el fet d'haver dividit per $\sqrt{n}$ per a fer les variables z^j unitàries, i si apliquéssim les fórmules de transició als vectors v_s solució de l'ACG de l'apartat anterior, obtindríem els mateixos factors u_s obtinguts clàssicament, és clar:

$$u_s = \frac{1}{\sqrt{t_s}} Z^t v_s = \frac{1}{\sqrt{h_s}} \sqrt{n} Z^t v_s = \frac{1}{\sqrt{h_s}} X^t v_s.$$

Fixem-nos que ara els individus-files de la matriu Z, són els individus de l'ACP però dividits per $\sqrt{n}$. La seva projecció sobre u_s és:

$$v_{si}\sqrt{t_s} = \frac{v_{si}\sqrt{h_s}}{\sqrt{n}} = \frac{F_s(i)}{\sqrt{n}},$$

essent $F_s(i)$ la projecció del corresponent individu-fila de la matriu X.

Diguem també, que un individu canònic

$$(0,\ldots,0,\overset{j)}{1},0,\ldots,0)^t$$

de l'espai $\mathcal{R}^p$ té una projecció sobre u_s igual a u_{sj} (j-èsima component del vector u_s). Observem que d'individus canònics n'hi ha p, identificables amb les p variables inicials. Aquestes variables, en la seva anàlisi, tenen projeccions sobre l'eix factorial v_s, $G_s(j)$. Resulta $G_s(j) = u_{sj}\sqrt{\frac{h_s}{n}}$.

Llevat, doncs, del factor $\sqrt{\frac{h_s}{n}}$, la projecció d'una variable j sobre l'eix v_s de l'anàlisi de les variables és igual a la projecció d'un individu canònic $(0,\ldots,0,\overset{j)}{1},0,\ldots,0)^t$ sobre l'eix u_s.

Igualment, tindríem els resultats recíprocs per a les variables canòniques

$$(0,\ldots,0,\overset{i)}{1},0,\ldots,0)^t$$

de l'espai $\mathcal{R}^n$, identificables amb els n individus.

Fins ara hem obtingut la següent relació entre les diferents anàlisis:

$$\begin{array}{ccc} & \mathbf{AFD}(X,M,D) & \\ & \downarrow & \\ & \mathbf{ACG}(X,M^{-1},I) & \\ \swarrow & & \searrow \\ \mathbf{AC}(\Delta,M^{-1},I) & & \mathbf{ACP}(Z,I,I) \end{array}$$

Anem a completar l'esquema amb l'Anàlisi de Correspondències, tant amb la Simple com amb la Múltiple.

Capítol 4

Anàlisi de Correspondències

Igual com hem fet al capítol anterior sobre l'ACP, tampoc no presentem aquí d'una manera exhaustiva l'Anàlisi de Correspondències; només n'exposem la conceptualització bàsica i detallem la seva relació amb l'AC i amb l'ACG, i, en última instància, amb l'AFD.

4.1 L'Anàlisi de Correspondències Simples (ACS)

4.1.1 Elements

Considerem una població d'N efectius repartits segons dues característiques qualitatives, I i J, que cadascuna té n i p modalitats respectivament.

Tenim les dues taules disjuntes completes per a cadascuna de les dues característiques, taules de dimensions, respectivament, Nxn i Nxp, i codificades amb uns i zeros

segons si l'individu o efectiu en qüestió tria o no tria la modalitat corresponent:

$$\begin{array}{c} \qquad\qquad I \qquad\qquad\qquad\qquad\qquad J \\ \left(\begin{array}{ccccccccccc} 0 & 1 & 0 & 0 & \ldots & 1 & 0 & 0 & 0 & \ldots \\ 0 & 0 & 1 & 0 & \ldots & 0 & 1 & 0 & 0 & \ldots \\ 0 & 0 & 0 & 1 & \ldots & 0 & 0 & 0 & 1 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{array}\right) \\ \qquad\quad \longleftrightarrow \qquad\qquad\qquad \longleftrightarrow \\ \qquad\quad n \qquad\qquad\qquad\quad p \end{array}$$

Considerem també la taula de creuament de contingències de la població segons les característiques I i J. El terme general l'indiquem k_{ij}. La taula de freqüències serà la de terme general $\frac{k_{ij}}{N} = f_{ij}$.

La taula de freqüències, amb les marginals, és:

$$\begin{array}{c} \qquad\quad J \\ I \left(\begin{array}{ccc} \vdots & & \\ \ldots & f_{ij} & \ldots \\ \vdots & & \\ \ldots & f_j & \ldots \end{array}\right) \begin{array}{c} \vdots \\ f_i \\ \vdots \\ 1 \end{array} \end{array}$$

Escrivim $f_j^i = \frac{f_{ij}}{f_i}$ i $f_i^j = \frac{f_{ij}}{f_j}$.

L'anàlisi tracta, per a aconseguir de poder emprar la mètrica euclidiana canònica en lloc de la mètrica χ^2[Vol 85] -mètrica adequada per a calcular distàncies entre distribucions-, els següents elements:

- Un núvol d'n punts x_i de l'espai $\mathcal{R}^p$, amb ponderació f_i, i de coordenada general:

$$x_{ij} = \frac{f_j^i}{\sqrt{f_j}}.$$

  El centre de gravetat és el punt G de coordenada general $\sqrt{f_j}$.

- Un núvol de p punts y^j de l'espai $\mathcal{R}^n$, amb ponderació f_j, i de coordenada general:

$$y_{ij} = \frac{f_i^j}{\sqrt{f_i}}.$$

  El centre de gravetat és el punt H de coordenada general $\sqrt{f_i}$.

La simetria juga un paper important en l'ACS, i simplifica notablement la seva presentació.

Fixem-nos que la inèrcia dels punts x_i respecte del seu centre de gravetat és

$$In(I) = \sum_i f_i(x_{ij} - \sqrt{f_j})^2 = \sum_{ij} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j},$$

expressió idèntica a la inèrcia dels punts y^j, $In(J)$, que no és res més que la distància χ^2 entre les distribucions $(f_{ij})_{ij}$ i $(f_i f_j)_{ij}$ centrada sobre aquesta última, i que ens dóna una mesura de la informació[Vol 85] aportada per la taula de freqüències.

Anem a presentar, en primer lloc, l'ACS com una Anàlisi Canònica de les taules disjuntes completes[Vol 85] i així sorgirà la mètrica χ^2 com l'adequada a l'estudi; posteriorment farem el plantejament i la solució clàssics.

4.1.2 L'ACS: cas particular de l'Anàlisi Canònica

Siguin le dues taules disjuntes completes de les característiques I i J que escrivim R i S respectivament:

$$R = \begin{array}{c} \overset{I}{} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix} \\ \overset{\longleftrightarrow}{n} \end{array} \quad S = \begin{array}{c} \overset{J}{} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix} \\ \overset{\longleftrightarrow}{p} \end{array}$$

R té dimensions Nxn i S Nxp.

Considerem les n columnes d'R, $R_i, i = 1, \dots, n$, i les p columnes de $S, S_j, j = 1, \dots, p$.

L'AC de les taules R i S busca els dos vectors unitaris per la norma canònica, Z del subespai engendrat pels n vectors R_i i T del subepai engendrat pels p vectors S_j, tals que $cos(Z, T)$ sigui màxim.

Així, l'ACS és l'AFD del triplet (Δ, M^{-1}, I) essent Δ la matriu $Nx(n+p)$ juxtaposició d'R i S, M^{-1} la matriu:

$$\begin{pmatrix} (R^tR)^{-1} & \vec{0} \\ \vec{0} & (S^tS)^{-1} \end{pmatrix},$$

i I la identitat en $\mathcal{R}^N$.

Fixem-nos que R^tR i S^tS són les respectives matrius diagonal de terme general el nombre d'individus posseint la modalitat corresponent.

Utilitzant les relacions:

$$Z = a_1R_1 + \ldots + a_nR_n = R\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

$$T = b_1S_1 + \ldots + b_pS_p = S\begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}$$

amb $a_i, b_j \in \mathcal{R}$, podem plantejar el problema de l'AC com:

Trobar els vectors $\vec{a} \in \mathcal{R}^n$ i $\vec{b} \in \mathcal{R}^p$ tals que

$$\vec{a}^tR^tR\vec{a} = \vec{b}^tS^tS\vec{b} = 1$$

fent l'expressió

$$\vec{a}^tR^tS\vec{b}$$

màxima.

Sigui n_i el nombre d'individus posseint la modalitat i de la característica I i n_j el nombre d'individus posseint la modalitat j de la característica J.

El terme general de la matriu R^tS és n_{ij}, que és el nombre d'individus posseint alhora la modalitat i d'I i la modalitat j de J.

Aleshores, estem buscant a_i i b_j tals que

$$\sum_{i=1}^{n} n_ia_i^2 = \sum_{j=1}^{p} n_jb_j^2 = 1$$

i

$$\sum_{i,j=1}^{n,p} n_{ij} a_i b_j$$

sigui màxim.

Si posem $f_{ij} = \frac{n_{ij}}{n}$ i $f_i = \sum_j f_{ij}, f_j = \sum_i f_{ij}$, tenim que

$$\sum_{i,j=1}^{n,p} \frac{f_{ij}}{\sqrt{f_i f_j}} a_i \sqrt{n_i} b_j \sqrt{n_j}$$

ha de ser màxim.

Sigui B la matriu nxp de terme general

$$\frac{f_{ij}}{\sqrt{f_i f_j}}.$$

Siguin els vectors $v \in \mathcal{R}^n$ de terme general $a_i\sqrt{n_i}$ i $u \in \mathcal{R}^p$ de terme general $b_j\sqrt{n_j}$.

Estem aleshores, buscant u i v tals que

$$u^t u = v^t v = 1$$

i $v^t B u$ sigui màxim.

Per multiplicadors de Lagrange[Vol 85], les solucions són els vectors propis associats als valors propis més grans de les matrius $B^t B$ i BB^t:

$$BB^t v = \lambda v$$
$$B^t B u = \lambda u$$

amb

$$Bu = \sqrt{\lambda} v$$
$$B^t v = \sqrt{\lambda} u.$$

Fixem-nos que estem davant dels resultats i les fórmules de transició d'una AFD del triplet (B, I, I).

La matriu B, veurem, serà la que des d'un punt de vista clàssic l'ACS ens portarà a analitzar.

Si ara considerem les files de B com a nous perfils de les modalitat d'I i calculem la distància euclídea usual entre dues modalitat així caracteritzades, resulta:

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{f_{ij}}{\sqrt{f_i f_j}} - \frac{f_{i'j}}{\sqrt{f_{i'} f_j}}\right)^2 = \sum_{j=1}^{p} \frac{1}{f_j}\left(\frac{f_{ij}}{\sqrt{f_i}} - \frac{f_{i'j}}{\sqrt{f_{i'}}}\right)^2,$$

essent aquesta última expressió, la distància χ^2 entre els perfils $\left(\frac{f_{ij}}{\sqrt{f_i}}\right)_{j=1,\ldots,p}$ i $\left(\frac{f_{i'j}}{\sqrt{f_{i'}}}\right)_{j=1,\ldots,p}$.

Igualment tindríem la distància χ^2 entre els perfils de les modalitats de J:

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_i}\left(\frac{f_{ij}}{\sqrt{f_j}} - \frac{f_{ij'}}{\sqrt{f_{j'}}}\right)^2.$$

En la següent secció veurem les mateixes caracteritzacions i els mateixos resultats obtinguts des del punt de vista clàsssic.

4.1.3 Plantejament i solució clàssics de l'ACS

Tornem ara a considerar els dos núvols de punts x_i i y^j presentats a l'apartat dels elements de l'ACS.

La inèrcia del núvol de punts x_i respecte del seu centre de gravetat G segons la direcció d'un vector u d'$\mathcal{R}^p$ és:

$$\sum_{i=1}^{n} f_i \| ((x_i^t - G^t)u)u \|_c^2 = u^t L u$$

essent L la matriu de terme general:

$$l_{jj'} = \sum_{i=1}^{n} f_i (x_{ij} - \sqrt{f_j})(x_{ij'} - \sqrt{f_{j'}}).$$

Resulta que G és vector propi d'L associat al valor propi zero.

Si considerem la matriu S de terme general:

$$s_{jj'} = \sum_{i=1}^{n} f_i x_{ij} x_{ij'},$$

resulta que G n'és vector propi associat al valor propi 1, i que L i S tenen els mateixos vectors propis ortogonals a G associats als mateixos valors propis.

Doncs bé, el problema clàssic de trobar u que maximitzi l'expressió u^tLu sota la restricció $u^tu = 1$, és el mateix que el de maximitzar u^tSu sota la restricció $u^tu = 1$, si no tenim en compte el vector G que és vector propi d'S associat al valor propi 1. Els altres valors propis d'S seran inferiors a la unitat[Vol 85] i iguals que els d'L.

Per tant, és possible de treballar amb S en lloc d'L. És a dir: és possible de calcular i maximitzar inèrcies respecte de l'origen en lloc de respecte del centre de gravetat i els resultats no varien.

Si volem integrar els pesos dins les pròpies dades, fixem-nos:

$$S = B^tB$$

essent B la matriu de terme general

$$b_{ij} = x_{ij}\sqrt{f_i} = \frac{f_{ij}}{\sqrt{f_i f_j}}.$$

Aleshores, estem davant de l'anàlisi d'un triplet (B, I, I) i ja coneixem perfectament tots els resultats.

Anomenem S matriu inèrcia del núvol de punts x_i.

Si fem l'anàlisi dels individus y^j, resulta igualment que cal només fer l'anàlisi de la matriu T de terme general:

$$t_{ii'} = \sum_{j=1}^{p} f_j y_{ij} y_{i'j},$$

que resulta $T = BB^t$. L'anomenem matriu inèrcia del núvol de punts y^j.[1]

[1]Fixem-nos que la introducció de la matriu B ens permet de tractar l'anàlisi dels punts x_i i la dels punts y^j com les anàlisis directa i dual del triplet (B, I, I).
Anàlisis equivalents serien:

- equivalent a l'anàlisi directa de (B, I, I) ho és l'anàlisi directa d'(F, M, D) essent $F = \left(\frac{f_{ij}}{f_i}\right)_{i,j}$, $D = diag(f_i)$, $M = diag(\frac{1}{f_j})$ -anàlisi que és l'enfocament clàssic;

Tenim, doncs, els eixos principals u_s del núvol de punts definit per les files de la matriu B, que no són sinó els punts x_i multiplicats per $\sqrt{f_i}$ (és a dir, amb la ponderació integrada), i tenim els eixos principals v_s del núvol de punts definit per les columnes de la matriu B (punts y^j multiplicats per $\sqrt{f_j}$), ambdós eixos associats als mateixos valors propis $0 < h_s < 1$; i tenim també les relacions de transició usuals:

$$u_s = \frac{1}{\sqrt{h_s}} B^t v_s$$
$$v_s = \frac{1}{\sqrt{h_s}} B u_s.$$

La projecció sobre u_s d'un punt x_i serà:

$$F_s(i) = \sum_{j=1}^{p} x_{ij} u_{sj}.$$

La projecció sobre v_s d' un punt y^j serà:

$$G_s(j) = \sum_{i=1}^{n} y_{ij} v_{si}.$$

Fixem-nos que les projeccions dels punts $x_i - G$ i $y^j - H$ són les mateixes, als ser els vectors propis u_s i v_s ortogonals respectivament a G i H.

Les relacions entre les projeccions[2] $G_s(j)$ i $F_s(i)$ i llur interpretació clàssica vénen donades per les fórmules (vegeu per exemple, Mardia[Mar 79], Greenacre[Gre 84]):

$$G_s(j) = \frac{1}{\sqrt{h_s}} \sum_{i=1}^{n} \frac{f_{ij}}{f_j} F_s(i)$$
$$F_s(i) = \frac{1}{\sqrt{h_s}} \sum_{j=1}^{p} \frac{f_{ij}}{f_j} G_s(j).$$

- equivalent a l'anàlisi dual de (B, I, I) ho és l'anàlisi directa de (G, M_G, D_G) essent $G = \left(\left(\frac{f_{ij}}{f_j}\right)_{i,j}\right)^t$, $D_G = diag(f_j)$, $M_G = diag(\frac{1}{f_i})$;
- equivalent a l'anàlisi directa de (B, I, I) també ho és la dual de $(G, , M_G, D_G)$;
- i finalment, equivalent a la dual de (B, I, I) ho és la dual del triplet (F, M, D).

Òbviament, en cada cas s'ha de tenir en compte què projectem i sobre quin vector, i traduir adequadament.

[2]Fixem-nos que la coordenada de la projecció de $\left(\frac{f_{ij}}{f_i}\right)_j$ sobre el factor principal u^* en l'anàlisi del triplet (F, M, D) coincideix amb la de la projecció del punt x_i sobre el vector u corresponent en l'anàlisi de (B, I, I); i la mateixa relació s'observa entre $\left(\frac{f_{ij}}{f_j}\right)_i$ i y^j.

4.1.4 Taules juxtaposades i taules lògiques

Acabem de presentar l'ACS com l'anàlisi d'una taula de creuament de contingències de terme general k_{ij}, de la qual, en primer lloc, se'n construeix la respectiva taula de freqüències, de terme general f_{ij}, i posteriorment els núvols de punts x_i i y^j dels espais $\mathcal{R}^p$ i $\mathcal{R}^n$ respectivament.

Doncs bé, podem juxtaposar dues taules formant-ne una d'única que pot ser analitzada de la mateixa manera -és a dir, analitzada segons el que en podem dir Anàlisi de Correspondències d'una taula de contingències o de la corresponent taula de freqüències.[3]

Sigui doncs, sobre una mateixa població d'N individus, el creuament d'una variable categòrica I, amb n modalitats, amb dues d'altres, J_1 i J_2, amb q_1 i q_2 modalitats respectivament. La taula juxtaposada té, doncs, dimensions nxp, essent $p = q_1 + q_2$:

$$I \begin{pmatrix} & J_1 & & & J_2 & \\ & \vdots & & & \vdots & \\ \cdots & k_{ij}^1 & \cdots & \cdots & k_{ij}^2 & \cdots \\ & \vdots & & & \vdots & \end{pmatrix}$$

La suma total d'efectius de la taula és $2N$. La taula de freqüències corresponent té per terme general

$$f_{ij}^s = \frac{k_{ij}^s}{2N},$$

amb $s = 1,2; i = 1,\ldots,n, j = 1,\ldots,q_1$ amb $s = 1$, i $j = q_1 + 1,\ldots,p$ amb $s = 2$.

Per a l'anàlisi d'aquesta taula de freqüències segons l'Anàlisi de Correspondències presentat a l'apartat anterior, ens remetem a Volle[Vol 85]. Els resultats són, però, clars -tal vegada val la pena de remarcar que la inèrcia del núvol de punts-fila de la taula juxtaposada, $In(I)$, resulta ser

$$1/2(In_1(I) + In_2(I)),$$

[3]A partir d'ara, doncs, per Anàlisi de Correspondències entendrem aquesta anàlisi.

amb $In_k(I)$ la inèrcia dels punts-fila en la respectiva taula de creuament IxJ_k, $k = 1, 2$.

Per altra banda, el creuament global d'm variables categòriques sobre N individus pot presentar-se com una taula lògica en què cadascuna de les m variables, $J_1, \ldots, J_m$ es desglossa en les seves modalitats i indicant per 1 i per 0, respectivament, si l'individu en qüestió posseeix o no la corresponent modalitat:

variables	J_1	J_2	...	J_m
modalitats	$1, 2, \ldots, m_1$	$1, 2, \ldots, m_2$	...	$1, 2, \ldots, m_m$
individus				
1	01.....	10.....	...	00.....
2	10.....	00.....	...	10.....
⋮				
N	00.....	10.....	...	00.....

Sigui $p = \sum_{s=1}^{m} m_s$. Les dimensions de la taula són Nxp. El terme general de la taula l'indiquem, simplement, $k_{ij}, i = 1, \ldots, N, j = 1, \ldots, p$. k_{ij} val 1 ó 0 segons el que l'individu i posseeix.

La corresponent taula Nxp de freqüències susceptible de ser analitzada segons l'Anàlisi de Correspondències té per terme general:

$$f_{ij} = \frac{k_{ij}}{mxN}.$$

Fixem-nos que no és res més que una juxtaposició de taules en què la variable I ara és el conjunt d'individus amb modalitats, per dir-ho així, cadascun dels individus, i per tant, amb $n = N$.

Bé, en definitiva, l'Anàlisi de Correspondències presentat a l'apartat 1.3 d'aquest capítol, el podem aplicar a simples taules de creuament de dues variables categòriques (ACS), però també a taules juxtaposades i a taules lògiques, amb tots els resultats i totes les interpretacions adequades.

4.2 L'Anàlisi de Correspondències Múltiples (*ACM*)

4.2.1 L'*ACM* com a cas particular de l'*ACG*

Suposem que sobre n individus avaluem m variables categòriques i que entenem la taula lògica corresponent com si tinguéssim m paquets de variables i en cada paquet tantes variables com modalitats té la variable que realment és el paquet:

$$\begin{array}{c} \qquad X_1 \qquad\qquad\qquad\qquad\qquad\qquad\qquad X_m \\ \left(\begin{array}{ccccc} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{array} \cdots \begin{array}{ccccc} 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{array}\right) \\ \quad \longleftrightarrow \qquad\qquad\qquad\qquad \longleftrightarrow \\ \quad m_1 \qquad\qquad\qquad\qquad\quad m_m \end{array}$$

Sigui X la matriu que juxtaposa aquests "paquets". Sigui $p = \sum_{s=1}^{m} m_s$.

Considerem el problema de l'*ACG* d'aquesta matriu X, i per tant l'operador:

$$\Pi = \sum_{k=1}^{m} \Pi_k$$

essent Π_k l'operador projecció canònica:

$$X_k(X_k^t X_k)^{-1} X_k^t.$$

La matriu M^{-1} pròpia de l'*ACG* formada per blocs $(X_k^t X_k)^{-1}$ és ara, la matriu pxp diagonal de terme general $\frac{1}{k_j}$ essent k_j el nombre d'individus posseint la modalitat j.

L'*ACM* és, doncs, l'*AFD* del triplet $(X, diag(1/k_j), I_n)$.

Coneixem l'*ACG* -és a dir, la recerca dels vectors v_s que solucionen el problema de maximitzar $v^t \Pi v$ sota $v^t v = 1$, vectors propis de Π associats als més grans valors propis h_s; i coneixem la seva interpretació geomètrica com a vectors, les projeccions canòniques dels quals sobre els subespais engendrats pels m paquets de variables, són "bons" representants d'aquests paquets -i igualment tenim els vectors obtinguts per

transició:

$$u_s = \frac{1}{\sqrt{h_s}} X^t v_s.$$

Doncs bé, anem a veure com tots aquests elements i resultats són els propis de l'*ACM* presentat clàssicament.

4.2.2 L'*ACM* com a anàlisi d'una taula lògica

La presentació de l'*ACM* més usual[Vol 85] és la de, simplement, una Anàlisi de Correspondències sobre una taula lògica[4] de creuament global de m variables categòriques $J_1, \ldots, Jm$ sobre una població d'n individus.[5] És a dir, la taula lògica que consideràvem anteriorment:

variables	J_1	J_2	...	J_m
modalitats	$1, 2, \ldots, m_1$	$1, 2, \ldots, m_2$	...	$1, 2, \ldots, m_m$
individus				
1	01.....	10.....	...	00.....
2	10.....	00.....	...	10.....
⋮				
n	00.....	10.....	...	00.....

Igualment, indiquem $p = \sum_{s=1}^{m} m_s$, i per tant la taula té dimensions nxp.

Aleshores, les projeccions $F_s(i)$ i $G_s(j)$ ho seran, és clar, respectivament, dels n individus i de les p modalitats que configuren les m variables. Òbviament, les fórmules de relació fonamental entre $F_s(i)$ i $G_s(j)$ són, ara:

$$F_s(i) = \frac{1}{\sqrt{h_s}} \sum_{j=1}^{p} \frac{1}{m} k_{ij} G_s(j)$$
$$G_s(j) = \frac{1}{\sqrt{h_s}} \sum_{i=1}^{n} \frac{1}{k_j} k_{ij} F_s(i)$$

essent $k_{ij} = 1$ ó 0 segons si l'individu i posseeix o no la modalitat j (és a dir, k_{ij} és el terme general de la taula lògica), i k_j és el nombre d'individus posseint la modalitat j.

Anem a veure la relació amb l'*ACM* com a una *ACG*.

[4] Taula disjunta completa.

[5] Per raons de notació, preferim aquí considerar n el nombre d'individus.

Resulta que l'operador Π corresponent és:

$$\Pi = mT_L,$$

essent T_L la matriu inèrcia del núvol de punts definit per les columnes de la taula lògica, i per tant, els vectors v_s que solucionen el problema de maximitzar $v^t \Pi v$ sota $v^t v = 1$, vectors propis de Π associats als més grans valors propis h_s, són els factors principals del l'anàlisi de les modalitats de l'ACM entesa com a anàlisi sobre la taula lògica.

Sabem $T_L = B_L B_L^t$, essent B_L la matriu nxp corresponent a la taula lògica de terme general:

$$\frac{f_{ij}}{\sqrt{f_i f_j}} = \frac{k_{ij}}{\sqrt{k_j m}},$$

essent k_{ij} el terme general de la pròpia taula lògica i k_j el nombre d'individus que posseeixen la modalitat j, com ja sabem.

Aleshores:

$$mT_L = mB_L B_L^t = \Pi = XM^{-1}X^t,$$

essent M^{-1} la matriu pxp de l'ACG corresponent de terme general $\frac{1}{k_j}$.

Resulta, és clar:

$$\sqrt{m}B_L = \left(\frac{k_{ij}}{\sqrt{k_j}}\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}} = (k_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}} \left(diag\left(\frac{1}{\sqrt{k_j}}; j = 1,\ldots,p\right)\right) = XM^{-\frac{1}{2}}.$$

Hem de tenir en compte aquesta relació a l'hora de considerar els eixos principals corresponents a l'espai dels individus en l'ACG i en l'ACM com a anàlisis d'una taula lògica.

En aquest segon punt de vista, tenim:

$$u_{s_L} = \frac{1}{\sqrt{h_{s_L}}} B_L^t v_s,$$

essent h_{s_L} els corresponents valors propis de l'anàlisi.

En el primer punt de vista:

$$u_s = \frac{1}{\sqrt{h_s}} X^t v_s.$$

Com que

$$h_{s_L} = \frac{h_s}{m},$$

resulta:

$$u_{s_L} = \frac{1}{\sqrt{h_s}} \sqrt{m} B_L^t v_s = \frac{1}{\sqrt{h_s}} M^{-\frac{1}{2}} X^t v_s = M^{-\frac{1}{2}} u_s.$$

Si bé amb la precaució de tenir en compte aquesta relació entre u_s i u_{s_L}, l'equivalència entre l'*ACM* com a cas particular de l'*ACG* i l'*ACM* com a Anàlisi de Correspondències sobre una taula lògica, és clara.

4.2.3 L'*ACM* com l'anàlisi d'una taula de contingència de múltiples creuaments (matriu de Burt)

L'*ACM* pot presentar-se també com l'Anàlisi de Correspondències d'una matriu, anomenada **matriu de Burt**, formada agrupant totes les taules J_r per J_l que podem construir creuant les variables de dues en dues:

$$\begin{array}{c|ccc} & J_1 & & J_m \\ J_1 & (\ldots) & \cdots & (\ldots) \\ & \vdots & \ddots & \vdots \\ J_m & (\ldots) & \cdots & (\ldots) \end{array}$$

Els creuaments diagonal d'una variable amb ella mateixa són blocs tot de zeros excepte en la pròpia diagonal en què hi ha el nombre d'individus de la població que posseeixen la modalitat. Aquesta matriu de Burt és una matriu quadrada de dimensió $p = \sum_{s=1}^{m} m_s$.

Aplicarem l'Anàlisi de Correspondències a la matriu de Burt. Fixem-nos que la simetria de la matriu fa que només tingui sentit considerar les files (o bé les columnes) i llurs projeccions i interpretacions.

De la matriu de Burt (considerada taula de contingències), passem a la corresponent taula de freqüències, també simètrica, de terme general:

$$f_{jj'} = \frac{k_{jj'}}{m^2 xn},$$

essent $k_{jj'}$ el terme general de la matriu de Burt.

Cada modalitat j ve representada per un vector de l'espai $\mathcal{R}^p$ de coordenada general:

$$\frac{f_{jj'}}{\sqrt{f_j f_{j'}}}.$$

És el vector que a l'Anàlisi de Correspondències anomenàvem y^j.I la matriu B, ara és la matriu simètrica pxp de terme general:

$$\frac{f_{jj'}}{\sqrt{f_j f_{j'}}}.$$

Els vectors propis, u_s, de la matriu B^2 associats als valors propis corresponents, h_s, també són vectors propis de la matriu B associats als valors propis corresponents, que són les arrels quadrades dels de la matriu B^2[Vol 85].

Es demostra que B és la matriu inèrcia, S_L, del núvol de punts definit per les files de la taula lògica en l'ACM presentat com a Anàlisi de Correspondències sobre aquesta taula lògica.

Per tant, l'equivalència entre les dues presentacions és clara, si bé hem de tenir en compte que abans, com a anàlisi d'una taula lògica, projectàvem les modalitats enteses com els corresponents vectors y^j de l'espai $\mathcal{R}^n$, que no són els vectors y^j d'ara, de l'espai $\mathcal{R}^p$, i que si abans projectàvem sobre els corresponents eixos principals v_s ara ho fem sobre u_s -que el que sí són és els duals dels v_s en l'anàlisi de la taula lògica. La relació entre les projeccions ve donada per:

$$G_{S_B}(j) = G_{S_L}(j)(h_s)^{\frac{1}{4}},$$

indicant amb subíndexs B o L segons si estem considerant l'ACM com a anàlisi de la matriu de Burt o de la taula lògica; els h_s són els valors propis en l'anàlisi de la matriu de Burt -valors propis de B^2, que són el quadrat dels corresponents a l'anàlisi de la taula lògica.

A més, en l'anàlisi de la taula lògica, tenim també les projeccions dels individus, $F_s(i)$, sobre els eixos u_s. Doncs bé, si afegim a la taula de Burt files, tantes com individus, de manera que cadascuna només tingui valors 1 en les modalitats posseïdes per l'individu en qüestió, i amb els altres valors sempre 0, les projeccions sobre u_s d'aquests individus com a files suplementàries de l'Anàlisi de Correspondències de la matriu de Burt (vegeu[Vol 85] pàgina 151 sobre elements suplementaris), resulta que valen igualmenmt $F_s(i)$.

L'equivalència, doncs, entre els dos punts de vista ho és tant per a les modalitats com per als individus.

4.2.4 L'ACS com a cas particular del'ACM

Hem vist que l'ACM, que no és sinó un cas particular de l'ACG, pot presentar-se com una Anàlisi de Correspondències d'una taula lògica i també com una Anàlisi de Correspondències de la matriu de Burt. N'hem vist l'equivalència.

L'ACS, que és un cas particular de l'AC i per tant de l'ACG, serà també, en aquest sentit, és clar, cas particular de l'ACM. A més, l'ACS també pot presentar-se com a Anàlisi de Correspondències sobre una taula de creuament de dues variables categòriques. Encara, finalment, si sobre la taula lògica definida per aquestes dues variables categòriques hi realitzem una Anàlisi de Correspondències (és a dir, fem com una ACM sobre la taula lògica amb $m = 2$ de les dues variables), els resultats que s'obtenen són també els mateixos. És a dir, des d'aquest segon punt de vista, l'ACS també és un cas particular de l'ACM.

Ens remetem, com altres vegades, a Volle[Vol 85] per a veure amb més detall totes aquestes relacions.[6]

L'estructura relacional completa de les anàlisis que ens interessen, resulta, en definitiva:

$$
\begin{array}{ccccc}
 & & \mathbf{AFD(X,M,D)} & & \\
 & & \downarrow & & \\
 & & \mathbf{ACG(X,M^{-1},I)} & & \\
 & \swarrow & \downarrow & \searrow & \\
\mathbf{AC(\Delta,M^{-1},I)} & & \mathbf{ACM(X,diag(\frac{1}{k_j}),I)} & & \mathbf{ACP(Z,I,I)} \\
\downarrow & & \downarrow & & \\
\mathbf{ACS(\Delta,M^{-1},I)} & = & \mathbf{ACS(X,diag(\frac{1}{k_j}),I)} & &
\end{array}
$$

[6]Igualment, Volle, en el seu capítol dedicat a l'*AC*, ens demostra que la Rgeressió Múltiple i l'Anàlisi Discriminant són també casos particulars de l'*AC*; aquestes conegudes relacions no les tractem però, en aquesta Tesi.

Capítol 5

Algunes extensions de les anàlisis presentades

5.1 Introducció

Els mètodes d'anàlisi de dades han demostrat la seva eficàcia en l'estudi de grans i complexes quantitats d'informacions. Permetent representacions simplificades de les grans taules s'han convertit en importants eines de síntesi que ens posen de manifest les tendències dominants, les jerarquies relacionals, tot eliminant efectes marginals que podrien pertorbar la percepció global de les dades.

Podem dir, segons el que hem vist fins aquí i de manera sintètica, que el principi d'aquests mètodes d'anàlisi és únic[Esc 88]: dos núvols de punts representant les files i les columnes de la taula estudiada són construïts i representats gràficament, amb representacions de files i columnes, val a dir, fortament relacionades.

Així, amb aquest únic principi, l'Anàlisi en Components Principals tracta taules que creuen individus i variables numèriques, l'Anàlisi de Correspondències Simples tracta taules de freqüències i l'Anàlisi de Correspondències Múltiple s'aplica a variables qualitatives.

Ara bé, l'eficàcia dels mètodes d'anàlisi de dades ha fet que s'apliquessin en estudis en què la complexitat de les dades a tractar difícilment es veu recollida en simples

codificacions per taules de files i columnes. Les extensions dels mètodes:

- en estudis en què domina l'aspecte evolutiu d'unes dades, el caràcter, diguem-ne, més múltiple de la informació i en els quals s'han de tractar hipertaules;
- en estudis preferentment dirigits a controlar efectes, a posar de manifest relacions i oposicions locals o parcials, només;
- en estudis que comporten rols de les variables a tractar diferents als que juguen en els mètodes clàssics, o que, per tenir variables de diferent naturalesa, cal aleshores tractar-les agrupades diversament;

tals extensions, dèiem, han configurat camins d'investigació i desenvolupament en l'anàlisi de dades dins dels quals s'han d'inserir els nous treballs.

Voldríem aquí ressenyar breument el mètode STATIS degut a C. Lavit, Y. Escoufier i C. Roux, indicat particularment en el tractament simultani de vàries taules en estudis d'evolució en què cada taula correspon a un instant de temps, i el mètode de B. Escofier i J. Pagès de l'Anàlisi Factorial Múltiple que tracta taules de dades en què un mateix conjunt d'individus està descrit per varis grups de variables, tant quantitatives com qualitatives; els dos mètodes tenen pràcticament els mateixos objectius diferenciant-se en la presentació i en prou punts del desenvolupament metodològic com per a arribar a considerar interessant d'exposar-los tots dos.

També, el mètode de comparació de taules binàries dels mateixos autors B. Escofier i J. Pagès igualment indicat en estudis d'evolució, ara de taules de contingència.

Després ressenyarem l'Anàlisi de Correspondències No Simètric de N. Lauro i L. D'Ambra en què les variables categòriques no juguen un paper simètric.

I finalment, introduirem les anàlisis parcials, locals o condicionals -que tracten la descripció d'aspectes parcials de les dades controlant-ne d'altres.[1] Parlarem

[1]Les relacions entre les diferents extensions presentades, si bé, com ja dèiem, algunes són clares, de moment no seran tractades en detall; és un interessant treball que deixem per a més endavant.

breument de l'*ACM* Condicional de B. Escofier[Esc 87] i de l'Anàlisi Local de L. Lebart i T. Aluja[Alu 85]. Aquesta última anàlisi -que aconsegueix de controlar els efectes mitjançant grafs no orientats definits entre els individus- la desenvoluparem exhaustivament en el capítol posterior, doncs és a partir de la seva metodologia que es comprèn l'anàlisi presentada al darrer capítol.

5.2 El mètode STATIS

5.2.1 Principis

Aquest mètode [Lav 84], que com ja hem dit és adequat en estudis d'evolució en què tenim vàries taules cadascuna corresponent a una unitat de temps, tracta sempre els mateixos n individus caracteritzats pels valors que, en l'instant k, prenen sobre ells p_k variables numèriques. En cada instant o situació k, doncs, tenim p_k variables, diferents, si cal, d'una situació a una altra. Variem les situacions de la manera $1 \leq k \leq K$.

Les anàlisis i les projeccions factorials (sempre en subespais de dimensió petita -com a màxim tres), es configuren en tres etapes:

- una anàlisi en què es comparen les diferents taules. L'anomenem *anàlisi d'interestructura.*

- una anàlisi que ens dóna una visió dels individus que és resum de les proporcionades per cada taula: *anàlisi resum-compromís* de totes les taules.

- una representació simultània dels individus segons cada situació i segons la visió sintètica donada pel compromís: *anàlisi d'intraestructura.*

5.2.2 Desenvolupament del mètode

Anàlisi d'interestructura

Seguint [Lav 84][Lav 88], suposem que la mètrica de l'espai dels individus és la identitat; altrament, transformaríem les dades per a poder igualment considerar-ho així.

Es tracta d'anar més enllà de veure les taules com a "files+columnes" per a copsar-ne l'evolució en les relacions, i això s'aconsegueix considerant cada taula -cada situació k- com a un sol element estadístic; de fet, com a una nova variable. Vegem-ho.

Sigui D la matriu nxn dels pesos dels individus.

A cada taula X_k de dimensió nxp_k li associem un operador $X_k X_k^t D$ la traça del qual representa la inèrcia del núvol dels individus en la situació k.[2] Aquests K operadors són matrius nxn i per tant els podem considerar vectors de l'espai $\mathcal{R}^{n^2}$ i si amb ells com a columnes construïm una matriu n^2xK tindrem la nova matriu de dades a analitzar.

Com és habitual a l'ACP, diagonalitzarem la matriu de correlacions[3] entre les noves variables-situació; definirem aquesta matriu a partir de l'operador *traça* -extensió natural de les mètriques usuals[Lav 84].

Tenim, doncs:

$$cov(X_k X_k^t D, X_l X_l^t D) = tr(X_k X_k^t D X_l X_l^t D).$$

I per tant, la matriu C_{KxK} de correlacions té terme general:

$$c_{kl} = \frac{tr(X_k X_k^t D X_l X_l^t D)}{\sqrt{tr(X_k X_k^t D)^2 tr(X_l X_l^t D)^2}}.$$

[2]Quan tractem l'Anàlisi Factorial Múltiple tornarem a parlar d'aquests operadors; el que direm completarà el que s'ha dit aquí.

[3]També podríem fer-ho amb la matriu de covariàncies, és clar, almenys en el cas que cada taula tingui les mateixes variables.

Fixem-nos, de cara a millor interpretar aquesta matriu, que si les dades estan centrades, resulta:

$$cov(X_k X_k^t D, X_l X_l^t D) = \sum_{i=1}^{p_k} \sum_{j=1}^{p_l} cov^2(X_{ki}, X_{lj}),$$

essent X_{ki} la variable i de la taula k.

Aleshores, doncs, és clar que tenim definida una matriu a analitzar que el que expressa és la relació entre dues situacions mitjançant les correlacions entre les variables d'una i altra situació, totes amb totes, per dir-ho així.

La representació factorial, que podem anomenar gràfica de la interestructura, projecta els operadors. Cada operador, cada situació, doncs, és un punt del pla principal (si projectem sobre el millor subespai de dimensó dos), i les interpretacions a fer són les pròpies de l'ACP.

Anàlisi resum-compromís

L'ACP de cada taula ens proporciona K visions diferents dels mateixos individus. Ara voldríem trobar individus-promig del conjunt de taules. Seran uns individus ficticis que resumiran de la millor manera tota la informació.

Escrivim $W_k = X_k X_k^t$, $1 \leq k \leq K$.

Anomenem *matriu del compromís* la matriu

$$W = \sum_{k=1}^{K} \alpha_k W_k,$$

combinació lineal de les matrius W_k; de fet, tenim una situació combinació lineal de les K situacions:

$$WD = \sum_{k=1}^{K} \alpha_k W_k D.$$

Els escalars α_k els agafarem de manera que $\alpha = (\alpha_1, \ldots, \alpha_K)$ sigui el primer vector propi unitari de la matriu C de correlacions entre els operadors-situació.[4]

[4]Per les propietats de C podem agafar $\alpha_k \geq 0; 1 \leq k \leq K$.

Ens remetem directament a [Lav 84][Lav 88] per a justificar α segons les propietats que el caracteritzen -en el sentit de les quals diem que WD és el millor operador-resum, que no és sinó el sentit amb què a l'ACP diem que és millor el primer factor principal, u, vector de l'espai dels individus.

Fixem-nos que podríem considerar una matriu fictícia X tal que $W = XX^t$ les files de la qual representarien individus-compromís.

Segons l'ACP de la matriu X, podríem projectar els individus-compromís (files d'X) en el pla[5] engendrat pels usuals dos primers factors principals.

Sabem que les coordenades dels individus-compromís són les files de la matriu $nx2$:

$$Y = V_2\Lambda^{1/2}$$

essent Λ la matriu diagonal dels dos primers valors propis de WD i V_2 la usual matriu $nx2$ les columnes de la qual són els dos primers vectors propis D-ortonormals de WD.

Tenim doncs, una representació de cada individu-promig o compromís.

Representació simultània: etapa d'intraestructura

Acabem d'obtenir la projecció de cada individu-promig en un pla creat en l'anàlisi de WD. Ara volem projectar en aquest mateix pla els individus tal i com són vistos per cadascuna de les taules. Així tindrem els individus representats una vegada per cada taula inicial i una com a individus-promig o compromís. La comparació entre les representacions és clarament interessant en l'estudi evolutiu que tractem.

Primer, anem a caracteritzar els individus segons que són vistos per cada taula. Es tracta de considerar les matrius Y_k, $1 \leq k \leq K$ de dimensió $nx2$, les files de les quals són les coordenades dels n individus sobre el pla factorial principal en l'anàlisi de la corresponent taula k. És a dir, cada matriu Y_k ens expressa els individus tal i com són vistos per la taula k, per la situació k.

[5]És clar que també podríem fer-ho en dimensió tres.

Les coordenades de les projeccions dels n individus expressats per les files de les matrius Y_k sobre el pla principal de l'anàlisi del compromís sobre el qual abans hem projectat els individus-compromís, són les línies de la matriu $nx2$:

$$\hat{Y}_k = AY_k,$$

essent A una matriu nxn l'expressió de la qual establí M.C.Place a la seva tesi doctoral(1980) [Lav 84].

Hi ha d'altres propostes que no projecten directament els individus-files d'Y_k sinó les relacions entre ells en cada situació k expressades per diferents elements estadístics[Lav 84].

En definitiva, es tracta de tenir en un mateix pla factorial les projeccions dels individus en cada situació k i les projeccions del individus-compromís i estudiar-ne convenientment les relacions.

Encara, en la mateixa línia de representacions simultànies, podem tenir en compte que tant les $p = \sum_{k=1}^{K} p_k$ variables, com les components principals[6] de l'anàlisi factorial de les variables en cada taula i també en el compromís, són vectors de l'espai $\mathcal{R}^n$, i que per tant poden ésser projectades en el pla principal de projecció de variables de l'anàlisi del compromís, pla engendrat pels dos primers vectors propis de WD. L' estudi relacional d'aquestes representacions ens ofereix la possibilitat d'establir les oposicions i evolucions de les dades d'una manera exhaustiva.

5.3 Anàlisi Factorial Múltiple

5.3.1 Principis

L'Anàlisi Factorial Múltiple (AFM) tracta taules en què un mateix conjunt d'individus està descrit per diferents grups de variables [Esc 88]. Aquestes variables poden ser

[6]Com a components principals podem considerar tant els factors principals de l'anàlisi de les variables com les coordenades dels n individus sobre els factors de l'anàlisi dels individus.

tant quantitatives com qualitatives, però cada grup només en conté d'un tipus. Pot analitzar-se també, l'evolució, sobre d'un mateix conjunt d'individus, d'una sèrie de variables en períodes diferents, cada període definint un grup de variables.

En [Esc 88], B. Escofier i J. Pagès exposen l'exemple de la caracterització de 21 vins negres en què s'analitzen 29 variables agrupades en diferents grups (olfactació en repòs, visió, olfactació després de remoure, gust, judici global). L'AFM té per objectiu l'anàlisi conjunt d'individus, variables i grups de variables per a aconseguir d'aprofundir en les relacions i oposicions observades evitant que un grup de variables prengui un rol preponderant que faci il.lusòria la consideració dels altres grups.

Seguirem bàsicament [Esc 88] en la breu exposició de l'AFM que farem, considerant, però, només variables numèriques. Ens remetem als mateixos autors (secció 7.6 de [Esc 88]) per al tractament de variables qualitatives en què aleshores l'AFM es presenta com a generalització de l'ACM -com també ho és de l'ACP en el cas de variables numèriques (en l'ACP cada grup es reduiria a una sola variable).

5.3.2 Desenvolupament de l'AFM

Elements

Sigui I un conjunt d'individus, i K el conjunt de variables que el descriuen.

Agrupem les variables en J grups, i notem per K_j el conjunt de variables del grup $j, j = 1, \ldots, J$.

Si indiquem per les mateixes lletres, I, K, K_j, el cardinal dels conjunts I, K, K_j respectivament, tenim:

- la taula IxK completa X.

- J taules X_j de dimensions IxK_j associada cadascuna a un grup de variables.

La matriu de pesos dels I individus és:

$$D = diag(p_i, p_i \geq 0, \sum_i p_i = 1),$$

i la matriu de pesos de les K variables:

$$M = diag(m_k, m_k \geq 0).$$

Els pesos m_k varien segons si fem una anàlisi separada per a cada grup en què és usual de considerar $m_k = 1$ (pesos inicials) per a totes les variables, o si fem una anàlisi global en què aleshores el pes inicial d'una variable del grup j es divideix per λ_j^1 (ponderació), essent λ_j^1 el primer valor propi de l'ACP usual separat del grup j -és a dir, de la taula X_j. Aquesta ponderació proposta per Escofier i Pagès (vegeu, per a millor comprensió, les seccions 7.2.1 i 7.2.2 de [Esc 88]), segons les seves pròpies paraules, té per objectiu d'equilibrar el rol dels grups en tots els aspectes de l'anàlisi. És essencial en aquesta anàlisi.

AFM en $\mathcal{R}^K$ dels individus

En termes d'Anàlisi Factorial, el que l'AFM dels individus fa és el següent:

- una representació gràfica del núvol d'individus caracteritzats pel conjunt de totes les variables. Es tracta només d'una ACP de la matriu X amb els pesos de la matriu D i mètrica donada per la matriu M segons la ponderació indicada anteriorment.

- una representació simultània dels J núvols de punts d'individus caracteritzats per cada grup de variables. Aquesta representació s'obté afegint a la taula X les j taules $\tilde{X}_j$ també de dimensions $I x K$ obtingudes completant amb $K - K_j$ zeros les representacions dels I individus segons les K_j variables definidores de cada grup j. El tractament com a individus suplementaris de les files de les j taules

$\tilde{X}_j$ i llurs projeccions com a tals en l'anàlisi de la taula X, ens proporciona la representació simultània que volíem dels j núvols d'individus.

AFM en $\mathcal{R}^I$ de les variables

La representació de les K variables s'obté directament de l'*ACP* de la taula completa X: és, doncs, la representació dual de la dels individus del núvol complet.

Com sempre en l'*ACP*, la representació de les variables pot considerar-se [Esc 88]:

- una ajuda a la interpretació de la representació del núvol d'individus
- una representació òptima de les correlacions entre variables.

A més, és interessant un estudi de les correlacions entre les components principals de cada grup. Les components principals de la taula X_j, recordem-ho, són les projeccions del núvol d'individus sobre els factors principals de l'anàlisi dels individus -o bé, també, els factors principals de l'anàlisi de les variables. Doncs bé, si introduïm les components principals com a variables suplementàries en l'anàlisi de la taula X, tindrem el que ens interessava.

Fixem-nos, doncs, que fins aquí podríem dir que la metodologia bàsica de l'*AFM* és el tractament com a suplementaris dels elements estadístics adequats a l'objectiu plantejat.

AFM en $\mathcal{R}^{I^2}$ dels grups de variables

La representació de cada grup de variables per un sol punt és, lògicament, una bona manera de tenir la possibilitat de comparar els grups. Introduïm l'espai $\mathcal{R}^{I^2}$ per a donar aquesta representació.

B. Escofier i J. Pagès proposen, per a ben definir els grups, els mateixos operadors que hem presentat en l'apartat del mètode STATIS: els operadors $W_j = X_j M_j X_j^t$.

Les matrius M_j són les matrius de pesos de les variables dels corresponents grups j i generalment són les matrius identitat I_{K_j}.[7]

Aleshores, cada operador és una columna d'una certa matriu I^2xJ i representa el grup corresponent. Es considera igualment, com fèiem en el mètode STATIS, el producte escalar:

$$< W_jD, W_lD >= tr(W_jDW_lD).$$

El producte escalar entre W_jD i W_lD és, és clar, una mesura de la relació entre els grups j i l. Per comparar globalment els grups busquem de descriure les relacions-oposicions dels operadors W_jD projectant-los en subespais d'$\mathcal{R}^{I^2}$ de dimensió reduïda. No es tracta sinó de fer l'anàlisi usual sobre la matriu I^2xJ dels operadors obtenint la projecció del núvol de grups sobre els factors principals.

Hi ha el problema de la bona interpretació d'aquests factors, vectors d'$\mathcal{R}^{I^2}$, i de les coordenades de les projeccions dels grups.

B. Escofier i J. Pagès (secció 7.4.4 de [Esc 88]) busquen els factors principals d'$\mathcal{R}^{I^2}$ de la forma $z_sz_s^tD$ associats a les variables z_s d'$\mathcal{R}^I$ components principals de l'anàlisi de la taula X de manera que la suma de les coordenades de les projeccions dels operadors W_jD sobre $z_sz_s^tD$ sigui:

$$\sum_j < W_jD, z_sz_s^t >=$$

inèrcia de les variables de tots els grups projectades sobre z_s.

És a dir, la projecció del núvol de punts de l'espai $\mathcal{R}^{I^2}$ amb el producte escalar de la traça en què cada punt del núvol representa un grup, sobre l'eix

$$z_sz_s^tD \in \mathcal{R}^{I^2}$$

s'interpreta de manera que un grup j té per coordenada sobre l'eix la inèrcia projectada de les variables del grup sobre $z \in \mathcal{R}^I$ [Pag 89].

[7] $M_j = I_{K_j}$ inicialment; en les anàlisis globals ponderem dividint per λ_j^1.

5.3.3 Conclusions

Fixem-nos com la idea que subjau en l'AFM és la de buscar variables lligades als grups de variables, tal com ho vèiem al final de l'apartat anterior; com a mesura de relació entre una variable z i un grup j hi hem definit la inèrcia del núvol projectat sobre z. En aquest sentit l'AFM es pot veure com una Anàlisi Canònica Generalitzada [Esc 88] en què el criteri de lligament de la inèrcia del núvol projectat substitueix el de projecció ortogonal (cosinus de l'angle, per dir-ho així)(secció 7.3 de[Esc 88]), criteri, aquest últim, poc adequat quan dintre dels grups les variables estan correlacionades (secció 7.3 de [Esc 88]); és a dir, diríem que l'AFM és indicat com a tècnica d'anàlisi canònica (recerca de variables lligades als grups) quan en els grups les variables estan correlacionades.

Diguem [Pag 89], finalment, que mitjançant certes manipulacions, la majoria d'aspectes de la problemàtica de les taules múltiples, són tractats per l'AFM, de fet, com a aspectes d'una anàlisi factorial simple.

5.4 Mètode de Comparació de Taules Binàries

5.4.1 Principis

Ens interessen estudis conjunts de vàries taules de contingència definides a partir d'una mateixa parella de variables categòriques sobre diferents poblacions -poblacions diferents perquè resulten de la categorització feta per una tercera variable, o perquè són avaluades en moments diferents.

Seguim [Esc 88] en el desenvolupament del mètode breument presentat aquí.

Notem per I, J, T tres variables categòriques i igualment el nombre respectiu de categories, avaluades sobre un conjunt d'individus.

Considerem el paral.lelepípede de freqüències relatives de terme general: $f_{ijt}, i, j, t = 1, \ldots, I, J, T$ respectivament, amb $\sum f_{ijt} = 1$.

És a dir, tenim les variables I, J creuades sobre T poblacions definides per les categories de T -si T representa el temps, les T poblacions no seran sinó la mateixa avaluada en moments diferents.

Fixem-nos que així una de les direccions, la direcció T, juga un paper diferent a les altres dues.

Els marges binaris del paral.lelepíde són les tres taules de contingència binàries obtingudes sumant sobre un dels tres índexs. Els termes generals d'aquests marges els escrivim, respectivament: $f_{ij.}, f_{i.t}, f_{.jt}$.

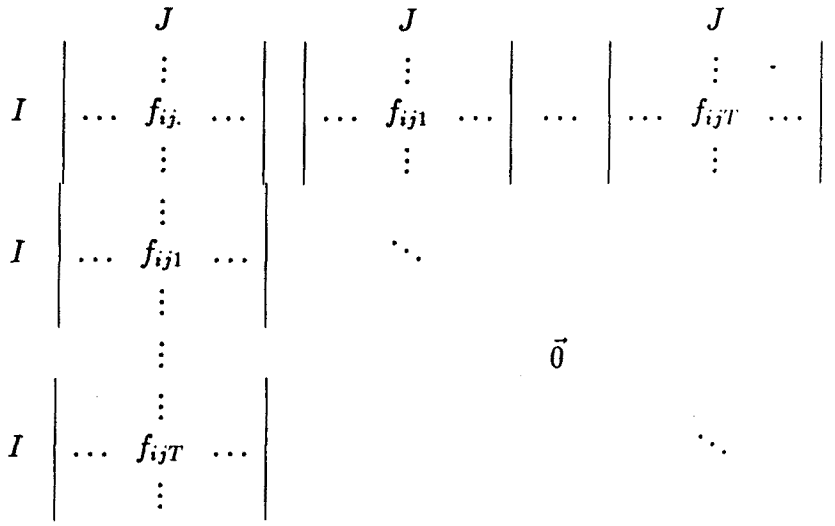
Parlem també dels tres marges monaris, per dir-ho així, obtinguts sumant sobre dos índexs: $f_{i..}, f_{.j.}, f_{..t}$.

Les anàlisis que B. Escofier i J. Pagès desenvolupen per a tractar la complexitat d'una taula ternària volen [Esc 88] orientar la reflexió sobre les pròpies taules ternàries i proposar eines adequades i adaptables als problemes que comporten.

5.4.2 Desenvolupament del mètode

El desenvolupament del mètode s'estructura en tres etapes:

- L'Anàlisi de Correspondències de la taula IxJ obtinguda sumant les T taules, posant-les aquestes, a més, com a línies i columnes suplementàries:



L'anàlisi de la suma de les T taules és l'anàlisi d'un núvol mitjana. Els factors principals posaran en evidència les tendències comunes a les T taules.

Posar les T taules com a suplementàries ens permet d'estudiar, a través de les projeccions, les relacions entre cada columna j de cada taula i el seu perfil mitjana obtingut de la taula suma. I igualment per a les files.

Fixem-nos, però, que en la construcció dels eixos principals només hi intervé la dispersió o la inèrcia entre grups; és a dir, del núvol mitjana només, i no gens la dispersió o inèrcia dintre dels grups determinats per cada taula t, $1 \leq t \leq T$.

- L'Anàlisi de Correspondències de la taula $Ix(JT)$ formada per la juxtaposició de les T taules:

$$\begin{array}{c} \\ I \end{array} \left| \begin{array}{c} J \\ \vdots \\ \cdots \; f_{ij1} \; \cdots \\ \vdots \end{array} \right| \left| \begin{array}{c} J \\ \vdots \\ \cdots \; f_{ij2} \; \cdots \\ \vdots \end{array} \right| \cdots \left| \begin{array}{c} J \\ \vdots \\ \cdots \; f_{ijT} \; \cdots \\ \vdots \end{array} \right|$$

Com a taula suplementària s'hi afegeix la taula suma tractada a l'anterior anàlisi i la taula de T columnes acumulant les categories de la variable J.[8]

La naturalesa d'aquesta taula és complexa i és necessari de completar els resultats amb nombrosos índexs d'ajuda a la interpretació [Esc 88](secc. 8.4).

Diguem, de totes maneres, que aquesta anàlisi representa bé les diferències quan les tendències comunes dels T grups no són predominants. L'anàlisi proporciona mesures adequades de la importància de les diferències entre, per exemple, la columna $(f_{ij_0t_1})_i$ i la columna $(f_{ij_0t_2})_i$; és a dir: a categoria j_0 igual, diferència en el comportament de la variable I entre el grup t_1 i el grup t_2.

L' anàlisi té un caràcter mixt que el configura la intervenció conjunta de dispersions i inèrcies entre i dintre dels grups.

[8]La juxtaposició pot fer-se també, diguéssim, encolumnant les taules. Els resultats d'una i altra juxtaposició són diferents.

- Finalment, una anàlisi en què només hi intervinguin les relacions i oposicions internes de cada grup. Es tracta, doncs, d'analitzar una taula en què s'hi hagin restat les d'entre grups.

  B. Escofier i J. Pagès [Esc 88] proposen de tractar la taula $IxJT$ de terme general:

  $$r_{ijt} = f_{ijt} - \frac{f_{ij.}f_{.jt}}{f_{i..}} + f_{i..}f_{.jt},$$

  que dóna com a distància al quadrat entre dues línies en l'Anàlisi de Correspondències usual[9]

  $$d^2(i,l) = \sum_{jt} \frac{1}{r_{.jt}} [(\frac{r_{ijt}}{r_{i..}}) - (\frac{r_{ljt}}{r_{i..}})]^2,$$

  l'expressió:

  $$\sum_{jt} (\frac{f_{ijt}}{f_{i..}} - \frac{f_{ljt}}{f_{l..}})^2 \frac{1}{f_{.jt}} - \sum_{j} (\frac{f_{ij.}}{f_{i..}} - \frac{f_{lj.}}{f_{l..}})^2 \frac{1}{f_{.j.}},$$

  en què es veu com a la distància només hi intervenen (en positiu, és clar), les diferències dintre de cada grup t.

L'anàlisi global que aquesta metodologia ens permet de fer posa de manifest la complexitat dels tractaments adequats en estudis com els que anem presentant.

5.5 Anàlisi de Correspondències No Simètrica

5.5.1 Principis

Anem a presentar breument la idea que ens porta a treballar amb anàlisis no simètriques -no simètriques en el sentit que les variables no juguen els papers intercanviables que clàssicament els els són assignats.

Les situacions que de fet ens porten a consideracions no simètriques sobre les variables són moltes i diverses.

[9] Anàlisi de la matriu $(\frac{r_{ijt}}{r_{i..}})_{i,jt}$, amb mètrica la matriu diagonal $JTxJT$ de terme general $\frac{1}{r_{.jt}}$ i matriu de pesos la matriu diagonal IxI de terme general $r_{i..}$.

Desenvoluparem l'Anàlisi de Correspondències No Simètrica($ACNS$) per a taules de freqüències [Dam 89] i només donarem indicacions per a l'anàlisi en tres variables.[10]

5.5.2 Desenvolupament de l'$ACNS$

Siguin dues variables qualitatives I i K, amb n i p categories respectivament. L'$ACNS$ pretén d'avaluar la influència de les p categories de K sobre I.

Sigui F_{IK} la taula usual de contingències de terme general:

$$f_{ik}, i = 1, \ldots, n, j = 1, \ldots, p,$$

tal que $\sum_{i=1}^{n} \sum_{k=1}^{p} f_{ik} = 1$, i siguin les distribucions marginals F_I i F_K amb termes generals, respectivament:

$$f_{i.} = \sum_{k} f_{ik} \;\; i \;\; f_{.k} = \sum_{i} f_{ik}.$$

L'$ACNS$ considera les p distribucions condicionades

$$\left(\frac{f_{ik}}{f_{.k}}\right)_{i=1,\ldots,n}$$

amb referència a la distribució F_I que representa l'absència d'influència de la variable K sobre la variable I. Vegem-ho.

Escrivim la matriu

$$F_{I/K} = F_{IK} D_K^{-1},$$

essent D_K la matriu diagonal pxp de terme general $f_{.k}$.

$F_{I/K}$ és una matriu les p columnes de la qual representen, cadascuna, una distribució

$$\left(\frac{f_{ik}}{f_{.k}}\right)_{i=1,\ldots,n}$$

[10] I encara, ens remetem directament a l'article citat per a anàlisis més complexes, com ara l'Anàlisi Parcial No Simètrica en què s'insereix la no simetria en l'estudi de relacions entre dues variables qualitatives controlant l'efecte d'una tercera. Igualment, ens remetem a Lauro i Siciliano (1988)[Lau 88] per a veure les relacions entre l'$ACNS$ i els models logístics.

sobre les categories de la variable I.

Aquestes columnes configuren un núvol de p elements d'$\mathcal{R}^n$ que és el que volem analitzar.

Considerem en $\mathcal{R}^n$ la mètrica identitat i considerem la matriu de pesos dels p elements del núvol D_K.

El centre de gravetat del núvol és el vector d'$\mathcal{R}^n$ de terme general

$$\sum_{k=1}^{p} f_{.k}\frac{f_{ik}}{f_{.k}} = \sum_{k=1}^{p} f_{.k} = f_{i.}$$

Així doncs, la inèrcia del núvol és:

$$\sum_{i=1}^{n}\sum_{k=1}^{p} f_{.k}\left(\frac{f_{ik}}{f_{.k}} - f_{i.}\right)^2.$$

L'$ACNS$ és l'anàlisi de correspondències usual respecte del centre de gravetat de la matriu $(F_{I/K})^t$ amb mètrica la identitat i matriu de pesos D_k.

Fixem-nos que l'anàlisi de correspondències usual d'una taula de freqüències

$$(f_{ik}) \begin{array}{l} i = 1,\ldots,n \\ k = 1,\ldots,p \end{array}$$

quan estudia les categories-columna, pot presentar-se com una anàlisi (de les files) de la matriu pxn $(F_{I/K})^t$ amb mètrica la matriu diagonal D_I^{-1} amb terme general $\frac{1}{f_{i.}}$, i amb matriu de pesos D_K.

Així doncs, la diferència entre l'$ACNS$ i l'usual està en la mètrica usada -que és la que dóna la simetria o no entre les variables.

Ens remetem a [Dam 89] per als càlculs i les interpretacions habituals en una Anàlisi Factorial Descriptiva. Només afegim que considerar la mètrica identitat en lloc de D_I^{-1} implica evitar de donar més importància al rol de categories amb freqüència $f_{i.}$ baixa.

5.5.3 Extensió a l'anàlisi de correspondències de tres factors

A l'apartat anterior acabem de veure que l'*ACNS* consisteix en l'anàlisi de la matriu *pxn*:

$$((\frac{f_{ik}}{f_{.k}})_{\substack{i=1,\ldots,n \\ k=1,\ldots,p}})^t,$$

amb mètrica per a les files la identitat I_n i matriu de pesos la matriu diagonal de terme general $f_{.k}, k = 1,\ldots,p$.

Òbviament, la distància entre dues files ve donada per:

$$d^2(k,k') = \sum_{i=1}^{n} (\frac{f_{ik}}{f_{.k}} - \frac{f_{ik'}}{f_{.k'}})^2.$$

Aleshores, l'extensió natural a l'estudi de tres variables I,K,J amb categories, respectivament, n,p,q, en el sentit d'analitzar l'estructura de dependència de la variable I respecte a les altres dues, ens porta a tractar la matriu de dimensions $(pq)xn$ amb files, una per a cada parell (k,j), els vectors d'$\mathcal{R}^n$:

$$\left(\frac{f_{ikj}}{f_{.kj}}\right)_{i=1,\ldots,n}$$

entenent f_{ikj} i $f_{.kj}$ de la manera usual a les taules de freqüències.

La mètrica considerada és la identitat I_n i la matriu de pesos la matriu diagonal $(pq)x(pq)$ de terme general $f_{.kj}, \forall (k,j), k = 1,\ldots,p; j = 1,\ldots,q$.

La distància entre dues files ve donada per:

$$d^2(kj,k'j') = \sum_{i=1}^{n} (\frac{f_{ikj}}{f_{.kj}} - \frac{f_{ik'j'}}{f_{.k'j'}})^2.$$

5.6 Les Anàlisis Condicionals o Locals

Anem a donar la idea dels estudis que aquestes anàlisis ens permeten de realitzar.

L'anàlisi descriptiva de dades consisteix en el tractament de grans quantitats de dades mitjançant, bàsicament, projeccions d'individus i variables en espais de dimensió reduïda, i per tant, mitjançant visualitzacions gràfiques de les relacions i oposicions.

Sovint hem de centrar la nostra atenció, però, sobre aspectes parcials o locals de les dades. Per exemple, en estudis de dades electorals, tal vegada ens interessin les observacions a part de la localització geogràfica dels electors, o ens interessi de mantenir fixades variables (que podem anomenar instrumentals [Rao 64]) com ara l'status sòcio-econòmic.

Són vàries les metodologies que ens permeten el control dels efectes de les variables que volem fixar -fins i tot anàlisis ja presentades ens ho permeten.

Volem, aquí, exposar breument l'Anàlisi de Correspondències Múltiples Condicional de B. Escofier [Esc 87], i tot seguit presentar la solució a través de grafs connectant els individus -solució que desenvoluparem de manera prou completa al capítol posterior.[11]

5.6.1 L'*ACM* Condicional

L'*ACM* Condicional[Esc 87] consisteix en:

Elements

- Una població I d'n individus.
- El conjunt J de les modalitats de Q variables qualitatives.
- La variable temps T amb un conjunt que també anomenem T de modalitats.
- La taula disjunta completa creuant els individus de la població I amb les modalitats J:

$$\begin{array}{c} \\ I \end{array} \begin{array}{c} \begin{array}{c} J \end{array} \\ \begin{pmatrix} \vdots \\ \ldots \quad k_{ij} \quad \ldots \\ \vdots \end{pmatrix} \end{array}$$

[11] Ja hem dit que ens interessa sobretot la metodologia per grafs de cara a comprendre millor les anàlisis del darrer capítol.

- La taula de creuament de la població I amb les modalitats de la variable T:

$$\begin{array}{c} \\ I \end{array} \begin{array}{c} T \\ \left(\begin{array}{ccc} \vdots & & \\ \ldots & k_{it} & \ldots \\ \vdots & & \end{array} \right) \end{array} \begin{array}{c} \vdots \\ \}I_t \\ \vdots \end{array}$$

essent I_t, $t = 1, \ldots, T$ les classes que la variable T indueix en la població I.

Fixem-nos:

$$\begin{array}{l} k_i = \sum_j k_{ij} = Q \\ k_j = \sum_i k_{ij} \\ k_t = card I_t = \sum_i k_{it} \\ f_i = \frac{k_i}{nQ} = \frac{1}{n} \\ \frac{f_{ij}}{f_i} = \frac{k_{ij}}{Q} \end{array}$$

El baricentre format pel núvol format pels k_t individus d'una classe I_t és:

$$\sum_{i \in I_t} \frac{Q}{Qk_t} \frac{k_{ij}}{Q} = \frac{b_{jt}}{Qk_t}$$

essent $b_{jt} = \sum_{i \in I} k_{ij} k_{it} = \sum_{i \in I_t} k_{ij}$.

Anàlisis

Les anàlisis que B. Escofier proposa són:

- Una anàlisi dels individus que, en lloc de la usual taula de terme general $\frac{f_{ij}}{f_i} = \frac{k_{ij}}{Q}$, tracta les línies de la taula IxJ de terme general:

$$\frac{k_{ij}}{Q} - \frac{b_{jt}}{Qk_t}$$

si $i \in I_t$.

Així s'aconsegueix d'eliminar en el núvol dels individus la dispersió deguda al temps, recentrant al baricentre de la classe I_t a la qual pertany l'individu.

- Una anàlisi de les modalitats que, en lloc d'analitzar, com és usual, la taula IxJ de terme general $\frac{f_{ij}}{f_j} = \frac{k_{ij}}{k_j}$, analitza:

$$\frac{k_{ij}}{k_j} - \frac{b_{jt}}{k_t k_j}$$

si $i \in I_t$,

que, al ser $\frac{b_{jt}}{k_j k_t}$ la coordenada de la projecció[12] de $\left(\frac{k_{ij}}{k_j}\right)_i$ sobre l'espai engendrat per $e_t \in \mathcal{R}^n$, indicador de la modalitat t de T, representa haver eliminat la part de distància induïda per T.

- Les dues anàlisis, com és habitual en l'Anàlisi de Correspondències, poden presentar-se com les anàlisis directa i dual d'una mateixa taula. En aquest cas, la taula IxJ de terme general:

$$k^*_{ij} = k_{ij} - \frac{b_{jt}}{k_t} + \frac{k_j}{n}$$

si $i \in I_t$.

És a dir:

dades − model + producte dels marges ponderat.

El model, és clar, tradueix la situació en què la relació entre I i J només es deu a la variable T[Esc 87]:

$$\begin{array}{c} \\ I \end{array} \begin{array}{c} J \\ \left(\begin{array}{ccc} & \vdots & \\ \cdots & \frac{b_{jt}}{k_t} & \cdots \\ & \vdots & \end{array} \right) \end{array} \begin{array}{c} \vdots \\ \}I_t \\ \vdots \end{array}$$

- Finalment, B. Escofier presenta l'ACM Condicinal com una anàlisi sobre una matriu de Burt IxJ[Esc 87] de terme general:

$$b^*_{jj'} = b_{jj'} - \sum_t \frac{b_{jt} b_{j't}}{k_t} + \frac{k_j k_{j'}}{n}$$

[12]Ortogonal per la mètrica nxn diagonal de terme general $1/f_i = n$.

essent $b_{jj'}$ el terme general de la matriu de Burt inicial associada a la taula $(k_{ij})_{ij}$.

5.6.2 La solució per grafs: l'Anàlisi Local

Aquesta solució consisteix en establir una relació a priori entre els individus de manera que connecti els que estiguin lligats pel factor a controlar. Això ho podrem fer a través d'un graf no orientat -l'estructura del qual detallarem al proper capítol.

Lebart [Leb 69] fou el primer a introduir aquesta anàlisi per a grafs de contigüitat en dades espaials i per això fou anomenada Anàlisi Factorial Local. Posteriorment, es generalitzà a grafs de similaritat entre individus [Al2 84] amb el nom d'Anàlisi Factorial Parcial perquè es basa en la mateixa idea de les variables instrumentals de Rao[Rao 64] i en l'anàlisi de correlacions parcials -sense haver de fer, però, hipòtesis probabilístiques[Alu 88]. Finalment, també ha estat generalitzada a grafs cronològics per Carlier[Car 85].

En definitiva, l'Anàlisi Local -que és aquest nom que finalment mantenim-, ens permet d'eliminar els efectes de factors que controlem i de visualitzar relacions i oposicions "a factor controlat", per dir-ho així, a través de la definició d'un graf entre

els individus.[13]

Anem a desenvolupar-la, aquesta anàlisi.

[13]B. Escofier generalitza l'ACM Condicional en[Esc 89], presentant també una solució que inclou grafs ponderats.

Sigui I el conjunt d'individus i $g_{ii'}$ el valor de veïnatge entre i i i' en un cert graf; $g_{ii'} = 0$ si i i i' no són veïns; $g_i = \sum_{i'} g_{ii'}$; en principi $\sum_{i'} g_{ii'} = \sum_i g_{ii'}$, però pot aplicar-se l'estudi a grafs més generals.

Sigui Z la taula de creuament usual IxJ, J conjunt de modalitats.

Aleshores, la taula:

$$\textbf{dades} - \textbf{model} + \textbf{producte dels marges ponderat},$$

que escrivim matricialment $R = Z - L + P$, considera ara la matriu model de terme_general:

$$l_{ij} = \sum_{i'} \frac{g_{ii'} k_{i'j}}{g_i},$$

que no és sinó la generalitzada de l'ACM Condicional (en què $g_{ii'} = 1$ si $i' \in$ {mateixa partició que i} i $g_i = k_t$).

Doncs bé, B. Escofier proposa les dues següents Anàlisis de Correspondències:

- de la taula L, que anomena ACM suau, i que al situar com a perfil de cada individu el del baricentre dels seus veïns, és una anàlisi entre classes, per dir-ho així, que elimina les variàncies locals;
- de la taula R, que anomena ACM de les diferències Locals, que estudia les variacions locals -és una anàlisi dintre de veïns.

Capítol 6

Anàlisi Local

6.1 Elements

Segons que dèiem al final del capítol anterior, secció 6.2, anem a definir els elements que ens permetran d'estudiar les relacions i oposicions locals o parcials entre individus i variables a través de grafs.

Sigui un conjunt d'n individus sobre els quals hi tenim avaluades p variables relacionats per un graf no orientat reflexiu i simètric, amb ells com a vèrtexs i eixos expressant la relació binària establerta que els connecta.

Sigui Q la matriu simètrica nxn associada al graf de manera que el seu terme general q_{ij} val 1 si i i j estan connectats per un eix del graf i 0 altrament.

Sigui R la matiu diagonal nxn dels graus de cada vèrtex de terme general $r_{ii} = \sum_j q_{ij}$.

Sigui $m = \sum_i r_{ii}$.

Sigui T la matriu n^2xn creuant eixos amb vèrtexs codificant amb un 1 en el lloc i i un -1 en el lloc j l'eix que connecta i amb j; si dos vèrtexs no estan connectats codifiquem amb zeros la fila corresponent en la matriu T, i igualment són files de tot zeros les corresponents als bucles.

Sigui, encara, E la matriu nxn tota plena d'uns. E és la matriu associada al graf

complet que connecta tots els individus.

Sigui B la matriu n^2xn creuant vèrtexs amb eixos del graf complet. Fixem-nos que la matriu dels graus d'aquest graf és nI_n, i aleshores el valor m corresponent és n^2.

Finalment, diguem que escrivim $\sum_{(i,i')\in G}$ quan sumem per a tots els eixos d'un graf G -entenent que una vegada és l'eix (i,i') i una altra l'eix (i',i) quan $i \neq i'$, i que, si cal, també hi ha l'eix (i,i).

Escrivim, com és usual, X la matriu nxp dels valors de les p variables sobre els n individus, M la mètrica en l'espai $\mathcal{R}^p$ i D la matriu diagonal nxn dels pesos p_i dels n individus.

Considerem ara, a més, la matriu diagonal n^2xn^2 L de pesos dels eixos, que no són sinó, aquests pesos, el producte dels corresponents als vèrtexs: $p_i p_{i'}$.

Enunciem el resultat que ens permetrà de fer les anàlisis adequades:

Teorema 1 *$R-Q$ és positiva simètrica i pot expressar-se $R-Q = 1/2\ T^tT$.*

En el graf complet resulta:

$$nI_n - U = \frac{1}{2}B^tB.$$

Anem a presentar, primerament, l'Anàlisi Local sobre variables numèriques; és a dir, l'Anàlisi en Components Principals Local($ACPL$), i des de dos punts de vista:

- com a una anàlisi de dades obtingudes per transformació de la matriu X original
- considerant una nova "mètrica" que sovint no serà sinó una semimètrica i mantenint la mateixa matriu inicial X.

Després presentarem l'Anàlisi de Cortrespondències Local(ACL).

6.2 L'*ACPL* com a transformació de la matriu inicial

Fixem-nos que T és l'operador que ens dóna les diferències entre els valors de les variables en individus connectats pel graf -és a dir, TX és la matriu n^2xp de les diferències; i T pot considerar-se una aplicació:

$$\mathcal{R}^n \xrightarrow{T} \mathcal{R}^{n^2}.$$

Aleshores, l'esquema de dualitat que ens formalitza la relació entre els espais adequats al nostre estudi local és el següent:

$$\begin{array}{ccc} \mathcal{R}^p & \xleftarrow{(TX)^t} & \mathcal{R}^{n^2*} \\ \downarrow M & & \uparrow L \\ \mathcal{R}^{p*} & \xrightarrow{TX} & \mathcal{R}^{n^2} \end{array}$$

Per tant, si considerem la matriu de pesos dels eixos, L, i aquests eixos com a "nous individus" a analitzar[1], podem dir que l'*ACPL* consisteix en l'*AFD* del triplet (TX, M, L).

La descomposició usual ens donarà:

$$TX = V\Lambda U^t$$

amb $V^tLV = U^tMU = I$.

L'*AFD* del triplet (TX, M, L) ens porta a la recerca del subespai H d'$\mathcal{R}^p$ que maximitzi[Alu 88]:

$$\| TXP_H^t \|_{LM}^2 = \sum_{(i,i')\in G} p_i p_{i'} d_{M\,H}^2(i,i'),$$

essent la matriu pxp, P_H, com és usual, l'operador projecció sobre H i $d_{M\,H}$ la M-distància dels individus projectats en H.

[1]Fixem-nos que ara les p variables-columnes de la matriu TX, per la definició de T, sempre estan centrades.

L'anàlisi, és clar, ens porta a diagonalitzar la matriu $A = M^{\frac{1}{2}} X^t T^t L T X M^{\frac{1}{2}}$.

Fixem-nos que aquesta matriu A, quan $M = I_p$ i $p_i = \frac{1}{\sqrt{m}}$ és $2V^l$ essent V^l la matriu de covariàncies local:

$$V^l = \frac{1}{m} X^t (R - Q) X$$

de terme general:

$$v^l_{jj'} = \frac{1}{2m} \sum_{(i,i') \in G} (x_{ij} - x_{i'j})(x_{ij'} - x_{i'j'}).$$

Quan el graf és complet

$$V^l = \frac{1}{n^2} X^t (nI - E) X$$

de terme general

$$v^l_{jj'} = \frac{1}{2n^2} \sum_{i,i'} (x_{ij} - x_{i'j})(x_{ij'} - x_{i'j'}) = \frac{1}{n} \sum_{i} (x_{ij} - \bar{x_j})(x_{ij'} - \bar{x_{j'}});$$

és a dir, V^l esdevé la matriu de covariàncies usual, és clar.

Encara, quan també $p_i = \frac{1}{\sqrt{m}}$ però $M = S^{-2}$, essent S^{-2}[Alu 80] la matriu diagonal pxp de terme general

$$\frac{1}{\frac{1}{n} \sum_i (x_{ij} - \bar{x_j})^2},$$

és a dir, la matriu diagonal de les inverses de les variàncies, s_j^2, de les p variables; aleshores, dèiem, la matriu A és $2C_{con}$ essent

$$C_{con} = \frac{1}{m} S^{-1} X^t (R - Q) X S^{-1}$$

la matriu de contigüitat de terme general

$$c_{jj'} = \frac{1}{2m} \sum_{(i,i') \in G} \frac{(x_{ij} - x_{i'j})}{s_j} \frac{(x_{ij'} - x_{i'j'})}{s_{j'}}.$$

La diagonal de la matriu C_{con} està formada pels coeficients de Geary [Gea 54] de les p variables

$$\frac{variància\ \ local\ \ \ sobre\ \ el\ \ graf\ \ de\ \ j}{variància\ \ (total)\ \ de\ \ j}.$$

Ara, quan el graf és complet,

$$C_{con} = \frac{1}{n^2} S^{-1} X^t (nI - E) X S^{-1},$$

és a dir, C_{con} no és sinó la usual matriu de correlacions.

Fixem-nos també que quan M és diagonal de terme general $m_j, j = 1, \ldots, p$,

$$tr(A) = \| TX \|^2_{LM} = \sum_j \sum_{(i,i')\in G} p_i p_{i'} m_j (x_{ij} - x_{i'j})^2,$$

que és la mesura del total de dispersió local, que es pot interpretar com una generalització del doble de la suma de les variàncies locals o de la suma dels coeficients de Geary de les variables.

En l'*ACPL*, és clar, les corresponents fórmules de transició esdevenen:

$$U = X^t T^t L V \Lambda^{-1}$$

$$V = TXMU\Lambda^{-1}.$$

Les coordenades dels eixos projectats[2], és a dir, de les files de la matriu TX projectades són $\Upsilon = TXMU = T\Psi$, on Ψ són les coordenades dels individus.

Les coordenades de les variables projectades són $\Phi = X^t T^t L V$.

6.3 L'*ACPL* com a semimètrica en $\mathcal{R}^n$

L'*ACPL* també pot presentar-se des del punt de vista de una Anàlisi Factorial Descriptiva del triplet $(X, M, T^t L T)$, però cal adonar-se que $T^t L T$ no és necessàriament una mètrica, encara que sí que sempre serà simètrica i positiva, és a dir, una semimètrica. Hem d'estructurar l'anàlisi de manera que puguem tractar amb aquesta semimètrica.

[2]Cal anar en compte en la interpretació de les projeccions dels eixos; s'ha de saber veure que precisament són això, "eixos connectant individus".

Si afegim al triplet (X, M, D) la matriu $n^2 x n$ T i la mètrica L en $\mathcal{R}^{n^2}$, tenim les següents relacions:

$$\begin{array}{ccccc} \mathcal{R}^p & \xleftarrow{X^t} & \mathcal{R}^{n*} & \xleftarrow{T^t} & \mathcal{R}^{n^2*} \\ \downarrow M & & \uparrow D & & \uparrow L \\ \mathcal{R}^{p*} & \xrightarrow{X} & \mathcal{R}^n & \xrightarrow{T} & \mathcal{R}^{n^2} \end{array}$$

Considerem que $rang(TX) = r$.

La descomposició de l'anàlisi del triplet (TX, M, L) ens dóna:

$$TX = V\Lambda U^t$$

amb $V^tLV = U^tMU = I$.

Sigui W el subespai generat per les files de la matriu TX; les columnes de la matriu U formen una base M-ortonormal d'aquest subespai.

Sigui $W_k =< U_k >$ el subespai generat per les k primeres columnes d'U.

Com ja sabem, el subespai W_k és óptim en el sentit que fa l'expressió

$$tr(P_S X^t T^t LTXM)$$

màxima entre tots els subespais S de W de dimensió k.

Considerem l'operador projecció sobre W, P_W, que pot escriure's com UU^tM; aquest operador ens permet de considerar la següent descomposició:

$$X = XP_W^t + X(I - P_W^t).$$

La projecció sobre W de qualsevol fila de TX és la mateixa fila; aleshores:

$$TXP_W^t = TX$$

i per tant

$$TX(I - P_W^t) = 0.$$

Considerem ara

$$XP_W^t = XMUU^t = \tilde{V}\Lambda U^t,$$

on $\tilde{V} = XMU\Lambda^{-1}$ i $\tilde{V}^t(T^tLT)\tilde{V} = I$.

És a dir, que tenim:

$$X = \tilde{V}\Lambda U^t + X(I - {}^tP_W)$$

amb $\tilde{V}^t\Delta\tilde{V} = U^tMU = I$, on Δ és la semimètrica T^tLT.

Usant la mateixa notació que en l'ACP, podem dir que

$$\hat{X}_k = \tilde{V}_k\Lambda_k{U^t}_k$$

és la matriu més a prop d'X en el sentit de la semimètrica Δ.

Remarquem que $\tilde{V}$ té *rang* r (si, com és usual, $r \leq n$).

Finalment, tenim $V = T\tilde{V}$; això ens posa de manifest que de fet les variables tenen les mateixes coordenades en els dos punts de vista que hem desenvolupat:

$$(TX)^tLV = X^t(T^tLT)\tilde{V}.$$

Òbviament, les coordenades dels individus són XMU mentre que les coordenades dels eixos són $TXMU$.

Bé, resumint, diguem que l'$ACPL$ consisteix en la descomposició del triplet (X, M, Δ), on $\Delta = T^tLT$ és la semimètrica induïda per la matriu T, obtinguda creuant eixos i vèrtexs, i $L = D \otimes D$ és la mètrica induïda per D sobre els eixos.

A més, recordem-ho, els dos desenvolupaments que hem presentat, ens porten a les mateixes projeccions.

6.4 Relació entre l'anàlisi local i la global

Sigui B la matriu n^2xn creuant eixos i vèrtexs en el graf complet. És fàcil de veure que l'$ACPL$ de BX és equivalent a l'ACP de la matriu X[Alu 88].

Concretament, en el cas $M = I_p$, l'anàlisi de (BX, M, D) amb D la matriu diagonal de terme general $1/n$, és equivalent a l'anàlisi de (X, M, L) essent L la matriu diagonal de terme general $1/n^2$.

Podem anomenar les columnes de TX variables locals i les columnes de BX variables globals, i referir-nos a l'$ACPL$ de BX com a l'anàlisi global.

Estem interessats en la relació entre les variables globals i les corresponents variables locals.

Per això definim la matriu de covariàncies entre les vaiables globals i les locals:

$$V_{g.l} = 1/2\frac{1}{n\sqrt{m}}X^tB^tTX = 1/2\frac{1}{n\sqrt{m}}X^tT^tTX = \frac{1}{n\sqrt{m}}X^t(R-Q)X = \frac{\sqrt{m}}{n}V_{local}$$

i, amb les notacions habituals, la matriu de correlacions:

$$C_{g.l} = S_g^{-1}V_{g.l}S_{local}^{-1}.$$

És a dir,

$$corr(j_g, j'_{local}) = \frac{cov(j_g, j'_{local})}{\sqrt{v(j_g)v(j'_{local})}} = \frac{\sqrt{m}}{m}\frac{cov(j_{local}, j'_{local})}{\sqrt{v(j_g)v(j'_{local})}}.$$

I per tant, la correlació entre una variable global i la seva corresponent local és:

$$corr(j_g, j_{local}) = \frac{\sqrt{m}}{n}\sqrt{\frac{v(j_{local})}{v(j_g)}} = \sqrt{\frac{\sum_{(i,i')\in G}(x_{ij}-x_{i'j})^2}{\sum_{i,i'}(x_{ij}-x_{i'j})^2}}$$

A més, ens interessa de visualitzar el canvi en les variables quan passem del tractament global a un tractament local projectant els dos tipus de variables en el mateix subespai de l'anàlisi global, considerant les locals com a suplementàries.

Les coordenades de les variables locals projectades sobre el factor principal $v_g \in \mathcal{R}^{n^2}$ de l'anàlisi global de les variables resulten les components de:

$$\phi_{local} = X^tT^tL_gv_g,$$

essent L_g la corresponent matriu de pesos dels eixos del graf complet:

$$L_g^{1/2} = diag(1/n) = \frac{\sqrt{m}}{n}diag(1/\sqrt{m}) = \frac{\sqrt{m}}{n}L_{local}^{1/2}.$$

Ens remetem a[Alu 88] per a veure interpretacions d'aquesta projecció simultània de variables locals i globals.

6.5 L'Anàlisi de Correspondències Simples Local

A l'ACS (secció 4.1.3), amb les notacions usuals, es tractava de diagonalitzar la matriu L de dimensions pxp de terme general:

$$l_{jj'} = \sum_{i=1}^{n} f_i(\frac{f_{ij}}{f_i\sqrt{f_j}} - \sqrt{f_j})(\frac{f_{ij'}}{f_i\sqrt{f_{j'}}} - \sqrt{f_{j'}}),$$

expressió que també pot escriure's:

$$1/2\sum_{i,i'} f_i f_{i'}(\frac{f_{ij}}{f_i\sqrt{f_j}} - \frac{f_{i'j}}{f_{i'}\sqrt{f_j}})(\frac{f_{ij'}}{f_i\sqrt{f_{j'}}} - \frac{f_{i'j'}}{f_{i'}\sqrt{f_{j'}}}).$$

Diagonalitzar L equival (secció 4.1.3) a diagonalitzar S_{pxp} de terme general:

$$s_{jj'} = \sum_{i=1}^{n} f_i \frac{f_{ij}}{f_i\sqrt{f_j}} \frac{f_{ij'}}{f_i\sqrt{f_{j'}}}.$$

Matricialment,

$$S = M^{1/2}F^tDFM^{1/2}$$

essent F la matriu de terme general $\frac{f_{ij}}{f_i}$, M la matriu diagonal de terme general $\frac{1}{f_j}$ i D la matriu diagonal de pesos de terme general f_i. És a dir, les matrius usuals del triplet que analitza l'ACS: (F, M, D).

La relació entre L i S ve donada per:

$$L = S - G^{1/2}G^{1/2^t}$$

essent $G^{1/2}$ el vector de dimensió p de terme general $\sqrt{f_j}$.

Doncs bé, l'Anàlisi de Correspondències Simples Local($ACSL$) analitza el triplet (TF, M, L^*), essent L^* la matriu diagonal n^2xn^2 de terme general $\frac{n^2}{m} f_i f_{i'}$.

Es tracta de diagonalitzar:

$$S_{local} = M^{1/2}F^tT^tL^*TFM^{1/2},$$

que té terme general:

$$\frac{n^2}{m}\sum_{(i,i')\in G} f_i f_{i'}\left(\frac{f_{ij}}{f_i\sqrt{f_j}} - \frac{f_{i'j}}{f_{i'}\sqrt{f_j}}\right)\left(\frac{f_{ij'}}{f_i\sqrt{f_{j'}}} - \frac{f_{i'j'}}{f_{i'}\sqrt{f_{j'}}}\right).$$

És a dir, resulta que S_{local} és dues vegades la matriu de covariàncies local.[3]

En definitiva, l'anàlisi local entesa com a anàlisi de les covariàncies locals, ens porta a l'anàlisi del triplet

$$(TX, M, diag(\frac{n^2}{m} f_i f_{i'})).$$

6.6 L'Anàlisi de Correspondències Múltiples Local

Finalment, presentem, breument, l'Anàlisi de Correspondències Múltiples Local ($ACML$) partint de les notacions i els resultats de l'ACM usual.

Recordem (secció 4.2.2) aquesta anàlisi des del punt de vista d'una Anàlisi de Correspondències d'una taula lògica o taula disjunta completa, Z, de q variables categòriques amb un total de $p = \sum_{s=1}^{q} q_s$ modalitats avaluades sobre n individus - taula formada per zeros i uns i el terme general de la qual escrivim k_{ij}.

Doncs bé, l'Anàlisi de Correspondències de la taula Z, ens porta a diagonalitzar (secció 4.1.3) la matriu usual

$$M^{1/2}F^tDFM^{1/2},$$

essent F, en la notació acostumada, la matriu nxp de terme general $\frac{f_{ij}}{f_i}$, que al nostre cas és $\frac{k_{ij}}{q}$; M la mètrica pxp diagonal de terme general $\frac{1}{f_j}$ que al nostre cas és $\frac{qn}{k_j}$, amb k_j el nombre d'individus que posseeixen la modalitat j; i essent, finalment, D, la matriu diagonal de pesos nxn de terme general f_i, que al nostre cas és $1/n$.

Si al fer l'anàlisi mantenim la inèrica respecte del centre de gravetat (secció 4.1.3),

[3]Surt dues vegades la matriu de covariàncies local perquè tenim la matriu T definida per a cada eix no orientat; és a dir, perquè $R - Q = 1/2T^tT$. Aquesta T, per altra banda, ens fa que les noves variables de la matriu TX sempre estiguin centrades, recordem-ho.

la matriu a diagonalitzar és

$$A = M^{1/2}F^tDFM^{1/2} - G^{1/2}G^{1/2^t},$$

essent G el vector de dimensió p de terme general f_j, és a dir, al nostre cas, $\frac{k_j}{qn}$.

Aquesta matriu A pot escriure's:

$$A = diag(\frac{1}{\sqrt{k_j}})\frac{1}{q}Z^tZdiag(\frac{1}{\sqrt{k_j}}) - G^{1/2}G^{1/2^t}.$$

Introduint les matrius associades al graf complet entre individus, resulta

$$A = \frac{1}{qn}diag(\frac{1}{\sqrt{k_j}})Z^t(nI - E)Zdiag(\frac{1}{\sqrt{k_j}}).$$

El terme general d'aquesta matriu és:

$$a_{jj'} = \sum_{i=1}^{n} \frac{k_{ij}k_{ij'}}{q\sqrt{k_jk_{j'}}} - \frac{\sqrt{k_jk_{j'}}}{nq},$$

que és igual a[Alu 86]:

$$\frac{1}{2nq\sqrt{k_jk_{j'}}} \sum_{i,i'} (k_{ij} - k_{i'j})(k_{ij'} - k_{i'j'}).$$

És clar, ara, en l'anàlisi local ens interessa, com sempre, no el tractament de la matriu d'inèrcia clàssica A, sinó la matriu que reculli només la inèrcia local, que ha de tenir terme general:

$$a_{jj'}^{local} = \frac{n}{2mq\sqrt{k_jk_{j'}}} \sum_{(i,i')\in G} (k_{ij} - k_{i'j})(k_{ij'} - k_{i'j'}).$$

Fixem-nos que, com ja hem vist a la secció anterior, la matriu local corresponent, a més de només sumar per als individus connectats pel graf, és clar, multiplica les variàncies per $\frac{n^2}{m}$ -valor que és igual a 1 en el graf complet.

En definitiva, com era d'esperar vista l'expressió matricial de la matriu A en funció de les matrius associades al graf complet, la matriu que l'$ACML$ tracta, amb tots els

resultats i totes les interpretacions habituals, és:

$$A^{local} = \frac{n}{qm} diag(\frac{1}{\sqrt{k_j}}) Z^t (R - Q) Z diag(\frac{1}{\sqrt{k_j}}),$$

essent, recordem-ho, R i Q les matrius de graus i associada, respectivament, al graf G que relaciona els individus segons el factor la influència del qual volem eliminar.

Capítol 7

Anàlisi Parcial Interna i Simultània

7.1 Introducció

Hem parlat de l'interès que tenen les anàlisis que tracten aspectes locals o parcials de les dades -anàlisis que eliminen els efectes de factors a controlar o fixar.

Aquests factors són exògens a l'anàlisi -és a dir, estudiem les relacions i oposicions entre certes variables controlant-ne d'altres externes a l'anàlisi.

Ara bé, tal vegada hi hagi relacions entre variables d'una anàlisi que siguin, per dir-ho així, redundants -que per exemple, dues variables tinguin una relació espúria[Sar 84] perquè es degui pràcticament del tot a la relació que tinguin amb una tercera variable, present també a l'anàlisi.

Volem detectar si tenim moltes relacions espúries; volem presentar una anàlisi que estudiï les relacions que hi ha entre variables eliminant els efectes que d'altres variables present també a l'anàlisi podrien causar provocant interpretacions enganyoses.

Tindríem una tria ideal del grup de variables a analitzar, si aquest grup presentés poques diferències entre una anàlisi usual (de la matriu de correlacions, per exemple, si es tracta de variables numèriques) i l'anàlisi que presentem -anàlisi en què les relacions són més fidedignes, més directes, estan més depurades; podríem trobar grups de variables ortogonals entre ells, i que inicialment no ho semblaven, d'ortogonals, i tal

vegada aconseguir de veure quines variables són les més representatives. Ens evitaríem relacions enganyoses i/o redundants.

I volem presentar aquesta anàlisi més adequada d'una manera simultània; és a dir, estudiant simultàniament totes les relacions entre variables eliminant els possibles efectes de les altres variables de l'anàlisi, en cada cas les que convinguin.

Presentem l'anàlisi primer per a variables numèriques, amb el tractament de correlacions parcials, però amb certa cura -cura que ens portarà a l'anàlisi d'una matriu transformada i a l'anàlisi equivalent dels residus d'adequats models lineals. Parlarem de l'Anàlisi en Components Principals Parcial Interna i Simultània.

I després presentarem l'Anàlisi de Correspondències Múltiples Parcial Interna i Simultània, que amb l'ajuda de tota una estructura de grafs -seguint doncs la idea de l'Anàlisi Local- ens porta a l'anàlisi d'una "nova matriu de Burt" depurada dels efectes interns que volem eliminar.

L'Anàlisi Parcial Interna i Simultània és una metodologia que, per comparació amb els resultats d'una anàlisi estàndar, ens permet de detectar relacions espúries i de millorar la tria de variables a analitzar. Presentem en aquesta Tesi la idea i els resultats bàsics, que en posteriors estudis podrem anar desenvolupant i millorant -i interpretant més acuradament.

7.2 L'*ACP* Parcial Interna i Simultània

7.2.1 Elements

L'*ACP*, tal i com l'hem presentada al capítol tercer d'aquesta Tesi, tracta la matriu X de dimensions nxp resultat de centrar i reduir una inicial matriu A dels valor que p

variables aleatòries prenen sobre n individus:

$$x_{ij} = \frac{a_{ij} - \frac{\sum_{i=1}^{n} a_{ij}}{n}}{\sqrt{\sum_{i=1}^{n} \frac{\left(a_{ij} - \frac{\sum_{i=1}^{n} a_{ij}}{n}\right)^2}{n}}}$$

Recordem que obteníem els factors (secció 3.2.3) principals u tals que fan màxima l'expressió

$$\sum_{i=1}^{n} \|(x_i^t u)u\|_c^2 = u^t X^t X u$$

diagonalitzant X^tX: u_1 és el vector propi de norma canònica igual a 1 associat al valor propi més gran d'X^tX, h_1; successivament tindríem els altres factors $u_s, s > 1$, ortonormals.

Recordem també que $(1/n)X^tX$ és la matriu de correlacions de les variables originals a^j, columnes de la matriu A.

De fet, l'anàlisi la presentàvem com l'AFD del triplet (Z, I_p, I_n) essent $Z = \frac{1}{\sqrt{n}}X$.

És clar que els vectors propis u_s d'X^tX associats a valors propis h_s són vectors propis de Z^tZ associats a valors propis h_s/n.

La mateixa relació hi ha amb els factors duals v_s, vectors propis ortonormals per la norma canònica d'XX^t i de ZZ^t.[1]

Recordem, finalment, que consideràvem les projeccions dels individus x_i (fila i de la matriu X) sobre els factors principals u_s:

- $F_s(i)$ és la coordenada de la projecció d'x_i sobre u_s;

i la projecció de les variables z^j (columna j de la matriu[2] Z) sobre els factors v_s:

- $G_s(j)$ és la coordenada de la projecció de z^j sobre v_s.

[1]Atenció que en l'anàlisi del triplet (Z, I_p, I_n) surten factors u_s i v_s, i en l'anàlisi de $(X, I_p, diag(1/n))$ vectors u_s i $v_s^* = \sqrt{n}v_s$. Ara bé, v_s també és vector propi d'XX^t.

[2]Consideràvem les variables $z^j = \frac{x^j}{\sqrt{n}}$ per a aconseguir bones interpretacions geomètriques de les variables. És clar que $corr(z^j, z^{j'}) = corr(x^j, x^{j'}) = corr(a^j, a^{j'})$.

Ara voldríem tractar la matriu de correlacions parcials en què cada terme $p_{jj'}$ fos la correlació parcial entre z^j i $z^{j'}$ respecte de les altres variables de l'anàlisi.

Anem a parlar, abans, una mica de la correlació parcial entre variables aleatòries.

- Siguin dues variables aleatòries v_1 i v_2; la correlació entre elles dues eliminant la influència que d'altres variables, $v_3, \ldots, v_m$, poden exercir damunt d'elles, s'obté calculant la correlació entre $\tilde{v}_1$ i $\tilde{v}_2$ essent $\tilde{v}_i = v_i - \hat{v}_i$, essent $\hat{v}_i, i = 1, 2$, la variable predicció de v_i per l'hiperplà de regressió amb variables explicatives $v_3, \ldots, v_m$. Aquesta correlació s'anomena correlació parcial entre v_1 i v_2 respecte de les altres variables, i s'escriu $r_{12/3\ldots m}$.

- Si només hi ha una variable pertorbadora, resulta[Cua 81]:

$$r_{12/3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}},$$

essent r_{ij} la correlació usual.

- La fórmula és recorrent, i per a $m - 2$ variables pertorbadores, resulta:

$$r_{12/34\ldots m} = \frac{r_{12/34\ldots(m-1)} - r_{1m/34\ldots(m-1)}r_{2m/34\ldots(m-1)}}{\sqrt{1 - r_{1m/34\ldots(m-1)}^2}\sqrt{1 - r_{2m/34\ldots(m-1)}^2}}.$$

- Si a més, $P^2_{1/234\ldots m}$ és el coeficient de determinació entre v_1 i les variables $v_2, v_3, \ldots, v_m$, resulta[Cua 81]:

$$1 - P^2_{1/234\ldots m} = (1 - r_{12}^2)(1 - r_{13/2}^2)(1 - r_{14/23}^2)\ldots(1 - r_{1m/234\ldots(m-1)}^2).$$

Bé, amb tots els elements fins ara presentats, anem a veure quina anàlisi podem realitzar sobre les correlacions parcials de les variables-columna de la matriu X (o Z), essent, és clar, cada vegada diferents les variables la influència de les quals es vol eliminar.

7.2.2 L'*ACP* Parcial Interna i Simultània: Transformació de la matriu inicial

Sigui Z la matriu nxp de les variables centrades i normalitzades per la norma canònica z^j.

La matriu de correlacions entre les variables z^j és[3] Z^tZ.

Considerem que $rang Z^tZ = p$, altrament, és conegut que el conjunt de punts que caracteritzen les variables (les p columnes de la matriu Z), estan continguts en un subespai de dimensió $r < p$ i aleshores, $p - r$ variables-columna són combinació lineal de les altres, la qual cosa no té cap interès.

Considerem la matriu pxp $C_{parcial}$ de correlacions parcials de terme general

$$p_{jj'} = corr(z^j - \hat{z}^j, z^{j'} - \hat{z}^{j'}),$$

essent $\hat{z}^j$ i $\hat{z}^{j'}$ les prediccions respectivament de z^j i $z^{j'}$ per les $p-2$ restants variables presents a l'anàlisi.

Resulta[Whi 88]:

$$C_{parcial} = -(D(Z^tZ)^{-1}D) + I_p,$$

essent D la matriu pxp diagonal construïda amb els inversos de les arrels quadrades dels elements de la diagonal de $(Z^tZ)^{-1}$ i I_p la identitat en $\mathcal{R}^p$.

Els elements diagonals de $(Z^tZ)^{-1}$ tenen la forma[Dil 84]:

$$\Pi_{jj} = \frac{1}{1 - P^2_{j/altres}},$$

essent $P^2_{j/altres}$ el coeficient de determinació del model lineal de la variable z^j en funció de les altres $p-1$ variables de l'anàlisi; $P^2_{j/altres} < 1$ doncs considerem que el $rang Z^tZ$ és màxim.

Si bé $C_{parcial}$ no és definida positiva ho és

$$\tilde{C}_{parcial} = D(Z^tZ)^{-1}D,$$

[3]Recordem que seria $1/nX^tX$ amb X la matriu de les variables centrades i reduïdes.

la qual només es diferencia de $C_{parcial}$ en el signe de les correlacions no trivials (no diagonals).

La interpretació de $\tilde{C}_{parcial}$ és evident.

Anem a veure que, efectivament, $\tilde{C}_{parcial} = T^tT$, essent T una matriu nxp de $rang = p$.[4]

Enunciem doncs:

Teorema 7.1

$$\tilde{C}_{parcial} = T^tT,$$

amb

$$T = ZU\Lambda^{-2}U^tD,$$

essent U la matriu pxp les columnes de la qual la formen els p vectors propis ortonormals per la norma canònica de la matriu Z^tZ -és a dir, la matriu dels factors principals de l'anàlisi usual de la matriu Z-; i Λ la matriu diagonal formada per les arrels quadrades dels valors propis de Z^tZ.

Efectivament, demostrem-ho amb tres punts:

1.De la descomposició en valors singulars de la matriu Z en l'AFD del triplet (Z, I_p, I_n), resulta:

$$Z = V\Lambda U^t,$$

- U matriu pxp de columnes formant una base ortonormal per la norma canònica de les files de la matriu Z;

[4]Sempre, és clar, $n \leq p = rangZ = rangZ^tZ$.

- V matriu nxp de columnes formant una base ortonormal per la base canònica de les columnes de la la matriu Z;
- Λ la matriu diagonal pxp dels valors singulars, és a dir, le arrels quadrades dels valors propis de la matriu Z^tZ.

2.Resulta:

$$(Z^tZ)^{-1} = (U\Lambda V^tV\Lambda U^t)^{-1} =$$

$$(U\Lambda\Lambda U^t)^{-1} = (U^t)^{-1}\Lambda^{-2}U^{-1} =$$

$$U\Lambda^{-2}U^t$$

doncs U és matriu ortogonal.

Per tant:

$$(Z^tZ)^{-1} = U\Lambda^{-1}V^{-1}V\Lambda^{-1}U^t = U\Lambda^{-1}V^tV\Lambda^{-1}U^t = (V\Lambda^{-1}U^t)^tV\Lambda^{-1}U^t = S^tS,$$

essent $S_{nxp} = V\Lambda^{-1}U^t$.

3.La matriu S s'obté de la matriu Z multiplicant per $U\Lambda^{-2}U^t$:

$$S = ZU\Lambda^{-2}U^t.$$

Per tant, tenim, si $T = SD$:

$$\tilde{C}_{parcial} = D(Z^tZ)^{-1}D = DS^tSD = (SD)^tSD = T^tT,$$

amb $T = ZU\Lambda^{-2}U^tD$.

Anem ara a estudiar la matriu S, i també la matriu T.

Fixem-nos que la matriu S representa una nova matriu de dades -p noves s^j variables avaluades sobre els n individus. Com que s'obté de la inicial Z multiplicant per $U\Lambda^{-2}U^t$, resulta:

- Agafar Z i considerar ZU vol dir considerar l'expressió dels n individus z_i en la base ortonormal d'$\mathcal{R}^p$ dels factors principals $u_1, \ldots, u_p$. És a dir, teníem $z_i \in \mathcal{R}^p$, amb la base canònica, i els transformem en les files de ZU.

  Òbviament, si els vectors z_i sumaven $\vec{0} \in \mathcal{R}^p$, també sumaran $\vec{0} \in \mathcal{R}^p$ les n files de ZU.

- Considerar $ZU\Lambda^{-2}$, és dividir cada columna s de la matriu ZU per $\frac{1}{\lambda_s}$, essent λ_s valor propi de Z^tZ.

- Considerar $ZU\Lambda^{-2}U^t$ vol dir retornar a expressar els individus respecte de la base canònica d'$\mathcal{R}^p$.

Fixem-nos, doncs, que $S = ZU\Lambda^{-2}U^t$ està formada per p variables-columna centrades.

Quant a les variàncies empíriques:

- Tenint en compte que la columna s-èsima de ZU està formada pels valors $\frac{F_s(i)}{\sqrt{n}}$ (secció 3.2.3), resulta:

$$\sum_{i=1}^{n} 1/n \frac{F_s^2(i)}{n} = \frac{\lambda_s}{n},$$

  que és la variància empírica d'aquesta variable-columna s-èsima.

- La variància de la variable-columna s-èsima de $ZU\Lambda^{-2}$ val $\frac{1}{n\lambda_s}$.

- Calculem la variància empírica de la variable s^j, columna j-èsima de $S = ZU\Lambda^{-2}U^t$:

$$1/n \sum_{i=1}^{n} s_{ij}^2 = 1/n \sum_{i=1}^{n} \left(\sum_{s=1}^{p} \frac{F_s(i)}{\sqrt{n}\lambda_s} u_{sj}\right)^2 = 1/n \sum_{s=1}^{p} u_{sj}^2 \left(\sum_{i=1}^{n} \frac{(F_s(i))^2}{n\lambda_s^2}\right) = 1/n \sum_{s=1}^{p} \frac{u_{sj}^2}{\lambda_s},$$

  essent u_{sj} la component j-èsima del vector u_s.

  Fixem-nos ara en el vector diagonal de la matriu $(Z^tZ)^{-1}$:

$$\vec{v}_{diag}((Z^tZ)^{-1}) = (\Pi_{11}, \ldots, \Pi_{pp})^t.$$

Sabem:

$$U^tZ^tZU = U^tU\Lambda V^tV\Lambda U^tU = \Lambda^2 = diag(\lambda_s).$$

Per tant:

$$U^t(Z^tZ)^{-1}U = diag(\frac{1}{\lambda_s}).$$

Per tant:

$$(Z^tZ)^{-1} = Udiag(\frac{1}{\lambda_s})U^t.$$

Per tant:

$$\Pi_{jj} = \sum_{s=1}^{p} \frac{u_{sj}^2}{\lambda_s}.$$

En definitiva, ha resultat que la variància empírica de la variable-columna s^j és:

$$\frac{\Pi_{jj}}{n}.$$

Tenim doncs la matriu S nxp amb columnes s^j centrades i de variància empírica $\frac{\Pi_{jj}}{n}$.

Si ara considerem la matriu S_N normalitzada de la S, amb columnes t_j de terme general:

$$\frac{s_j}{\sqrt{\frac{\Pi_{jj}}{n}}\sqrt{n}} = \frac{s_j}{\sqrt{\Pi_{jj}}},$$

resulta $S_N = T = SD$, i, aleshores, és clar:

$$S_N^tS_N = T^tT = diag(\frac{1}{\sqrt{\Pi_{jj}}})S^tSdiag(\frac{1}{\sqrt{\Pi_{jj}}}) = D(Z^tZ)^{-1}D = \tilde{C}_{parcial}.$$

En definitiva, $\tilde{C}_{parcial}$ és definida positiva i s'obté fent T^tT on les columnes t_j de la matriu T representan els valors d'unes noves variables normalitzades sobre els n individus; la matriu de correlacions de les variables t_j és $T^tT = \tilde{C}_{parcial}$.

Proposem doncs, de tractar T en lloc de la matriu inicial Z; és a dir, proposem de fer l'anàlisi del triplet (T, I_p, I_n) en lloc del triplet (Z, I_p, I_n) i així, de fet, analitzarem les correlacions parcials de les variables inicials.

7.2.3 L'*ACP* Parcial Interna i Simultània com a Anàlisi de Residus

Acabem de veure que l'anàlisi que presentem tracta les dades inicials transformades de la següent manera:

- Canvi de coordenades per a expressar les dades en funció dels factors principals dels individus.
- Ponderació pels valors propis.
- Retornar als eixos inicials canònics.
- Normalitzar per la norma canònica.

Anem a veure més clarament què representa la matriu T, nova matriu de dades, obtenint-la a partir dels adequats models lineals.

Sigui ara $\hat{z}^j$ la predicció lineal de z^j per les altres variables presents també en l'anàlisi -totes les altres $p-1$ variables.

Considerem els residus

$$\varepsilon^j = z^j - \hat{z}^j,$$

noves variables avaluades sobre els n individus.

Considerem E la matriu nxp d'aquestes variables ε^j.

Aleshores, tenim el resultat:

Teorema 7.2

$$S = ED^{-2}$$

Efectivament, demostrem-ho també en tres punts:

1.En un model lineal d'y en funció de $p-1$ regressores

$$y = a_1x_1 + a_2x_2 + \ldots + a_px_p + e,$$

amb $y, x_i \in \mathcal{R}^n, a_i \in R$, i una, qualsevol, $x_{i_0} = (1, \ldots, 1) \in \mathcal{R}^n$ que ens proporciona el terme independent a_{i_0}, tenim els residus empírics:

$$e = y - (a_1x_1 + a_2x_2 + \ldots + a_px_p) = y - Xa,$$

essent X la matriu nxp de columnes x_i i a el vector $(a_1, \ldots, a_p)^t$.

Resulta[Le2 85]:

$$a = (X^tX)^{-1}X^ty,$$

i per tant:

$$e = y - X(X^tX)^{-1}X^ty.$$

2.Anem, al nostre cas, a calcular els residus d'ajust d'una variable z^j, columna de la matriu Z, amb les altres columnes com a regressores.

Sigui E_j una matriu pxp tota de zeros menys l'element (j,j), que val 1.

Sigui e_j el vector d'$\mathcal{R}^p$ tot de zeros menys la component j que val 1.

Sigui U la matriu nxp tota d'uns.

Resulta:

- $z^j = Ze^j$.
- $Z-(Z-U)E_j$ és una matriu nxp formada substituint la columna j de la matriu Z per un vector tot d'uns.

Per tant, els residus de l'ajust considerat s'obtenen:

$$Ze_j - (Z-(Z-U)E_j)(Z-(Z-U)E_j)^t(Z-(Z-U)E_j)^{-1}(Z-(Z-U)E_j)^tZe_j.$$

3.Anem a veure que aquesta última epressió val

$$Z(Z^tZ)^{-1}D^2e_j,$$

és a dir,

$$SD^2e_j.$$

És a dir, la variables-columna s^j de la matriu S dividida per Π_{jj} (secció 7.2.2).

Efectivament:

- Fixem-nos que pel fet de tenir les variables centrades,

$$(Z-(Z-U)E_j)^t(Z-(Z-U)E_j)$$

és la matriu Z^tZ substituint els elements (j,k) i (k,j), $k=1,\ldots,p$, $k\neq j$, per zeros, i l'element (j,j) per n.

És a dir, $(Z-(Z-U)E_j)^t(Z-(Z-U)E_j)$ és la matriu A_j de correlacions de les variables $z^k, k=1,\ldots,p, k\neq j$, intercalant com a fila j el vector d'$\mathcal{R}^p$

$$(0,\ldots,0,n,0,\ldots,0),$$

n en el lloc j; i, igualment, intercalant com a columna j el mateix vector.

Fixem-nos que si escrivim c_{sj} la correlació entre les variables z^s i z^j, tenim:

$$A_j=\begin{pmatrix} c_{11} & \cdots & c_{1j-1} & c_{1j+1} & \cdots & c_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ c_{j-11} & \cdots & c_{j-1j-1} & c_{j-1j+1} & \cdots & c_{j-1p} \\ c_{j+11} & \cdots & c_{j+1j-1} & c_{j+1j+1} & \cdots & c_{j+1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ c_{p1} & \cdots & c_{pj-1} & c_{pj+1} & \cdots & c_{pp} \end{pmatrix}$$

- Aleshores, es veu fàcilment que $(Z-(Z-U)E_j)^t(Z-(Z-U)E_j)^{-1}$ és la matriu A_j^{-1} intercalant com a fila j i com a columna j el vector d'$\mathcal{R}^p$

$$(0,\ldots,0,1/n,0,\ldots,0),$$

$1/n$ en el lloc j.

- També resulta:

$$(Z-(Z-U)E_j)^tZe_j = \begin{pmatrix} c_{1j} \\ \vdots \\ c_{j-1j} \\ 0 \\ c_{j+1j} \\ \vdots \\ c_{pj} \end{pmatrix}$$

- Per tant,

$$(Z-(Z-U)E_j)^t(Z-(Z-U)E_j)^{-1}(Z-(Z-U)E_j)^tZe_j =$$

$$(Z-(Z-U)E_j)^t(Z-(Z-U)E_j)^{-1}\begin{pmatrix} c_{1j} \\ \vdots \\ c_{j-1j} \\ 0 \\ c_{j+1j} \\ \vdots \\ c_{pj} \end{pmatrix}$$

Expressió que resulta ser un vector d'$\mathcal{R}^p$ amb component j-èsima igual a zero i les altres

$$-\frac{A_{kj}}{detA_j}, k = 1,\ldots,p, k \neq j,$$

essent A_{kj} l'adjunt de l'element (k,j) en la matriu de correlacions Z^tZ.

Efectivament, vegem aquesta última afirmació:

Sigui $k \in \{1,\ldots,p\}, k \neq j$.

Considerem la fila k-èsima del producte que estem calculant:

$$\sum_{s=1,s\neq j}^{p} \frac{\alpha_{ks}}{detA_j} c_{sj}$$

essent α_{ks} els adjunts dels valors c_{ks} en la matriu A_j -atenció! que el valor c_{ks} no ocupa sempre el lloc (k,s) en aquesta matriu A_j.

Anem a veure que

$$\sum_{s=1,s\neq j}^{p} \alpha_{ks} c_{sj} = -A_{kj}.$$

Efectivament, siguin les matrius:[5]

$$A_j = \begin{pmatrix} c_{11} & \cdots & c_{1j-1} & c_{1j+1} & \cdots & c_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ c_{j-11} & \cdots & c_{j-1j-1} & c_{j-1j+1} & \cdots & c_{j-1p} \\ c_{j+11} & \cdots & c_{j+1j-1} & c_{j+1j+1} & \cdots & c_{j+1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ c_{p1} & \cdots & c_{pj-1} & c_{pj+1} & \cdots & c_{pp} \end{pmatrix}$$

$$Z^t Z = \begin{pmatrix} c_{11} & \cdots & c_{1j-1} & c_{1j} & c_{1j+1} & \cdots & c_{1p} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ c_{j-11} & \cdots & c_{j-1j-1} & c_{j-1j} & c_{j-1j+1} & \cdots & c_{j-1p} \\ c_{j1} & \cdots & c_{jj-1} & c_{jj} & c_{jj+1} & \cdots & c_{jp} \\ c_{j+11} & \cdots & c_{j+1j-1} & c_{j+1j} & c_{j+1j+1} & \cdots & c_{j+1p} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ c_{p1} & \cdots & c_{pj-1} & c_{pj} & c_{pj+1} & \cdots & c_{pp} \end{pmatrix}$$

Fixem-nos que, amb $k \neq j$,

$$A_{kj} = (-1)^{k+j} \begin{vmatrix} c_{11} & \cdots & c_{1j-1} & c_{1j+1} & \cdots & c_{1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ c_{k-11} & \cdots & c_{k-1j-1} & c_{k-1j+1} & \cdots & c_{k-1p} \\ c_{k+11} & \cdots & c_{k+1j-1} & c_{k+1j+1} & \cdots & c_{k+1p} \\ \vdots & & \vdots & \vdots & & \vdots \\ c_{p1} & \cdots & c_{pj-1} & c_{pj+1} & \cdots & c_{pp} \end{vmatrix}$$

Desenvolupant el determinant per la fila de les correlacions amb la variable j, queda

$$(-1)^{k+j}(c_{j1}(-1)^{\eta_1}|\alpha_{k1}| + \ldots + c_{jj-1}(-1)^{\eta_{j-1}}|\alpha_{kj-1}| +$$

$$+c_{jj+1}(-1)^{\eta_{j+1}}|\alpha_{kj+1}| + \ldots + c_{jp}(-1)^{\eta_p}|\alpha_{kp}|)$$

essent $\eta_s, s = 1, \ldots, p, s \neq j$, la paritat en A_{kj} de l'element c_{js}.

Anem a estudiar la paritat:

[5]Cal considerar, quan sigui necessari, que $c_{sk} = c_{ks}, \alpha_{sk} = \alpha_{ks}, A_{sk} = A_{ks}$.

- La fila k té la paritat del valor k en Z^tZ.
- La columna j té la paritat del valor j en Z^tZ.
- L'element $c_{js}, s = 1, \ldots, p, s \neq j$, té paritat-fila en A_{kj} igual a la del valor j si $j < k$; si $j > k$ igual a la del valor $j - 1$.
- L'element $c_{ks}, s = 1, \ldots, p, s \neq j$, té paritat-fila en A_j igual a la del valor k si $j > k$; si $j < k$ igual a la del valor $k - 1$.
- La paritat-columna de c_{js} en A_{kj} i de c_{ks} en A_j és la mateixa.

  Per tant:

  * si els valors k i j tenen igual paritat, resulta:

$$\sum_{s=1, s\neq j}^{p} (-1)^{\eta_s} |\alpha_{ks}| \, c_{js} = A_{kj},$$

    essent η_s la paritat de c_{js} en A_{kj} que sempre és diferent a la paritat de c_{ks} en A_j, i per tant:

$$A_{kj} = \sum_{s=1, s\neq j}^{p} (-1)\alpha_{ks} c_{js};$$

  * si els valors k i j tenen diferent paritat, resulta:

$$\sum_{s=1, s\neq j}^{p} (-1)^{\eta_s} |\alpha_{ks}| \, c_{js} = -A_{kj},$$

    i com que la paritat de c_{js} en A_{kj} és aleshores igual a la paritat de c_{ks} en A_j, resulta:

$$A_{kj} = -\sum_{s=1, s\neq j}^{p} \alpha_{ks} c_{js}.$$

En definitiva, sempre tenim:

$$A_{kj} = -\sum_{s=1, s\neq j}^{p} \alpha_{ks} c_{js}.$$

Per tant, és clar:

$$(Z-(Z-U)E_j)^t(Z-(Z-U)E_j)^{-1}(Z-(Z-U)E_j)^tZe_j = \begin{pmatrix} -\frac{A_{1j}}{detA_j} \\ \vdots \\ -\frac{A_{j-1j}}{detA_j} \\ 0 \\ -\frac{A_{j+1j}}{detA_j} \\ \vdots \\ -\frac{A_{pj}}{detA_j} \end{pmatrix}$$

que era el que volíem demostrar.

- Resumint, hem vist que el residu:

$$\varepsilon^j = z^j - \hat{z}^j$$

val

$$Ze_j - (Z-(Z-U)E_j)\begin{pmatrix} -\frac{A_{1j}}{detA_j} \\ \vdots \\ -\frac{A_{j-1j}}{detA_j} \\ 0 \\ -\frac{A_{j+1j}}{detA_j} \\ \vdots \\ -\frac{A_{pj}}{detA_j} \end{pmatrix}$$

Aquesta expressió, tenint en compte què és la matriu $Z-(Z-U)E_j$, és igual a:

$$Z\begin{pmatrix} \frac{A_{1j}}{detA_j} \\ \vdots \\ \frac{A_{j-1j}}{detA_j} \\ 1 \\ \frac{A_{j+1j}}{detA_j} \\ \vdots \\ \frac{A_{pj}}{detA_j} \end{pmatrix}$$

- Finalment, vegem que també $Z(Z^tZ)^{-1}D^2e_j$ val la mateixa expressió.

Recordem que D^2 és la matriu pxp diagonal amb els inversos dels elements de la diagonal de $(Z^tZ)^{-1}$, elements que escrivim Π_{jj}.

Resulta:

$$\Pi_{jj} = \frac{detA_j}{det(Z^tZ)},$$

per la pròpia definició de matriu inversa.

Per tant,

$$(Z^tZ)^{-1}D^2e_j = \begin{pmatrix} \frac{A_{11}}{det(Z^tZ)} & \cdots & \frac{A_{1p}}{det(Z^tZ)} \\ \vdots & & \vdots \\ \frac{A_{p1}}{det(Z^tZ)} & \cdots & \frac{A_{pp}}{det(Z^tZ)} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{det(Z^tZ)}{detA_j} \\ 0 \\ \vdots \\ 0 \end{pmatrix} =$$

$$\begin{pmatrix} \frac{A_{1j}}{detA_j} \\ \vdots \\ \frac{A_{j-1j}}{detA_j} \\ 1 \\ \frac{A_{j+1j}}{detA_j} \\ \vdots \\ \frac{A_{pj}}{detA_j} \end{pmatrix}$$

doncs $A_{jj} = detA_j$.

I per tant,

$$Z(Z^tZ)^{-1}D^2e_j = Z \begin{pmatrix} \frac{A_{1j}}{detA_j} \\ \vdots \\ \frac{A_{j-1j}}{detA_j} \\ 1 \\ \frac{A_{j+1j}}{detA_j} \\ \vdots \\ \frac{A_{pj}}{detA_j} \end{pmatrix}$$

com volíem demostrar.

Hem vist, doncs, que efectivament

$$\varepsilon^j = z^j - \hat{z}^j = Z(Z^tZ)^{-1}D^2e_j,$$

i que per tant, com enunciava el teorema:

$$E = SD^2.$$

Tenim doncs, E una matriu nxp de columnes ε^j, variables avaluades sobre els n individus, centrades i de variància empírica:

$$\sum_{i=1}^{n}(1/n)\varepsilon_{ij}^2 = \sum_{i=1}^{n} 1/n\frac{s_{ij}^2}{\Pi_{jj}^2} =$$

$$\sum_{i=1}^{n} 1/n\frac{s_{ij}^2}{\Pi_{jj}}\frac{1}{\Pi_{jj}} = \frac{1}{n}\frac{1}{\Pi_{jj}}.$$

Normalitzant E, resulta:

$$E_{normalitzada} = \frac{1}{\sqrt{n}}Ediag(\sqrt{n\Pi_{jj}}) = ED^{-1} = SD^2D{-}1 = SD = T.$$

És a dir, la matriu T és la matriu normalitzada de les dades -residu E i la matriu de correlacions de les variables ε^j és $T^tT = \tilde{C}_{parcial}$.

El triplet (T, I_p, I_n) té, doncs, aquesta altra interpretació.[6]

Com dèiem, és la comparació de les anàlisis de (Z, I_p, I_n) i de (T, I_p, I_n) que ens ha de portar a unes més depurades i fidedignes interpretacions de les variables que estudiem.

[6]Hem vist, de fet, una generalitzció de $corr(x, y) = -corr(x - \hat{x}, y - \hat{y})$, esent $\hat{x}$ l'ajust lineal d'x per y i $\hat{y}$ el d'y per x, igualtat que es demostra fàcilment a partir de les definicions i que és certa quan $corr^2(x, y) \neq 1$, és clar.

7.3 L'*ACM* Parcial Interna i Simultània

7.3.1 Elements

L'*ACM* usual pot presentar-se (secció 4.2.2) com l'anàlisi de la taula disjunta completa Z de q variables categòriques amb un total de p modalitats, avaluades sobre n individus.

L'anàlisi desemboca en la diagonalització de la matriu:

$$\frac{1}{q}M^{1/2}Z^tZM^{1/2} - G^{1/2}(G^{1/2})^t,$$

essent $M^{1/2}$ la matriu diagonal pxp de terme general l'invers de l'arrel quadrada del nombre d'individus posseint la modalitat j; és a dir: $\frac{1}{\sqrt{k_j}}$, i G el vector d'$\mathcal{R}^p$ de terme general $\frac{k_j}{qn}$.

Precisament, l'*ACM* es presenta també (secció 4.2.3) com l'anàlisi de la matriu B pxp de Burt, matriu de creuament de totes les variables amb totes. Fixem-nos que $B = Z^tZ$, i que pot escriure's:

$$B = \begin{bmatrix} & \vdots & \\ \cdots & (Z_s^tZ_r) & \cdots \\ & \vdots & \end{bmatrix}_{s,r=1,\ldots,q}$$

essent $Z_s, s = 1,\ldots,q$, la corresponent matriu disjunta completa sobre els n individus de la variable categòrica s.

És a dir, $Z_s^tZ_r$ és la taula de contingència de les característiques s i r..

Recordem que la matriu a diagonalitzar pot escriure's (secció 6.6):

$$\frac{n}{qn^2}M^{1/2}Z^t(nI - E)ZM^{1/2},$$

essent I la identitat en $\mathcal{R}^n$ i E una matriu nxn tot d'uns; i recordem que això ens permetia d'introduir fàcilment l'Anàlisi Local (secció 6.6) que diagonalitza:

$$\frac{n}{mq}M^{1/2}Z^t(R - Q)ZM^{1/2},$$

essent R i Q les matrius de graus i associada corresponents al graf que relaciona els individus adequats amb la finalitat pròpia de l'Anàlisi Local (damunt d'un graf) de fixar certs factors externs a les variables de l'anàlisi; m és el doble del nombre d'arestes del graf més el nombre de bucles.

Fixem-nos que la matriu $Z^t(R-Q)Z$ pot llegir-se[7] com una nova matriu de Burt amb l'estructura:

$$B_L = \left[\begin{array}{c} \vdots \\ \ldots \quad (Z_s^t(R-Q)Z_r) \quad \ldots \\ \vdots \end{array} \right]_{s,r=1,\ldots,q}$$

És a dir, podem dir que de l'ACM usual[8] a l'ACM Local damunt d'un graf hi ha el pas de tractar la taula usual de Burt:

$$(Z_s^t Z_r)_{s,r=1,\ldots,q}$$

a tractar la taula de Burt "havent fixat els efectes d'una variable exògena":

$$(Z_s^t(R-Q)Z_r)_{s,r=1,\ldots,q}$$

Doncs bé, ara voldríem arribar a una nova taula de Burt en què el creuament de dues característiques, s i r, es fes fixant els efectes de les altres característiques de l'anàlisi; és a dir, cada creuament fos semblant als creuaments locals $Z_s^t(R-Q)Z_r$ però amb R i Q corresponents a un graf, cada vegada diferent, que ens permetés de controlar les característiques de l'anàlisi $Z_k, k \neq s$ i $k \neq r$.

7.3.2 Estructura de grafs com a solució del problema

El nostre propòsit, és, doncs, aconseguir:

[7] A més, recordem, $Z^t(R-Q)Z = 1/2Z^tT^tTZ = 1/2(TZ)^tTZ$, essent T la matriu n^2xn de creuament eixos-vèrtexs del graf.

[8] La qual anàlisi, de fet, per equivalència, tractava inèrcies respecte de l'origen i no respecte del centre de gravetat, i per tant, no considerava el vector G.

- blocs $Z_s^t(R-Q)Z_r$ recollint en diferents grafs el fixament de totes les característiques de l'anàlisi diferents a Z_si Z_r; i també
- aconseguir que la matriu global tingui la forma ususal, ja sigui de la matriu de Burt $B = Z^tZ$, com de la local $B_L = 1/2(TZ)^tTZ$, de X^tX.

La solució que proposem consisteix en l'estructura de grafs i la seva configuració global següents:

- Grafs l, $l = 1,\ldots,q$, amb matrius R_l, Q_l, T_l, fixant totes les variables menys la Z_l -és a dir, que connecten els individus posseint les mateixes modalitats en totes les característiques $Z_k, k \neq l$-, en els quals tenim:

$$R_l - Q_l = 1/2T_l^tT_l.$$

- Grafs (r,s), $r,s = 1,\ldots,q, r \neq s$, amb matrius R_{rs}, Q_{rs}, T_{rs}, fixant totes les variables menys Z_r i Z_s -és a dir, connectant els individus posseint les mateixes modalitats en totes les característiques $Z_k, k \neq r$ i $k \neq s$-, en els quals tenim:

$$R_{rs} - Q_{rs} = 1/2T_{rs}^tT_{rs}.$$

- Construir la taula conjunta nxn:

$$\sum_{l=1}^{q} T_l^tT_l + \sum_{r,s=1,r<s}^{q} T_{rs}^tT_{rs},$$

que és igual a:

$$(T_1^t \ldots T_q^tT_{12}^t \ldots T_{rs(r<s)}^t \ldots T_{(q-1)q}^t) \begin{pmatrix} T_1 \\ \vdots \\ T_q \\ T_{12} \\ \vdots \\ T_{rs}, r < s \\ \vdots \\ T_{(q-1)q} \end{pmatrix}$$

Anomenem T la matriu $(n^2(q+\frac{q(q-1)}{2}))xn$:

$$\begin{pmatrix} T_1 \\ \vdots \\ T_q \\ T_{12} \\ \vdots \\ T_{rs}, r<s \\ \vdots \\ T_{(q-1)q} \end{pmatrix}$$

Aleshores,

$$1/2T^tT = \sum_{l=1}^{q} 1/2T_l^tT_l + \sum_{r,s=1,r<s}^{q} 1/2T_{rs}^tT_{rs} =$$

$$\sum_{l=1}^{q}(R_l - Q_l) + \sum_{r,s=1,r<s}^{q}(R_{rs} - Q_{rs}).$$

- Considerar $1/2Z^tT^tTZ$, taula que, anem a veure-ho, està formada pels següents blocs:

  - $q(q-1)$ blocs no diagonals de la forma:[9]

$$Z_s^t(R_{rs} - Q_{rs})Z_r, r, s = 1, \ldots, q, r \neq s.$$

  - q blocs diagonal de la forma:

$$Z_l^t(R_l - Q_l + \sum_{r,s=1;r<s,(r=l)\vee(s=l)}^{q}(R_{rs} - Q_{rs}))Z_l,$$

    amb $l = 1, \ldots, q$.

Efectivament:

$$Z^tT^tTZ =$$

[9]Fixem-nos que $R_{rs} = R_{sr}, Q_{rs} = Q_{sr}T_{rs} = T_{sr}$.

$$\begin{pmatrix} Z_1^t \\ \vdots \\ Z_q^t \end{pmatrix} (T_1^t \dots T_q^t T_{12}^t \dots T_{rs(r<s)}^t \dots T_{(q-1)q}^t) \begin{pmatrix} T_1 \\ \vdots \\ T_q \\ T_{12} \\ \vdots \\ T_{rs}, r < s \\ \vdots \\ T_{(q-1)q} \end{pmatrix} (Z_1 \dots Z_q)$$

Que resulta:

$$\begin{pmatrix} Z_1^t T_1^t & \dots & Z_1^t T_{(q-1)q}^t \\ Z_2^t T_1^t & \dots & Z_2^t T_{(q-1)q}^t \\ \vdots & & \vdots \\ Z_q^t T_1^t & \dots & Z_q^t T_{(q-1)q}^t \end{pmatrix} \begin{pmatrix} T_1 Z_1 & \dots & T_1 Z_q \\ T_2 Z_1 & \dots & T_2 Z_q \\ \vdots & & \vdots \\ T_{(q-1)q} Z_1 & \dots & T_{(q-1)q} Z_q \end{pmatrix}$$

Ara bé, els blocs $T_h Z_k$, pel que representen com a matrius $T_h Z_k$, matrius de diferències de les modalitats de Z_k en individus connectats pel graf corresponent, resulten nuls si, amb $h = 1, \dots, q$, és $h \neq k$, o bé si, amb $h = (r, s)$, és $r \neq k$ i $s \neq k$.

Per tant, efectivament, els $q(q-1)$ blocs no diagonals de $Z^t T^t T Z$ són:

$$Z_s^t T_{rs}^t T_{rs} Z_r, r, s = 1, \dots, q, r \neq s,$$

i els q blocs diagonal són:

$$Z_l^t (T_l^t T_l + \sum_{r,s=1; r<s, (r=l) \vee (s=l)}^{q} T_{rs}^t T_{rs}) Z_l,$$

amb $l = 1, \dots, q$, que era el que volíem demostrar.

Doncs bé, tenim pràcticament configurada la nova matriu de Burt a tractar, amb blocs no diagonals senzills, mitjançant un graf, cada vegada diferent però, en cada cas, el que és clarament intuïtiu, i amb blocs diagonals menys senzills, obtinguts, cadascun, mitjançant tot un conjunt de grafs de fixament -no només el que fixa totes les variables

menys la corresponent a l'ordre del bloc diagonal, sinó també tots els que, deixant-ne de fixar dues, de variables, una d'elles és aquesta corresponent a l'ordre del bloc.

Ens cal, encara, ponderar pels valors m_l i m_{rs} de cada graf, tal i com ho fèiem també a l'Anàlisi Local.

En l'Anàlisi Local tractàvem:

$$\frac{n}{mq} M^{1/2} Z^t (R - Q) Z M^{1/2},$$

és a dir, la matriu de Burt damunt del graf

$$B_L = Z^t (R - Q) Z,$$

o bé la matriu ponderada

$$Z^t (1/m(R - Q)) Z.$$

Anem a fer el mateix al nostre cas.

Considerem:

$$\sum_{l=1}^{q} \frac{1}{m_l} (R_l - Q_l) + \sum_{r,s=1, r<s}^{q} \frac{1}{m_{rs}} (R_{rs} - Q_{rs}),$$

que pot escriure's:

$$R_{pon} - Q_{pon}$$

essent

$$R_{pon} = \sum_{l=1}^{q} \frac{1}{m_l} R_l + \sum_{r,s=1, r<s}^{q} \frac{1}{m_{rs}} R_{rs},$$

i

$$Q_{pon} = \sum_{l=1}^{q} \frac{1}{m_l} Q_l + \sum_{r,s=1, r<s}^{q} \frac{1}{m_{rs}} Q_{rs}.$$

Aleshores, si considerem la taula T ponderada:

$$T_{pon} = \begin{pmatrix} \frac{1}{\sqrt{m_1}}T_1 \\ \vdots \\ \frac{1}{\sqrt{m_q}}T_q \\ \frac{1}{\sqrt{m_{12}}}T_{12} \\ \vdots \\ \frac{1}{\sqrt{m_{rs}}}T_{rs}, r < s \\ \vdots \\ \frac{1}{\sqrt{m_{(q-1)q}}}T_{(q-1)q} \end{pmatrix}$$

és clar:

$$1/2T_{pon}^t T_{pon} = 1/2\sum_{l=1}^{q}\frac{1}{m_l}T_l^l T_l + 1/2\sum_{r,s=1,r<s}^{q}\frac{1}{m_{rs}}T_{rs}^t T_{rs} =$$

$$\sum_{l=1}^{q}\frac{1}{m_l}(R_l - Q_l) + \sum_{r,s=1,r<s}^{q}\frac{1}{m_{rs}}(R_{rs} - Q_{rs}) =$$

$$R_{pon} - Q_{pon}.$$

Resumint, l'ACM Parcial Interna i Simultània, ens porta a diagonalitzar la matriu:

$$\frac{n}{q}M^{1/2}Z^t(R_{pon} - Q_{pon})ZM^{1/2},$$

essent $R_{pon} - Q_{pon} = 1/2T_{pon}^t T_{pon}$, de manera que $Z^t(R_{pon} - Q_{pon})Z$ no és sinó una nova matriu de Burt ponderada formada per:

- $q(q-1)$ blocs no diagonals de la forma:

$$\frac{1}{m_{rs}}Z_s^t(R_{rs} - Q_{rs})Z_r, r, s = 1, \ldots, q, r \neq s.$$

- q blocs diagonal de la forma:

$$Z_l^t(\frac{1}{m_l}(R_l - Q_l) + \sum_{r,s=1;r<s,(r=l)\vee(s=l)}^{q}\frac{1}{m_{rs}}(R_{rs} - Q_{rs}))Z_l.$$

amb $l = 1, \ldots, q$.

L'Anàlisi Local damunt d'un graf tractava:[10]

$$\frac{n}{q} M^{1/2} Z^t (R_{pon} - Q_{pon}) Z M^{1/2},$$

essent $R_{pon} - Q_{pon} = 1/2 T_{pon}^t T_{pon}$, de manera que $Z^t(R_{pon} - Q_{pon})Z$ és la matriu de Burt ponderada damunt d'un graf formada per q^2 blocs de la forma:

$$1/m Z_s^t (R - Q) Z_r, r, s = 1, \ldots, q.$$

La semblança, doncs, formal, de les dues anàlisis és evident.

Les projeccions d'individus i modalitats, són, evidentment, en l'ACM Parcial Interna i Simultània, les traduccions exactes de les de l'Anàlisi Local.

Com ja hem anat dient al llarg d'aquest capítol, la comparació entre les anàlisis usuals -ara l'ACM usual- i l'anàlisi presentada, és el que pot ajudar-nos a aconseguir millors estudis i millors interpretacions. Anem a veure en el capítol següent i últim de la Tesi, dues aplicacions de l'Anàlisi Parcial Interna i Simultània, una amb variables numèriques i l'altra amb variables categòriques.

[10] Ara, aquí, R, Q i T fan referència a l'únic graf de l'AL; $R_{pon} = 1/mR, Q_{pon} = 1/mQ, T_{pon} = \frac{1}{\sqrt{m}} T$.

Capítol 8

Aplicació de l'Anàlisi Parcial Interna i Simultània

8.1 Aplicació de l'*ACP* Parcial Interna i Simultània

Anem a tractar la matriu[Alu 92] dels valors que sobre 43 gossos i llops prenen les següents sis variables de mides del crani:

- x_1=locb=llargada còndilo-basal
- x_2=loms=llargada de la mandíbula superior
- x_3=labm=amplada bi-maxil.lar
- x_4=locs=llargada caní superior
- x_5=lopm=llargada primer molar superior
- x_6=lapm=amplada primer molar superior

A la taula 8.1.1 hi tenim les dades, sempre amb les columnes ordenades segons tal i com acabem de presentar les variables:

129	64	95	17.5	11.2	13.8
154	74	76	20.0	14.2	16.5
170	87	71	17.9	12.3	15.9

188	94	73	19.5	13.3	14.8
161	81	55	17.1	12.1	13.0
164	90	58	17.5	12.7	14.7
203	109	65	20.7	14.0	16.8
178	97	57	17.3	12.8	14.3
212	114	65	20.5	14.3	15.5
221	123	62	21.2	15.2	17.0
183	97	52	19.3	12.9	13.5
212	112	65	19.7	14.2	16.0
220	117	70	19.8	14.3	15.6
216	113	72	20.5	14.4	17.7
216	112	75	19.6	14.0	16.4
205	110	68	20.8	14.1	16.4
228	122	78	22.5	14.2	17.8
218	112	65	20.3	13.9	17.0
190	93	78	19.7	13.2	14.0
212	111	73	20.5	13.7	16.6
201	105	70	19.8	14.3	15.9
196	106	67	18.5	12.6	14.2
158	71	71	16.7	12.5	13.3
255	126	86	21.4	15.0	18.0
234	113	83	21.3	14.8	17.0
205	105	70	19.0	12.4	14.9
186	97	62	19.0	13.2	14.2
241	119	87	21.0	14.7	18.3
220	111	88	22.5	15.4	18.0
242	120	85	19.9	15.3	17.6
199	105	73	23.4	15.0	19.1
227	117	77	25.0	15.3	18.6
228	122	82	24.7	15.0	18.5
232	123	83	25.3	16.8	15.5
231	121	78	23.5	16.5	19.6
215	118	74	25.7	15.7	19.0
184	100	69	23.3	15.8	19.7
175	94	73	22.2	14.8	17.0
239	124	77	25.0	16.8	27.0
203	109	70	23.3	15.0	18.7
226	118	72	26.0	16.0	19.4
226	119	77	26.5	16.8	19.3
210	103	72	20.5	14.0	16.7

Taula 8.1.1

A la taula 8.1.2 hi tenim les mateixes dades però normalitzades:

-0.428781	-0.436818	0.380018	-0.211317	-0.353716	-0.188577
-0.287648	-0.333953	0.058616	-0.062745	-0.010896	-0.019382
-0.197323	-0.200228	-0.025964	-0.187545	-0.228015	-0.056981

-0.095708	-0.128223	0.007868	-0.092460	-0.113742	-0.125912
-0.248131	-0.261947	-0.296618	-0.235088	-0.250870	-0.238709
-0.231195	-0.169369	-0.245871	-0.211317	-0.182306	-0.132179
-0.011028	0.026075	-0.127459	-0.021145	-0.033750	-0.000583
-0.152161	-0.097363	-0.262787	-0.223202	-0.170878	-0.157245
0.039780	0.077508	-0.127459	-0.033031	0.000532	-0.082047
0.090588	0.170086	-0.178207	0.008569	0.103378	0.011950
-0.123934	-0.097363	-0.347366	-0.104345	-0.159451	-0.207376
0.039780	0.056935	-0.127459	-0.080574	-0.010896	-0.050715
0.084942	0.108367	-0.042880	-0.074631	0.000532	-0.075780
0.062361	0.067221	-0.009048	-0.033031	0.011959	0.055815
0.062361	0.056935	0.041700	-0.086517	-0.033750	-0.025649
0.000263	0.036362	-0.076712	-0.015203	-0.022323	-0.025649
0.130105	0.159800	0.092447	0.085826	-0.010896	0.062082
0.073652	0.056935	-0.127459	-0.044917	-0.045178	0.011950
-0.084417	-0.138509	0.092447	-0.080574	-0.125169	-0.176044
0.039780	0.046648	0.007868	-0.033031	-0.068032	-0.013116
-0.022319	-0.015071	-0.042880	-0.074631	0.000532	-0.056981
-0.050545	-0.004784	-0.093628	-0.151888	-0.193733	-0.163511
-0.265067	-0.364812	-0.025964	-0.258859	-0.205160	-0.219909
0.282528	0.200946	0.227775	0.020454	0.080523	0.074615
0.163977	0.067221	0.177027	0.014512	0.057668	0.011950
0.000263	-0.015071	-0.042880	-0.122174	-0.216588	-0.119646
-0.106998	-0.097363	-0.178207	-0.122174	-0.125169	-0.163511
0.203494	0.128940	0.244691	-0.003317	0.046241	0.093414
0.084942	0.046648	0.261606	0.085826	0.126232	0.074615
0.209139	0.139227	0.210859	-0.068688	0.114805	0.049549
-0.033609	-0.015071	0.007868	0.139311	0.080523	0.143546
0.124459	0.108367	0.075532	0.234397	0.114805	0.112214
0.130105	0.159800	0.160111	0.216568	0.080523	0.105947
0.152686	0.170086	0.177027	0.252225	0.286215	-0.082047
0.147041	0.149513	0.092447	0.145254	0.251933	0.174878
0.056716	0.118654	0.024784	0.275997	0.160514	0.137279
-0.118289	-0.066504	-0.059796	0.133369	0.171942	0.181145
-0.169097	-0.128223	0.007868	0.067997	0.057668	0.011950
0.192203	0.180373	0.075532	0.234397	0.286215	0.638597
-0.011028	0.026075	-0.042880	0.133369	0.080523	0.118480
0.118814	0.118654	-0.009048	0.293826	0.194796	0.162345
0.118814	0.128940	0.075532	0.323540	0.286215	0.156079
0.028489	-0.035644	-0.009048	-0.033031	-0.033750	-0.006849

Taula 8.1.2

La matriu de correlacions d'aquestes sis variables és:

1.00000	0.95874	0.34818	0.61295	0.71794	0.58725
0.95874	1.00000	0.20033	0.66100	0.73596	0.59465
0.34818	0.20033	1.00000	0.36996	0.35028	0.35478

0.61295	0.66100	0.36996	1.00000	0.89351	0.76264
0.71794	0.73596	0.35028	0.89351	1.00000	0.78922
0.58725	0.59465	0.35478	0.76264	0.78922	1.00000

Taula 8.1.3

Ara bé, com dèiem al capítol setè de la Tesi, ens interessa de tractar la matriu de correlacions parcials:[1]

1.00000	0.94993	0.62475	-0.40129	0.21781	0.03750
0.94993	1.00000	-0.61717	0.38554	-0.08530	-0.02238
0.62475	-0.61717	1.00000	0.33043	-0.10473	0.06622
-0.40129	0.38554	0.33043	1.00000	0.67987	0.19141
0.21781	-0.08530	-0.10473	0.67987	1.00000	0.30464
0.03750	-0.02238	0.06622	0.19141	0.30464	1.00000

Taula 8.1.4

Fixem-nos que entre les dues matrius de correlacions s'observen els següents principals canvis:

- Deixen de ser "altes" les correlacions entre x_1 i x_5, x_2 i x_5, x_4 i x_5 i x_6 i x_5, amb la qual cosa, se'ns revela x_5 com a una variable "espúriament" correlacionada amb les altres.
- Augmenten les correlacions entre x_1 i x_3 i entre x_2 i x_3.
- La correlació entre x_1 i x_2 es manté molt alta.

Com hem demostrat al capítol setè, les noves dades a tractar -la matriu de correlacions de les quals és la matriu de les correlacions parcials de les variables inicials canviada de signe excepte en la diagonal-, s'obtenen mitjançant transformacions i no són res més que els residus de l'ajustament de cada variables inicial pel corresponent model lineal de la resta de variables.[2]

[1] A l'apèndix A es mostra el càlcul d'aquesta matriu emprant el sistema MINITAB[Min 91].

[2] A l'apèndix B es mostra el càlcul d'aquestes noves variables tant mitjançant transformacions com obtenint els residus dels ajustaments -ajudant-nos sempre del sistema MINITAB.

Aquestes noves dades, normalitzades, són les de la taula 8.1.5:

-0.362280	0.275762	0.600615	-0.004795	-0.175921	0.067242
0.053237	-0.198922	-0.004674	-0.086909	0.254170	0.016023
-0.049672	0.031017	0.051510	-0.027676	-0.146446	0.196079
0.105007	-0.118121	-0.026423	0.067692	-0.036304	-0.071717
0.203855	-0.235403	-0.305340	0.095729	-0.046648	-0.040076
-0.201794	0.169602	-0.024835	-0.165092	0.069838	0.073517
-0.059764	0.075893	-0.061015	-0.000969	-0.050973	0.053182
-0.158801	0.158577	-0.052791	-0.183113	0.079545	0.023153
-0.080691	0.110407	-0.030401	-0.065688	0.071740	-0.109266
-0.240483	0.267425	0.005615	-0.229855	0.206768	-0.049798
0.162978	-0.153790	-0.342897	0.201562	-0.107517	-0.124352
-0.030850	0.057059	-0.058128	-0.136172	0.097621	-0.026785
-0.124270	0.169233	0.079608	-0.182382	0.118066	-0.091300
-0.047017	0.070776	0.026437	-0.124074	0.021973	0.096742
-0.066708	0.103672	0.107511	-0.152988	0.035353	0.017180
-0.096660	0.114162	0.007991	-0.019884	-0.017585	-0.006631
-0.103735	0.176982	0.135914	0.123834	-0.261384	0.057567
0.179712	-0.133570	-0.203335	0.059035	-0.121478	0.083327
0.152370	-0.160379	0.020910	0.135423	-0.064842	-0.163095
-0.023677	0.067212	0.041014	0.031219	-0.139943	0.046584
-0.070804	0.060752	0.029471	-0.163941	0.167881	-0.054035
-0.162734	0.217674	0.092637	-0.017619	-0.141227	-0.026491
0.301196	-0.381698	-0.149422	-0.021208	0.121783	-0.067428
0.186971	-0.104952	0.077531	-0.061443	-0.030758	0.000369
0.290704	-0.258314	-0.032981	0.030946	0.009506	-0.071986
0.117926	-0.045878	-0.051298	0.175235	-0.324028	0.036835
0.054211	-0.056823	-0.136963	0.034277	-0.000327	-0.088796
0.121917	-0.064996	0.129243	-0.088973	-0.037857	0.075946
-0.032834	0.017769	0.209375	-0.085592	0.124575	-0.053061
0.028549	0.006697	0.173832	-0.355947	0.272922	-0.019099
-0.012481	-0.025817	-0.043347	0.116491	-0.065216	0.112018
0.189360	-0.169545	-0.115108	0.341816	-0.220443	-0.034968
-0.087585	0.131697	0.143035	0.239070	-0.251327	-0.012300
-0.115380	0.100206	0.179845	0.029758	0.318865	-0.517168
-0.078775	0.059713	0.068272	-0.191832	0.262660	-0.000219
-0.116740	0.108419	0.018209	0.231740	-0.115645	-0.032451
-0.179770	0.076437	-0.014336	-0.110775	0.226300	0.118993
-0.186489	0.098123	0.088554	-0.025440	0.154504	-0.044231
0.048573	-0.069942	-0.112615	-0.144055	-0.070195	0.704158
-0.052567	0.030296	-0.051883	0.100146	-0.056077	0.077899
0.200022	-0.205521	-0.218672	0.338193	-0.135240	-0.034694
0.056731	-0.091803	-0.068701	0.190161	0.078906	-0.140779
0.289242	-0.280088	-0.181963	0.104095	-0.075598	0.023912

Taula 8.1.5

A la Fig. 8.1.1 hi tenim les projeccions[3] en el pla principal de les variables en l'ACP de les variables inicials, i a la Fig 8.1.2 les obtingudes en l'ACP de les noves dades-transformació de les inicials.

En aquesta segona anàlisi, les interpretacions també són les usuals, però tenint en compte:

- Els angles indiquen les correlacions parcials (canviades de signe) entre les variables originals -i és interessant de detectar ortogonalitats que inicialment no es donaven.

- Les variables -com a configuradores dels eixos-, s'interpreten com les inicials "depurades" per l'ajust lineal permès per les altres.

Així, s'observa una millor caracterització del primer eix per les tres variables x_1, x_2, x_3 , que estan parcialment més correlacionades, i del segon eix per, bàsicament, x_4 i x_5; i, de fet, s'observen dos conjunts de variables: x_1, x_2 i x_3 per una banda, i x_4, x_5 i x_6 per l'altra -resultat ben diferent de l'observat a l'anàlisi primera en què totes les variables es confonien força en les seves projeccions. Val a dir que les tres primeres variables es refereixen a mides de tot el crani mentre que les altres tres a mides de dents.

Apuntem, només, el dit fins aquí. Les interpretacions de les dues anàlisis són clares i deixem oberta la seva perllongació.[4]

[3]Calculades amb el sistema SPAD[Spa 87].

[4]A l'apèndix C hi ha el pla principal de les projeccions variables+individus per a l'anàlisi inicial obtingudes amb el sistema SPAD; i a l'apèndix D, igualment per a l'anàlisi de les dades transformades.

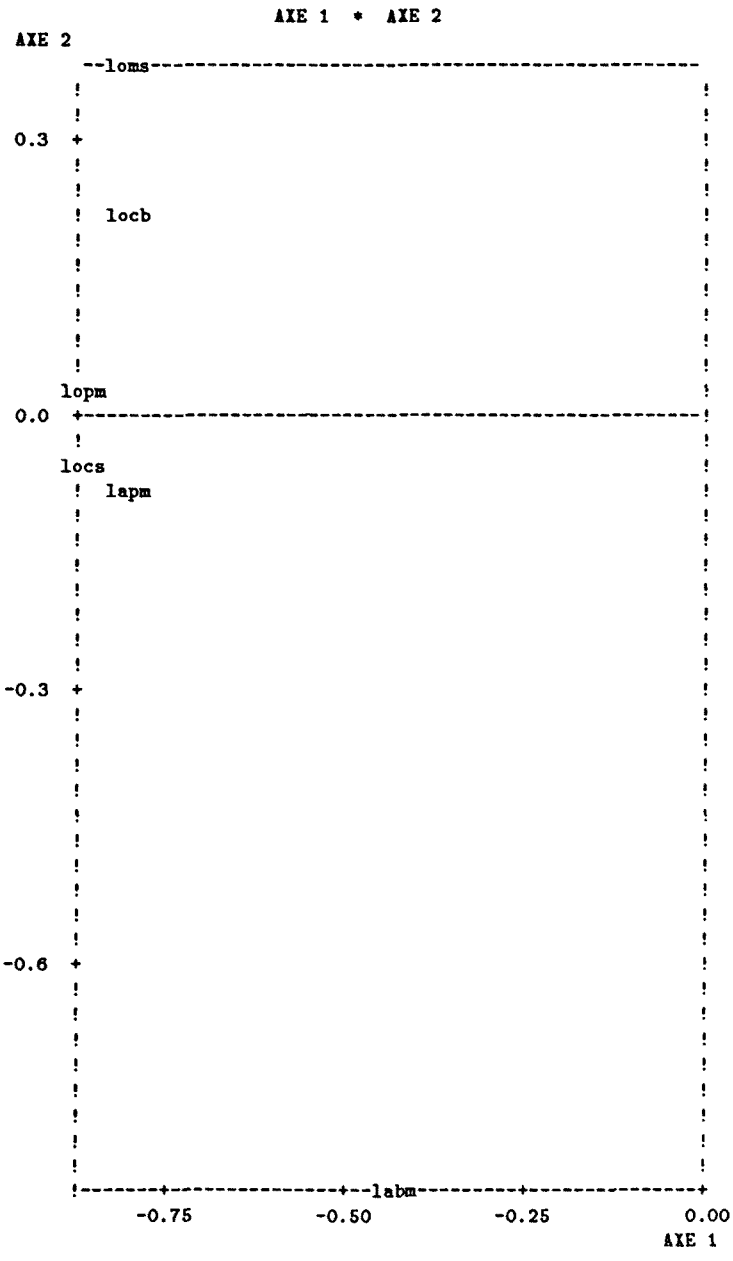

AXE 1 * AXE 2
AXE 2
loms
0.3
locb
lopm
0.0
locs
lapm
-0.3
-0.6
labm
-0.75
-0.50
-0.25
0.00
AXE 1

Fig. 8.1.1

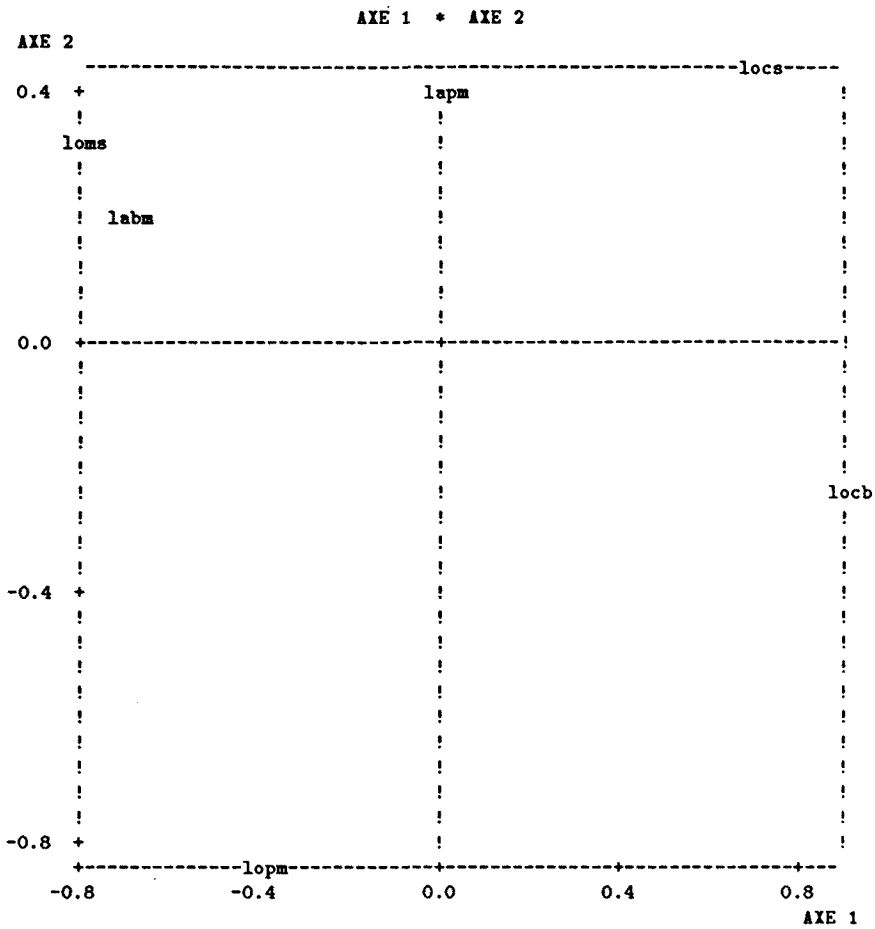

AXE 1 * AXE 2
AXE 2
locs
0.4
lapm
loms
labm
0.0
locb
-0.4
-0.8
lopm
-0.8
-0.4
0.0
0.4
0.8
AXE 1

Fig. 8.1.2

8.2 Aplicació de l'*ACM* Parcial Interna i Simultània

Anem a tractar un exemple acadèmic[5] en què tenim tres variables[6] categòriques A, B, C, amb dues modalitats cadascuna, configurant la següent matriu Z disjunta completa:

A1	A2	B1	B2	C1	C2
1	0	0	1	1	0
1	0	0	1	1	0
0	1	1	0	0	1
0	1	1	0	0	1
1	0	1	0	0	1
1	0	0	1	1	0
0	1	0	1	1	0
0	1	1	0	0	1
0	1	1	0	0	1
1	0	1	0	0	1

La matriu de Burt usual és:

5	0	2	3	3	2
0	5	4	1	1	4
2	4	6	0	0	6
3	1	0	4	4	0
3	1	0	4	4	0
2	4	6	0	0	6

La matriu de Burt ponderada per n^2, essent n el nombre d'individus, és:

0.05	0.00	0.02	0.03	0.03	0.02
0.00	0.05	0.04	0.01	0.01	0.04
0.02	0.04	0.06	0.00	0.00	0.06
0.03	0.01	0.00	0.04	0.04	0.00

[5]Els tractaments menys acadèmics per l'*ACM* Parcial Interna i Simultània requereixen un bon nombre d'individus i no massa modalitats. En estudis posteriors s'han de donar les eines interpretatives més adequades i específiques d'aquesta anàlisi -creiem que aquesta és la línia a continuar. També, segurament cladrà que l'anàlisi sigui selectiva quant als grafs a considerar -i no tan exhaustiva com la que presentem aquí. Tanmateix, les interpretacions són les habituals de l'*ACM* Local[Alu 84][Alu 91] i sempre en la direcció de mostrar quines són les relacions més fidedignes.

[6]A partir d'ara sempre referenciades en aquest ordre.

0.03	0.01	0.00	0.04	0.04	0.00
0.02	0.04	0.06	0.00	0.00	0.06

La matriu $M^{1/2}$ de les inverses de les arrels de les freqüències de cada modalitat és:

0.4472140	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
0.0000000	0.4472140	0.0000000	0.0000000	0.0000000	0.0000000
0.0000000	0.0000000	0.4082480	0.0000000	0.0000000	0.0000000
0.0000000	0.0000000	0.0000000	0.5000000	0.0000000	0.0000000
0.0000000	0.0000000	0.0000000	0.0000000	0.5000000	0.0000000
0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.4082480

Compararem l'ACM usual de Z amb l'Anàlisi Local fixant una quarta variable, D, també amb dues categories, amb matriu disjunta:

D1	D2
1	0
0	1
1	0
0	1
1	0
0	1
1	0
0	1
0	1
0	1

La matriu $R - Q$ del graf resultat de fixar D (segons el capítol sisè secció 1.6) és:

3	0	-1	0	0	-1	-1	0	0	0
0	5	0	-1	-1	0	0	-1	-1	-1
-1	0	3	0	0	-1	-1	0	0	0
0	-1	0	5	-1	0	0	-1	-1	-1
0	-1	0	-1	5	0	0	-1	-1	-1
-1	0	-1	0	0	3	-1	0	0	0
-1	0	-1	0	0	-1	3	0	0	0
0	-1	0	-1	-1	0	0	5	-1	-1
0	-1	0	-1	-1	0	0	-1	5	-1
0	-1	0	-1	-1	0	0	-1	-1	5

La matriu de Burt fixada la variable D -és a dir, $Z^t(R-Q)Z$-, resulta:

13	-13	-6	6	6	-6
-13	13	6	-6	-6	6
-6	6	12	-12	-12	12
6	-6	-12	12	12	-12
6	-6	-12	12	12	-12
-6	6	12	-12	-12	12

Anem a interpretar una mica aquesta matriu:

El graf resultat de fixar D dóna dues classes, formada la primera pels individus 1,3,6 i 7, i la segona per la resta. Aleshores:

- hi ha 13 parelles d'individus que varien en A tenint igual la variable D
- hi ha 12 parelles que varien en B tenint igual D
- hi ha 12 parelles que varien en C tenint igual D
- hi ha 6 parelles que varien alhora i en el mateix sentit A i B tenint igual D
- hi ha 6 parelles que varien alhora i en el mateix sentit A i C tenint igual D
- hi ha 12 parelles que varien alhora i en el mateix sentit B i C tenint igual D

La matriu de Burt fixada D i ponderada per les arestes del graf resulta:

0.250000	-0.250000	-0.115385	0.115385	0.115385	-0.115385
-0.250000	0.250000	0.115385	-0.115385	-0.115385	0.115385
-0.115385	0.115385	0.230769	-0.230769	-0.230769	0.230769
0.115385	-0.115385	-0.230769	0.230769	0.230769	-0.230769
0.115385	-0.115385	-0.230769	0.230769	0.230769	-0.230769
-0.115385	0.115385	0.230769	-0.230769	-0.230769	0.230769

Seguint els resultats del anteriors capítols de la Tesi i segons les projeccions usuals[Alu 86], tant en l'ACM usual com en la Local, es tracta de diagonalitzar

$$\frac{n}{q} M^{1/2} B_{pon} M^{1/2},$$

essent, com dèiem, n el nombre d'individus, q el nombre de variables o qüestions i p el nombre total de modalitats; B_{pon} la corresponent matriu de Burt ponderada, i $M^{1/2}$ la matriu diagonal pxp de terme general l'invers de l'arrel quadrada del nombre d'individus posseint cada modalitat.

Busquem doncs, els vectors i valors u_α, λ_α, tals que:

$$\frac{n}{q} M^{1/2} B_{pon} M^{1/2} u_\alpha = \lambda_\alpha u_\alpha$$

amb $u_\alpha^t u_\alpha = 1$.

Aleshores, les coordenades de les p modalitats sobre els corresponents vectors principals de l'anàlisi de variables, v_α, les dóna el vector p dimensional

$$\phi_\alpha = \sqrt{\lambda_\alpha} M^{1/2} u_\alpha.$$

L'ACM usual ens dóna els dos[7] valors propis significatius

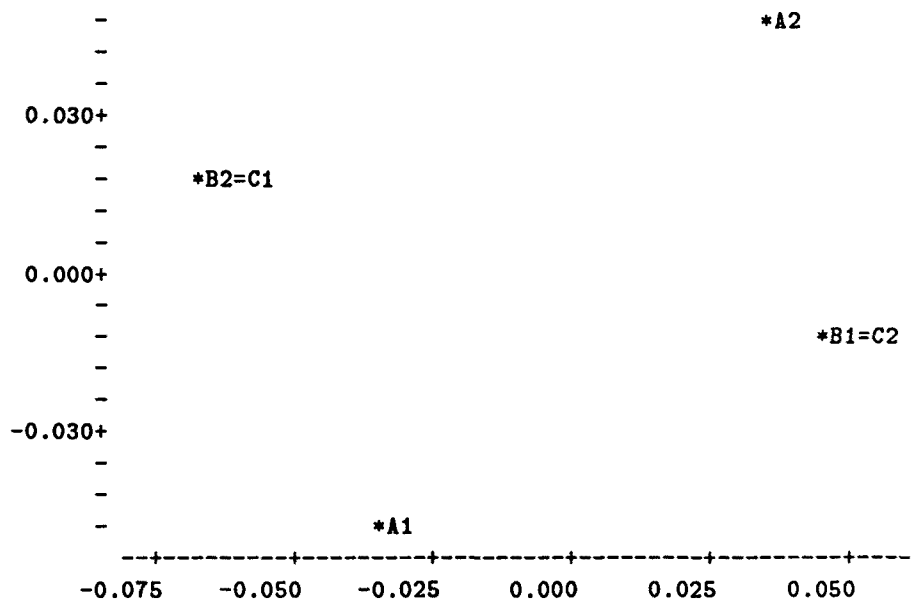
```
0.075459    0.024541
```

i les coordenades sobre el pla principal de les sis modalitats

```
-0.0360967   -0.0450596
 0.0360969    0.0450596
 0.0456177   -0.0118851
-0.0684263    0.0178276
-0.0684263    0.0178276
 0.0456177   -0.0118850
```

[7]Sense tenir en compte el primer, que de fet és $\lambda = 1$.

Així, tenim les següents projeccions en el pla principal de les sis modalitats:



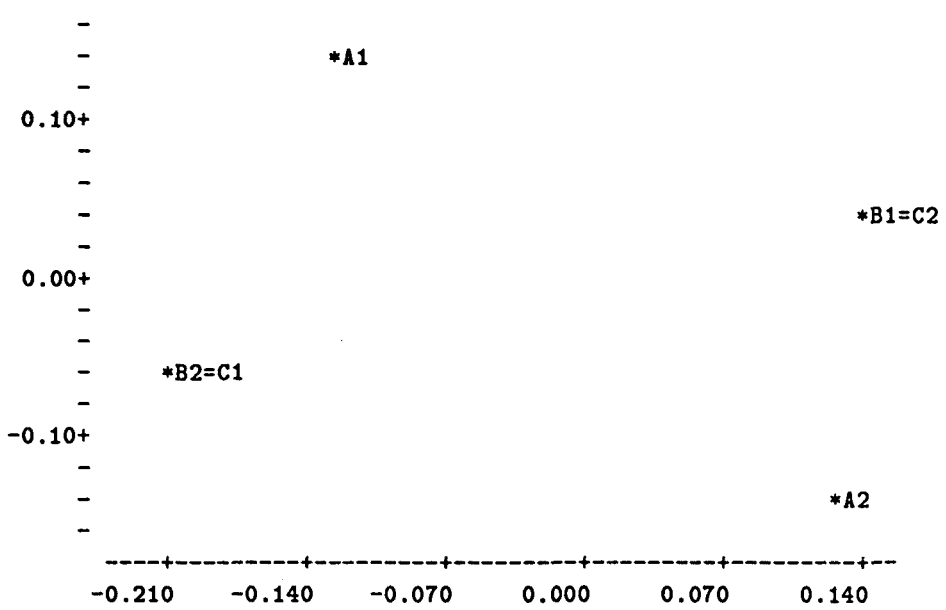
L'Anàlisi Local ens dóna els dos valors propis significatius

```
0.757324   0.217035
```

i les coordenades sobre el pla principal

```
-0.127678   0.130506
 0.127678  -0.130506
 0.140749   0.039462
-0.211123  -0.059193
-0.211123  -0.059193
 0.140749   0.039462
```

resultant una projecció en aquest pla principal:



Observem que, degut a la naturalesa de la variable D que fixem, no observem relacions diferents en una i altra anàlisi -precisament perquè així hem volgut la variable D.

Anem, finalment, a fer l'ACM Parcial Interna i Simultània sobre les dades de la matriu Z.

Segons que sembla, les relacions entre A i B i entre A i C són les que podem posar en dubte. Anem a veure-ho.

Sigui doncs la matriu de Burt construïda mitjançant tota l'estructura de grafs detallada al capítol setè:

33	-33	0	0	0	0
-33	33	0	0	0	0
0	0	10	-10	-10	10
0	0	-10	10	10	-10
0	0	-10	10	10	-10
0	0	10	-10	-10	10

Comentem aquesta matriu[8]:

- hi ha 11 parelles d'individus que canvien la categoria d'A tenint les mateixes tries en B i C
- a més, s'acumulen també al primer quadrant 11 parelles que canvien en A amb C fixada, i 11 que canvien en A amb B fixada
- en canvi, no n'hi ha cap que canviï B amb A i C fixades, ni que canviï C amb A i B fixades
- no hi ha cap parella que canviï alhora A i B posseint la mateixa modalitat en C
- tampoc n'hi ha cap que canviï alhora A i C posseint la mateixa modalitat en B
- hi ha 10 parelles que canvien alhora i amb igual orientació B i C tenint la mateixa modalitat en A.[9]

En definitiva, si reflexionem, tenim clara relació entre B i C, però, en canvi, entre A i B i entre A i C no hi ha tal relació.

Anem a fer el tractament gràfic. La matriu de Burt ponderada,[10] B_{pon}, d'aquesta anàlisi és:

[8]La lectura la faig observant la pròpia construcció de la matriu

[9]Atenció que aquest fet, en la matriu de Burt Parcial Interna i Simultània, també influeix en els corresponents blocs-diagonal.

[10]A l'apèndix E hi ha el programa en Fortran creat per al càlcul d'aquesta nova matriu de Burt.

```
 0.634615 -0.634615  0.000000  0.000000  0.000000  0.000000
-0.634615  0.634615  0.000000  0.000000  0.000000  0.000000
 0.000000  0.000000  0.200000 -0.200000 -0.200000  0.200000
 0.000000  0.000000 -0.200000  0.200000  0.200000 -0.200000
 0.000000  0.000000 -0.200000  0.200000  0.200000 -0.200000
 0.000000  0.000000  0.200000 -0.200000 -0.200000  0.200000
```

La matriu a diagonalitzar és, com sempre,

$$\frac{n}{q} M^{1/2} B_{pon} M^{1/2},$$

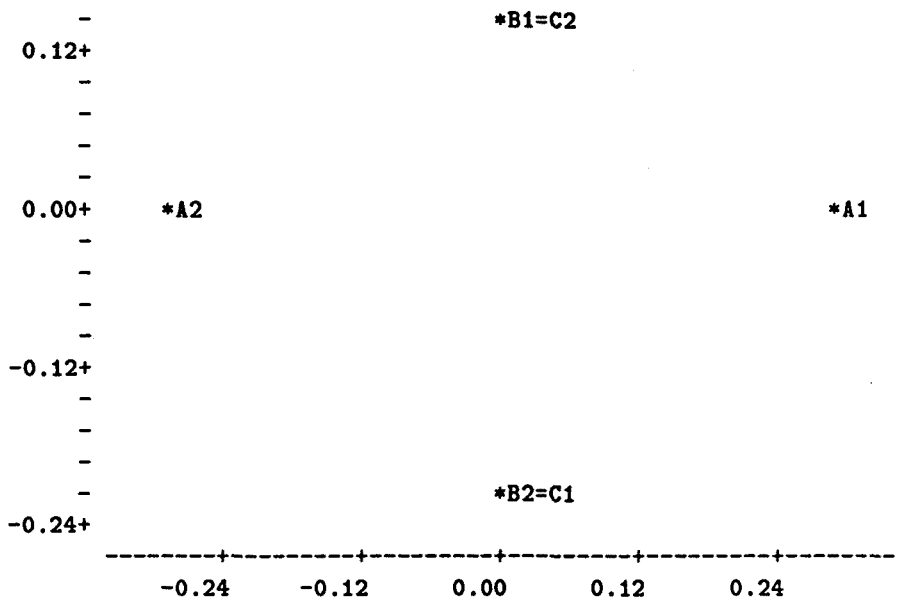
que dóna els valors propis significatius

```
0.846155   0.555555
```

i les coordenades de les sis modalitats en el pla principal:

```
 0.290888   0.000000
-0.290888   0.000000
 0.000000   0.136083
 0.000000  -0.204124
 0.000000  -0.204124
 0.000000   0.136083
```

Finalment, la projecció de les modalitats en aquest pla, resulta:



La diferència que observem respecte de l'*ACM* usual és, doncs, que la variable A no està tan relacionada com semblava amb les altres dues -les quals, òbviament, entre elles sí que ho estan.

Sempre serà la comparació entre l'*ACM* usual i la Parcial Interna i Simultània -i evidentment, la comparació també amb diverses Anàlisis Locals-, el que ens permetrà d'arribar a resultats que matisen els obtinguts només amb l'anàlisi usual.

A l'apèndix F hem tractat un exemple també acadèmic però amb 100 individus i variables A, B, C amb 3, 2 i 2 categories respectivament, conservant bàsicament la mateixa estructura que en aquest exemple que acabem de comentar.

Apèndix A

Correlacions Parcials

Càlcul de la matriu de correlacions parcials de les variables a tractar per l'ACP Parcial Interna i Simultània.

La matriu de correlacions parcials entre les sis variables és:

```
 1.00000   0.94993   0.62475  -0.40129   0.21781   0.03750
 0.94993   1.00000  -0.61717   0.38554  -0.08530  -0.02238
 0.62475  -0.61717   1.00000   0.33043  -0.10473   0.06622
-0.40129   0.38554   0.33043   1.00000   0.67987   0.19141
 0.21781  -0.08530  -0.10473   0.67987   1.00000   0.30464
 0.03750  -0.02238   0.06622   0.19141   0.30464   1.00000
```

Anem a calcular, per exemple, la correlació parcial entre la primera i la segona variables:

```
*dades de les sis variables: MTB >print c1-c6

 ROW     C1     C2     C3     C4     C5     C6

   1    129     64     95   17.5   11.2   13.8
   2    154     74     76   20.0   14.2   16.5
   3    170     87     71   17.9   12.3   15.9
   4    188     94     73   19.5   13.3   14.8
   5    161     81     55   17.1   12.1   13.0
   6    164     90     58   17.5   12.7   14.7
   7    203    109     65   20.7   14.0   16.8
   8    178     97     57   17.3   12.8   14.3
   9    212    114     65   20.5   14.3   15.5
  10    221    123     62   21.2   15.2   17.0
  11    183     97     52   19.3   12.9   13.5
```

```
 12    212    112    65   19.7   14.2   16.0
 13    220    117    70   19.8   14.3   15.6
 14    216    113    72   20.5   14.4   17.7
 15    216    112    75   19.6   14.0   16.4
 16    205    110    68   20.8   14.1   16.4
 17    228    122    78   22.5   14.2   17.8
 18    218    112    65   20.3   13.9   17.0
 19    190     93    78   19.7   13.2   14.0
  .      .      .     .      .      .      .

*residu de la regr.: MTB >regress c1 4 c3-c6;
                      SUBC> resids c91.

The regression equation is
C1 = - 21.0 + 0.352 C3 - 2.02 C4 + 16.2 C5 + 0.66 C6

Predictor       Coef      Stdev    t-ratio       p
Constant      -20.99      37.47      -0.56   0.579
C3            0.3517     0.3600       0.98   0.335
C4            -2.018      2.672      -0.76   0.455
C5            16.215      5.379       3.01   0.005
C6             0.661      2.061       0.32   0.750

s = 19.63       R-sq = 53.3%     R-sq(adj) = 48.4%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    4     16739.5      4184.9     10.86   0.000
Error        38     14638.4       385.2
Total        42     31377.9

*l'altre residu: MTB > regress c2 4 c3-c6;
                  SUBC> resids c92.

The regression equation is
C2 = - 4.3 - 0.117 C3 + 0.14 C4 + 7.80 C5 + 0.29 C6

Predictor       Coef      Stdev    t-ratio       p
Constant       -4.30      20.27      -0.21   0.833
C3           -0.1172     0.1948      -0.60   0.551
C4             0.135      1.446       0.09   0.926
C5             7.801      2.911       2.68   0.011
C6             0.292      1.115       0.26   0.795

s = 10.62       R-sq = 54.6%     R-sq(adj) = 49.9%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
```

```
Regression     4       5164.5       1291.1      11.45     0.000
Error         38       4286.2        112.8
Total         42       9450.7

*la corr. parcial entre c1 i c2: MTB > corr c91 c92 m10

Correlation of C91 and C92 = 0.950

MTB > print m10
 MATRIX M10

  1.00000   0.94993
  0.94993   1.00000
```

Òbviament, de la mateixa manera aniríem calculant les altres correlacions parcials.

Apèndix B

Noves Variables

Càlculs de les noves variables que l'anàlisi ha de tractar.

- Primer, segons transformacions.

  Anem a calcular la primera nova variable normalitzada:

```
*a la matriu m3 hi tenim les dades originals normalitzades

*la matriu m2 de les correlacions:

 MATRIX M2

  1.00000   0.95874   0.34818   0.61295   0.71794   0.58725
  0.95874   1.00000   0.20033   0.66100   0.73596   0.59465
  0.34818   0.20033   1.00000   0.36996   0.35028   0.35478
  0.61295   0.66100   0.36996   1.00000   0.89351   0.76264
  0.71794   0.73596   0.35028   0.89351   1.00000   0.78922
  0.58725   0.59465   0.35478   0.76264   0.78922   1.00000

*valors i vectors propis de la matriu de correlacions

MTB > eigen m2 c100 m4

*matriu de les inverses dels valor propis

MTB > let c101=1/c100
MTB > diag c101 m5

*transformacions

MTB > trans m4 m10
MTB > multi m3 m4 m9
```

```
MTB > multi m9 m5 m9
MTB > multi m9 m10 m7
MTB > copy m7 c71-c76

*normalitzem la primera nova variable

MTB > let c81=(c71-mean(c71))/sqrt(ssq(c71-mean(c71)))
MTB > print c81

C81(primera nova variable normalitzada)

 -0.362281   0.053240  -0.049671   0.105007   0.203855  -0.201791  -0.059764
 -0.158802  -0.080693  -0.240480   0.162980  -0.030851  -0.124271  -0.047016
 -0.066711  -0.096660  -0.103735   0.179712   0.152370  -0.023676  -0.070806
 -0.162735   0.301195   0.186968   0.290707   0.117929   0.054212   0.121918
 -0.032835   0.028547  -0.012480   0.189358  -0.087586  -0.115380  -0.078773
 -0.116741  -0.179768  -0.186489   0.048570  -0.052566   0.200021   0.056730
  0.289241
```

- Segon, com a residus.

  Anem a calcular el primer residu normalitzat:

```
*residu de c1 amb totes les altres

MTB > regress c1 5 c2-c6;
SUBC> resids c51.

The regression equation is
C1 = - 13.4 + 1.76 C2 + 0.557 C3 - 2.26 C4 + 2.52 C5 + 0.149 C6

Predictor       Coef       Stdev    t-ratio       p
Constant      -13.44       11.87      -1.13   0.265
C2           1.75550     0.09493      18.49   0.000
C3            0.5575      0.1145       4.87   0.000
C4           -2.2554      0.8463      -2.66   0.011
C5             2.521       1.857       1.36   0.183
C6            0.1491      0.6531       0.23   0.821

s = 6.215       R-sq = 95.4%     R-sq(adj) = 94.8%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    5     29948.6      5989.7    155.05   0.000
Error        37      1429.3        38.6
Total        42     31377.9
```

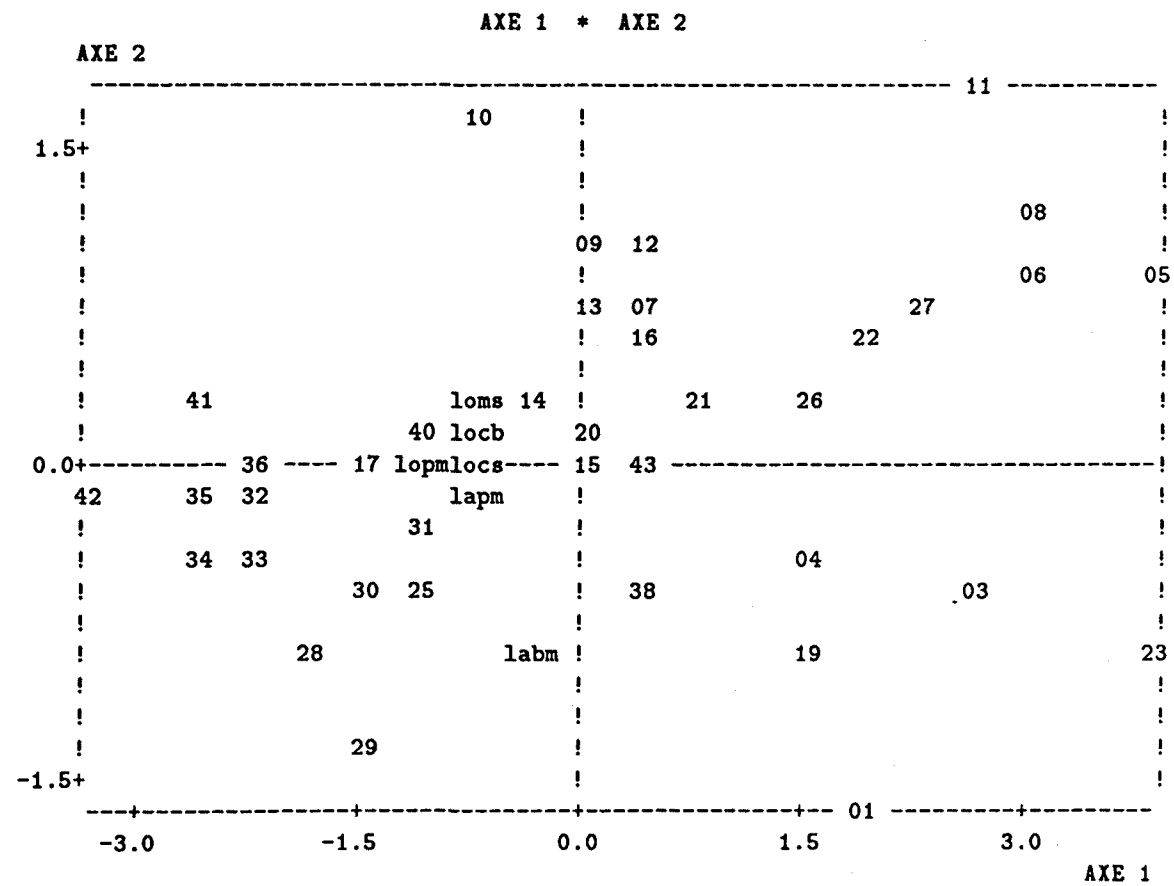
```
*normalitzem aquest primer residu

MTB > let c56=(c51-mean(c51))/sqrt(ssq(c51-mean(c51)))

MTB > print c56

C56(primera nova variable normalitzada)

 -0.362280   0.053238  -0.049672   0.105007   0.203855  -0.201793  -0.059764
 -0.158801  -0.080691  -0.240483   0.162979  -0.030850  -0.124270  -0.047017
 -0.066708  -0.096660  -0.103735   0.179712   0.152369  -0.023677  -0.070805
 -0.162734   0.301196   0.186971   0.290704   0.117926   0.054211   0.121917
 -0.032835   0.028549  -0.012480   0.189360  -0.087585  -0.115380  -0.078775
 -0.116740  -0.179770  -0.186489   0.048573  -0.052567   0.200022   0.056731
  0.289242
```

Observem com, llevat dels petits error de precisió del MINITAB, obtenim la mateixa nova variable normalitzada.

Apèndix C

Projeccions

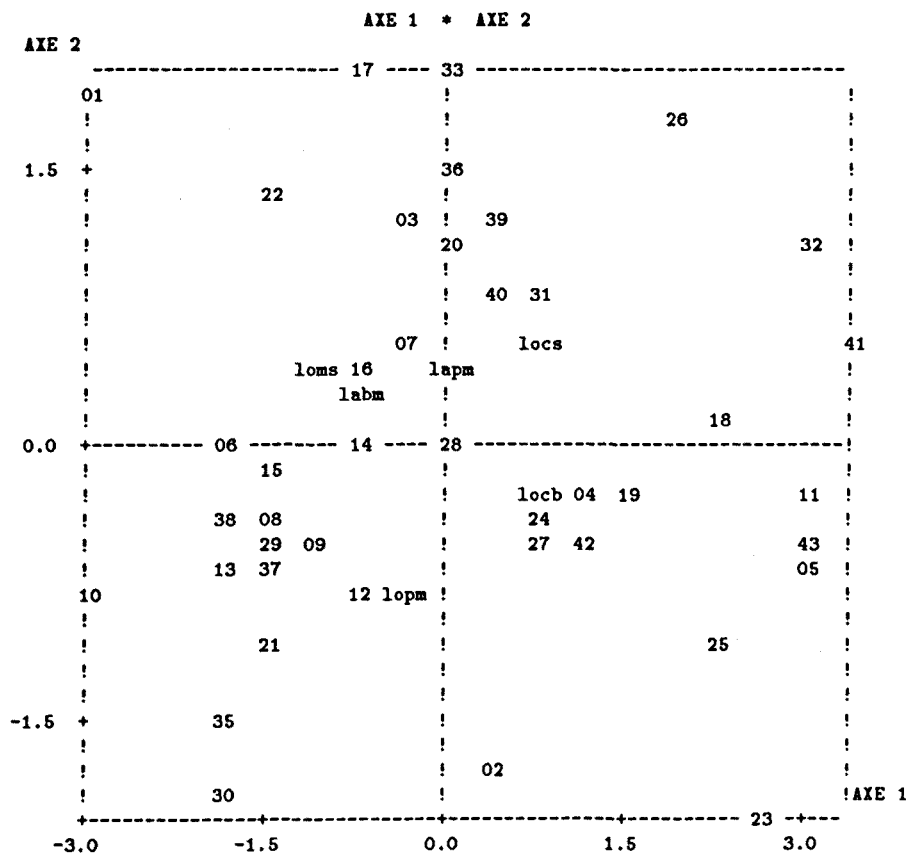
Projeccions de les variables i dels individus a l'ACP usual de les variables a tractar:



Apèndix D

Projeccions

Projeccions de les variables i dels individus en l'anàlisi de les variables transformades -és a dir, en l'Anàlisi Parcial Interna i Simultània:



Apèndix E

Algorisme

Programa en FORTRAN per al càlcul de la nova matriu de Burt ponderada en l'Anàlisi de Correspondències Múltiples Parcial Interna i Simultània:

```
program tesi

integer*4 icard(700),inici(700),iseg(700),iqus(100)
integer*4 imc1(700,100),imc2(700,100),ientr(700,100)
integer*4 iz(700,200),ia(700,200),ib(700,200)
real*8    r(700,200),c(200),burt(200,200)
integer*4 iburt(200,200)

real*8    temp1,temp2,temp3,temp4,temp5

data icert/1/,ifals/0/
        data r/140000*0.0/

        open(unit=2, file='sortida.dat', type='new')

        call llegir(ientr,iqus,n,iq)
        call modalitats(iqus,iq,isq)
call crearz(ientr,iqus,iz,n,iq)
        call calculard(iz,c,isq,n)
        m=iq-1
        n1=iq*m/2
        m1=iq-2
  call combinacions(iq,m,imc1,imc2,n)
        do ig=1,iq
   call inicialitzar(icard,inici,iseg,ir,n,isq)
           call estructuradades(imc1,iq,icard,inici,iseg,n,ientr,ig)
           call sumaarestes(icard,n,is)
           call calculmatriu(icard,inici,iseg,iz,ia,ib,isq,n)

           do i= 1,n
```

```
                do j=1,isq
      temp1 = ib(i,j)
      temp2 = ia(i,j)
      temp3 = temp1 - temp2
      temp4 = is
      temp5 = temp3 / temp4
                      r(i,j) = r(i,j) + temp5
                enddo
            enddo
          enddo
          do ig=1,n1
   call inicialitzar(icard,inici,iseg,ir,n,isq)
            call estructuradades(imc2,m,icard,inici,iseg,n,ientr,ig)
            call sumaarestes(icard,n,is)
            call calculmatriu(icard,inici,iseg,iz,ia,ib,isq,n)
            do i=1,n
                do j=1,isq
      temp1 = ib(i,j)
      temp2 = ia(i,j)
      temp3 = temp1 - temp2
      temp4 = is
      temp5 = temp3 / temp4
                      r(i,j) = r(i,j) + temp5
                enddo
            enddo
          enddo

call matburt(r,iz,burt,n,isq)
call matburt1(iz,iburt,n,isq)
call escmatburt(burt,isq)
call escmatburt1(iburt,isq)

          stop
          end



          subroutine llegir(ientr,iqus,n,iq)

integer*4 ientr(700,100),iqus(100)

          open(unit=1, file='entrada.dat', type='old')
          read(1,9000) n
          read(1,9000) iq

          do j=1,n
             read(1,2000)(ientr(j,i), i=1,iq)
          enddo

          read(1,2000) (iqus(i),i=1,iq)
```

```
        return
9000    format(i5)
2000    format(4X,100i2)
        end


        subroutine modalitats(iqus,iq,isq)

  integer*4 iqus(100)

        isq=0
        do i=1,iq
           isq=isq+iqus(i)
        enddo

        return
        end



subroutine crearz(ientr,iqus,iz,n,iq)

integer*4 ientr(700,100),iqus(100)
integer*4 iz(700,200)

data icert/1/,ifals/0/

do i=1,n
  icolini = 0
  do j = 1,iq
    iz(i,icolini+ientr(i,j)) = icert
    icolini = icolini + iqus(j)
  enddo
enddo

return
end

subroutine calculard(iz,c,isq,n)

integer*4 iz(700,200)
real*8    c(200)

real*8    temp

do j = 1,isq
  iw = 0
  do i = 1,n
    iw = iw + iz(i,j)
  enddo
```

```
  temp = iw
  c(j) = 1 / sqrt(temp)
enddo

return
end

subroutine combinacions(iq,m,imc1,imc2,n)

integer*4 imc1(700,100),imc2(700,100)

call combi1(iq,m,imc1,n)
call combi2(iq,m,imc1,imc2)

return
end


subroutine combi1(iq,m,imc1,n)

integer*4 imc1(700,100)

do j=1,m
    imc1(1,j) = j
    imc1(2,j) = j + 1
enddo
do i=3,iq
    do j=1,iq-(i-1)
imc1(i,j) = i + (j - 1)
    enddo
    do j=iq-(i-2),iq - 1
imc1(i,j) = iq - j
    enddo
enddo

return
end


subroutine combi2(iq,m,imc1,imc2)

integer*4 imc1(700,100),imc2(700,100)

if (iq .eq. 3) then
    imc2(1,1) = 2
    imc2(2,1) = 3
    imc2(3,1) = 1
else
    do j=1,iq-2
imc2(1,j) = imc1(2,j)
imc2(2,j) = imc1(2,j) + 1
```

```
    enddo
    is = 1
    do k=3,iq-1
do j=1,iq-2-is
    imc2(k,j) = imc2(k-1,j) + 1
enddo
do j=iq-2-is+1,iq-2
    imc2(k,j) = iq-j
enddo
is = is + 1
    enddo
    do i=3,iq-2
l = (i - 2) * iq - (i * (i - 3)) / 2
do j=1,iq-i
    imc2(l,j) = imc1(i,j)
      imc2(l + 1,j) = imc1(i,j) + 1
enddo
do j=iq-i+1,iq-2
    imc2(l,j) = iq - (j + 1)
    imc2(l + 1,j) = iq - (j + 1)
enddo
is = 1
m = (i - 1) * iq - (i * (i - 1)) / 2
do k=l+2,m
    do j=1,iq-i-is
imc2(k,j) = imc2(k-1,j) + 1
    enddo
    do j=iq-i-is+1,iq-i
imc2(k,j) = iq-j
    enddo
    do j=iq-i+1,iq-2
imc2(k,j) = imc2(l,j)
    enddo
    is = is + 1
enddo
    enddo
    l = iq * (iq - 1) / 2
    m = imc1(iq - 1,1)
    imc2(l - 2,1) = m
    imc2(l - 1,1) = m + 1
    do j=2,iq-2
imc2(l - 2,j) = iq - (j + 1)
imc2(l - 1,j) = iq - (j + 1)
    enddo
    do j=1,iq-2
imc2(l,j) = iq - (j + 1)
    enddo
endif

return
end
```

```
subroutine inicialitzar(icard,inici,iseg,ir,n,isq)

integer*4 icard(700),inici(700),iseg(700),ir(700,200)

do i=1,n
  icard(i) = 0
  inici(i) = 0
  iseg(i) = 0
enddo

return
end

subroutine estructuradades(imc,k0,icard,inici,iseg,n,ientr,ig)

integer*4 imc(700,100),ientr(700,100)
integer*4 icard(700),inici(700),iseg(700)

data icert/1/,ifals/0/

do k=1,n
  itrobat = ifals
  i = 1
  do while ((inici(i) .ne. 0) .and.
     *    (itrobat .eq. ifals))
    call comparar(idif,k,inici(i),inici,ientr,imc,k0,ig)
    if (idif .eq. ifals) then
      itrobat = icert
    else
      i = i + 1
    endif
  enddo
  icard(i) = icard(i) + 1
  if (itrobat .eq. icert) then
    iseg(k) = inici(i)
    inici(i) = k
  else
    inici(i) = k
  endif
enddo

return
end


subroutine comparar(idif,k,kk,inici,ientr,imc,k0,ig)

integer*4 inici(700),ientr(700,100)
integer*4 imc(700,100)
```

```
data icert/1/,ifals/0/

idif = ifals
is = 1
do while ((idif .eq. ifals) .and. (is .lt. k0))
  index = imc(ig,is)
  if (ientr(k,index) .eq. ientr(kk,index)) then
    idif = ifals
  else
    idif = icert
  endif
  is = is + 1
enddo

return
end


subroutine sumaarestes(icard,n,is)

integer*4 icard(700)

i = 1
is = 0
do while ((icard(i) .ne. 0) .and. (i .lt. n))
  is = is + icard(i) * icard(i)
  i = i + 1
enddo

return
end


subroutine calculmatriu(icard,inici,iseg,iz,ia,ib,isq,n)

integer*4 icard(700),inici(700),iseg(700)
integer*4 iz(700,200),ia(700,200),ib(700,200)

data icert/1/,ifals/0/

do j=1,isq
  l = 1
  ifi = ifals
  do while (ifi .eq. ifals)
    i = inici(l)
    iseguent = iseg(i)
    kl = iz(i,j)
    do while (iseguent .ne. 0)
      kl = kl + iz(iseguent,j)
      iseguent = iseg(iseguent)
    enddo
```

```
    i = inici(l)
    iseguent = iseg(i)
    ia(i,j) = kl
    ib(i,j) = icard(l) * iz(i,j)
    do while (iseguent .ne. 0)
      ia(iseguent,j) = kl
      ib(iseguent,j) = icard(l) * iz(iseguent,j)
      iseguent = iseg(iseguent)
    enddo
    if (l .eq. n) then
      ifi = icert
    else
      if (inici(l+1) .eq. 0) then
ifi = icert
      else
l = l + 1
      endif
    endif
  enddo
enddo

return
end

subroutine matburt(r,iz,burt,n,isq)

real*8    r(700,200),burt(200,200)
integer*4 iz(700,200)

real*8    temp

do ip=1,isq
  do it=1,isq
    burt(ip,it) = 0
    do i=1,n
      temp = iz(i,ip)
      burt(ip,it) = burt(ip,it) + (r(i,it) * temp)
            enddo
          enddo
        enddo

return
end



subroutine matburt1(iz,iburt,n,isq)

integer*4 iz(700,200),iburt(200,200)

do ip=1,isq
```

```
  do it=1,isq
    iburt(ip,it) = 0
    do i=1,n
      iburt(ip,it) = iburt(ip,it) + (iz(i,it) * iz(i,ip))
           enddo
         enddo
       enddo

return
end


subroutine escmatburt(burt,isq)

real*8 burt(200,200)

do ip=1,isq
  write(2,4444) (burt(ip,it),it=1,isq)
enddo
 4444 format(10(1X,f6.3))

return
end

subroutine escmatburt1(iburt,isq)

integer*4 iburt(200,200)

do ip=1,isq
  write(2,4444) (iburt(ip,it),it=1,isq)
enddo
 4444 format(20(1X,i3))

return
end
```

Apèndix F

Exemple

Sigui la matriu de les categories que posseeixen 100 individus en tres variables, A, B, C, amb 3,2 i 2 modalitats respectivament:

	A	B	C
1	1	2	1
2	3	1	2
3	2	2	1
4	3	1	2
5	1	2	1
6	3	1	2
7	2	1	2
8	1	2	1
9	1	1	2
10	2	1	2
11	1	2	1
12	2	1	2
13	2	2	1
14	3	1	2
15	1	2	1
16	1	1	2
17	2	1	2
18	3	2	1
19	3	1	2
20	3	1	2
21	1	2	1
22	2	1	2
23	2	2	1
24	2	1	2
25	1	2	1
26	1	1	2
27	2	1	2
28	3	2	1
29	1	1	2

```
30  3 1 2
31  1 2 1
32  2 1 2
33  2 2 1
34  3 1 2
35  1 2 1
36  3 1 2
37  2 1 2
38  1 2 1
39  1 1 2
40  3 1 2
41  1 2 1
42  2 1 2
43  2 2 1
44  2 1 2
45  1 2 1
46  3 1 2
47  2 1 2
48  1 2 1
49  3 1 2
50  2 1 2
51  1 2 1
52  3 1 2
53  2 2 1
54  2 1 2
55  1 2 1
56  3 1 2
57  2 1 2
58  3 2 1
59  1 1 2
60  2 1 2
61  1 2 1
62  3 1 2
63  3 2 1
64  2 1 2
65  1 2 1
66  1 1 2
67  2 1 2
68  3 2 1
69  1 1 2
70  2 1 2
71  3 2 1
72  2 1 2
73  2 2 1
74  3 1 2
75  1 2 1
76  1 1 2
77  2 1 2
78  3 2 1
79  1 1 2
80  2 1 2
```

81	1	2	1
82	2	1	2
83	2	2	1
84	3	1	2
85	3	2	1
86	1	1	2
87	2	1	2
88	3	2	1
89	1	1	2
90	2	1	2
91	3	2	1
92	2	1	2
93	2	2	1
94	3	1	2
95	1	2	1
96	3	1	2
97	2	1	2
98	1	2	1
99	1	1	2
100	3	1	2

Considerem la matriu de Burt usual:

34	0	0	13	21	21	13
0	36	0	27	9	9	27
0	0	30	20	10	10	20
13	27	20	60	0	0	60
21	9	10	0	40	40	0
21	9	10	0	40	40	0
13	27	20	60	0	0	60

Considerem la nova matriu de Burt ponderada de l'Anàlisi Parcial Interna i Simultània:

0.583	-0.312	-0.271	0.000	0.000	0.000	0.000
-0.312	0.675	-0.363	0.000	0.000	0.000	0.000
-0.271	-0.363	0.635	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.214	-0.214	-0.214	0.214
0.000	0.000	0.000	-0.214	0.214	0.214	-0.214
0.000	0.000	0.000	-0.214	0.214	0.214	-0.214
0.000	0.000	0.000	0.214	-0.214	-0.214	0.214

Considerem la matriu $M^{1/2}$:

```
0.17   0.00   0.00   0.00   0.00   0.00   0.00
0.00   0.17   0.00   0.00   0.00   0.00   0.00
0.00   0.00   0.18   0.00   0.00   0.00   0.00
0.00   0.00   0.00   0.13   0.00   0.00   0.00
0.00   0.00   0.00   0.00   0.16   0.00   0.00
0.00   0.00   0.00   0.00   0.00   0.16   0.00
0.00   0.00   0.00   0.00   0.00   0.00   0.13
```

Els tres valors propis significatius de l'anàlisi són:

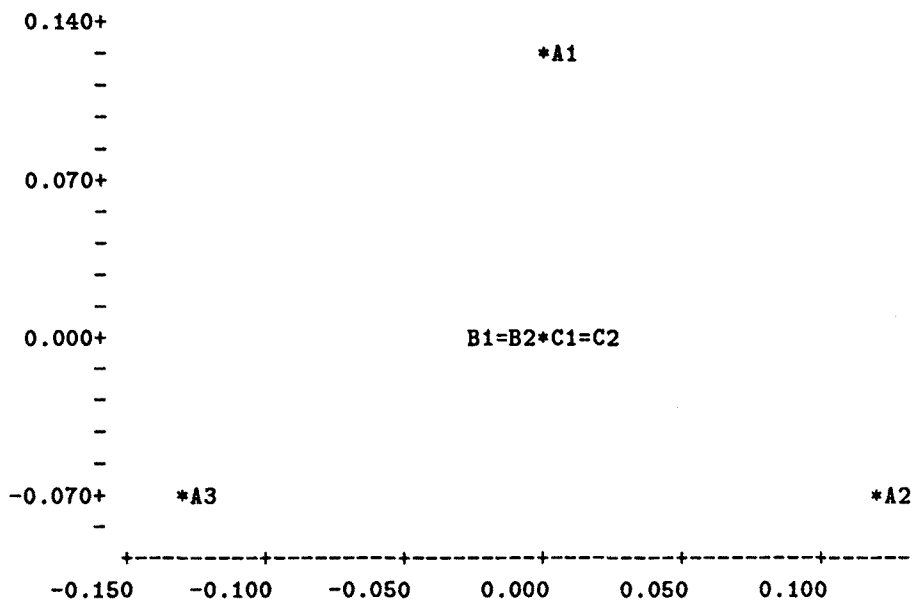
```
1.03892   0.85842   0.60633
```

Les coordenades, obtingudes com sempre, de les set modalitats en els tres primers eixos són:

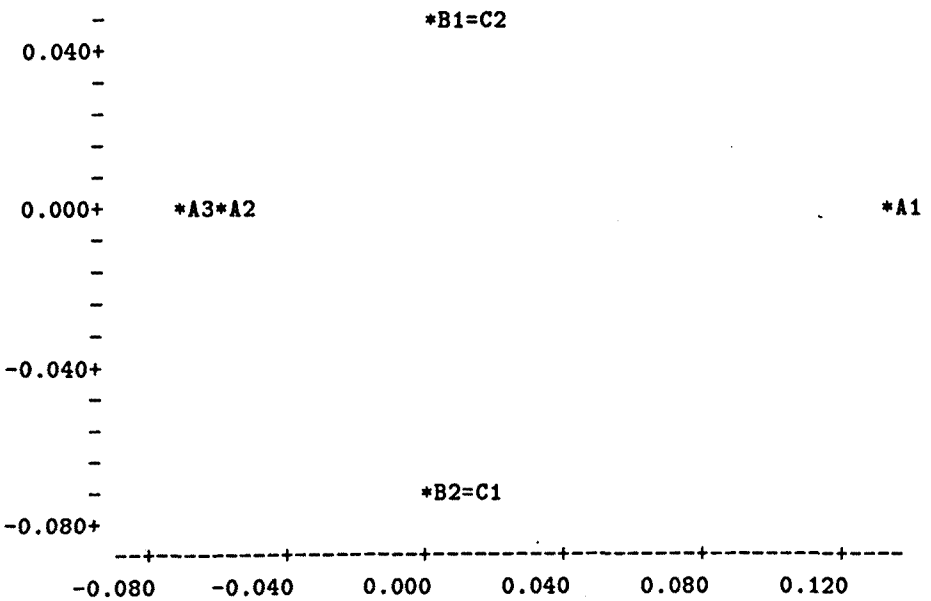
```
   1er eix     2on eix     3er eix

-0.004993    0.127290    0.0000000
 0.121476   -0.063501    0.0000000
-0.130727   -0.071604    0.0000000
 0.000000    0.000000    0.0451370
 0.000000    0.000000   -0.0683732
 0.000000    0.000000   -0.0683732
 0.000000    0.000000    0.0451370
```

Anem a considerar les projeccions en aquests tres eixos principals.

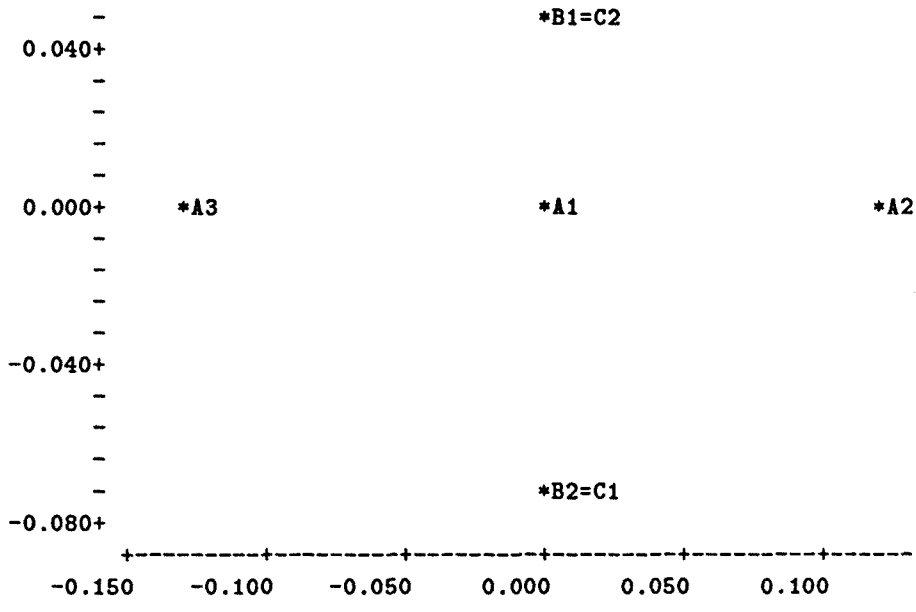
- Sobre els eixos primer i segon:



- Sobre els eixos tercer i segon:



- Sobre els eixos primer i tercer:



Com es pot veure, al tenir tres categories en A, la no relació entre aquesta variable i les altres dues s'observa també amb aquestes tres projeccions de clara interpretació.

Bibliografia

[Alu 80] T. Aluja. Local and Partial Correspondence Analysis. Report de Recerca RR86103. FIB. UPC. Barcelona. 1980.

[Alu 84] T. Aluja. Mètodes de Classificació i Anàlisi Factorial sobre un Graf. *Tesi Doctoral.* ETSEIB. Barcelona. 1984.

[Al2 84] T. Aluja, L. Lebart. Local and Partial Principal Components Analysis and Correspondence Analysis. Proceedings COMPSTAT-84, pp. 113-118. Physica-Verlag. Vienna. 1984.

[Alu 85] T. Aluja, L. Lebart. Factorial Analysis upon a Graph. *Bulletin Technique du CESIA.* Vol. 3. pp. 4-34. Paris. 1985.

[Alu 86] T. Aluja, M. Martí. Análisis de Correspondencias Múltiples sobre un grafo. Report de Recerca. FIB. UPC. Barcelona. 1986.

[Alu 88] T. Aluja. Local and Partial Correspondence Analysis. Application to the Analysis of Electoral Data. *Computational Statistics Quarterly, 2.* pp. 89-103. Physica-Verlag, Heidelberg. 1988.

[Alu 91] T. Aluja, R. Nonell. Local Principal Components Analysis. *Qüestió.* Vol. 15, 3. Barcelona. (Acceptada la seva publicació).

[Alu 92] T. Aluja. Apunts assignatura Anàlisi de Dades FIB. UPC. Barcelona 1992.

[Ami 79] M. Amirchahy, D. Néel, editors. Cours INRIA. Fontainebleau, 1979.

[And 84] T.W. Anderson. An Introduction in Multivariate Statistical Analysis. Wiley. 2a. ed. N.Y. 1984.

[Ben 76] J.P. Benzécri. Histoire et prehistoire de l'Analyse des données. *Les cahiers de l'Analyse des données.* Vol. 1, 1 à 4. Dunod. Paris. 1976.

[Ben 82] J.P. Benzécri. L'Analyse des Données, 2. L'Analyse des Correspondances. Dunod. Paris. 1982. 4a. éd.

[Bou 80] J M.. Bouroche, G. Saporta. L'analyse des données. *Que sais-je?* Paris. 1980.

[Cai 76] F. Caillez, J.-P. Pages. Introduction à l'Analyse des Données. SMASH. Paris. 1976.

[Cai 84] F. Caillez. Analyse des Données. SMS. Montréal. 1984.

[Car 85] A. Carlier. Analyse des évolutions sur tables de contingence: Quelques aspects operationels. *4èmes. Journées Internationales sur Analyse de Données et Informatique.* INRIA, Versailles. 1985.

[Car 90] A. Carlier. Tutorial on some Factorial Data Analysis Methods. Toulouse. 1990.

[Cas 86] Ph. Casin, J.C. Turlot. Une présentation de l'Analyse Canonique Généralisée dans l'Espace des Individus. *Revue de Statistiques Appliquées.* Vol. XXV, n. 3. Paris. 1986.

[Cli 81] A.D. Cliff, J.K. Ord. Spatial processes. Models and applications. Pion Limited. London. 1981.

[Cua 81] C.M. Cuadras. Métodos de Análisis Multivariante. EUNIBAR. Barcelona. 1981.

[Dam 85] L.d' Ambra. Alcune estensioni dell'analisi in componenti principali per lo studio di sistemi evolutivi. *Richerche Economiche*, XXXIX, 2. pp. 233-260. 1985.

[Dam 89] L.d' Ambra, N. Lauro. Non Symmetrical Analysis of Three-way contingency tables. Napoli. 1989.

[Dil 84] W.R.. Dillon, M. Goldstein. Multivariate Analysis. Wiley. N.Y. 1984.

[Esc 87] B. Escofier. Analyse des Correspondances Multiples Conditionelle. INRIA, 2. Versailles. 1987.

[Esc 88] B. Escofier, J. Pagès. Analyses Factorielles Simples et Multiples. Dunod. Paris. 1988.

[Esc 89] B. Escofier. Multiple Correspondence Analysis and Neighboring Relation. *Data Analysis, Learning Symbolic and Numeric Knowledge.* Ed. E. Diday. INRIA. Antibes. 1989.

[Esc 80] Y. Escoufier. Exploratory Data Analysis when Data are Matrices. *Recent Development in Statistical Inference and Data Analysis.* Ed. Matusita, pp. 45-53. 1980.

[Fic 87] B. Fichet, C. Lauro, editors. Methods for Multidimensional Data Analysis. ECAS. Lectures Notes. Napoli. 1987.

[Fou 84] F. Foucart. Aspects of Multivariate Statistical Theory. Wiley. N.Y. 1982.

[Gea 54] R.C. Geary. The contiguity ratio and statistical mapping. *The Inc. Statistician*, pp. 115-145. 1954.

[Gre 84] M.J. Greenacre. Theory and Applications of Correspondence Analysis. Academic Press. London. 1984.

[Ken 67] M.G. Kendall, A. Stuart. The advanced theory of statistics. 2 vols. Griffin and Company. London. 1967.

[Lau 88] N. Lauro, R. Siciliano. Correspondence analysis and modelling for contingency tables: symmetrical and non-symmetrical approaches. *Third International Workshop on Statistical Modelling.* Vienna. 1988.

[Lav 84] C. Lavit. STATIS. Rapport. Montpeller. 1984.

[La2 84] C. Lavit, C. Roux. Analyse conjointe de plusieurs tableaux de données par les méthodes STATIS. *Rapport Technique 8402.* Montpeller. 1984.

[La3 84] C. Lavit, C. Roux. Simultaneous Analysis of Data Matrices by STATIS. Version of STATIS for COMPSTAT-84. Montpeller. 1984.

[Lav 88] C. Lavit. Analyse conjointe des Tableaux Quantitatifs. Masson. Paris. 1988.

[Leb 69] L. Lebart. Analyse statistique de la contigüité. *Publ. ISUP, 18*, pp. 81/112. 1969.

[Leb 75] L. Lebart. Validité des resultats en Analyse des Données. CREDOC. Paris. 1975.

[Leb 84] L. Lebart. Correspondence analysis of graph structures. *Bulletin technique du CESIA.* Vol. 2, n. 1-2, pp. 5-19. Paris. 1984.

[Leb 85] L. Lebart. Quelques progrès récents dans la pratique de l'Analyse des Données. CNRS-CREDOC. Paris. 1985.

[Le2 85] L. Lebart, A. Morineau, J.P. Fénelon. Tratamiento Estadístico de Datos. Marcombo s.a. Barcelona. 1985.

[Mar 79] K.V. Mardia, J.T. Kent, J.M. Bibby. Multivariate Analysis. Academic Press, N.Y. 1979.

[Min 91] Minitab Inc. Minitab Reference Manual. Boston. Versió 1991.

[Mor 91] A. Morineau. *Seminari Qualitat de les Enquestes.* I.C.E. Barcelona. 1991.

[Mui 82] R.J. Muirhead. Aspects of Multivariate Statistical Theory. Wiley N.Y. 1982.

[Nak 81] J.P. Nakache, A. Cheralier, V. Morice. Exercices commentés de mathématiques pour l'analyse statistique des données. Dunod. Paris. 1981.

[Pag 89] J. Pagès. Introduction à l'Analyse Factorielle de Tableaux Multiples. *Secondes Journées Internationales: Analyse Statistique de grands tableaux el données d'enquête.* Pointre-à-Pitre. 1989.

[Rao 64] C.R. Rao. The use and interpretation of principal components analysis in applied research. *Sankhya, 26*, pp. 329-357. 1964.

[Sap 78] G. Saporta. Théories et Méthodes de la Statistique. Éditions Technip. Paris. 1978.

[Sar 84] V. Saris, H. Stronkhorst. Causal Modelling in Nonexperimental Research. SRF. Amsterdam. 1984.

[Spa 87] SPAD. Manuel de Référence. CISIA. 1987.

[Vol 85] M. Volle. Analyse des Données. 3ème édition. Economica. Paris.

[Whi 88] J. Whittaker. Graphical Modelling. ESRC Workshop. University of Lancaster, U.K. 1988.