Universitat Politècnica de Catalunya

Departament d'Estadística i Investigació Operativa

Estudi de les inèrcies estructurals en anàlisis de correspondències. Aportacions per a una millora de les anàlisis

Memòria presentada per en Josep Daunis i Estadella a la Universitat Politècnica de Catalunya amb la finalitat d'obtenir el grau de **doctor en Estadística**

Director: Tomàs Aluja i Banet

Barcelona, 20 d'octubre del 2004.

Agraïments

Durant aquests molts anys de treball en la tesi he tingut la gran sort de treballar amb tota una sèrie de persones que sempre que ha calgut m'han ajudat, d'una manera o altra, a tirar endavant. No és senzill agrair utilitzant només paraules tot el suport rebut, però ho provaré...

- D'entrada, l'agraïment més gran per a en Tomàs Aluja, director d'aquesta tesi, tot el suport que m'ha donat durant tot el temps que ha durat (i que és molt!) aquest treball i, especialment, per haver-me ofert la seva ajuda, dedicació, amistat, temps,... des d'aquell setembre de 1992 que va atendre a un nou estudiant de doctorat que arribava de Girona més perdut que un esquimal en el desert.

- Al meu *cap* i amic, en Carles Barceló, que va ser el que em va dirigir cap a un departament de la UPC, d'on ell en tenia bones referències i on jo hi podria fer un bon treball de recerca. Gràcies per totes les estones dedicades, el mestratge estadístic i els *collonsos* dedicats.

- A la Mei i en Narcís, primers companys de despatx, per haver compartit la major part dels anys de durada d'aquesta tesi, aguantant-me en el mals dies i compartint els bons, i per totes les estones que hem anat compartint.

- A en Santi, pel temps dedicat i compartit, emprenyant-lo en el seu despatx, compartint els primers congressos, i per ser una mica el germà gran, ja que es pot dir que hem begut de la mateixa font.

- A tota la gent del Grup del Símplex per la seva amistat. Els dies de reunions, congressos, workshops,... costaran d'oblidar. Gràcies, Vera, per la teva empenta. Martin pel bon humor i les teves trencades de gel. Glòria per la teva amistat i alegria. Raimon per la teva frescor i a la resta, que ja han estat citats abans.

- He d'agrair també a tots els companys del departament d'Informàtica i Matemàtica Aplicada a la UdG l'ajuda que m'han ofert sempre que ha fet falta. L'Esther Barrabés, pels bons moments compartits durant la carrera i ara. L'Imma Boada per les estones disteses i les xerrades que hem tingut. En David Juher, com si fos un company de despatx més, amb tot el que això comporta. En Miguel Sainz per tenir sempre un moment per solucionar qualsevol dubte que tinguessis i, en general, a tota la gent d'estadística i de matemàtiques: en Jordi Poch, l'Anna Pérez, en Joan Serarols, en Paco Martín, i els que encara no he citat: l'Àngel Calsina, en Josep Maria Humet, en Joan Saldanya, en Jordi Ripoll i en general als temps parcials, tant els que han deixat una empremta més profunda, com aquells que ha tingut un pas més discret. Sense oblidar a la resta de companys de la secció informàtica del departament (Joans, Joseps, Maries i ...).

- Pel que fa al suport tècnic i administratiu aquests anys a Girona, em cal agrair al present en Robert Valentí i al passat, l'Albert Vergés i en Pep Blanes, i tot el reguitzell d'ajudants que han tingut, pel seu esforç per tal que les "màquines"del departament estiguin sempre a punt, i a la Mercè Bautista i en Jordi Fontrodona per tota l'ajuda administrativa rebuda.

També mirant cap enfora del Departament, cal agrair moltes ajudes i bons records durant aquests anys de treball.

- Un record especial per la gent del «Laboratoire de Statistique et Probabilités» de Tolosa, que em varen acollir, ensenyar i fer veure que sempre es pot anar més lluny: l'Antoine de Falguerolles, en Carlier, en Caussinus, ...

- Als matemàtics i no matemàtics: començant pel "mestre"en Xevi i la Shashi, en Pitu i l'Anna, en Toni i l'Assumpta, en Ramon i l'Odina, l'Esther (valgui la redundància) i en Quim, na Bel i l'Andreas, en Reus i l'Elena, i finalment na Maria i en Toni i tota la patoleia que en segueix, per totes les trobades, tesis i moltes altres coses compartides.

- No puc oblidar-me de la meva etapa a la direcció de l'Escola Politècnica Superior de Girona, i agrair des d'aquí totes les bones (i també les crítiques) estones compartides amb la Carmen Andrés, la Montserrat Armengol, la Rosa Mundet, en Jaume Masó,... per un costat i la Carmen Carretero, en Jordi Farjas, sotsdirectores i sotsdirectors, coordinadores i coordinadors, i la resta d'equip directiu de l'EPS per l'altre.

- Encara que podria estar en l'apartat anterior o molt abans, li dedico unes ratlles per ell sol. Gràcies Xavier Pueyo, encara que m'has ajudat molt, segur que si no fos per tu que m'has fet participar com a secretari en la direcció del departament primer i com a sotsdirector acadèmic en la direcció de l'Escola després, amb tot l'enriquiment que això ha comportat, hauria acabat abans la Tesi.

- Tampoc no voldria deixar sense esment la gent del departament d'EIO, en Manel Martí, en Ramon Nonell, la Karina Gibert,... i l'inestimable suport de la Celia.

- I guardant-lo pel final, l'agraïment a la meva família. Primer de tot l'agraïment més gran i en tots els aspectes per la Montserrat, per aguantar-me i per suportar la tercera jornada laboral que t'ha robat tant de temps. Gràcies a la Clara, en Bernat, la Maria i la Bruna, que han fet d'aquests més de 10 anys, si més no en un altre aspecte, l'etapa més productiva i rica de la meva vida. Gràcies al pare Quim i la mare Dolors, a les germanes Olga i Mai, cosines i cosins, cunyats i cunyada, nebots, uncles i ties, tietes,...(potser tindran raó al final els del despatx que tinc parents a sota cada pedra), pel suport rebut i per ajudar-me a tirar endavant. I pels que ja no hi són, que també se n'alegrarien.

Espero no haver-me deixat a ningú, perquè amb els anys que hi porto, és molt fàcil fer-ho. Si algú es troba a faltar en aquesta llista, li demano que no dubti a afegir-s'hi i que m'estiri les orelles.

Pepus

Girona, setembre del 2004

Índex

Capítol 1

Pròleg

"Quan surts a fer el viatge cap a Ítaca,
has de pregar que el camí sigui llarg,
ple d'aventures, ple de coneixences.
Has de pregar que el camí sigui llarg,
que siguin moltes les matinades d'estiu,
que, amb quina delectança, amb quina joia!
entraràs en un port que els teus ulls ignoraven;
que vagis a ciutats d'Egipte, a moltes,
per aprendre i aprendre dels que saben.
Sempre tingues al cor la idea d'Ítaca.
Has d'arribar-hi, és el teu destí,
però no forcis gens la travessia.
És preferible que duri molts anys
i que ja siguis vell quan fondegis a l'illa,
ric de tot el que hauràs guanyat fent el camí,
sense esperar que t'hagi de dar riqueses Ítaca"

Konstantinos P. Kavafis (traducció Carles Riba)

1.1 Introducció al treball de recerca

Davant de la realitat multidimensional que ens envolta, és cada vegada més freqüent voler destriar allò realment important d'allò superflu. Però tots els paquets estadístics estàndards que tenim al nostre abast, encara no poden abordar aquest problema per la complexitat que suposa. Tot i això, és un camp que ha estat i està sent explorat

per tal de poder treure profit de les dades que ens proporcionen informació i d'aquí n'han sortit moltes línies d'investigació. Per aquests motius s'ha escollit com a tema de la tesi l'adaptació i desenvolupament de mètodes que ens permetin escollir on rauen les informacions importants de ser analitzades. El títol d'aquesta tesi, «Estudi de les inèrcies estructurals en anàlisis de correspondències. Aportacions per a una millora de les anàlisis», recull no només aquest propòsit, sinó que expressa el camp on ens mourem, **les anàlisis de correspondències**, tant si són correspondències simples (ACS), com múltiples (ACM), de grans taules de dades. També recull una problemàtica inherent a elles, **les inèrcies estructurals**, que marcarà el nostre enfocament de per on afrontarem la problemàtica, l'estudi de les inèrcies i la seva descomposició.

Finalment, el propòsit de millorar aquestes anàlisis, comportarà desenvolupar i estudiar propostes de millora, fent servir el modelatge loglineal i el condicionament com a dues eines per al seu estudi.

1.2 Objectius de la tesi

Amb el propòsit inicial de mirar de simplificar les anàlisis de correspondències múltiples, per tal de no analitzar la totalitat de la taula de Burt, ja que aquesta inclou l'anàlisi de les taules diagonals, cas particular de les anàlisis de correspondències simples de matrius quadrades, ens hem marcat els objectius que a continuació presentem de manera esquemàtica.

1. Objectius relacionats amb l'anàlisi de correspondències simples de matrius quadrades:
    - Analitzar la problemàtica de la influència de les diagonals.
    - Estudiar les diferents propostes de solució i analitzar-les per veure'n semblances i diferències.
    - Elaborar una nova proposta d'anàlisi per a eliminar la influència de la diagonal.
2. Objectius relacionats amb l'anàlisi de correspondències múltiples:
    - Analitzar la problemàtica de la influència de les subtaules de la diagonal.
    - Estudiar les diferents propostes de solució i analitzar-les per veure les diferents estratègies emprades.

- Estudiar l'aplicació de la metodologia emprada en ACS per tal d'aplicar-la a les subtaules de la diagonal de la taula de Burt.
- Implementar la metodologia i fer un estudi comparatiu amb les altres metodologies emprades.

3. Objectius relacionats amb la simplificació de les ACM:

   - Estudiar l'aplicació de les anàlisis condicionals per la simplificació de l'estudi de relacions.
   - Estudiar l'ús de de la modelització loglineal i la seva relació amb el condicionament.
   - Analitzar la problemàtica del condicionament múltiple i trobar metodologies alternatives.
   - Utilitzar la metodologia de condicionament múltiple per tal de buscar les relacions bàsiques entre les nostres variables.
   - Aplicar la metodologia emprada per anul·lar la influència d'una subtaula en aquelles subtaules que ens donin relacions espúries entre les variables

1.3 Organització del treball de recerca

La memòria d'aquesta tesi doctoral s'estructura, després d'un **Capítol 1: Pròleg** on es realitza la introducció referent al que són els objectius de la tesi i l'organització del treball de recerca, en els capítols Anàlisis factorials de dades, Models loglineals i models gràfics, Estudi de les inèrcies en anàlisis de correspondències, ACM respecte un model i ACM condicional, ACM multicondicional i Conclusions que seran descrits a continuació.

L'estructura d'aquesta memòria respon a la persecució dels objectius que acabem d'exposar en la secció anterior. Cada capítol s'inicia amb una introducció en la qual es motiva el seu contingut i finalitza amb una secció que emfasitza els aspectes aportats en el present treball en el capítol. S'ha procurat de fer un treball aplicat, la qual cosa ha comportat que hàgim procurat, en la mesura d'allò possible, acompanyar les teories amb dades, taules i gràfics il·lustratius, essencials moltes vegades en aquest afany de visualitzar la multidimensionalitat.

El **Capítol 2: Anàlisis factorials de dades** es destina gairebé exclusivament a presentar les eines utilitzades en les anàlisis factorials de dades, com són la descomposició en

valors singulars i la generalització d'aquesta. A continuació, s'inicia la presentació de les metodologies, on es detalla una anàlisi de tipus general, l'anàlisi canònica per a la comparació de dos grups de variables, i la seva generalització, l'anàlisi canònica generalitzada, per a més de dos grups. Presentarem les anàlisis de correspondències, simples i múltiples, com un cas particular de les anàlisis canòniques, però també les presentarem des d'altres perspectives més clàssiques, com són l'anàlisi per perfils en correspondències simples i l'anàlisi de la taula de Burt per correspondències múltiples. S'introdueixen, també, les diferents tècniques de representació gràfiques associades - representacions simètriques i Biplots - i les eines d'ajuda a la interpretació. A més, s'introdueix el concepte d'inèrcia associat al núvol de punts, conjuntament amb les formulacions bàsiques sobre inèrcies que utilitzarem en els capítols posteriors.

En el **Capítol 3: Models loglineals i models gràfics** es desenvolupen els models loglineals i els models gràfics, amb el seu estudi i descripció. S'inicia el capítol amb el concepte d'independència condicional i la formulació de models i l'expressió mitjançant ells de la independència condicional, eines que s'utilitzaran més endavant per al modelatge de les relacions entre les variables. Es desenvolupen, a continuació, la formulació dels models loglineals per a taules de dues, tres i més de tres variables i les restriccions que els caracteritzen - suma zero o còrner zero- i les relacions de transició entre les dues possibles formulacions del model, exemplificant-ho amb la generació d'un model i l'estimació dels paràmetres. Es realitza un estudi de la influència dels paràmetres en la generació de models, sobretot de la importància del termes de les interaccions sobre els termes independents i els efectes principals, ja que amb l'ús de models necessitàvem tenir acotat l'efecte d'aquestes interaccions. En aquest capítol s'introdueix la deviància, com a raó de versemblança entre dos models, la seva expressió i relació amb l'estadístic χ^2 i altres indicadors de divergència de models. Més endavant, es relacionarà la deviància amb la inèrcia de les anàlisis factorials.

Aquests dos primers capítols són capítols metodològics on el que es fa bàsicament és un recull de la metodologia existent, amb algunes aportacions pròpies en l'estudi de la influència dels paràmetres.

En el **Capítol 4: Estudi de les inèrcies en anàlisis de correspondències** s'inicia amb la relació entre la inèrcia, el coeficient de contingència χ^2 i la deviància, veient que suposat el model d'independència per a la distribució de dues variables categòriques els estadístics són equivalents. A continuació, s'estudien les descomposicions de la inèrcia

com a l'estudi de les informacions aportades pels individus o variables. En aquest capítol s'estudien les problemàtiques existents en l'anàlisi de correspondències simples de matrius quadrades - simètriques i antisimètriques-, on la inèrcia de la diagonal sol tenir un paper preponderant, exemplificant-ho. Pel que fa referència a les anàlisi de correspondències múltiples de la taula de Burt, es fa la descomposició de les inèrcies per blocs i s'estudia la problemàtica dels blocs diagonals, blocs al seu torn diagonals que s'obtenen per construcció de la matriu per creuament de totes les variables categòriques amb totes.

Es fa un repàs d'algunes de les metodologies per al tractament de les problemàtiques en aquestes anàlisis, basades en el tractament de les inèrcies estructurals, ja sigui per imputació de valors o per reconstitució, o en descomposició en simetria-no simetria. Es fa una proposta nova de metodologia per al tractament de matrius quadrades no simètriques. Aquesta proposta, basada en una doble descomposició, per una part en l'anàlisi de la simetria-no simetria i per altra banda utilitzant la reconstitució factorial de la part simètrica, basada en un algorisme k-EM, on k és l'ordre de reconstitució. Aquesta nova proposta descompon l'anàlisi en tres parts: anomenades anàlisi de la part intra, anàlisi de la part entre i recíproca i anàlisi dels saldos. El treball ha generat un article que ha estat publicat a la revista **Qüestiió. Volum 26, n.3, pàgs. 483-501, 2002**. S'ha exemplificat en el cas de correspondències simples, sobre els moviments deguts al treball entre les comarques catalanes. Per a la taula de Burt, s'ha aplicat la metodologia a la reconstitució k-EM de les taules de la diagonal. Això ens ha portat a una anàlisi que ens dóna un resultat equivalent al que es coneix com a JCA (Joint Correspondence Analysis).

En el **Capítol 5: ACM respecte un model i ACM condicional** la primera part es dedica a presentar les ACM sobre un model, veient-ne com a un cas particular l'anàlisi de correspondències simples, i l'ACM condicional, on s'estudia l'ACM on hi ha un lligam a una variable qualitativa externa. En aquest darrer cas, i ja com a treball original, la nostra aportació es basa en realitzar l'estudi de la inèrcia i la seva descomposició, en dues parts lligades a la variable condicionadora externa, la inèrcia inter i la inèrcia intra. A continuació per tal de determinar l'afectació o no del condicionament en els resultats de l'anàlisi, un cop descomposada la inèrcia de l'ACM condicional, es troba la formulació de la seva distribució i mitjançant aquesta, s'interpreta la importància o no del condicionament. Els resultats s'acompanyen de simulacions per a diferents tipus de relacions entre les variables, generades per models loglineals, així com per a diferents nivells d'interacció. Així doncs, usant l'ACM condicional i els models loglineals introduïts prèviament, estudiem el comportament de la inèrcia, o equivalentment de la deviància, en relació al model i

a diferents nivells d'intensitat de la dependència entre les variables. Aquests resultats han estat acceptats per a la seva publicació en la revista **Computational Statistics** de Springer.

El **Capítol 6: ACM multicondicional** ens permet entrar a treballar en les anàlisis condicionals múltiples amb les problemàtiques que genera la seva implementació. Aquesta part és totalment nova i en ell es desenvolupament aproximacions possibles, ja que no es poden generalitzar trivialment a partir de l'anàlisi condicional, per la no ortogonalitat de les variables condicionadores. Un cop estudiades les possibilitats d'anàlisi, com són l'aproximació mitjançant la inversa generalitzada, es continua amb possibles condicionaments pel cas de dues variables condicionadores. Mitjançant l'estudi de les inèrcies condicionals i els models loglineals es desenvolupa la que presentem com la nostra proposta d'anàlisi multicondicional. No es recorre al procés d'ortogonalització de variables, la solució més fàcil per evitar el problema de la no ortogonalitat, ja que en molts casos objecte d'estudi no és possible la seva aplicació directa per la manera que són proporcionades les dades. Aquests resultats són comparats amb els que s'obtenen en un procés de modelització loglineal. Com a finalització en el Capítol 7, s'aplica la proposta d'anàlisi desenvolupada a un exemple de dades.

El treball de recerca finalitza amb un **Epíleg** en les quals es fa un resum de les principals aportacions del treball i s'indiquen quines podrien ser algunes de les línies de recerca futures en aquest camp. Després trobem la Bibliografia i s'inclouen en forma d'apèndix algunes de les macros que s'han programat en el paquet estadístic Minitab per tal d'obtenir les anàlisis, les descomposicions de la inèrcia i els models loglineals desitjats.

1.4 Notació

La notació que s'ha emprat es pot veure que en el transcurs del treball de recerca ha estat modificada, no havent-hi una notació més o menys estàndard en aquesta branca d'investigació. Aquesta modificació progressiva és deguda al fet que en uns primers estadis quan es treballa amb poques variables es pot detallar molt la notació, però a mesura que condicionàvem i aplicàvem condicionaments múltiples, no ens podíem permetre aquest detall. Per altra banda alguna de la terminologia utilitzada no ha estat trobada en els mitjans al nostre abast, per tant hem cregut convenient detallar-la aquí, indicant la terminologia equivalent en altres llengües, aquesta és: deviància - *deviance* i clica *clique*

Capítol 2

Anàlisis Factorials de Dades

En aquest capítol i sota el títol d'anàlisis factorials, tractarem les anàlisis multivariants de dades categòriques, que sota aquest o d'altres noms com anàlisis factorials descriptives, anàlisis multivariants no lineals,... s'han vingut desenvolupant des dels anys trenta del segle XX. Farem una primera introducció històrica per a passar després a descriure els procediments emprats i a una descripció dels mètodes més usuals.

2.1 Orígens Històrics

L'Anàlisi de Correspondències, s'originà a França durant la dècada del 60 del treball de Jean-Paul Benzécri [Ben64] i col·laboradors, entre els quals destaquem el treball exposat en la tesi doctoral de l'any 1969 de Brigitte Escofier [EC65], i fou en el camp de la lingüística, en l'estudi de taules del modern llenguatge xinès on apareixen en les fileres d'aquestes taules les consonants i en les columnes les vocals finals. En aquestes taules de doble entrada, taules de contingència, les dades consistien en les indicacions de quines combinacions filera-columna eren permeses en el llenguatge o no. D'aquí sorgeix el terme **correspondència** per denotar l'associació entre els dos elements dels dos conjunts de fileres i columnes.

L'Anàlisi de Correspondències és una tècnica multidimensional que està basada en una àlgebra i ens uns procediments numèrics i que té la finalitat d'obtenir una representació gràfica i unes eines d'ajut a la interpretació. Els principis subjacents d'aquesta metodologia algèbrica tenen, però, uns antecedents força antics: es remunten a treballs de Hirschfeld [Hir35](més tard conegut com H.O. Hartley) on dóna una formulació algèbrica de *correlació* entre les fileres i les columnes d'una taula de contingència. Horst [Hor35], independentment, va suggerir idees similars des d'un punt de vista no matemàtic en la

literatura psicomètrica, usant el terme de *method of reciprocal averages*.

Més tard, R.A. Fisher l'any 1940 [Fis40], modificà la mateixa teoria, en el camp biomètric, derivant-la cap a una anàlisi discriminant d'una taula de contingència i l'aplicà al conegut exemple del color dels ulls i dels cabells d'un grup d'escolars sota el nom de *anàlisi canònica de correlacions*.

Mentrestant, Louis Guttman [Gut35] mentre treballava en l'escalatge de dades categòriques, independentment arribava als mateixos resultats en un camp diferent, així com tractava el cas general de més de dues variables categòriques, en el que ara anomenem Anàlisi de Correspondències Múltiples.

Aquest origen independent ha portat al desenvolupament paral·lel de dues escoles: l'escola biomètrica, Fisher, i l'escola psicomètrica, Guttman, on el mètode ha estat anomenat *reciprocal averaging* i *optimal scaling* respectivament. Aquest últim el podem trobar sota el nom de *dual scaling* a l'obra de Nisishato [Nis80].

Al Japó encapçalats per Chikio Hayashi [Hay52] desenvoluparen la idea proposada per Guttman d'escalatge de variables categòriques en el que ells anomenaren *quantificació de dades qualitatives -quantification method-*.

Aquestes tècniques i moltes d'altres com *homogeneity analysis, dual scaling, scalogram analysis*, biplots, anàlisi canònica, anàlisi canònica generalitzada, anàlisis discriminants, anàlisi en components principals són tècniques que comparteixen una mateixa teoria i uns mateixos procediments computacionals que l'anàlisi de correspondències, podent ser agrupades sota el nom d'**anàlisis factorials descriptives** (AFD).

Amb el desenvolupament del ordinadors i sobretot amb la gran empenta de l'escola francesa, on trobem les tècniques descriptives basades en la geometria i en una rica notació algèbrica el seu desenvolupament continua.

Aquesta notació, que en els orígens era una notació tensorial, distintiva de l'escola francesa, modernament ha estat substituïda per la notació matricial, a la qual la majoria dels estadístics estan més acostumats i en permet una major difusió, puix que aquesta notació tensorial fou assenyalada com una de les causes de la seva poca difusió inicial.

Un altre fet que influí en la seva poca difusió, fou el fet de la barrera lingüística que va representar el francès per a les escoles anglosaxones. Aquesta barrera fou trencada per la difusió d'aquestes tècniques per autors com Gifi, Ramsay i de Leeuw i per l'aparició de publicacions en llengua anglesa tals com són l'obra de Lebart i altres [LMW84] i Greenacre [Gre84] on introduí una altra interpretació de l'anàlisi de correspondències.

Actualment les AFD són presents en els més importants paquets estadístics d'anàlisi multivariant de dades.

Posteriorment podríem destacar l'obra de molts altres autors que han treballat i estan treballant encara en aquesta ciència que encara està en constant evolució, on en podem destacar l'apropament a tècniques de tractament de dades basades en l'ús de models [HL85] [Goo86] [CdF87] [HFL89], les anàlisis no simètriques [LB94], les anàlisis a tres vies [CK96] o les anàlisis parcials [Non92], etc.

2.2 Introducció a les Anàlisis Factorials de Dades

Un problema estàndard en l'Estadística, davant la complexitat de la multidimensionalitat, és *trobar procediments, tècniques per analitzar-ne els resultats, amb la finalitat de fer aquestes anàlisis més fàcils* com ja deia J. Tukey [Tuk62], i aquesta és la recerca essencial de l'Anàlisi Factorial Descriptiva (AFD). Això la majoria de vegades ha conduït a cercar una reducció de la dimensionalitat, és a dir, a recercar un subespai S, de dimensió menor a la dimensió de l'espai de les dades originals, on projectarem les nostres dades, per analitzar en aquest subespai les dades de manera que ens resulti més fàcil l'anàlisi.

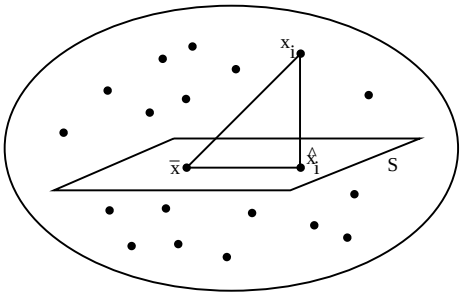


Figura 2.1: Projeccions de les dades sobre el subespai S

El nostre objectiu el tindrem, moltes vegades, en la recerca del subespai S de dimensió K on les nostres dades projectades guardin la 'major relació' amb les dades originals. Per tant, haurem d'establir els procediments per:

- trobar el subespai de projecció
- decidir criteri de 'millor' subespai
- projectar dades
- mesurar 'relacions' entre les dades originals i les dades projectades

Una expressió de 'major relació' la podem tenir en, per exemple, que siguin mínimes les distàncies ponderades al quadrat entre els punts i llurs projeccions, tenint present que no totes les dades han de tenir el mateix pes, és a dir, la mateixa importància:

$$\min_S \left(\sum_i \omega_i d_M^2(x_i - \hat{x}_i) \right)$$

on x_i és la dada original, $\hat{x}_i$ és la seva projecció en S i w_i és el pes de la unitat i i M és la mètrica usada per a definir la distància.

Podem veure que aquesta condició és equivalent a la recerca del subespai on la projecció del núvol conserva al màxim l'estructura del núvol original, és a dir, la recerca del subespai on la projecció del núvol de dades conservi la màxima inèrcia possible del núvol original:

$$\max_S \left(\sum_i \omega_i d_M^2(x_i - \bar{x}) \right)$$

on $\bar{x}$ és el centre de gravetat:

$$\bar{x} = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

La prova d'això la tenim en que si suposem que $\bar{x} \in S$, fàcilment comprovable, llavors per a cada dada x_i

$$d_M^2(\hat{x}_i - \bar{x}) + d_M^2(x_i - \hat{x}_i) = d_M^2(x_i - \bar{x})$$

i com que $\sum d_M^2(x_i - \bar{x})$ és constant per a cada núvol de punts, imposar que $\sum d_M^2(\hat{x}_i - \bar{x})$ sigui màxim, és equivalent a imposar que $\sum d_M^2(x_i - \hat{x}_i)$ sigui mínim.

I acabem d'introduir un dels criteris que ens portarà la nostra selecció del subespai **la inèrcia**, com a criteri de selecció del 'millor' subespai.

Tenim doncs els següents elements: una matriu de dades, una matriu de pesos per ponderar les dades i una mètrica que ens permet calcular distàncies, fer projeccions i calcular inèrcies.

Siguin:

Matriu de dades: $X_{I*J} = \begin{pmatrix} x_{11} & \dots & x_{1J} \\ \dots & & \dots \\ x_{I1} & \dots & x_{IJ} \end{pmatrix}$

Matriu de pesos: W_{I*I} matriu diagonal

Mètrica: M_{J*J} matriu simètrica i definida positiva

Sovint la mètrica pot ser una mètrica euclídea o euclídea ponderada, és a dir, $M_{J*J} = D_q$, matriu diagonal, on les elements de la diagonal, $q_1, \ldots, q_I$ són valors positius, que ens donen els pesos assignats a cada dimensió. Així:

$$d^2_M(x,y) = d^2_{D_q}(x,y) = (x-y)^T D_q (x-y) = \sum_j q_j (x_j - y_j)^2$$

Per altra banda, la solució al problema de trobar el subespai S de dimensió K que ens minimitzi la suma dels quadrats ponderats entre les observacions i les seves projeccions, o el que és el mateix la pèrdua d'inèrcia, està relacionat amb l'anomenada descomposició en valors singulars (DVS) o *estructura bàsica*, eina ben coneguda en àlgebra, que inclou el concepte de descomposició en valors i vectors propis.

2.3 Metodologia de la DVS

Introduïm a continuació les tècniques equivalents de la descomposició en valors singulars (DVS) i la descomposició en valors singulars generalitzada (DVSG), on veurem que la diferència es troba en la introducció de les mètriques en les dades.

2.3.1 Descomposició en Valors Singulars

Donada una matriu A_{I*J} qualsevol de rang K, la descomposició en valors singulars (DVS) és la possibilitat de poder expressar aquesta matriu de la forma:

$$A_{I*J} = U_{I*K} D_{K*K} V^T_{K*J}$$

és a dir

$$A = \sum_k \alpha_k u_k v_k^T$$

on

$$U^T U = V^T V = I \qquad \text{i} \qquad \alpha_1 \geq \ldots \geq \alpha_K \geq 0$$

Els K vectors ortonormals $u_1, \ldots, u_K \in \mathbb{R}^I$ d'U són anomenats els vectors singulars esquerra i els vectors ortonormals $v_1, \ldots, v_K \in \mathbb{R}^J$ de V són anomenats els vectors singulars dreta. Els $u_1, \ldots, u_K$ són una base ortonormal per les columnes d'A i són vectors propis de la matriu AA^T amb valors propis $\alpha_1^2, \ldots, \alpha_K^2$.

Anàlogament els $v_1, \ldots, v_K$ són una base ortonormal per les fileres d'A i són vectors propis de la matriu $A^T A$ amb valors propis $\alpha_1^2, \ldots, \alpha_K^2$.

Si nosaltres despreciem els darrers termes de la DVS d'A, llavors una aproximació d'A serà

$$A^* = \sum_{k=1}^{K^*} \alpha_k u_k v_k^T$$

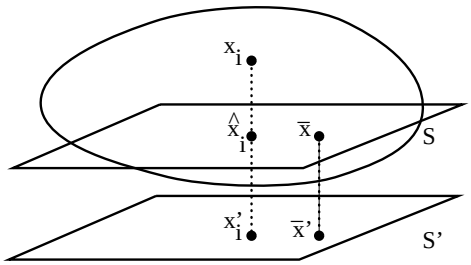
i s'anomena l'aproximació mínim quadràtica d'A de rang K^*, en el sentit que minimitza

$$\sum\sum (a_{ij} - l_{ij})^2 = \text{traça}((A-L)(A-L)^T)$$

entre totes les matrius L de rang K^* (Teorema d'Eckart-Young [EY36]).

Una propietat és que l'aproximació mínim quadràtica conté $\bar{x}$, la qual cosa gràficament la hem suposat.

Proposició 2.3.1.1 *Siguin $x_1, \ldots, x_I \in \mathbb{R}^J$ amb masses respectives $\omega_1, \ldots, \omega_I$ Podem considerar sense pèrdua de generalitat que $\sum \omega_i = 1$. Llavors S, l'aproximació mínim quadràtica, conté $\bar{x}$*



Demostració: Considerem S' la millor aproximació i S l'aproximació que conté $\bar{x}$. Si S' no contingués $\bar{x}$, llavors existiria $\bar{x}'$ que seria la millor aproximació de $\bar{x}$.

Llavors qualsevol $\hat{x}_i \in S$, podria ser expressat com $\hat{x}_i = x_i' + v$ on x_i' és la millor aproximació en S' i $v = \bar{x} - \bar{x}'$ és el vector constant que va de $\bar{x}'$ a $\bar{x}$.

Llavors en S' tindríem:

$$\sum_i \omega_i (x_i - x_i')^T D_q (x_i - x_i') = \sum_i \omega_i (x_i - \hat{x}_i + \hat{x}_i - x_i')^T D_q (x_i - \hat{x}_i + \hat{x}_i - x_i') =$$

$$\sum_i \omega_i (x_i - \hat{x}_i)^T D_q (x_i - \hat{x}_i) + \sum_i \omega_i (\hat{x}_i - x_i')^T D_q (\hat{x}_i - x_i') + \sum_i \omega_i (x_i - \hat{x}_i)^T D_q (\hat{x}_i - x_i')$$

Com $\hat{x}_i - x_i' = v$ és un vector constant, $\sum_i \omega_i (\hat{x}_i - x_i')^T D_q (\hat{x}_i - x_i') \geq 0$ i també $D_q(\hat{x}_i - x_i')$ és constant, llavors:

$$\sum_i \omega_i(x_i - \hat{x}_i) = \sum_i \omega_i x_i - \sum_i \omega_i \hat{x}_i = \bar{x} - \sum_i \omega_i(x'_i - \bar{x} + \bar{x}') =$$

$$= \bar{x} - \bar{x}' - \bar{x} + \bar{x}' = 0$$

i per tant

$$\sum_i \omega_i(x_i - x'_i)^T D_q(x_i - x'_i) \geq \sum_i \omega_i(x_i - \hat{x}_i)^T D_q(x_i - \hat{x}_i)$$

□

2.3.2 Descomposició en Valors Singulars Generalitzada

Anem ara a introduir una generalització de la DVS, la Descomposició en Valors Singulars Generalitzada (DVSG) on s'introdueixen dues mètriques noves, no restringint-ho a la mètrica euclídea clàssica I.

Si considerem dues matrius simètriques definides positives Ω_{I*I} i Φ_{J*J}, llavors qualsevol matriu A_{I*J} de rang K, pot ser expressada com:

$$A_{I*J} = N_{I*K} D_{K*K} M^T_{K*J} = \sum_k \alpha_k n_k m_k^T$$

on les columnes d'N i M són ortonormals respecte de Ω i Φ respectivament

$$N^T \Omega N = M^T \Phi M = I$$

Aquesta descomposició s'anomena la descomposició en valors singulars generalitzada (DVSG).

La DVSG és fàcilment demostrable a partir de la DVS. Si considerem la matriu

$$\Omega^{1/2} A \Phi^{1/2}$$

on per ser Φ simètrica, $\Phi = W D_\mu W^T$ i per tant $\Phi^{1/2} = W D_\mu^{1/2} W^T$ (anàlogament per Ω), llavors, la DVS d'aquesta matriu és:

$$\Omega^{1/2} A \Phi^{1/2} = U D_\alpha V^T$$

amb

$$U^T U = V^T V = I$$

si fem

$$N \equiv \Omega^{-1/2} U \qquad \text{i} \qquad M \equiv \Psi^{-1/2} V$$

llavors

$$I = U^T U = N^T \Omega^{1/2} \Omega^{1/2} N = N^T \Omega N$$

$$I = V^T V = M^T \Psi^{1/2} \Psi^{1/2} M = M^T \Psi M$$

Per tant, A pot ser descomposada com:

$$A = NDM^T = \sum_k \alpha_k n_k m_k^T \qquad \text{on} \qquad N^T \Omega N = M^T \Phi M = I$$

i tenim la descomposició volguda.

Si prenem com a Φ la matriu diagonal de les masses, $\Phi = D_w$, i com a Ω la matriu diagonal de la mètrica euclídea ponderada, $\Omega = D_q$, llavors la millor aproximació de rang K^* de la matriu A serà

$$A^* = N^* D^* M^{*T} = \sum^{K^*} \alpha_k n_k m_k^T$$

en el sentit que serà la que ens minimitza l'expressió de la norma de Hilbert-Schmidt segons D_q i D_w, segons l'expressió donada a continuació:

$$\| A - L \|^2_{D_q, D_w} = \sum_i \sum_j w_i q_j (a_{ij} - l_{ij})^2 =$$

$$= \sum_i w_i \sum_j q_j (a_{ij} - l_{ij})^2 = \sum_i w_i (a_i - l_i)^T D_q (a_i - l_i) = \sum_i w_i \| a_i - l_i \|^2_{D_q}$$

entre totes les matrius L de rang K^*.

Per tant el nostre problema metodològic, consistirà en realitzar la descomposició en valors singulars generalitzada de la tripleta (X, D_w, D_q) formada per les nostres matriu de dades, matriu de pesos i mètrica i escollir l'aproximació de la dimensió volguda.

Introduïm ara una proposició que ens relaciona la DVSG de la matriu de dades centrada i la DVSG de la matriu sense centrar, ja que l'operació de centrat serà habitual en el tractament previ de les dades.

Proposició 2.3.2.1 *Si tenim una matriu de dades X_{I*J}, on la suma de les fileres és constant, amb masses pels elements filera D_w i on la mètrica D_q compleix que q és inversament proporcional al centre de gravetat dels elements filera, llavors la DVSG de la matriu centrada X^c pot ser obtinguda de la DVSG de la matriu sense centrar X.*

Demostració:

Sigui

$$X^c = X - 1\bar{x}^T \quad \text{on} \quad \bar{x} = \frac{\sum \omega_i x_i}{\sum \omega_i} = \frac{X^T\omega}{1^T\omega}$$

i considerem la DVSG de X^c

$$X^c = ND_\mu M^T = \sum_{k=1}^{K} d_k n_k m_k^T \quad \text{amb} \quad N^T D_w N = M^T D_q M^T = I$$

Llavors

$$X = ND_\mu M^T + 1\bar{x}^T$$

Si vegem que

$$1 \perp_{D_w} N \quad \text{i} \quad \bar{x}^T \perp_{D_q} M$$

tindrem la descomposició volguda.

Prenem $X - 1\bar{x}^T = ND_\mu M^T$ si ho premultipliquem per $N^T D_w$

$$N^T D_w(X - 1\bar{x}^T) = N^T D_w N D_\mu M^T \iff N^T D_w(X - 1\bar{x}^T) = D_\mu M^T$$
$$\iff D_\mu^{-1} N^T D_w(X - 1\bar{x}^T) = M^T \iff M = (X^T - 1^T\bar{x})D_w N D_\mu^{-1}$$

si premultipliquem per $\bar{x}^T D_q$

$$\bar{x}^T D_q M = \bar{x}^T D_q(X^T - \bar{x}1^T)D_w N D_\mu^{-1} \tag{2.1}$$

com les hipòtesis de l'enunciat són que la mitjana és inversament proporcional als pesos de la mètrica

$$\bar{x} = \alpha D_q^{-1} 1 \Leftrightarrow D_q\bar{x} = \alpha 1 \Leftrightarrow \bar{x}^T D_q = \alpha 1^T$$

i que la suma de les fileres ha de ser constant (diem-ne c)

$$X.1 = c.1 \Leftrightarrow 1^T.X^T = c.1^T$$

si substituïm aquestes hipòtesis a (2.1)

$$\alpha 1^T(X^T - \bar{x}1^T)D_w N D_\mu^{-1} = (\alpha 1^T X^T - \alpha 1^T\bar{x}1^T)D_w N D_\mu^{-1} = 0$$

per ser $1^T.X^T = c.1^T$ i $1^T\bar{x} = 1^T X^T w/(1^T w) = c1^T w/(1^T w) = c$

Anàlogament podrem expressar N com:

$$N = (X - 1\bar{x}^T)D_q M D_\mu^{-1} \iff 1^T D_w N = 1^T D_w(X - 1\bar{x}^T)D_q M D_\mu^{-1}$$

$$1^T D_w(X - 1\bar{x}^T) = w^T(X - 1\bar{x}^T) = w^T X - w^T 1\bar{x}^T = w^T X - w^T 1\frac{(X^T w)^T}{(1^T w)^T}$$

$$= w^T X - w^T 1 w^T X/(w^T 1) = w^T X - w^T X = 0^T$$

per tant hem vist l'ortogonalitat.

Si prenem n_0 l'ortonormalització de l'1, m_0^T l'ortonormalització de $\bar{x}^T$ i d_0 el valor propi associat

$$X^c = X - 1\bar{x}^T = \sum_{k=1}^{K} d_k n_k m_k^T + d_0 n_0 m_0^T = \sum_{k=0}^{K} d_k n_k m_k^T$$

on com $1^T D_w 1 = 1^T w$ i $\bar{x}^T D_q \bar{x} = \alpha \bar{x}^T D_q D_q^{-1} 1 = \alpha \bar{x}^T 1 = \alpha c$ les ortonormalitzacions són:

$$n_0 = \frac{1}{(1^T w)^{1/2}} \quad m_0 = \frac{\bar{x}}{(\alpha c)^{1/2}} \quad \text{i} \quad d_0 = (\alpha c 1^T w)^{1/2}$$

i a més el valor propi d_0 és el més gran ja que $d_0 n_0 m_0^T = 1\bar{x}^T$ conté el centre de gravetat i per tant és la millor aproximació mínim quadràtica. □

2.3.3 La maximització de la inèrcia projectada

Anem a veure des d'un altre punt de vista el problema de trobar la millor projecció sobre un eix, en el sentit de maximitzar la inèrcia de la projecció. Sigui $u \in \mathbb{R}$ vector unitari, $u^t M u = 1$, que ens genera un eix qualsevol, llavors la projecció del núvol de punts sobre l'eix serà:

$$v = XMu$$

i la inèrcia de les projeccions serà igual a:

$$v^t N v \tag{2.2}$$

Voldrem ara fer màxima la quantitat (2.2) amb la restricció $u^t M u = 1$, és a dir:

$$max \left\{ u^t M X^t N X M u \quad | \quad u^t M u = 1 \right\} \tag{2.3}$$

Per a la solució de l'equació (2.3) apliquem el mètode dels multiplicadors de Lagrange

$$u^t M X^t N X M u - \lambda(u^t M u - 1) = 0$$

$$u^t M X^t N X M u - \lambda u^t M u + \lambda = 0$$

Derivant parcialment:

$$\frac{\partial}{\partial u}\left(u^t M X^t N X M u - \lambda u^t M u + \lambda\right) = 0$$

$$2MX^tNXMu - 2\lambda Mu = 0$$

$$X^tNXMu = \lambda u \tag{2.4}$$

D'on s'obté que u és un vector propi de valor propi λ de X^tNXM

Aplicant (2.4) a la condició de màxim de (2.3) obtenim:

$$u^tMX^tNXMu = u^tM\lambda u = \lambda u^tMu = \lambda 1 = \lambda$$

Per tant, la direcció que ens maximitzarà la inèrcia es correspon al vector propi associat al major valor propi -màx(λ)- de la matriu X^tNXM.

I es pot demostrar successivament que les direccions que ens maximitzen les inèrcies són les determinades pels vectors propis de X^tNXM associats als valors propis ordenats de major a menor (vegeu, per exemple [LMW84] I.5).

Així doncs, el procés serà diagonalitzar X^tNXM, obtenir-ne la seqüència de valors propis ordenats i els vectors associats a aquests.

I tornem a trobar el problema original de diagonalitzar un matriu, donades una mètrica i uns pesos.

2.4 L'Anàlisi Canònica i la comparació de grups de variables

Fins ara hem vist com fer la síntesi d'una taula, en aquesta secció introduirem l'anàlisi canònica **(AC)** i l'anàlisi canònica generalitzada **(ACG)** com a metodologies per a estudiar les relacions entre dos o més de dos grups de variables que ens descriuen un mateix grup d'individus. Ens interessarà especialment l'existència de factors comuns o factors associats a només una de les taules. Aquestes relacions ens permetran examinar els lligams entre aquests grups de variables i veure si mesuren o no les mateixes propietats i, per tant, si podem prescindir-ne d'alguna d'elles. Veurem en primer lloc l'anàlisi canònica, per a dos grups de variables, i després el generalitzarem a més de dos grups de variables.

2.4.1 L'Anàlisi Canònica

Suposem que tenim un grup d'n individus descrit per dos conjunts de p i q variables, i volem conèixer si aquestes dos grups de variables ens mesuren el mateix o no.

Les taules de dades per analitzar són:

$$X_1 = \begin{array}{c} \\ 1 \\ 2 \\ \cdot \\ \cdot \\ \cdot \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & p \end{array} \\ \left[\begin{array}{c} \\ \\ \\ x_{ij} \\ \\ \\ \\ \end{array}\right] \end{array} \qquad X_2 = \begin{array}{c} \\ 1 \\ 2 \\ \cdot \\ \cdot \\ \cdot \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & q \end{array} \\ \left[\begin{array}{c} \\ \\ \\ y_{ij} \\ \\ \\ \\ \end{array}\right] \end{array}$$

Considerem a més una mètrica diagonal D en l'espai dels individus $\mathbb{R}^n$ que ens pondera aquests segons uns pesos p_i, $p_i > 0$ amb $\sum p_i = 1$.

Si les variables estan centrades ($\sum_i p_i x_{ij} = 0 \forall j$), on x_{ij} representa a X_1 o X_2 indistintament, la covariància entre dues d'elles s'obté com:

$$Cov(x^j, x^k) = (x^j)^t D x^k$$

i $X^t D X$ és la matriu de covariàncies. A més, l'angle entre x^j i x^k, θ_{jk}, compleix:

$$\cos\theta_{jk} = \frac{< x^j, x^k >_D}{\parallel x^j \parallel \parallel x^k \parallel} = \frac{(x^j)^t D x^k}{\sqrt{(x^j)^t D x^j}\sqrt{(x^k)^t D x^k}}$$

expressió que no és sinó la correlació entre les variables x^j i x^k, quan aquestes estan centrades. Tenim a més com a matriu d'angles:

$$\cos(XY) = (X^t DY)(X^t DX)^{-1/2}(Y^t DY)^{-1/2}$$

Per a la tasca de veure si X_1 i X_2 ens mesuren o no el mateix, considerem els subespais generats per les columnes de X_1 i X_2: W_1 i W_2

$$W_1 = \{x | x = X_1 a\} \qquad i \qquad W_2 = \{y | y = X_2 b\}$$

En els casos extrems, si $W_1 \equiv W_2$, llavors X_1 i X_2 ens mesuren el mateix. Si $W_1 \perp W_2$, llavors X_1 i X_2 ens mesuren coses totalment diferents. Per tant, anem a estudiar la posició geomètrica relativa de W_1 i W_2, cercant els elements més propers d'aquests dos subespais, la qual cosa ens permetrà, en particular, conèixer la $dim(W_1 \cap W_2)$.

Aquesta anàlisi també és coneguda com a l'Anàlisi Canònica de Correlacions **(ACC)**, donada per Hotelling([Hot36]), ja que el que busquem és trobar la màxima correlació entre dos elements de W_1 i W_2.

L'**ACC** es pot veure també com una extensió de la regressió múltiple, ja que en aquesta X_2 seria un conjunt amb una única variable ($q = 1$).

2.4.2 Recerca de les variables canòniques

La tècnica emprada és la següent: trobar la parella (ξ_1, η_1) de vectors normalitzats, $\xi_1 \in W_1$ i $\eta_1 \in W_2$, que formen l'angle més petit. ξ_1 i η_1, anomenades variables canòniques, són combinacions lineals respectivament de les variables del primer i segon grup.

A continuació cercarem una parella (ξ_2, η_2), amb ξ_2 D-ortogonal a ξ_1 i η_2 D-ortogonal a η_1, tal que l'angle sigui mínim, i així anar fent. Obtindrem finalment p parelles de variables canòniques (suposem, sense pèrdua de generalitat, que $p = dim(W_1)$ i $q = dim(W_2)$ amb $p \leq q$).

Siguin A_1 i A_2 els operadors de projecció D-ortogonal sobre W_1 i W_2. És fàcil verificar que:

$$A_1 = X_1(X_1^t D X_1)^{-1} X_1^t D$$
$$A_2 = X_2(X_2^t D X_2)^{-1} X_2^t D$$

Geomètricament, el problema és de trobar les dues direccions dels subespais que són més properes (Figura 2.2).

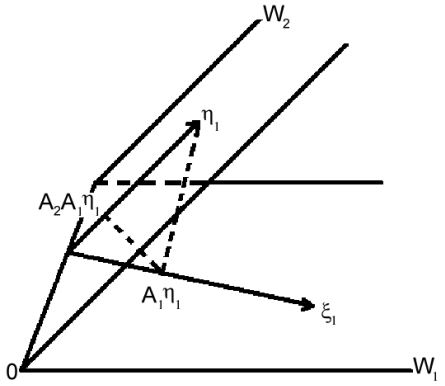


Figura 2.2: Variables canòniques

Solució dins $\mathbb{R}^n$

Volem trobar ξ_1 i η_1 tals que $cos(\xi_1, \eta_1)$ sigui màxim. Si $\xi_1 \neq \eta_1$, llavors es pot veure que $A_1\eta_1$ ha de ser colineal a $\xi_1 \in W_1$ i recíprocament, $\eta_1 \in W_2$ ha de ser el més proper a ξ_1, per tant $A_2\xi_1$ ha de ser colineal a η_1, i al seu torn η_1 ha de ser colineal a $A_2A_1\eta_1$:

$$A_2A_1\eta_1 = \lambda_1\eta_1$$

En resum: el nostre problema esdevé trobar els valors propis i vectors propis de la matriu A_2A_1.
Observacions:

- Es pot demostrar que A_2A_1 és diagonalitzable, amb λ_i valors propis reals positius, car A_1 i A_2 són matrius positives.
- Aquesta descomposició és equivalent a la de A_1A_2.
- λ_1 representa el quadrat del cosinus de l'angle format per η_1 i ξ_1, per tant $\lambda_1 \leq 1$, ja que com:

  $A_2A_1\eta_1 = \lambda_1\eta_1 = X_2(X_2^tDX_2)^{-1}X_2^tDX_1(X_1^tDX_1)^{-1}X_1^tD\eta_1$

  i com $\eta_1 \in W_2 \Rightarrow \eta_1 = X_2b$,

  llavors $X_2(X_2^tDX_2)^{-1}X_2^tDX_1(X_1^tDX_1)^{-1}X_1^tDX_2b = \lambda_1X_2b$

  i si aquesta expressió la premultipliquem per X_2^tD,

  $X_2^tDX_2(X_2^tDX_2)^{-1}X_2^tDX_1(X_1^tDX_1)^{-1}X_1^tDX_2b = \lambda_1X_2^tDX_2b$ i per tant

  llavors ens surt l'expressió $cos^2(X_1, X_2) = \lambda_1$.
- Si $\lambda = 1$, $\Rightarrow \eta_1 = \xi_1$ d'on obtenim $\eta_1 \in W_1 \cap W_2$
- L'ordre de multiplicitat del valor propi 1 és la dimensió de $W_1 \cap W_2$
- Els vectors propis associats al valor propi 0 ens generen la part de W_2 ortogonal a W_1

Es pot veure que aquesta resolució és anàloga a la de trobar els vectors a i b que ens proporcionen les combinacions lineals $\xi_1 = X_1a$ i $\eta_1 = X_2b$ tal que el cosinus de l'angle $cos(\xi_1, \eta_1)$ o la correlació $\rho(\xi_1, \eta_1)$ sigui màxima, sota la restricció de tenir els vectors a i b normalitzats.

Com a conclusió tenim les relacions següents:

$$\begin{aligned} A_2A_1\eta_i = \lambda_i\eta_i \quad & \sqrt{\lambda}_i\eta_i = A_2\xi_i \\ A_1A_2\xi_i = \lambda_i\xi_i \quad & \sqrt{\lambda}_i\xi_i = A_1\eta_i \\ \xi_i'D\xi_j = 0 \quad & \eta_i'D\eta_j = 0 \quad i \neq j \\ \text{i a més:} \quad & \eta_i'D\xi_j = 0 \quad i \neq j \end{aligned}$$

Solució dins $\mathbb{R}^p$ i $\mathbb{R}^q$

Les variables canòniques ξ_1 i η_1 es poden expressar com a combinacions lineals de les columnes de X_1 i X_2:

$$\xi_i = X_1a_i \qquad i \qquad \eta_i = X_2b_i$$

Els a_i i b_i són els factors canònics que es poden obtenir directament de la següent manera:

$$A_1A_2\xi_i = \lambda_i\xi_i \Leftrightarrow A_1A_2X_1a_i = \lambda_iX_1a_i$$

si reemplacem els projectors per llur expressió:

$$X_1(X_1^tDX_1)^{-1}X_1^tDX_2(X_2^tDX_2)^{-1}X_2^tDX_1a_i = \lambda_iX_1a_i$$

on si premultipliquem per $(X_1^tDX_1)^{-1}X_1^t$, podem simplificar i obtenir:

$$(X_1^tDX_1)^{-1}X_1^tDX_2(X_2^tDX_2)^{-1}X_2^tDX_1a_i = \lambda_ia_i$$

i anàlogament:

$$(X_2^tDX_2)^{-1}X_2^tDX_1(X_1^tDX_1)^{-1}X_1^tDX_2b_i = \lambda_ib_i$$

En el cas que les variables siguin centrades, $X_1^tD1 = X_2^tD1 = 0$, les matrius $X_i^tDX_j$ són les matrius de covariàncies.

Amb les notacions:

$$\begin{aligned} V_{11} = X_1^tDX_1 \qquad & V_{12} = X_1^tDX_2 \\ V_{22} = X_2^tDX_2 \quad & V_{21} = X_2^tDX_1 = V_{12}^t \end{aligned}$$

podem escriure les equacions dels factors canònics com:

$$V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}a_i = \lambda_ia_i \tag{2.5}$$

$$V_{22}^{-1}V_{21}V_{11}^{-1}V_{12}b_i = \lambda_ib_i \tag{2.6}$$

on el producte de matrius són els coeficients de correlació al quadrat i els vectors propis λ_i són els quadrats dels coeficient de correlació canònica entre les variables canòniques. Com que tenim $\xi_i = X_1 a_i$ i $\eta_i = X_2 b_i$, per tal que les variables canòniques siguin de variància unitària, les normalitzarem de la manera següent:

$a_i^t V_{11} a_i = 1$, és a dir V_{11} és una mètrica en l'espai $\mathbb{R}^p$ i la qual cosa és equivalent a tenir la mètrica D sobre els η_i, $a_i^t V_{11} a_i = a_i^t X_1^t D X_1 a_i = \eta_i^t D \eta_i$

Anàlogament, $b_i^t V_{22} b_i = 1$ és a dir V_{22} és una mètrica en l'espai $\mathbb{R}^q$ i podem obtenir les relacions de transició:

$$b_i = \frac{1}{\sqrt{\lambda_i}} V_{22}^{-1} V_{21} a_i \qquad i \qquad a_i = \frac{1}{\sqrt{\lambda_i}} V_{11}^{-1} V_{12} b_i$$

2.4.3 Representació de les variables

Són possibles dues classes de representacions segons s'escullin les variables canòniques de W_1 o de W_2. Si escollim W_1, representarem les variables inicials (columnes de X_1 i X_2), D-normades, projectades sobre la base D-ortonormal formada pels ξ_i.

En particular, la projecció sobre el pla engendrat per ξ_1 i ξ_2 dóna la figura anomenada cercle de correlacions, ja que si les columnes d'X_1 són D-normades, així com les d'X_2, les components en la base dels ξ_i són els coeficients de correlació entre les variables inicials i les variables canòniques.

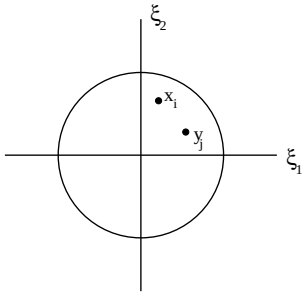


Figura 2.3: Representació cercle de correlacions

Si x_k és la k-èsima columna d'X_1, tenim que la projecció sobre ξ_1 és:

$$x_k' D \xi_1 = x_k' D X_1 a_1.$$

És a dir, com:

$$X_1^t D X_1 a_1 = V_{11} a_1 \tag{2.7}$$

la projecció és la k-èsima component de l'expressió anterior (2.7), és dir el coeficient de correlació.

Anàlogament, si y_l és la l-èsima columna d'X_2:

$$y_l' D \xi_1 = \delta_l' X_2' D X_1 a_1$$

és la l-èsima component de $V_{21} a_1$.

Taula 2.1: Coordenades de les variables en l'anàlisi canònica

Element representat	Representació sobre X_1	Representació sobre X_2
columnes X_1	$V_{11} a_1$	$V_{12} b_1$
columnes X_2	$V_{21} a_1$	$V_{22} b_1$

2.4.4 Representació dels individus

Podem, aquí també, representar els individus de dues maneres diferents segons les variables canòniques emprades.

Si escollim el pla definit per (ξ_1, ξ_2), les coordenades del j-èsim punt són les components j-èsimes de les variables canòniques ξ_1 i ξ_2.

Taula 2.2: Coordenades dels individus en l'anàlisi canònica

Element representat	Representació sobre X_1	Representació sobre X_2
Individus	ξ_1	η_1

Una propietat de l'anàlisi canònica que farem servir posteriorment és la següent:

Propietat 2.4.4.1 *Si tenim X_1 i X_2 centrades, les variables canòniques ξ i η vectors propis de $A_1 A_2$ i $A_2 A_1$ respectivament, tenen la següent propietat: $\xi + \eta$ és vector propi de $A_1 + A_2$*

Demostració:
suposem z, tal que $(A_1 A_2) z = \mu z$. Si premultipliquem per A_1 o A_2

$$A_1(A_1 + A_2) z = \mu A_1 z \qquad \Leftrightarrow \qquad A_1 z + A_1 A_2 z = \mu A_1 z$$

$$A_1 A_2 z = (\mu - 1) A_1 z \qquad \text{i} \qquad A_2 A_1 z = (\mu - 1) A_2 z$$

$$\begin{array}{l} A_1A_2A_1z = (\mu-1)^2A_1z \\ A_2A_1A_2z = (\mu-1)^2A_2z \end{array} \Rightarrow \begin{array}{l} A_1z = \xi \\ A_2z = \eta \end{array}$$

$$A_1z + a_2z = \mu Z \Rightarrow \mu z = \xi + \eta$$

$$(A_1 + A_2)(\xi + \eta) = k(\xi + \eta)$$

$$z = \frac{1}{\mu}(\xi + \eta)$$

□

La variable z posseeix la propietat de ser la més relacionada amb les dues taules X_1 i X_2 en el sentit que té màxima la suma de quadrats dels coeficients de correlació múltiple amb X_1 i X_2 (Figura 2.4).

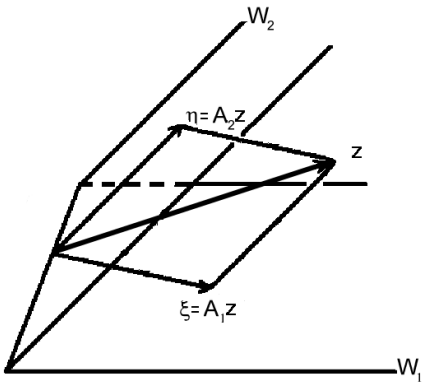


Figura 2.4:

Per veure-ho, com que el coeficient de correlació múltiple de z amb X_1 és:

$$R_1^2 = \frac{z^tDA_1z}{z^tDz} = \frac{||A_1z||^2}{||z||^2}$$

i és el cosinus de l'angle entre z i W_1, ja que les variables estan centrades com ja havíem vist.

Obtenim el mateix anàlogament per X_2 i amb la suma tenim el resultat mencionat.

2.5 Correspondències Simples

Aquest mètode fou proposat per J.P. Benzécri, amb el propòsit d'estudiar els lligams entre dues variables qualitatives, relacions anomenades *correspondències*.

En el pla matemàtic tant podem presentar l'anàlisi de correspondències simples **ACS** com una variant de l'anàlisi canònica, com a una anàlisi de components principals amb

una mètrica χ^2, aquesta darrera formulació molt més difosa, o també de moltes altres formes: reciprocal averaging, dual scaling, ... ([Gre84]). Donarem però una preferència a l'anàlisi canònica, ja que aquesta té l'avantatge de respectar la simetria existent en l'anàlisi i la de generalitzar l'anàlisi a diverses variables qualitatives, l'anàlisi de correspondències múltiples (**ACM**).

2.5.1 Taula de contingència i taules associades

Quan volem analitzar dues variables qualitatives, cadascuna d'elles ens pot ser donada de forma independent en una taula o totes dues conjuntament agrupats els seus valors en una única taula. Anem a veure aquestes taules i les relacions entre elles.

Podem tenir, com ja hem dit, primerament les dades presentades per dues **taules en forma disjuntiva completa**. Una variable categòrica qualsevol X pot ser representada mitjançant una taula on es creuen individus i categories i si l'individu i pren la categoria j llavors a la casella (i, j) hi trobarem un 1 i zeros a la resta de la filera. Obtindríem una taula de la forma següent:

$$X = \begin{array}{c} \\ 1 \\ 2 \\ \cdot \\ \cdot \\ \cdot \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & \ldots & m \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & & & \\ & & & & \\ & & & & \\ 0 & 1 & 0 & \ldots & 0 \end{array}\right] \end{array}$$

A dues variables qualitatives Q_1 i Q_2, els correspondran dues taules X_1 i X_2.

En correspondències simples ens trobem també amb **taules de contingència**, taules de doble entrada on cadascuna de les variables és una variables categòrica amb un cert nombre de modalitats.

Sigui la taula N producte del creuament de dues qüestions Q_1 amb I modalitats i Q_2 amb J modalitats on:

Taula original de dades: $N_{I*J} \equiv (n_{ij})$

- n_{ij} el nombre d'individus amb la modalitat i de Q_1 i j de Q_2
- $n_{i.} = \sum_j n_{ij}$ la marginal de les fileres, suma total dels elements de les fileres
- $n_{.j} = \sum_i n_{ij}$ la marginal de les columnes, suma total dels elements de les columnes

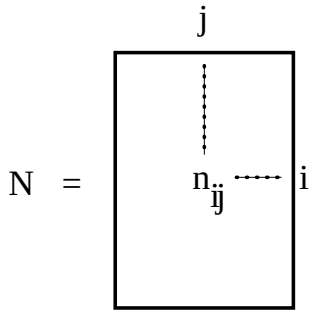


Figura 2.5: Taula de contingència

- $n = n_{..} = \sum_{i,j} n_{ij}$ el nombre total d'individus de la taula

Podem comprovar que les següents fórmules ens relacionen les taules disjuntives, la taula de contingència i les marginals.

$$N = X_1^T X_2$$
$$D_1 = X_1^T X_1$$
$$D_2 = X_2^T X_2$$

Podem definir una nova taula, **taula de freqüències relatives**, obtinguda a partir de l'anterior dividint l'efectiu de cada casella pel nombre d'individus totals.

Taula de freqüències relatives:

$$F_{I*J} = \frac{1}{n} N \qquad \text{on} \quad n = n.. = \sum_{ij} n_{ij} = 1^T.N.1$$

amb:

- $f_{ij} = \dfrac{n_{ij}}{n}$ la freqüència d'individus amb la modalitat i de Q_1 i j de Q_2
- $f_{i.} = \dfrac{n_{i.}}{n} = \sum_j f_{ij}$ la marginal de les fileres
- $f_{.j} = \dfrac{n_{.j}}{n} = \sum_i f_{ij}$ la marginal de les columnes
- $\sum_{ij} f_{ij} = 1$ el sumatori total de les freqüències de la taula

Com podem veure, en l'estudi de taules de contingència, hi haurà una dualitat entre l'estudi de la taula de dades per fileres o per columnes.

Sumes de fileres i columnes, marginals:

$$(f_{i.})_{i=1,\ldots,I} = \left(\sum_{j=1}^{J} f_{ij}\right)_{i=1,\ldots,I} \equiv F.1 \equiv f$$

$$(f_{.j})_{j=1,\ldots,J} = \left(\sum_{i=1}^{I} f_{ij}\right)_{j=1,\ldots,J} \equiv F^T.1 \equiv c$$

Matrius diagonals:

Notem per D_1 i D_2 les matrius diagonals dels efectius marginals de les dues variables d'efectius absoluts:

$$D_1 = \begin{bmatrix} n_{1.} & & 0 \\ & n_{2.} & \\ & & \ddots \\ 0 & & n_{I.} \end{bmatrix} \quad D_2 = \begin{bmatrix} n_{.1} & & 0 \\ & n_{.2} & \\ & & \ddots \\ 0 & & n_{.J} \end{bmatrix}$$

I anàlogament D_I i D_J les matrius diagonals dels marginals dels efectius relatius:

$$D_I \equiv diag(f_{i.}) \equiv diag(f) \qquad D_J \equiv diag(f_{.j}) \equiv diag(c)$$

Taules de perfils filera P_I i perfils columna P_J:

Definim els **perfils filera** com les freqüències de cada filera sobre el marginal de la filera i anàlogament definim els **perfils columna**:

Les taules de perfils filera es poden obtenir com:

$$P_I = D_I^{-1}F = D_1^{-1}N = (X_1^T X_1)^{-1} X_1^T X_2$$

i anàlogament les dels perfils columna com:

$$P_J = D_J^{-1}F^T = N^T D_2^{-1} = X_1^T X_2 (X_2^T X_2)^{-1}$$

Tenim per tant:

$$\left(\left\{\frac{f_{ij}}{f_{i.}}\right\}_{j=1,\ldots,J} = \left\{\frac{n_{ij}}{n_{i.}}\right\}_{j=1,\ldots,J}\right)_{i=1,\ldots,I} \equiv {D_1}^{-1}N \equiv D_I^{-1}F \equiv P_I$$

$$\left(\left\{\frac{f_{ij}}{f_{.j}}\right\}_{i=1,\ldots,I} = \left\{\frac{n_{ij}}{n_{.j}}\right\}_{i=1,\ldots,I}\right)_{j=1,\ldots,J} \equiv {D_2}^{-1}N^T \equiv D_J^{-1}F^T \equiv P_J$$

Podem observar que les taules de perfils filera, tal com han estat definits, tenen marginal filera constant igual a 1 (anàlogament amb els perfils columna i la marginal columna).

2.6 L'anàlisi canònica de dues variables categòriques

L'estudi de les relacions de dependència entre dues variables categòriques X_1 i X_2 en forma disjuntiva és l'estudi de les relacions entre les modalitats de cadascuna, per tant, podem aplicar l'anàlisi canònica presentada en la secció anterior.

Usant les equacions que ens donen els factors canònics (2.5) on tenim que:

$$\begin{aligned} V_{11} &= X_1^T D X_1 = \frac{1}{n} X_1^T X_1 = \frac{1}{n} D_1 = D_I \\ V_{22} &= X_2^T D X_2 = \frac{1}{n} X_2^T X_2 = \frac{1}{n} D_2 = D_J \\ V_{12} &= X_1^T D X_2 = \frac{1}{n} X_1^T X_2 = \frac{1}{n} N = F \\ V_{21} &= X_2^T D X_1 = V_{12}^T = F^T \end{aligned}$$

Llavors les equacions canòniques ens queden com:

$$\begin{aligned} V_{11}^{-1} V_{12} V_{22}^{-1} V_{21} \tilde{u}_i = D_I^{-1} F D_J^{-1} F^T \tilde{u}_i = \lambda_i \tilde{u}_i \\ V_{22}^{-1} V_{21} V_{11}^{-1} V_{12} \tilde{v}_i = D_J^{-1} F^T D_I^{-1} F \tilde{v}_i = \lambda_i \tilde{v}_i \end{aligned}$$

i les tindrem normalitzades per $\tilde{u}^T D_I \tilde{u} = 1$ i $\tilde{v}^T D_J \tilde{v} = 1$.

Podem també obtenir les relacions de transició:

$$\tilde{v}_i = \frac{1}{\sqrt{\lambda_i}} D_J^{-1} F^T \tilde{u}_i \qquad i \qquad \tilde{u}_i = \frac{1}{\sqrt{\lambda_i}} D_I^{-1} F \tilde{v}_i$$

La no simetria de les matrius $D_I^{-1} F D_J^{-1} F^T$ i $D_J^{-1} F^T D_I^{-1} F$ ens portarà mitjançant els canvis de variable: $u = D_I^{1/2} \tilde{u}$ i $v = D_J^{1/2} \tilde{v}$ a les expressions més fàcilment diagonalitzables

$$D_I^{-1/2} F D_J^{-1} F^T D_I^{-1/2} u_i = \lambda_i u_i \tag{2.8}$$

$$D_J^{-1/2} F^T D_I^{-1} F D_J^{-1/2} v_i = \lambda_i v_i \tag{2.9}$$

o el que és equivalent:

$$\begin{aligned} D_J^{-1/2} F^T D_I^{-1/2} = V \Lambda^{1/2} U^T \\ D_I^{-1/2} F D_J^{-1/2} = U \Lambda^{1/2} V^T \end{aligned}$$

normalitzades amb $u^T u = 1$ i $v^T v = 1$ i amb les següents relacions de transició:

$$v_i = \frac{1}{\sqrt{\lambda_i}} D_J^{-1/2} F^T D_I^{-1/2} u_i \qquad \text{i} \qquad u_i = \frac{1}{\sqrt{\lambda_i}} D_I^{-1/2} F D_J^{-1/2} v_i$$

Si definim la taula C com

$$C = D_I^{-1/2} F D_J^{-1/2} \tag{2.10}$$

taula de terme general $\frac{f_{ij}}{\sqrt{f_{i.}}\sqrt{f_{.j}}}$, que no és res més que la taula F ponderada per les marginals de les fileres i de les columnes, les expressions anteriors (2.8 i 2.9) es poden reescriure de la forma:

$$CC^T u_i = \lambda_i u_i \tag{2.11}$$

$$C^T C v_i = \lambda_i v_i \tag{2.12}$$

i el que és equivalent:

$$C^T = V \Lambda^{1/2} U^T$$

$$C = U \Lambda^{1/2} V^T$$

amb les relacions de transició $\Lambda^{1/2} V = C^T U$ i $\Lambda^{1/2} U = CV$.

Seguint l'esquema donat en l'anàlisi canònica podríem trobar les representacions per a les variables i per als individus, però ho farem després de la presentació de l'anàlisi d'una taula de contingència per perfils (secció 2.7).

Representació simultània òptima

Podrem obtenir, per tant, les dues representacions possibles en l'anàlisi canònica sobre cadascun dels conjunts de variables canòniques de X_1 i X_2. Ara bé, podríem trobar un eix z sobre el qual tinguéssim una representació optimal dels dos conjunts de variables?

Les projeccions de les categories de X_1 sobre z serien:

$$a = (X_1^T X_1)^{-1} X^T z = D_I^{-1} X_1^T z$$

i anàlogament per les categories de X_2:

$$b = n D_J^{-1} X_2^T z$$

Ens interessarà que l'eix z ens doni la variància més gran de les projeccions de X_1, és a dir $a^T D_I a$ màxim. Si ho volem per les dues variables qualitatives, voldrem que la mitjana de les dues sigui màxima:

$$\frac{1}{2}\left(a^T D_I a + b^T D_J b\right)$$

com $a^T D_I a = z^T X_1 D_I^{-1} D_I D_I^{-1} X_1^T z = n^2 z^T X_1 D_I^{-1} X_1^T z = n^2 z^T A_1 z$, on A_1 és el projector sobre el primer espai, obtenim que el màxim que haurem de cercar és el de:

$$\frac{1}{2}\left(z^T(A_1+A_2)z\right)$$

Si tenim la restricció $z^Tz=\lambda$, on λ és la inèrcia de l'eix sobre el qual volem la millor representació simultània òptima, llavors per maximitzar entrem la condició mitjançant un multiplicador de Lagrange i ens queda l'equació:

$$\frac{1}{2}\left(z^T(A_1+A_2)z\right)-k(z^Tz-\lambda)=\frac{1}{2}\left(z^T(A_1+A_2)z\right)-kz^Tz+k\lambda=0$$

Si derivem parcialment respecte z, obtenim:

$$(A_1+A_2)z-2kz=0$$

on veiem que no depèn del valor λ de la restricció, podent ser z^Tz igual a qualsevol valor constant en la restricció sense afectar això a la maximització. Per tant, la solució podem dir que es troba en els vectors propis de (A_1+A_2):

$$(A_1+A_2)z=2\mu z$$

o el que és equivalent:

$$X_1a+X_2b=2\mu z$$

Si premultipliquem aquesta expressió per $D_I^{-1}X_1^T$ i per $D_J^{-1}X_2^T$:

$$a+D_I^{-1}X_1^TX_2b=2\mu a \qquad \text{i} \qquad D_J^{-1}X_2^TX_1a+b=2\mu b$$

és a dir:

$$a+D_I^{-1}Nb=2\mu a \qquad \text{i} \qquad D_J^{-1}N^Ta+b=2\mu b$$

i per tant:

$$D_I^{-1}Nb=(2\mu-1)a \qquad \text{i} \qquad D_J^{-1}N^Ta=(2\mu-1)b$$

Obtenim les fórmules de transició i per substitució:

$$D_I^{-1}ND_J^{-1}N^Ta=(2\mu-1)^2a$$
$$D_J^{-1}N^TD_I^{-1}Nb=(2\mu-1)^2b$$

són les equacions de l'anàlisi canònica amb $\lambda=(2\mu-1)^2$ $a=\tilde{u}=\varphi_s$ i $b=\tilde{v}=\psi_s$.

Són la representació simultània òptima de fileres i columnes.

2.7 L'anàlisi d'una taula de contingència per perfils

2.7.1 Mètrica i ponderació

Anem a considerar els perfils filera i perfils columna com a taules de la nostra anàlisi. El nostre espai estarà dotat d'una mètrica i una ponderació. Anem a veure quina és la mètrica més usual, la que s'acostuma a prendre, i algunes de les seves propietats. També donarem a continuació la ponderació.

Definició de la mètrica

Havent escollit els perfils per construir la configuració de punts, la distància que s'adopta, usualment, és la distància xi-quadrat χ^2, que és una mètrica euclídea ponderada. Així, doncs, la distància entre dos perfils filera i i i' ve donada per:

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

I de manera simètrica, la distància entre dos perfils columna vindrà donada per:

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

Si ens fixem en aquesta distància, podem observar que només difereix de la distància euclídea en el fet que cada terme és ponderat per l'invers de la freqüència corresponent a aquest terme.

A més a més de la seva relació amb l'estadístic χ^2, una altra de les raons per les quals s'escull la distància χ^2 és que verifica la propietat de l'*equivalència distribucional*. I, a més, es correspon amb la mètrica euclídea de l'anàlisi canònica generalitzada amb variables categòriques.

Mètrica i estadístic χ^2

Usualment, per contrastar en una taula de contingència si les desviacions són significatives respecte el supòsit d'independència, és calcula l'estadístic χ^2:

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

on o són les freqüències observades i e són les freqüències esperades.

Si ho considerem des d'un punt de vista vectorial, podem escriure-ho com:

$$\chi^2 = (o-e)^T D_e^{-1}(o-e)$$

Si considerem els vectors de freqüències relatives:

$$x = \frac{o}{n} \qquad \text{i} \qquad y = \frac{e}{n}$$

llavors l'estadístic χ^2 es pot expressar com:

$$\chi^2 = n(x-y)^T D_y^{-1}(x-y) = n\sum \frac{1}{y}(x-y)^2 = n d^2_{y^{-1}}(x,y)$$

Com que aquesta distància euclídea ponderada és proporcional a l'estadístic χ^2, s'anomena distància χ^2, on la constant de proporcionalitat és el total d'individus n.

Equivalència distribucional

Una altra de les propietats de la mètrica escollida, és la de l'equivalència distribucional, que podríem enunciar de la següent manera:

Proposició 2.7.1.1 *(Equivalència distribucional)*

Si dos perfils fileres són idèntics, llavors les dues fileres corresponents de la matriu de dades originals poden ser substituïdes per una única filera, amb llur suma, sense que això afecti la geometria dels perfils columna

Demostració:

Suposem que agreguem els individus i_1 i i_2 que tenen perfils idèntics i els agreguem en un nou individu i_0, llavors $f_{i_0.} = f_{i_1.} + f_{i_2.}$. Si ens fixem en l'expressió $d^2(j,j')$ veurem que només dos termes contenen i_1 i i_2, aquests són:

$$\frac{1}{f_{i_1.}}\left(\frac{f_{i_1 j}}{f_{.j}} - \frac{f_{i_1 j'}}{f_{.j'}}\right)^2 + \frac{1}{f_{i_2.}}\left(\frac{f_{i_2 j}}{f_{.j}} - \frac{f_{i_2 j'}}{f_{.j'}}\right)^2 \tag{2.13}$$

Anem a veure que la suma d'aquests dos termes és el mateix que:

$$\frac{1}{f_{i_0.}}\left(\frac{f_{i_0 j}}{f_{.j}} - \frac{f_{i_0 j'}}{f_{.j'}}\right)^2 \tag{2.14}$$

Si prenem (2.14), ho podem reescriure com:

$$\frac{1}{f_{i_0.}}\left(\frac{f_{i_0 j}}{f_{.j}} - \frac{f_{i_0 j'}}{f_{.j'}}\right)^2 = f_{i_0.}\left(\frac{f_{i_0 j}}{f_{i_0.} f_{.j}} - \frac{f_{i_0 j'}}{f_{i_0.} f_{.j'}}\right)^2$$

i anàlogament podem expressar (2.13) com:

$$f_{i_1.}\left(\frac{f_{i_1j}}{f_{i_1.}f_{.j}}-\frac{f_{i_1j'}}{f_{i_1.}f_{.j'}}\right)^2+f_{i_2.}\left(\frac{f_{i_2j}}{f_{i_2.}f_{.j}}-\frac{f_{i_2j'}}{f_{i_2.}f_{.j'}}\right)^2 \tag{2.15}$$

com que els perfils filera són idèntics, tenim:

$$\frac{f_{i_1j}}{f_{i_1.}}=\frac{f_{i_2j}}{f_{i_2.}}=\frac{f_{i_0j}}{f_{i_0.}}=k$$

$$\frac{f_{i_1j'}}{f_{i_1.}}=\frac{f_{i_2j'}}{f_{i_2.}}=\frac{f_{i_0j'}}{f_{i_0.}}=k'$$

Llavors de (2.15):

$$f_{i_1.}\left(\frac{k}{f_{.j}}-\frac{k'}{f_{.j'}}\right)^2+f_{i_2.}\left(\frac{k}{f_{.j}}-\frac{k'}{f_{.j'}}\right)^2=(f_{i_1.}+f_{i_2.})\left(\frac{k}{f_{.j}}-\frac{k'}{f_{.j'}}\right)^2=$$

$$=f_{i_0.}\left(\frac{k}{f_{.j}}-\frac{k'}{f_{.j'}}\right)^2$$

que és la expressió (2.14) tal com volíem demostrar. □

Ponderació

En la secció anterior hem introduït la distància entre perfils on hem ponderat per l'invers de les freqüències, ara donarem a cada perfil filera o columna una ponderació o pes.

Aquesta ponderació serà donar a cada perfil un pes que sigui proporcional a la seva freqüència, per tal de no sobredimensionar, si consideréssim tots els pesos iguals, aquelles categories amb efectius molt petits. Així:

El perfil filera i tindrà un pes $f_{i.}$ i anàlogament el perfil columna j tindrà un pes $f_{.j}$

Centre de gravetat

Per a cadascun dels perfils filera i columna podem calcular el seu centre de gravetat, ponderats segons els pesos corresponents, ja definits. Així la coordenada j-èsima del centre de gravetat dels perfils filera serà:

$$\bar{x}=\frac{\sum_i \omega_i x_i}{\sum \omega_i}=\frac{\sum_i f_{i.}\frac{f_{ij}}{f_{i.}}}{\sum_i f_{i.}}=f_{.j}$$

I anàlogament la coordenada i-èsima del centre de gravetat dels perfils columna:

$$\bar{x}=\frac{\sum_j \omega_j x_j}{\sum \omega_j}=\frac{\sum_j f_{.j}\frac{f_{ij}}{f_{.j}}}{\sum_j f_{.j}}=f_{i.}$$

És a dir, el centre de gravetat dels perfils filera, té per coordenades les marginals de les columnes de la taula de freqüències relatives i les coordenades del centre de gravetat dels perfils columna són les de les marginals de les fileres de la taula de freqüències relatives.

Equivalència entre ponderació i mètrica

Tal com hem pogut veure els perfils filera tenen pesos $\{f_{i.}\}$ i una mètrica euclídea ponderada per $\{\frac{1}{f_{.j}}\}$, mentre que els perfils columna tenen pesos $\{f_{.j}\}$ i una mètrica euclídea ponderada per $\{\frac{1}{f_{i.}}\}$, per tant els pesos d'uns perfils són l'invers de la ponderació de la mètrica dels altres i viceversa. Això fa que alguns autors parlin només de pesos de les fileres i les columnes, sense dir explícitament quina és la ponderació de la mètrica, o simplement parlin de dues ponderacions.

A més veiem que el centre de gravetat dels perfils filera és exactament l'invers de la ponderació de la mètrica dels perfils filera i l'anàleg succeeix amb els perfils columna, propietat que farem servir més endavant.

2.7.2 Anàlisi

Com hem estat veient, l'estructura de la taula de contingència ens permet fer dues anàlisis, des de la perspectiva de les fileres i des de la perspectiva de les columnes.

Anem ara, un cop ja introduïdes les taules de dades, les mètriques i les ponderacions, a donar els elements per a cadascuna de les dues anàlisis i, a més, veurem que aquestes dues anàlisis estan plenament relacionades.

Dades originals de les anàlisis

Podem trobar esquematitzats en la Figura 2.6 els elements originals de la taula de freqüències que ens proporcionen els principals elements que intervenen en la definició dels elements de l'anàlisi.

Elements de les anàlisis

Els elements que utilitzarem en la nostra anàlisi són:

- Taules de perfils filera i perfils columna: P_I i P_J
- Masses del perfils filera i columna : $f_{i.} \equiv D_I$ i $f_{.j} \equiv D_J$

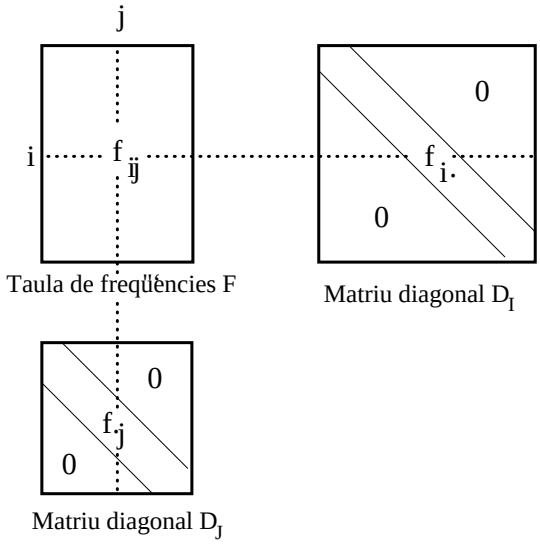


Figura 2.6: Elements originals de l'anàlisi

- Mètriques pels perfils fileres i columnes:

  Euclídea ponderada amb pesos D_J^{-1} Euclídea ponderada amb pesos D_I^{-1}

Donats els perfils, en la Taula 2.3 trobarem els elements de les anàlisis en $\mathbb{R}^J$ i en $\mathbb{R}^I$, els perfils com a taula de dades, les seves masses i la distància χ^2 com a mètrica, resumits en les tripletes de les anàlisis.

Taula 2.3: Elements de l'anàlisi: dades, pesos i mètriques

Núvol de fileres a $\mathbb{R}^J$	Núvol de columnes a $\mathbb{R}^I$
Els punts són els perfils filera $\frac{f_{ij}}{f_{i.}}$ $j = 1, \ldots, J$ La massa del punt i és $f_{i.}$ La distància entre dos punts és la euclídea amb ponderació $\frac{1}{f_{.j}}$	Els punts són els perfils columna $\frac{f_{ij}}{f_{.j}}$ $i = 1, \ldots, I$ La massa del punt j és $f_{.j}$ La distància entre dos punts és la euclídea amb ponderació $\frac{1}{f_{i.}}$
Elements (P_I, D_I, D_J^{-1})	Elements (P_J, D_J, D_I^{-1})

Per tant ara ja podem fer la DVSG dels perfils filera, anàlogament columna, amb les ponderacions i les mètriques corresponents.

DVSG dels perfils

Tal com hem vist a la Secció 2.3.2, la DVSG d'una matriu qualsevol A és

$$A_{I*J} = N_{I*K} D_{K*K} M^T_{K*J} = \sum_k \alpha_k n_k m_k^T \quad \text{on} \quad N^T \Omega N = M^T \Phi M = I$$

tenim, doncs, que la descomposició dels perfils filera serà:

$$P_I = N D_\eta M^T \qquad \text{on} \qquad N^T D_I N = M^T D_J^{-1} M = I$$

i la dels perfils columna:

$$P_J = U D_\rho V^T \qquad \text{on} \qquad U^T D_J U = V^T D_I^{-1} V = I$$

i a més, com estem en les hipòtesis de la Proposició 2.3.2.1, podem trobar equivalentment els valors i vectors propis de la DVSG de les respectives taules centrades:

$$P_I - 1c^T = N D_\eta M^T \qquad \text{on} \qquad N^T D_I N = M^T D_J^{-1} M = I$$

i la dels perfils columna:

$$P_J - 1f^T = U D_\rho V^T \qquad \text{on} \qquad U^T D_J U = V^T D_I^{-1} V = I$$

aquesta duplicitat quedarà solventada amb la següent proposició:

Proposició 2.7.2.1 *Podem trobar les DVSG dels perfils filera i dels perfils columna a partir de la DVSG de la taula de freqüències F amb masses D_I^{-1} i D_J^{-1}*

Demostració:
Sigui la DVSG de F

$$F = \tilde{A} D_\mu \tilde{B}^T \quad \text{on} \quad \tilde{A}^T D_I^{-1} \tilde{A} = \tilde{B}^T D_J^{-1} \tilde{B} = I \quad \text{amb} \quad \mu_1 \geq \ldots \mu_K \geq 0 \tag{2.16}$$

Si prenem la descomposició dels perfils filera

$$P_I = D_I^{-1} F = N D_\eta M^T \qquad \text{on} \qquad N^T D_I N = M^T D_J^{-1} M = I$$

i la premultipliquem per D_I

$$F = D_I N D_\eta M^T \quad \text{on} \quad (D_I N)^T D_I^{-1} D_I N = M^T D_J^{-1} M = I$$

tenim una primera equivalència. I anàlogament pels perfils columna centrats:

$$P_J = D_J^{-1} F^T = L D_\rho S^T \quad \text{on} \quad L^T D_J L = S^T D_I^{-1} S = I$$

premultiplicant-ho per D_J

$$F^T = D_J L D_\rho S^T \quad \text{on} \quad (D_J L)^T D_J^{-1} D_J L = S^T D_I^{-1} S = I$$

i, per tant, obtenim les relacions

$$D_\eta = D_\rho = D_\mu \qquad N = D_I^{-1}\tilde{A} \quad M = \tilde{B} \qquad L = D_J^{-1}\tilde{B} \quad S = \tilde{A}$$

Si a més en l'expressió (2.16) de la DVSG d'F, introduïm les mètriques en la matriu d'anàlisi, és a dir fem una DVS, obtenim:

$$D_I^{-1/2} F D_J^{-1/2} = A D_\mu B^T \quad \text{on} \quad A^T A = B^T B = I$$

amb $A = D_I^{-1/2}\tilde{A}$ i $B = D_J^{-1/2}\tilde{B}$

Expressió que és la descomposició de la taula (2.10), definida com a C, en l'anàlisi canònica, on

$$C = U\Lambda^{1/2} V^T \quad \text{amb} \quad U^T U = I \quad \text{i} \quad V^T V = I$$

□

Per tant, des d'un punt de vista factorial, les anàlisis dels perfils fileres i columnes i l'anàlisi de la taula F són equivalents a l'anàlisi canònica de dues variables categòriques. amb les relacions $A = U$ i $B = V$ i anàlogament $\tilde{A} = \tilde{U}$ i $\tilde{B} = \tilde{V}$ en les respectives DVSG i on $D_\lambda = D_\mu^2$.

2.7.3 Coordenades. Tipus de coordenades

Una vegada trobats els vectors propis que ens defineixen el subespai i als quals anomenarem **eixos factorials**, ja que la nostra taula de dades es pot descompondre com a una combinació d'aquest eixos, una factorització, anem a trobar les coordenades sobre aquests eixos projectant-hi les dades. Aquestes coordenades les tenim a la Taula 2.4.

Per les relacions d'ortogonalitat dels vectors propis donades en (2.16) tenim que aquestes dues expressions es poden simplificar en les dues donades a (2.17).

Taula 2.4: Coordenades principals

Coordenades dels perfils filera:	Coordenades dels perfils columna
$\psi = P_I D_J^{-1} \tilde{B}$	$\varphi = P_J D_I^{-1} \tilde{A}$

Coordenades fileres Coordenades columnes

$$\psi = D_I^{-1} \tilde{A} D_\mu \qquad \varphi = D_J^{-1} \tilde{B} D_\mu \tag{2.17}$$

Vegem aquesta igualtat per a les coordenades dels perfils fileres:

$$\psi = (P_I) D_J^{-1} \tilde{B} = (D_I^{-1} F) D_J^{-1} \tilde{B} = D_I^{-1} (F) D_J^{-1} \tilde{B}$$

Com:

$$F = \tilde{A} D_\mu \tilde{B}^T$$

Si postmultipliquem per $D_J^{-1} \tilde{B}$

$$F D_J^{-1} \tilde{B} = A D_\mu \tilde{B}^T D_J^{-1} \tilde{B} = \tilde{A} D_\mu$$

i, per tant,

$$\psi = D_I^{-1} (F) D_J^{-1} \tilde{B} = D_I^{-1} \tilde{A} D_\mu$$

Observació:

En les expressions que trobem a (2.17) hi tenim les coordenades per als perfils filera i columna sobre tots els eixos factorials, les coordenades d'una modalitat es troben a les files de ψ i φ respectivament. Si volem les coordenades en un espai optimal K^*-dimensional, haurem de prendre les K^* primeres columnes de ψ i φ respectivament.

Relació de les coordenades en $\mathbb{R}^I$ i $\mathbb{R}^J$

Les coordenades en els dos espais de perfils estan plenament relacionades mijançant les formules que es coneixen amb el nom de **Fórmules de Transició**.

Proposició 2.7.3.1 *Els conjunts de coordenades estan relacionats els uns amb els altres per les fórmules*

$$\varphi = D_J^{-1} F^T \psi D_\mu^{-1} \qquad i \qquad \psi = D_I^{-1} F \varphi D_\mu^{-1}$$

Demostració:

Com a conseqüència directa de (2.16) i de (2.17) tenim que:

$$\psi = D_I^{-1} A D_\mu \Longleftrightarrow D_J^{-1} F^T \psi = D_J^{-1} F^T D_I^{-1} A D_\mu = \varphi D_\mu$$

$$\varphi D_\mu = D_J^{-1} F^T \psi \Longleftrightarrow \varphi = D_J^{-1} F^T \psi D_\mu^{-1}$$

i anàlogament l'altre conjunt de coordenades. □

2.7.4 Propietats

Anem a veure algunes de les propietats que tenen les coordenades de les anàlisis, entre elles l'anomenada fórmula de reconstitució.

Proposició 2.7.4.1 *Les coordenades dels perfils fileres i columnes centrades, sobre els eixos factorials respectius, tenen el seu centre de gravetat en l'origen i la inèrcia en un eix és igual al valor propi al quadrat de l'eix.*

Demostració:

Anem a veure quin és el centre de gravetat de les coordenades dels perfils filera, aquest serà el sumatori del producte entre les masses de les fileres i les coordenades: $f^T\psi =$

$$= f^T(D_I^{-1}F - 1c^T)D_J^{-1}B = (f^T D_I^{-1} F - f^T 1 c^T) D_J^{-1} B = (1^T F - c^T) D_J^{-1} B =$$

$$= (c^T - c^T) D_J^{-1} B = 0$$

I anàlogament pels perfils columna.

Com que tenim el centre de gravetat en el zero, llavors la inèrcia és igual a la suma ponderada de quadrats de les coordenades:

$$\psi^T D_I \psi = \left(D_I^{-1} A D_\mu\right)^T D_I \left(D_I^{-1} A D_\mu\right) = D_\mu^T B^T D_J^{-1} B D_\mu = D_\mu^2$$

tal com volíem demostrar.

I això també és igual per a les coordenades de les columnes:

$$\varphi^T D_J \varphi = \left(D_J^{-1} B D_\mu\right)^T D_J \left(D_J^{-1} B D_\mu\right) = D_\mu^2$$

per tant les inèrcies dels núvol filera i columna són iguals. □

Definició 2.7.4.1 Anomenarem *k-èsima inèrcia principal* i la notarem per λ_k al valor de la inèrcia de k-èsim eix factorial. És a dir:$\lambda_k = \mu_k^2$.

Proposició 2.7.4.2 *La DVSG de la taula de freqüències centrada $F - fc^T$, amb la qual hem vist l'equivalència entre les DVSG dels perfils filera i perfils columna, es pot trobar de la DVSG de la taula de freqüències sense centrar F*

Demostració: Per a això, només hem de veure que si $F = AD_\mu B^T$, f és ortogonal a A i normalitzat segons la mètrica D_I^{-1} i c és ortogonal a B i normalitzat segons la mètrica D_J^{-1}.

Hem de veure, doncs,

$$f^T D_I^{-1} A = 0 \qquad f^T D_I^{-1} f = 1$$

i

$$c^T D_J^{-1} B = 0 \qquad c^T D_J^{-1} c = 1$$

Per l'equació (2.17) i la proposició (2.7.4.1)

$$f^T D_I^{-1} A = f^T \psi D_\mu^{-1} = 0 D_\mu^{-1} = 0$$

I encara més senzill és demostrar la normalitat:

$$f^T D_I^{-1} f = 1 f = 1$$

I com a conseqüència d'aquesta normalitat, el valor propi que li pertoca és 1, i, a més, és el més gran perquè conté el centre de gravetat. □

2.7.5 Coordenades estandarditzades

Per tal de normalitzar les coordenades, ja que hem vist que no estan estandarditzades, sinó que $\psi^T D_I \psi = \lambda = \mu^2$, es procedeix a l'estandardització de les coordenades principals en relació a la inèrcia del seu eix, les anomenades **coordenades estandarditzades**.

Taula 2.5: Coordenades estandarditzades

Coordenades estandarditzades fileres	Coordenades estandarditzades columnes
$\psi_s = \psi D_\mu^{-1}$	$\varphi_s = \varphi D_\mu^{-1}$

Algunes vegades, es sacrifica la simetria de la representació per tal de tenir guanys d'interpretació.

Definició 2.7.5.1 Anomenarem gràfic *asimètric* aquell on l'estandardització s'ha realitzat només en un dels dos conjunts de punts. És a dir, apliquem estandarditzacions diferents als dos conjunts de punts.

En el sistema de coordenades dels perfils i usant les relacions de transició i els vectors propis de la diagonalització podem obtenir:

$$\psi = P_I D_J^{-1} \tilde{B} = D_I^{-1} F \varphi D_\mu^{-1}$$
$$\varphi = P_J D_I^{-1} \tilde{A} = D_J^{-1} F^T \psi D_\mu^{-1}$$

de les dues darreres igualtats de cada expressió podem obtenir:

$$D_J^{-1}\tilde{B} = \varphi D_\mu^{-1} \qquad \tilde{B} = D_J \varphi_s$$
$$D_I^{-1}\tilde{A} = \psi D_\mu^{-1} \qquad \tilde{A} D_\mu = D_I \psi$$

com $A = D_I^{-1/2}\tilde{A}$, $B = D_J^{-1/2}\tilde{B}$, $\psi_s = \psi D_\mu^{-1}$ i $\varphi_s = \varphi D_\mu^{-1}$, substituint a $F = \tilde{A} D_\mu \tilde{B}^T$ obtenim:

$$F = \tilde{A} D_\mu \tilde{B}^T = D_I \psi \varphi_s^T D_J$$

per tant, el producte de les coordenades ψ φ_s ens donen la taula F ponderada.

$$D_J^{-1} F^T D_I^{-1} = \varphi \psi_s^T \tag{2.18}$$
$$D_I^{-1} F D_J^{-1} = \psi \varphi_s^T \tag{2.19}$$

Obtenim desenvolupant les fórmules de transició entre els vectors propis $\tilde{u}$ i $\tilde{v}$, les relacions anomenades fórmules baricèntriques, següents:

$$\psi = D_J^{-1} F^T \varphi_s \qquad \text{i} \qquad \varphi = D_I^{-1} F \psi_s$$

2.7.6 Biplots en AC

D'acord amb la definició de Gabriel [Gab71, Gab02]: un biplot és un gràfic pla d'aproximació dels elements d'una matriu Y, així com de la configuració de les fileres i de les columnes d'aquesta matriu. Un biplot consisteix en dos elements vectors a^T i b^T, els quals serveixen d'indicadors o marcadors per a les fileres i les columnes, respectivament, i el producte dels quals aproxima els elements d'Y.

$$y_{ij} \sim a_i^T b_j$$

Tal com acabem de veure (Fórmula 2.19) la taula $D_I^{-1}FD_J^{-1}$ pot ser descomposada com el producte de ψ i φ_s, és a dir, tenim un biplot. Més exactament en tenim dos de biplots

$$D_J^{-1}F^TD_I^{-1} = \varphi\psi_s^T$$
$$D_I^{-1}FD_J^{-1} = \psi\varphi_s^T$$

Un dels debats més fervorosos ha estat què era millor representar, el biplot o la representació simultània?

Tal com hem vist, la representació simultània té una sèrie de propietats, però, a més, tal com hem vist en l'anàlisi canònica (2.6) és la que ens dóna la millor representació de les fileres i les columnes simultàniament. Ara bé, tal com diu K.R Gabriel [Gab02], *el gràfic de Benzécri -representació simultània-, ha estat construït per fer òptimes les estimacions de la forma i de la variància, però no ajusta les dades òptimament.* Nogensmenys, preserva més del 95% de la bondat d'ajust de les dades, siguin quins siguin els valors propis i, per tant, hi ha una forta justificació, per la pràctica comuna de considerar el producte de les components com a una aproximació proporcional dels elements de la matriu d'estudi.

2.7.7 Fórmula de reconstitució

Donada l'escriptura de la taula de freqüències com

$$F = fc^T + AD_\mu B^T$$

si substituïm en ella les relacions de (2.17) llavors

$$F = fc^T + (D_I\psi D_\mu^{-1})D_\mu(D_\mu^{-1}\varphi D_J) = fc^T + D_I\psi D_\mu^{-1}\varphi D_J \simeq$$
$$\simeq f_{i.}f_{.j} + \sum_K f_{i.}f_{.j}\frac{1}{\mu}\psi\varphi = f_{i.}f_{.j}(1 + \sum_K \frac{1}{\mu}\psi\varphi)$$

puc reconstituir la taula de freqüències F a partir de les coordenades factorials. I el que és equivalent, si només prenem les coordenades en els primer K^* eixos factorials, tindrem una aproximació de rang K^* i serà la més bona en el sentit de mínims quadrats ponderats.

2.7.8 Qualitat de l'aproximació i eixos a retenir

Per una dimensió donada K^* $(1 \leq K^* \leq K = rang(F) - 1)$, la qualitat global de les representacions gràfiques de dimensió K^* es mesura per la relació entre la suma del K^* primers valors propis de l'AFC i llur suma completa d'1 a K.

$$Q = \frac{\sum_{j=1}^{K^*} \lambda_j}{\sum_{i=1}^{K} \lambda_i}$$

La qualitat de l'aproximació ens valora la inèrcia de la projecció en relació a la inèrcia total.

Si el que volem és saber quins eixos són els importants a retenir, hi ha diversos criteris: el primer d'ells és la qualitat de la informació, però òbviament com més eixos més qualitat i, a més a la pràctica aquest criteri és modificat per la simplicitat, com que fem representacions bàsicament bidimensionals, agafem els dos primers i mirem quina qualitat tenim. Un segon criteri podria ser basat en el nombre de valors propis significatius, és a dir, si tenim D eixos significatius, escollir tots els valors propis més grans que $1/D$. Finalment el darrer d'ells és mirar de tots els eixos, quan la quantitat d'informació aportada deixa de ser significativa. Per això, el que haurem de fer és un test per veure si els valors propis són o no zero. Podem aplicar un test per a anàlisis canòniques ([Sap90], p. 192) per veure si λ_1, λ_2 ... λ_k són significativament diferents de zero, utilitzant l'estadístic:

$$-\left[n-1-k-\frac{1}{2}(p+q+1)+\sum_{i=1}^{k}\frac{1}{\lambda_i}\right]\ln\left(\prod_{k+1}^{min(p,q)}(1-\lambda_i)\right)$$

que segueix aproximadament una llei $\chi^2_{(p-k)(q-k)}$ si els valors teòrics de λ_{k+1}, ... són zero.

2.7.9 Eines d'ajuda a la interpretació: contribucions i cosinus quadrats

Hem vist en la proposició 2.7.4.1 que la inèrcia del núvol d'individus és

$$\psi^T D_I \psi = D_\mu^2 = D_\lambda$$

això ens permet fer la descomposició en l'eix k per $k = 1, \ldots, K$:

$$\lambda_k = \sum_{i=1}^{I} \psi_{ik}^2 f_i$$

i anàlogament per a les columnes:

$$\varphi^T D_J \varphi = D_\mu^2 = D_\lambda$$

que descompon en:

$$\lambda_k = \sum_{j=1}^{J} \varphi_{jk}^2 c_j$$

Això ens permetrà saber quina és l'aportació de cada filera o columna a la inèrcia de l'eix.

Definició 2.7.9.1 La *contribució absoluta d'un punt i en un eix k* és la proporció amb la qual un punt contribueix a l'inèrcia de l'eix.

Com hem vist que en un eix k la inèrcia era $\lambda_k = \sum_{i=1}^{I} \psi_{ik}^2 f_i$, on $\psi_{ik}^2 f_i$ és la quantitat aportada per cada individu, llavors la contribució absoluta serà:

$$cnt(i,k) = \frac{\psi_{ik}^2 f_i}{\lambda_k}$$

Aquells individus amb contribucions absolutes grans seran els que ens determinaran la posició de l'eix i, per tant, ens ajuden a interpretar-lo.

Tenim per definició que $\sum_i cnt(i,k) = 1$. Si expressem les contribucions en percentatge, llavors aquesta suma és igual a 100.

Definició 2.7.9.2 Els *cosinus quadrats d'un punt i en un eix k* o la *contribució relativa d'un eix k sobre l'individu i* és la qualitat de representació d'un punt en un eix.

Com que la distància al quadrat d'un individu al centre de gravetat del núvol de punts la podem descompondre com a la suma de totes les projeccions al quadrat sobre els eixos $d^2(i,G) = \sum_k \psi_{ik}^2$, llavors al contribució relativa d'un eix a la representació del punt serà:

$$cos^2(i,k) = \frac{\psi_{ik}^2}{\sum_k \psi_{ik}^2}$$

s'anomena *cosinus quadrat* perquè aquesta quantitat obtinguda és exactament el cosinus quadrat de l'angle que hi ha entre la semirecta entre de l'origen a l'individu i i la semirecta de l'origen a la projecció sobre un eix k de l'individu i.

Si el cos^2 és proper a zero l'individu està mal representat en aquell eix, en canvi si els cos^2 és proper a 1, l'individu està ben representat en aquell eix.

Tenim per definició que $\sum_k cos^2(i,k) = 1$. Si expressem aquestes contribucions en percentatge, aleshores la suma és igual a 100.

Podem definir anàlogament les contribucions absolutes des del punt de vista de les columnes canviant les coordenades i els pesos dels individus per les coordenades i pesos de les columnes.

$$cnt(j,k) = \frac{\varphi_{jk}^2 c_j}{\lambda_k} \qquad cos^2(j,k) = \frac{\varphi_{jk}^2}{\sum_k \varphi_{jk}^2}$$

2.7.10 Elements suplementaris

Qualsevol element pot ser projectat en suplementari fent la projecció sobre els vectors propis de l'anàlisi. Això tant ho podem fer per individus filera com per individus columna, fent els corresponents centrats, si escau.

2.7.11 Exemple

Anem a presentar un exemple d'aplicació de les anàlisis de correspondències simples a la composició segons les diferents categories funcionarials del personal dels departaments d'una universitat. Podem trobar a la Taula 2.6 les dades d'aquest exemple.

Taula 2.6: Distribució per categories del personal dels departaments

nomdept	cu	tu-ceu	teu	ajud-atc	assoc	becaris
AEC	0	1	14	7	22	0
BIO	3	7	0	11	6	4
CAMB	1	10	0	11	7	11
DIDACT	2	2	18	2	9	0
DRETPRIV	5	3	2	11	6	2
DRETPUB	6	6	0	8	11	0
ECON	1	7	7	6	4	1
EIA	1	4	10	13	13	4
EMPR	1	5	15	5	9	0
EI	3	3	11	15	24	3
EQATA	3	8	12	13	15	5
FILOFILO	5	28	4	10	6	1
GEOHISAR	9	27	1	11	8	0
INFER	0	0	9	0	11	0
IMA	2	5	17	5	11	9
PEDA	2	10	7	3	9	1
PSIC	2	11	3	7	8	3
QUIM	2	11	0	13	3	4

En un paquet estadístic estàndard, la primera sortida que obtindrem és la de la descomposició de la inèrcia total $I_{Total} = 0.468872$ en cadascun dels eixos segons els valors propis associats, obtenint majoritàriament la inèrcia, la inèrcia acumulada i el percentatge acumulat d'inèrcia en els eixos referits. En aquest exemple podem veure que tenim 5 valors propis, és a dir un menys que el nombre de columnes de la taula, que és la dimensió més petita de les dues de la nostra taula. Podem observar que amb els dos primers

valors propis tenim una inèrcia acumulada del 78%, per tant, una representació utilitzant el primer pla factorial ens dóna una informació d'una qualitat equivalent al 78% inicial. Analitzant detalladament la taula de valors propis, podem veure que n'hi ha tres de grans i dos de petits, aquests darreres aporten només un 4% cadascun aproximadament.

Taula 2.7: Valors propis, percentatges d'inèrcia i percentatges acumulats

VALOR PROPI	INÈRCIA	INER.ACUMULAT
0.263743	56.2505	56.250
0.104865	22.3654	78.616
0.062880	13.4108	92.027
0.021616	4.6101	96.637
0.015769	3.3632	100.000

En segon lloc, una de les sortides més habituals dels paquets estadístics sol ser presentar les coordenades de les columnes i les fileres així com les eines d'ajuda a la interpretació que són les contribucions i els cosinus quadrats. Això ens permetrà visualitzar els gràfics de les coordenades i estudiar la qualitat de la representació de cadascun dels punts.

Taula 2.8: Coordenades, contribucions i cosinus quadrats de les columnes

categor	cor.1	cor.2	cor.3	cor.4	cor.5	cnt.1	cnt.2	cnt.3	cnt.4	cnt.5	cos.1	cos.2	cos.3	cos.4	cos.5
cu	-0,48	0,28	0,32	0,30	0,32	6,0	4,9	10,8	28,6	42,8	0,39	0,13	0,17	0,15	0,17
tu-ceu	-0,63	0,32	-0,23	-0,10	-0,02	31,5	20,0	18,0	8,9	0,8	0,71	0,18	0,10	0,02	0,00
teu	0,80	0,17	-0,25	0,13	-0,02	44,3	4,8	18,3	13,6	0,7	0,86	0,04	0,08	0,02	0,00
aj-atc	-0,25	-0,22	0,19	0,11	-0,17	5,3	10,0	12,9	11,7	38,8	0,34	0,25	0,20	0,06	0,15
assoc	0,34	-0,01	0,23	-0,18	0,05	11,6	0,0	21,8	37,1	3,8	0,58	0,00	0,26	0,15	0,01
becaris	-0,23	-0,96	-0,41	-0,02	0,17	1,4	60,3	18,2	0,2	13,1	0,05	0,78	0,14	0,00	0,03

Com en totes les aplicacions de la metodologia de l'anàlisi de correspondències, l'expert en la matèria hauria de fer l'anàlisi dels resultats, no només en base als gràfics sinó tenint presents les eines d'ajuda a la interpretació, ja que per exemple, el primer eix ve determinat des del punt de vista de les categories per *teu* i *tu-ceu* amb un 44.3 i un 31.5, amb un total del 75.8%, i a més amb coordenades oposades *tu-ceu* negativa i *teu* positiva, per tant definint aquest eix una oposició entre aquestes categories. Aquesta mateixa anàlisi del primer eix la podem fer des del punt de vista dels departaments on veiem que les grans contribucions són dels departaments de GEOHISAR i FILOFILO per un costat, amb coordenades negatives, i DIDACT, INFER i AEC per l'altre, coordenades positives. Per una banda podem analitzar aquesta oposició, de departaments clàssics de lletres per un costat (geografia, història, art, filologia, filosofia) oposats als altres (didàctica,

Taula 2.9: Coordenades, contribucions i cosinus quadrats de les fileres

nomdept	cor.1	cor.2	cor.3	cor.4	cor.5	cnt.1	cnt.2	cnt.3	cnt.4	cnt.5	cos.1	cos.2	cos.3	cos.4	cos.5
AEC	0,72	0,06	0,25	-0,22	-0,09	12,3	0,2	6,0	14,4	3,1	0,81	0,01	0,09	0,08	0,01
BIO	-0,47	-0,33	0,16	0,06	-0,03	3,7	4,6	1,7	0,8	0,2	0,62	0,30	0,07	0,01	0,00
CAMB	-0,48	-0,75	-0,28	-0,16	0,09	4,8	30,1	6,9	6,7	3,1	0,25	0,62	0,09	0,03	0,01
Didact	0,87	0,34	-0,23	0,27	0,06	13,3	5,1	3,8	15,9	1,0	0,76	0,12	0,05	0,07	0,00
DretPriv	-0,26	-0,19	0,42	0,37	0,06	1,1	1,4	11,8	25,7	1,1	0,16	0,08	0,42	0,32	0,01
DretPub	-0,31	0,17	0,59	0,04	0,24	1,6	1,2	24,4	0,3	15,6	0,18	0,05	0,66	0,00	0,11
ECON	0,02	0,16	-0,21	0,12	-0,21	0,0	0,9	2,6	2,3	9,9	0,00	0,20	0,36	0,11	0,34
EIA	0,22	-0,25	0,07	0,03	-0,16	1,2	3,9	0,5	0,3	10,2	0,35	0,44	0,03	0,01	0,17
EMPR	0,56	0,28	-0,18	0,13	-0,13	6,0	3,6	2,4	3,9	5,5	0,69	0,17	0,07	0,04	0,04
EI	0,30	-0,15	0,32	-0,08	-0,03	2,9	1,8	13,6	2,2	0,6	0,41	0,10	0,46	0,03	0,01
EQATA	0,13	-0,14	0,00	0,04	-0,02	0,5	1,4	0,0	0,5	0,2	0,45	0,50	0,00	0,04	0,01
FiloFilo	-0,63	0,44	-0,22	-0,08	-0,06	11,6	14,0	6,0	2,4	1,9	0,61	0,29	0,08	0,01	0,01
GeoHisAr	-0,72	0,48	0,02	0,01	0,10	15,4	17,3	0,1	0,0	4,7	0,68	0,30	0,00	0,00	0,01
INFER	1,07	0,21	0,06	-0,27	0,12	12,2	1,2	0,1	9,8	2,8	0,89	0,04	0,00	0,06	0,01
IMA	0,39	-0,31	-0,40	0,09	0,22	4,0	6,4	18,0	2,8	21,3	0,32	0,21	0,35	0,02	0,10
PEDA	0,02	0,30	-0,15	-0,16	0,08	0,0	4,0	1,6	5,4	1,8	0,00	0,63	0,15	0,17	0,04
PSIC	-0,30	0,00	-0,08	-0,16	0,00	1,6	0,0	0,5	5,5	0,0	0,74	0,00	0,05	0,20	0,00
QUIM	-0,66	-0,26	-0,04	0,07	-0,24	7,6	2,9	0,1	1,1	17,0	0,77	0,12	0,00	0,01	0,10

infermeria, arquitectura i enginyeria de la construcció) d'estudis de primer cicle. A més podem relacionar fileres i columnes, ja que en estudis de primer cicle el professorat sol ser diplomat o tècnic i per tant de categoria *teu* i en canvi en els estudis de lletres sol ser llicenciat i per tant *tu*.

Pel que fa a la representació, els departaments de PEDA o ECON, o els becaris(*bfdr*), estan molt mal representats en el primer eix i pel que fa a les categories i departaments que contribuïen en el primer eix, tots ells estan ben representats.

Per a completar l'exemple presentem els gràfics factorials d'aquestes dades:

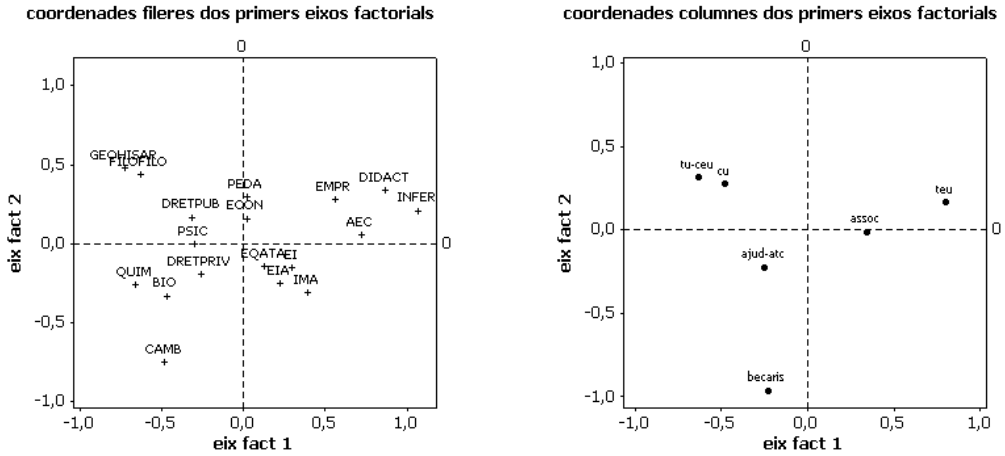


Figura 2.7: Gràfics factorials de les fileres i de les columnes

Aquests poden ser superposats en un gràfic comú per tal de veure els lligams entre modalitats de les fileres i de les columnes. Podem troba a la Figura 2.8 el gràfic conjunt.

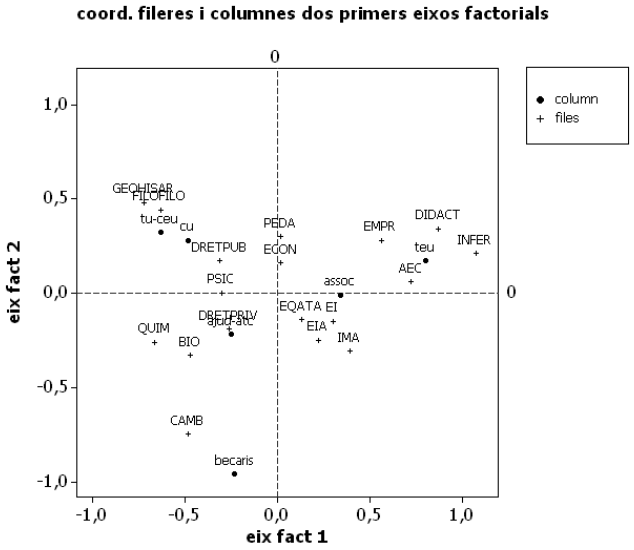


Figura 2.8: Gràfic factorial conjunt de les fileres i de les columnes

Per a obtenir una anàlisi més detallada caldria repetir el procés d'anàlisi per a cadascun dels eixos que hàgim escollit significatius, amb la informació que en pugui treure'n un especialista en el tema.

2.8 L'anàlisi canònica generalitzada

Hem vist a la secció (2.4) l'anàlisi canònica de dues variables, anem ara a estendre-ho a més de dues variables i posteriorment en veurem la seva aplicació per a la generalització per a més de dues variables de l'anàlisi de correspondències múltiples.

Per a fer-ho, necessitarem establir alguna manera de mesurar les correlacions múltiples entre els elements dels Q espais W_i que ens generaran les Q variables objectes d'estudi.

La generalització de l'anàlisi canònica (deguda a J.D. Carroll [Car68]), en comptes de buscar directament les variables canòniques, tal com havíem fet en el cas de dues variables, en cadascun dels subespais W_i associats a les taules X_i, busca una variable auxiliar z pertanyent a la suma dels W_i tal que la seva correlació amb les variables sigui màxima. Aquesta condició s'expressa com:

$$\sum_{i=1}^{Q} R^2(z, X_i) \quad \text{sigui màxim}$$

que no és res més que l'extensió d'una propietat que té l'anàlisi canònica.

Per la propietat que hem vist de l'anàlisi canònica (2.4.4.1), z és llavors valor propi de

$A_1 + \ldots + A_Q$.

$$(A_1 + \ldots + A_Q)z = \mu z$$

I podem obtenir les variables canòniques ξ_i projectant z sobre cadascun del W_i:

$$\xi_i = A_i z$$

Com $X = (X_1|X_2|\ldots|X_Q)$ és una taula formada per Q subtaules, d'n fileres i un total de $J = \sum_i J_i$ columnes, podem escriure z de la forma, $z = Xb$ i podem treballar en components de b a l'hora de buscar els vectors propis.

Com que $A_i = X_i(X_i^T D X_i)^{-1} Z_i^T D$, si anomenem $V_{ii} = X_i^T D X_i$ matriu de variància covariància del grup i-èsim i M, matriu bloc diagonal dels inversos dels V_{ii}:

$$M = \begin{pmatrix} V_{11}^{-1} & & & 0 \\ & V_{22}^{-1} & & \\ & & \ddots & \\ 0 & & & V_{pp}^{-1} \end{pmatrix}$$

llavors podem reescriure

$$A_i = X_i(V_{ii})^{-1} Z_i^T D$$

i

$$\sum_{i=1}^{Q} A_i = \sum_{i=1}^{Q} X_i(V_{ii})^{-1} Z_i^T D$$

que al seu torn es pot escriure amb la notació introduïda com:

$$\sum_{i=1}^{Q} A_i = X M X^T D$$

la qual cosa ens porta a que z és valor propi de $XMX^T D$ i com $z = Xb$ llavors:

$$XMX^T Dz = \mu z \Longleftrightarrow XMX^T DXb = \mu Xb \Longleftrightarrow MX^T DXb = \mu b \qquad (2.20)$$

2.9 Correspondències Múltiples

L'anàlisi de correspondències múltiples (**ACM**) és una tècnica de descripció de dades qualitatives: considerarem una població I formada per n individus que prenen valors en Q variables $X_1, X_2, \ldots, X_Q$ amb $J_1, J_2, \ldots, J_Q$ modalitats cadascuna i $J = \sum J_i$ modalitats totals.

És un mètode que és molt utilitzat a l'anàlisi exploratòria d'enquestes amb respostes múltiples per la seva bona adaptació.

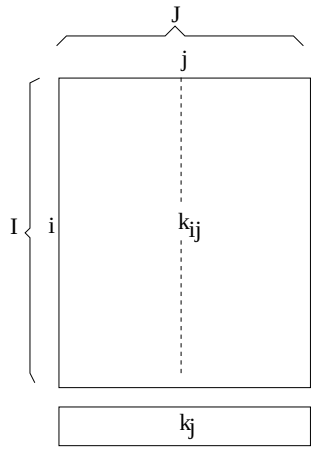
Veurem que en l'anàlisi d'aquestes dades es pot veure des d'una perspectiva de generalització de l'anàlisi canònica aplicada a més de dues variables categòriques o com a una aplicació simple de l'anàlisi de correspondències a la taula del creuament de totes les variables categòriques (**taula de Burt**).

2.9.1 Introducció. Notacions. Taula de Burt

Notacions

Denotem per I la població que està composada per n individus; J com el conjunt de totes les modalitats de les Q variables qualitatives.

Les dades poden ser codificades en una **taula disjuntiva completa**: creuant I i J



$k_{ij} = 1$ o 0

$Q = \sum_{j \in J} k_{ij} \quad \forall i$

n= card I

$k_j = \sum_{i \in I} k_{ij}$

En aquesta taula, individus per variables, la suma dels elements de cada filera de Z és constant i igual al nombre de variables Q i la suma de cada columna ens dóna l'efectiu marginal de cada modalitat de cada variable k_j. A més, tal com ha estat definida, la suma dels valors que pren cada individu restringit a cadascuna de les variables és constant i igual a 1. Finalment, la suma total de la taula és nQ.

Si prenem com a exemple les dades de nou individus ($I = 9$) sobre tres variables categòriques, amb tres, dues i tres modalitats cadascuna respectivament ($J = 8$), ens podria dona una taula Z d'I fileres i J columnes, juxtaposició de les taules Z_1, Z_2 i Z_3, següent:

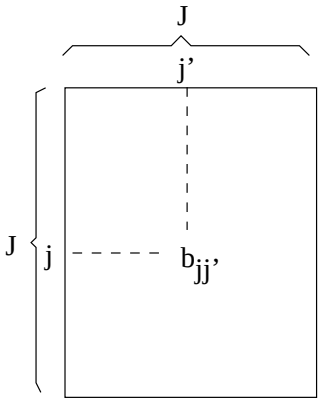
$$Z = (Z_1|Z_2|Z_3) = \left(\begin{array}{ccc|cc|ccc} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{array}\right)$$

Aquesta mateixa informació també se'ns pot donar en forma condensada on una sola columna per cada variable ens codificarà el fet de tenir la modalitat primera, segona, etc. de la variable en qüestió, en una taula també d'individus per variables d'I fileres i Q columnes.

Així, doncs, la taula de l'exemple anterior la podem trobar com una **taula en forma condensada** de la manera següent:

$$Y = (Y_1|Y_2|Y_3) = \left(\begin{array}{ccc} 2 & 1 & 1 \\ 1 & 1 & 3 \\ 2 & 2 & 2 \\ 2 & 2 & 3 \\ 1 & 1 & 1 \\ 3 & 1 & 3 \\ 2 & 2 & 1 \\ 1 & 2 & 2 \\ 3 & 1 & 2 \end{array}\right)$$

Finalment, també podem trobar la informació de la taula $I * J$ reduïda en la **taula de Burt** associada. Taula simètrica, de variables per variables, de dimensions $J * J$, on els individus queden confosos, de terme general $b_{jj'}$, obtinguda com $B = Z^T.Z$.



$$b_{jj'} = \sum_{i \in I} k_{ij} k_{ij'}$$

En el nostre exemple, tenim la taula de Burt amb 8 fileres i 8 columnes, següent:

$$B = \left(\begin{array}{cccc|cc|ccc} 3 & 0 & 0 & 2 & 1 & 1 & 1 & 1 \\ 0 & 4 & 0 & 1 & 3 & 2 & 1 & 1 \\ 0 & 0 & 2 & 2 & 0 & 0 & 1 & 1 \\ \hline 2 & 1 & 2 & 5 & 0 & 2 & 1 & 2 \\ 1 & 3 & 0 & 0 & 4 & 1 & 2 & 1 \\ \hline 1 & 2 & 0 & 2 & 1 & 3 & 0 & 0 \\ 1 & 1 & 1 & 1 & 2 & 0 & 3 & 0 \\ 1 & 1 & 1 & 2 & 1 & 0 & 0 & 3 \end{array}\right)$$

Com podem veure és una matriu simètrica amb estructura de Q^2 blocs, formada per submatrius i on les submatrius de la diagonal són, al seu torn, matrius diagonals $D_q = Z_q^T Z_q$, ja que són els creuaments de les diferents modalitats d'una variable amb elles mateixes. Això ens dóna que la diagonal de la matriu de Burt, conté els efectius marginals de cada modalitat k_j. Aquesta diagonal l'escriurem com a matriu D en la forma:

$$D = \begin{pmatrix} D_1 & & & 0 \\ & D_2 & & \\ & & \ddots & \\ 0 & & & D_Q \end{pmatrix} = \left(\begin{array}{cccc|c|ccc} k_1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & k_2 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 0 & \dots & k_{J_1} & \dots & 0 & \dots & 0 \\ \hline & & & & \ddots & & & \\ \hline 0 & 0 & \dots & 0 & \dots & k_l & \dots & 0 \\ 0 & 0 & \dots & 0 & \dots & 0 & \dots & k_J \end{array}\right)$$

A les taules Z i D els hi podem fer correspondre les taules de freqüències relatives associades

$$F = \frac{Z}{nQ} \qquad\qquad \tilde{D} = \frac{D}{nQ}$$

i les matrius de pesos de les modalitats D_p i dels individus D_n:

$$D_p = \frac{D}{nQ} \qquad\qquad D_n = \frac{1}{n} I_n$$

2.9.2 Anàlisi

Per a realitzar l'anàlisi per a Q variables començarem, com és habitual, pel cas particular de $Q = 2$ on veurem l'equivalència de les anàlisis de les taula Z, de la matriu de Burt B i de la matriu $Z_1^T Z_2$ i després ho estendrem a més de dues variables.

Anem doncs a analitzar l'equivalència de:

1. Anàlisi de correspondències de la taula Z
2. Anàlisi de correspondències de la taula B
3. Anàlisi de correspondències de la taula $Z_1^T Z_2$
4. Anàlisi canònica de Z_1 i Z_2

Per a veure aquesta equivalència, usant la notació que hem adoptat per l'anàlisi de correspondències, veurem que l'anàlisi de les diferents matrius ens porta als mateixos factors.

Realitzarem l'anàlisi de $D_J^{-1} F^T D_I^{-1} F$ per a la primera taula. En aquest cas, tenim: $F = \frac{Z}{nQ}$, $D_J = D_p = \frac{D}{nQ}$ i $D_I = D_n = \frac{1}{n} I_n$, per tant:

$$D_J^{-1} F^T D_I^{-1} F = \left(\frac{D}{nQ}\right)^{-1} \left(\frac{Z}{nQ}\right)^T \left(\frac{1}{n} I_n\right)^{-1} \frac{Z}{nQ} =$$
$$= nQ D^{-1} \frac{1}{nQ} Z^T n I_n \frac{1}{nQ} Z = \frac{1}{Q} D^{-1} Z^T Z$$

Per tant, haurem de trobar els valors i vectors propis de $\frac{1}{Q} D^{-1} Z^T Z$:

$$\frac{1}{Q} D^{-1} Z^T Z \psi_\alpha = \mu_\alpha \psi_\alpha \tag{2.21}$$

Si ara apliquem l'anàlisi de $D_J^{-1} F^T D_I^{-1} F$ per a la taula de Burt, tenim:
$F = \frac{B}{nQ^2}$, $D_J = D_I = \frac{D}{nQ}$, i per tant:

$$D_J^{-1} F^T D_I^{-1} F = \left(\frac{D}{nQ}\right)^{-1} \left(\frac{B}{nQ^2}\right)^T \left(\frac{D}{nQ}\right)^{-1} \frac{B}{nQ^2} =$$
$$= nQ D^{-1} \frac{1}{nQ^2} B^T nQ D^{-1} \frac{B}{nQ^2} = \frac{1}{Q^2} D^{-1} B^T D^{-1} B$$

Per tant, en aquest segon cas, haurem de trobar els valors i vectors propis de:

$$\frac{1}{Q^2}D^{-1}B^TD^{-1}B \tag{2.22}$$

Si prenem l'equació (2.21) i la premultipliquem per $\frac{1}{Q}D^{-1}B$:

$$\frac{1}{Q}D^{-1}B\left(\frac{1}{Q}D^{-1}Z^TZ\psi_\alpha\right) = \frac{1}{Q}D^{-1}B\mu_\alpha\psi_\alpha = \mu_\alpha\frac{1}{Q}D^{-1}Z^TZ\psi_\alpha = \mu_\alpha^2\psi_\alpha$$

per tant:

$$\frac{1}{Q^2}D^{-1}BD^{-1}B\psi_\alpha = \mu_\alpha^2\psi_\alpha$$

i, si tenim present que B és simètrica $B^T = B$, l'expressió a diagonalitzar (2.22), és l'expressió que acabem de trobar. Per tant les factoritzacions de Z i de B són les mateixes únicament que el valors propis de B són els de Z al quadrat.

Realitzarem l'anàlisi de $D_J^{-1}F^TD_I^{-1}F$ per a la tercera taula $Z_1^TZ_2$. Tenim: $F = \frac{Z_1^TZ_2}{n}$, $D_J = D_2 = \frac{Z_2^TZ_2}{n}$ i $D_I = D_1 = \frac{Z_1^TZ_1}{n}$, on D_1 i D_2 són les dues submatrius de la diagonal de D, per tant analitzem:

$$D_J^{-1}F^TD_I^{-1}F = \left(\frac{Z_2^TZ_2}{n}\right)^{-1}\left(\frac{Z_1^TZ_2}{n}\right)^{T}\left(\frac{Z_1^TZ_1}{n}\right)^{-1}\left(\frac{Z_1^TZ_2}{n}\right) = D_2^{-1}N^TD_1^{-1}N$$

Així doncs, hem de trobar els valors i vectors propis de:

$$D_2^{-1}N^TD_1^{-1}N \tag{2.23}$$

Per fer-ho partirem de l'equació (2.21) i de l'especial estructura de la matriu D:

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

$$\frac{1}{Q}D^{-1}Z^TZ\psi = \mu\psi = \frac{1}{2}\begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}^{-1}\begin{pmatrix} D1 & N \\ N^T & D_2 \end{pmatrix}\psi = \mu\psi$$

si diem:

$$\psi = \begin{pmatrix} a \\ b \end{pmatrix}$$

podem escriure llavors:

$$\frac{1}{2}\begin{pmatrix} D_1^{-1} & 0 \\ 0 & D_2^{-1} \end{pmatrix}\begin{pmatrix} D1 & N \\ N^T & D_2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \mu\begin{pmatrix} a \\ b \end{pmatrix}$$

$$\frac{1}{2}\begin{pmatrix} I_{J_1} & D_1^{-1}N \\ D_2^{-1}N^T & I_{J_2} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \mu \begin{pmatrix} a \\ b \end{pmatrix}$$

d'on podem obtenir les equacions:

$$\begin{cases} a + D_1^{-1}Nb = 2\mu a \\ D_2^{-1}N^T a + b = 2\mu b \end{cases} \quad \text{i anàlogament} \quad \begin{cases} D_1^{-1}Nb = (2\mu - 1)a \\ D_2^{-1}N^T a = (2\mu - 1)b \end{cases}$$

Són expressions de transició, que per substitució ens donen:

$$\begin{cases} D_1^{-1}ND_2^{-1}N^T a = (2\mu - 1)^2 a \\ D_2^{-1}N^T D_1^{-1}Nb = (2\mu - 1)^2 b \end{cases}$$

On obtenim la factorització de l'expressió (2.23) i de la seva transposada i la relació dels valors propis d'aquesta amb els de l'expressió primera (2.21). Per tant, les tres anàlisis són equivalents i la darrera equivalència ja l'havíem vista en la secció anterior. En la següent taula en fem un resum de les equivalències:

Taula 2.10: Equivalències de les anàlisis en ACM de dues qüestions

Taula a analitzar	dimensió	Taula que factoritzem	Valors propis
$Z = [Z_1 Z_2]$	$n * J$ on $J = J_1 + J_2$	$\frac{1}{2}D^{-1}Z^T Z$	μ
$B = Z^T Z$	$J * J$	$\frac{1}{4}D^{-1}B^T D^{-1}B$	μ^2
$N = Z_1^T Z_2$	$J_1 * J_2$	$D_2^{-1}N^T D_1^{-1}N$	$(2\mu - 1)^2$

Com a exemple, si prenem les variables Z_1 i Z_2 de l'exemple que hem anat elaborant, els valors propis de les seves anàlisis els podem trobar a la Taula 2.11.

Taula 2.11: Taula de valors propis de l'exemple

Taula a analitzar	Valors propis
$Z = [Z_1 Z_2]$	$1, 0.8010, 0.5, 0.1989, 0$
$B = Z^T Z$	$1, 0.6417, 0.25, 0.0396, 0$
$N = Z_1^T Z_2$	$1, 0.3625$

Aquests valors propis compleixen les igualtats prèviament descrites. A més, podem veure que ens surt a tot arreu el valor propi 1 corresponent al fet que les dades no estan

centrades i en els dos primers casos el valor propi 0, corresponent als lligams lineals entre modalitats de les diferents variables.

Per a realitzar l'anàlisi per a més de dues variables, podem simplement trobar els valors i vectors propis de la taula disjuntiva $Z = (Z_1|Z_2|\ldots|Z_Q)$, ja que això ens permetria obtenir una representació conjunta de les $J = J_1 + \ldots + J_Q$ categories:

$$\frac{1}{Q}D^{-1}Z^TZ\varphi_\alpha = \frac{1}{Q}D^{-1}B\varphi_\alpha = \mu_\alpha\varphi_\alpha \tag{2.24}$$

ara bé, aquesta extensió formal de dues a més de dues variables del mètode, no ens el justificaria matemàticament, aquesta justificació la tindrem en la secció següent.

2.9.3 L'anàlisi canònica generalitzada i l'ACM

Donada la nostra taula disjuntiva completa $Z = (Z_1|Z_2|\ldots|Z_Q)$, si apliquem la factorització (2.20), la matriu M no és res més que la inversa de la matriu D definida com a la diagonal de la matriu de Burt dividida per n, $M^{-1} = D/n$, i la matriu $D = \frac{1}{n}I_n$ per tant:

$$(D/n)^{-1}Z^T\frac{1}{n}I_nZb = \mu b \Longleftrightarrow nD^{-1}Z^T\frac{1}{n}Zb = \mu b$$

$$D^{-1}Z^TZb = \mu b \Longleftrightarrow D^{-1}Bb = \mu b$$

que exceptuant una constant és l'expressió que havíem de factoritzar en l'equació (2.24). El fet d'haver-hi la diferència d'aquesta constant $\frac{1}{Q}$, podríem evitar-lo si en comptes d'agafar la generalització deguda a Carrol, agaféssim com a generalització la mitjana de la suma de quadrats i no la suma mateixa:

$$\frac{1}{Q}\sum_{i=1}^{Q}R^2(z, X_i) \quad \text{sigui màxim}$$

A més, es pot veure que aquesta anàlisi ens porta a la mateixa descomposició factorial que ens dóna les components i els factors de l'anàlisi de components principals (**ACP**) d'una taula X amb mètrica M. Així doncs, la generalització de l'anàlisi canònica és equivalent a una ACP, que té més en compte la relació entre modalitats que no pas entre variables.

Tenim doncs, en resum que, en l'estudi des de la perspectiva d'$\mathbb{R}^J$ que el que cal és fer la descomposició de l'expressió:

$$\boxed{\frac{1}{Q}D^{-1}Bb = \mu b \quad \text{amb la normalització} \quad b^TD_pb = b^T\frac{D}{nQ}b = 1}$$

Però, com la matriu $\frac{1}{Q}D^{-1}B$ no és simètrica, la solució habitual per tal de resoldre-ho, sol ser premultiplicar-la per $D^{-1/2}$, així doncs:

$$\frac{1}{Q}D^{-1}Bb = \mu b \Longleftrightarrow \frac{1}{Q}D^{-1/2}D^{-1}Bb = \mu D^{-1/2}b$$

si anomenem: $\phi = D^{1/2}b$ o equivalentment $b = D^{-1/2}\phi$, llavors:

$$\frac{1}{Q}D^{-1/2}D^{-1}BD^{1/2}\phi = \mu\phi$$

on $\frac{1}{Q}D^{1/2}D^{-1}BD^{-1/2} = \frac{1}{Q}D^{-1/2}BD^{-1/2}$ és simètrica i per tant fàcilment diagonalitzable. Llavors per a obtenir b desfem el canvi.

Des de la perspectiva d'$\mathbb{R}^n$, ens caldrà fer la descomposició de:

$$\boxed{\frac{1}{Q}ZD^{-1}Z^Ta = \mu a \quad \text{amb la normalització} \quad a^TD_na = 1 \Leftrightarrow a^Ta = n}$$

Aquestes dues descomposicions són equivalents, ja que podem trobar unes **relacions de transició** d'una descomposició amb l'altra.

Sabem que:

$$\frac{1}{Q}D^{-1}Bb = \mu b \quad \text{amb} \quad b^TD_pb = b^T\frac{D}{nQ}b = 1$$

$$\frac{1}{Q}ZD^{-1}Z^Ta = \mu a \quad \text{amb} \quad a^TD_na = 1 \Leftrightarrow a^Ta = n$$

Si premultipliquem la primera d'aquestes dues expressions per Z i la segona per $D^{-1}Z^T$, obtenim:

$$\frac{1}{Q}ZD^{-1}Z^TZb = \mu Zb$$

$$\frac{1}{Q}D^{-1}Z^TZD^{-1}Z^Ta = \mu D^{-1}Z^Ta$$

i d'aquí obtenim que Zb és vector propi de la primera i $D^{-1}Z^Ta$ de la segona, per tant, múltiples dels vectors que teníem:

$$k_1Zb = a \quad \text{i} \quad k_2D^{-1}Z^Ta = b$$

Si ara els apliquem la condició de normalització, obtenim per a la primera expressió que:

$$(k_1Zb)^T(k_1Zb) = n \Leftrightarrow k_1^2b^TZ^TZb = n \Leftrightarrow k_1^2b^TBb = n$$

$$b^TBb = \frac{n}{k_1^2} \Leftrightarrow b^TQD\mu b = \frac{n}{k_1^2} \Leftrightarrow b^TDb = \frac{n}{Q\mu k_1^2} \Leftrightarrow b^T\frac{D}{nQ}b = \frac{1}{Q^2\mu k_1^2}$$

$$k_1^2 = \frac{1}{Q^2\mu} \Leftrightarrow k_1 = \frac{1}{Q\sqrt{\mu}}$$

I anàlogament per a la segona expressió:

$$(k_2 D^{-1} Z^T a)^T \frac{D}{nQ}(k_2 D^{-1} Z^T a) = 1 \Leftrightarrow k_2^2 a^T Z D^{-1} \frac{D}{nQ} D^{-1} Z^T a = 1$$

$$k_2^2 \frac{1}{n} a^T \frac{1}{Q} Z D^{-1} Z^T a = 1 \Leftrightarrow k_2^2 \frac{1}{n} a^T \mu a = 1 \Leftrightarrow \frac{k_2^2}{n} a^T \mu a = 1$$

$$k_2^2 \mu = 1 \Leftrightarrow k_2 = \frac{1}{\sqrt{\mu}}$$

Així doncs:

$$a = \frac{1}{Q\sqrt{\mu}} Z b \tag{2.25}$$

$$b = \frac{1}{\sqrt{\mu}} D^{-1} Z^T a \tag{2.26}$$

2.9.4 Coordenades

Tenim que les coordenades de cadascuna de les modalitats de les variables (en l'espai $\mathbb{R}^n$) són la projecció d'aquestes sobre els vectors propis i anàlogament per les coordenades dels individus (en l'espai $\mathbb{R}^J$) sobre els vectors propis respectius. Anem doncs a projectar, per a obtenir les coordenades:

Coordenades de les modalitats: hem de projectar sobre $\mathbb{R}^n$, per tant sobre els vectors propis a la nostra matriu de dades de l'anàlisi $D^{-1}Z^T$, obtenim: $\varphi = D^{-1}Z^T a$

Coordenades dels individus: hem de projectar sobre $\mathbb{R}^J$, per tant sobre els vectors propis b la nostra matriu de dades de l'anàlisi $\frac{1}{Q}Z$, per tant, tenim: $\psi = \frac{1}{Q}Zb$

Coordenades fileres	Coordenades columnes
$\psi = \frac{1}{Q}Zb$	$\varphi = D^{-1}Z^T a$

A més, podem veure que les expressions de les coordenades, són molt semblants a les relacions de transició dels vectors propis. Això ens donarà una de les següents propietats.

2.9.5 Propietats

Propietat 2.9.5.1 *La coordenada d'una modalitat en un eix és proporcional a la mitjana de les coordenades sobre el mateix eix dels individus que han escollit aquesta modalitat i anàlogament, sobre un eix qualsevol la coordenada d'un individu qualsevol és proporcional al punt mitjà de les coordenades que ell ha escollit.*

Demostració: Si prenem les coordenades de les columnes -modalitats- $\varphi = D^{-1}Z^T a$ i li apliquem les relacions de transició $a = \frac{1}{Q\sqrt{\mu}}Zb$, obtenim que:

$$\varphi = D^{-1}Z^T\frac{1}{Q\sqrt{\mu}}Zb = \frac{1}{\sqrt{\mu}}D^{-1}Z^T\frac{1}{Q}Zb = \frac{1}{\sqrt{\mu}}D^{-1}Z^T\psi$$

on l'expressió $D^{-1}Z^T\psi$ ens dóna la mitjana dels individus que han escollit aquesta modalitat. I anàlogament passa per les coordenades d'un individu sobre un eix, mitjançant les relacions de transició obtenim:

$$\psi = \frac{1}{\sqrt{\mu}}\frac{1}{Q}Z\varphi$$

on l'expressió $\frac{1}{Q}Z\varphi$ ens dóna la mitjana de les coordenades de les Q modalitats escollides per l'individu. □

A més d'aquesta propietat referida a les coordenades, tenim aquesta altre referida a les inèrcies de les modalitats i de les qüestions.

2.9.6 Qualitat de l'aproximació

Per una dimensió donada K^* $(1 \leq K^* \leq K = rang(B))$, la qualitat global de les representacions gràfiques de dimensió K^* es mesura per la relació entre la suma del K^* primers valors propis de l'AFC i llur suma completa d'1 a K.

$$Q = \frac{\sum_{j=1}^{K^*}\lambda_j}{\sum_{i=1}^{K}\lambda_i}$$

La qualitat de l'aproximació ens valora la inèrcia de la projecció en relació a la inèrcia total. En ACM les taxes d'inèrcia $\frac{\lambda_j}{\sum_{i=1}^{K}\lambda_i}$ són més dèbils que en ACS, ja que hi ha molts més eixos factorials, deguts a la presència de moltes més modalitats i, tot sovint, al fet que no hi ha modalitats amb efectius molt dèbils, que es donarien taxes més grans com veurem més endavant.

2.9.7 Eines d'ajuda a la interpretació: contribucions i cosinus quadrats

Anàlogament com hem vist en correspondències simples, les eines d'ajuda a la interpretació ens mesuren, per una banda, la qualitat de la representació d'un individu o modalitat en un eix factorial i per altra banda la influència d'un individu o modalitat en la determinació de l'eix. Definim doncs:

Definició 2.9.7.1 La *contribució absoluta d'un punt i en un eix k* és la proporció amb la qual un punt contribueix a l'inèrcia de l'eix.

Donat un eix qualsevol, k, la inèrcia d'aquest eix es pot trobar com $\lambda_k = \sum_{i=1}^{I} \psi_{ik}^2 f_i$, on $\psi_{ik}^2 f_i$ és la quantitat aportada per cada individu, distància al quadrat per la massa de l'individu, llavors la contribució absoluta serà:

$$cnt(i,k) = \frac{\psi_{ik}^2 f_i}{\lambda_k}$$

Aquells individus amb contribucions absolutes grans seran els que ens determinaran la posició de l'eix i, per tant, ens ajuden a interpretar-lo.

Tenim per definició que $\sum_i cnt(i,k) = 1$. Si expressem les contribucions en percentatge, llavors aquesta suma és igual a 100.

Definició 2.9.7.2 Els *cosinus quadrats d'un punt i en un eix k* o la *contribució relativa d'un eix k sobre l'individu i* és la qualitat de representació d'un punt en un eix.

Com que la distància al quadrat d'un individu al centre de gravetat del núvol de punts la podem descompondre com a la suma de totes les projeccions al quadrat sobre els eixos $d^2(i,G) = \sum_k \psi_{ik}^2$, llavors al contribució relativa d'un eix a la representació del punt serà:

$$cos^2(i,k) = \frac{\psi_{ik}^2}{\sum_k \psi_{ik}^2}$$

s'anomena *cosinus quadrat* perquè aquesta quantitat obtinguda és exactament el cosinus quadrat de l'angle que hi ha entre la semirecta entre de l'origen a l'individu i i la semirecta de l'origen a la projecció sobre un eix k de l'individu i.

Si el cos^2 és proper a zero l'individu està mal representat en aquell eix, en canvi si els cos^2 és proper a 1, l'individu està ben representat en aquell eix.

Tenim per definició que $\sum_k cos^2(i,k) = 1$. Si expressem aquestes contribucions en percentatge, aleshores la suma és igual a 100.

Podem definir anàlogament les contribucions absolutes des del punt de vista de les columnes canviant les coordenades i els pesos dels individus per les coordenades i pesos de les columnes.

$$cnt(j,k) = \frac{\varphi_{jk}^2 c_j}{\lambda_k} \qquad cos^2(j,k) = \frac{\varphi_{jk}^2}{\sum_k \varphi_{jk}^2}$$

2.9.8 Exemple

Anem a presentar un exemple d'aplicació de les anàlisis de correspondències múltiples a les notes obtingudes per 86 estudiants de dos grups diferents en 6 matèries de primer curs de carrera. Aquestes matèries són:

- Ecologia (ECO)
- Física (FIS)
- Matemàtiques (MAT)
- Química (QUI)
- Dibuix (DIB)
- Biologia (BIO)

que juntament amb la variable Grup són les 7 variables a analitzar.

Per a cadascun dels alumnes, s'han categoritzat les qualificacions obtingudes en 4 modalitats: no presentat(np), suspens(sus), aprovat(ap) i notable(not) que fan un total de 26 modalitats amb les dues corresponents als grup (ea i ia). Podem trobar a la Taula 2.13 les dades d'aquest exemple.

Tenim, doncs, una anàlisi de 7 qüestions, amb 26 modalitats totals i 86 individus. L'anàlisi de correspondències múltiples ens donarà la descomposició en valors propis de la inèrcia total, que en aquest cas és igual a $I_{Tot} = 2.71429$.

Aquesta descomposició la podem trobar en la Taula 2.12. Podem veure que ens dóna inèrcies corresponents a 19 eixos factorials, que corresponen al total de modalitats menys una unitat per qüestió $(6 * 3 + 1 * 1)$, a més les inèrcies aportades són petites, la més gran en aquest exemple és del 18% sobre el total. Amb els dos primers eixos només tenim una qualitat de 30.5% i per tenir una qualitat del 75% hem de recórrer a 9 eixos factorials, com ja havíem comentat anteriorment.

Presentem a continuació les coordenades de les columnes (modalitats) i de les fileres (individus) així com les eines d'ajuda a la interpretació que són les contribucions i els cosinus quadrats. Això ens permetrà visualitzar els gràfics de les coordenades i estudiar la qualitat de la representació de cadascun dels punts.

Com en totes les aplicacions de la metodologia de l'anàlisi de correspondències, s'ha de fer l'anàlisi dels resultats, no només en base als gràfics sinó tenint presents les eines d'ajuda a la interpretació. En aquest cas vegem en la Taula 2.15 que el primer eix ve

Taula 2.12: Valors propis, percentatges d'inèrcia i percentatges acumulats

Valor propi	% Inèrcia	% Iner.Acum.
0.510668	18.8141	18.814
0.317144	11.6842	30.498
0.276668	10.1930	40.691
0.187250	6.8987	47.590
0.174965	6.4461	54.036
0.157556	5.8047	59.841
0.146932	5.4133	65.254
0.137481	5.0651	70.319
0.131523	4.8456	75.165
0.110885	4.0853	79.250
0.102255	3.7673	83.017
0.090573	3.3369	86.354
0.080460	2.9643	89.318
0.077772	2.8653	92.184
0.062678	2.3092	94.493
0.051938	1.9135	96.406
0.039671	1.4616	97.868
0.033467	1.2330	99.101
0.024399	0.8989	100.000

determinat des del punt de vista de les modalitats per la qualificació de *Notable* en totes les 6 qüestions, ja que aquestes 6 modalitats contribueixen amb un 83.4 % de la inèrcia total del primer eix i, a més, totes amb coordenada positiva, per tant definint aquest eix com a l'eix dels notables.

Aquesta mateixa anàlisi del primer eix la podem fer des del punt de vista dels individus, en la Taula 2.16, on observem que les grans contribucions són dels individus 53 i 76, ja que ells sols contribueixen a explicar, des del punt de vista dels individus, el 61.4 % de la inèrcia de l'eix. Analitzats aquests dos individus són alumnes amb totes les qualificacions de notable, la qual cosa concorda amb el fet que estiguin situats prop del centre de gravetat de les modalitats *notable*.

Una anàlisi més detallada caldria repetir aquest procés per a cadascun dels eixos. Per a finalitzar l'exemple presentem els gràfics factorials d'aquestes dades:

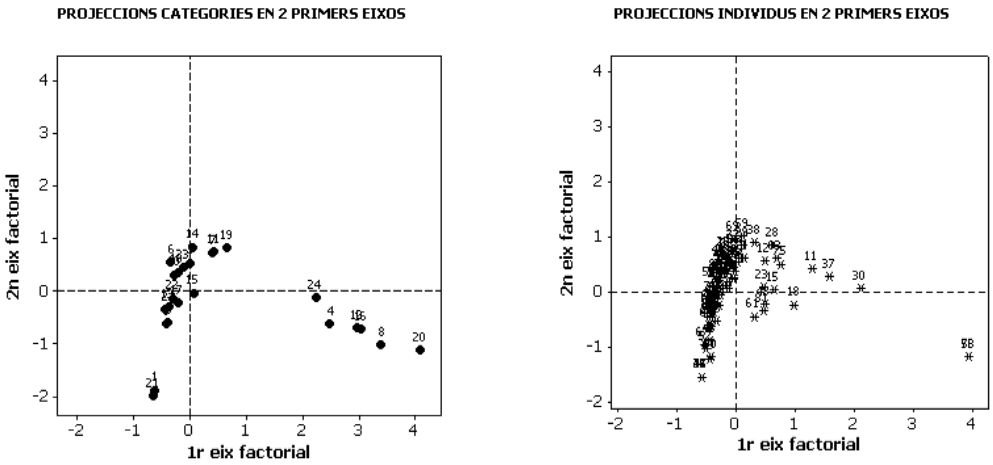


Figura 2.9: Gràfics factorials de les fileres i de les columnes

Aquests poden ser superposats en un gràfic comú per tal de veure els lligams entre modalitats de les fileres i de les columnes:

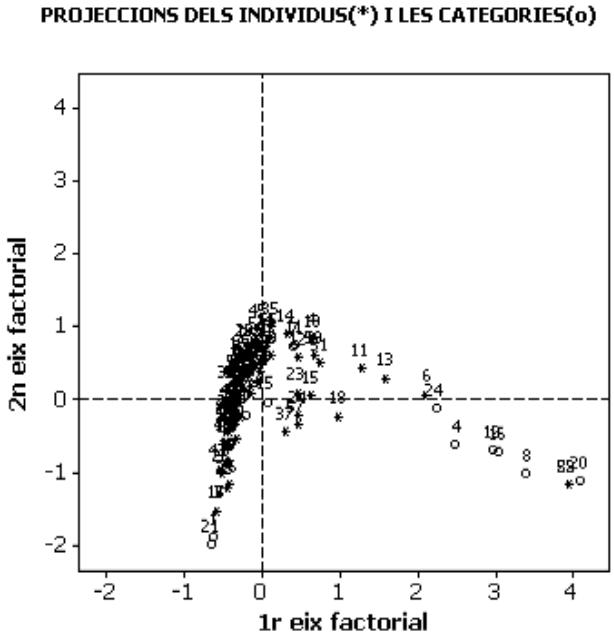


Figura 2.10: Gràfic factorial conjunt de les fileres i de les columnes

Taula 2.13: Qualificacions de les matèries de primer curs

ECO	FIS	MAT	QUI	DIB	BIO	GRUP
2	1	1	2	1	2	1
3	1	1	1	1	3	1
3	3	3	3	1	3	1
2	2	2	2	2	2	1
2	2	2	1	2	2	1
3	1	1	1	1	1	1
3	1	2	3	1	3	1
3	1	2	2	1	3	1
3	1	3	1	1	3	1
3	3	1	2	1	3	1
3	3	3	4	2	4	1
3	3	2	2	1	4	1
3	2	2	2	1	3	1
3	1	2	1	2	3	1
3	3	4	3	1	2	1
1	1	1	1	1	1	1
3	2	1	1	2	2	1
4	3	1	3	1	4	1
3	1	2	1	1	3	1
3	2	2	2	1	3	1
2	1	1	1	1	3	1
3	2	3	1	2	3	1
3	1	3	3	1	4	1
3	3	1	3	1	3	1
1	1	2	3	1	3	1
3	1	1	1	3	3	1
1	1	1	1	1	1	1
3	3	3	2	2	4	1
3	3	3	2	1	3	1
3	4	3	4	3	4	1
2	2	3	3	3	3	1
3	1	1	1	1	2	1
3	2	1	3	2	3	1
2	1	1	1	2	2	1
3	1	3	3	1	3	1
2	2	1	2	1	3	1
4	3	3	2	4	3	1
3	3	3	3	3	3	1
3	1	2	2	1	3	1
3	2	2	1	2	3	1
1	1	1	1	1	1	1
2	1	1	1	1	2	1
2	1	1	3	2	1	1

Taula 2.14: Qualificacions de les matèries de primer curs (Cont.)

ECO	FIS	MAT	QUI	DIB	BIO	GRUP
1	1	2	1	1	3	1
2	1	1	1	2	3	1
2	1	2	1	1	3	1
3	1	1	1	1	3	1
3	4	2	1	1	2	1
1	1	2	3	1	1	2
1	3	1	1	1	1	2
3	3	2	2	1	3	2
1	1	1	1	1	2	2
4	4	4	4	4	4	2
2	1	3	1	1	2	2
2	2	2	1	2	2	2
3	2	1	2	1	3	2
2	2	2	1	2	2	2
3	2	3	1	1	3	2
3	3	3	2	2	3	2
2	3	2	2	1	2	2
4	1	1	3	1	3	2
2	2	1	3	1	2	2
2	1	1	1	1	3	2
2	1	1	1	1	2	2
3	3	3	1	1	2	2
3	1	1	2	1	2	2
2	1	1	1	2	1	2
2	2	1	1	1	2	2
3	2	3	2	2	3	2
3	1	1	1	2	3	2
2	2	1	1	1	3	2
3	3	3	1	1	2	2
3	2	1	1	2	2	2
3	3	3	1	2	3	2
4	3	3	1	3	3	2
4	4	4	4	4	4	2
3	2	2	2	3	2	2
3	1	2	2	2	2	2
3	1	1	1	1	2	2
3	2	1	1	2	3	2
2	1	3	4	1	2	2
2	1	1	1	2	2	2
3	3	4	2	2	3	2
3	1	2	1	2	3	2
3	3	3	1	1	3	2
3	3	2	1	2	3	2

Taula 2.15: Coordenades, contribucions i cosinus quadrats de les columnes

modal	cor.1	cor.2	cor.3	cnt.1	cnt.2	cnt.3	cos.1	cos.2	cos.3
ECO-np	-0.63	-1.84	1.24	1.0	14.2	7.4	0.04	0.35	0.16
ECO-sus	-0.44	-0.37	-0.91	1.4	1.7	11.5	0.07	0.05	0.30
ECO-ap	0.00	0.55	0.21	0.0	7.8	1.2	0.00	0.40	0.06
ECO-not	2.50	-0.60	0.17	12.2	1.1	0.1	0.47	0.03	0.00
FIS-np	-0.40	-0.59	0.20	2.0	7.2	1.0	0.13	0.29	0.03
FIS-sus	-0.33	0.53	-0.85	0.8	3.1	9.1	0.04	0.09	0.23
FIS-ap	0.41	0.73	0.57	1.2	6.2	4.3	0.06	0.18	0.11
FIS-not	3.40	-1.03	-0.64	15.0	2.2	1.0	0.56	0.05	0.02
MAT-np	-0.41	-0.61	-0.10	2.0	7.0	0.2	0.12	0.27	0.01
MAT-sus	-0.27	0.32	-0.14	0.6	1.3	0.3	0.03	0.04	0.01
MAT-ap	0.43	0.77	0.42	1.3	6.9	2.3	0.06	0.21	0.06
MAT-not	3.00	-0.71	-0.57	11.7	1.1	0.8	0.44	0.02	0.02
QUI-np	-0.38	-0.31	-0.27	2.1	2.3	1.9	0.16	0.11	0.08
QUI-sus	0.04	0.84	-0.05	0.0	7.7	0.0	0.00	0.23	0.00
QUI-ap	0.05	0.02	1.04	0.0	0.0	9.7	0.00	0.00	0.23
QUI-not	3.05	-0.73	-0.53	15.1	1.4	0.8	0.57	0.03	0.02
DIB-np	-0.22	-0.21	0.33	0.8	1.1	3.2	0.07	0.06	0.15
DIB-sus	-0.20	0.34	-0.69	0.3	1.6	7.4	0.02	0.05	0.21
DIB-ap	0.65	0.84	0.51	0.8	2.2	0.9	0.03	0.05	0.02
DIB-not	4.11	-1.14	-0.58	16.5	2.0	0.6	0.61	0.05	0.01
BIO-np	-0.66	-1.96	0.98	1.1	16.1	4.6	0.04	0.39	0.10
BIO-sus	-0.29	-0.17	-0.97	0.7	0.4	14.6	0.04	0.01	0.41
BIO-ap	-0.12	0.47	0.33	0.2	5.2	2.8	0.01	0.23	0.11
BIO-not	2.23	-0.09	0.37	12.9	0.0	0.7	0.51	0.00	0.01
GRUP-ia	-0.07	0.05	0.45	0.1	0.1	5.9	0.01	0.00	0.26
GRUP-ea	0.09	-0.06	-0.57	0.1	0.1	7.5	0.01	0.00	0.26

Taula 2.16: Coordenades, contribucions i cosinus quadrats de les fileres

ind.	cor.1	cor.2	cor.3	cnt.1	cnt.2	cnt.3	cos.1	cos.2	cos.3
1	-0.36	-0.27	-0.28	0.3	0.3	0.3	0.07	0.04	0.05
2	-0.32	-0.17	0.31	0.2	0.1	0.4	0.11	0.03	0.10
3	0.10	0.61	0.91	0.0	1.3	3.5	0.00	0.19	0.42
4	-0.31	0.39	-0.86	0.2	0.6	3.1	0.04	0.06	0.30
5	-0.39	0.10	-0.92	0.4	0.0	3.5	0.07	0.00	0.40
6	-0.43	-0.78	0.49	0.4	2.2	1.0	0.08	0.28	0.11
7	-0.20	0.16	0.65	0.1	0.1	1.8	0.02	0.01	0.26
8	-0.20	0.36	0.36	0.1	0.5	0.5	0.03	0.09	0.09
9	-0.15	0.19	0.45	0.1	0.1	0.9	0.02	0.03	0.17
10	-0.07	0.46	0.47	0.0	0.8	0.9	0.00	0.14	0.15
11	1.17	0.41	0.22	3.1	0.6	0.2	0.27	0.03	0.01
12	0.43	0.56	0.47	0.4	1.1	0.9	0.06	0.11	0.08
13	-0.19	0.65	0.07	0.1	1.5	0.0	0.02	0.25	0.00
14	-0.29	0.21	0.02	0.2	0.2	0.0	0.06	0.03	0.00
15	0.58	0.07	0.29	0.8	0.0	0.3	0.07	0.00	0.02
16	-0.55	-1.39	0.77	0.7	7.1	2.5	0.09	0.55	0.17
17	-0.33	0.10	-0.60	0.3	0.0	1.5	0.07	0.01	0.22
18	0.90	-0.18	0.77	1.8	0.1	2.5	0.17	0.01	0.12
19	-0.29	0.07	0.30	0.2	0.0	0.4	0.07	0.00	0.08
20	-0.19	0.65	0.07	0.1	1.5	0.0	0.02	0.25	0.00
21	-0.41	-0.40	0.01	0.4	0.6	0.0	0.13	0.13	0.00
22	-0.13	0.61	-0.11	0.0	1.4	0.0	0.01	0.22	0.01
23	0.41	0.13	0.82	0.4	0.1	2.8	0.06	0.01	0.23
24	-0.07	0.25	0.77	0.0	0.2	2.5	0.00	0.04	0.34
25	-0.33	-0.45	0.94	0.2	0.7	3.7	0.04	0.07	0.30
26	-0.14	0.10	0.36	0.0	0.0	0.6	0.01	0.00	0.05
27	-0.55	-1.39	0.77	0.7	7.1	2.5	0.09	0.55	0.17
28	0.57	0.81	0.35	0.7	2.4	0.5	0.10	0.20	0.04
29	0.10	0.81	0.61	0.0	2.4	1.6	0.01	0.38	0.22
30	1.94	0.09	0.22	8.6	0.0	0.2	0.41	0.00	0.01

Taula 2.17: Coordenades, contribucions i cosinus quadrats de les fileres (Cont.)

ind.	cor.1	cor.2	cor.3	cnt.1	cnt.2	cnt.3	cos.1	cos.2	cos.3
31	0.03	0.59	0.27	0.0	1.3	0.3	0.00	0.08	0.02
32	-0.35	-0.33	-0.04	0.3	0.4	0.0	0.11	0.09	0.00
33	-0.21	0.34	0.10	0.1	0.4	0.0	0.02	0.06	0.01
34	-0.44	-0.42	-0.62	0.4	0.7	1.6	0.11	0.11	0.23
35	-0.06	0.27	0.81	0.0	0.3	2.7	0.00	0.04	0.38
36	-0.31	0.18	-0.22	0.2	0.1	0.2	0.05	0.02	0.03
37	1.46	0.28	0.36	4.9	0.3	0.5	0.29	0.01	0.02
38	0.27	0.87	0.96	0.2	2.8	3.8	0.02	0.20	0.24
39	-0.20	0.36	0.36	0.1	0.5	0.5	0.03	0.09	0.09
40	-0.27	0.50	-0.26	0.2	0.9	0.3	0.05	0.15	0.04
41	-0.55	-1.39	0.77	0.7	7.1	2.5	0.09	0.55	0.17
42	-0.44	-0.56	-0.34	0.4	1.2	0.5	0.13	0.22	0.08
43	-0.42	-0.79	0.26	0.4	2.3	0.3	0.05	0.19	0.02
44	-0.42	-0.54	0.58	0.4	1.1	1.4	0.07	0.12	0.14
45	-0.40	-0.26	-0.27	0.4	0.2	0.3	0.11	0.05	0.05
46	-0.38	-0.16	0.00	0.3	0.1	0.0	0.10	0.02	0.00
47	-0.32	-0.17	0.31	0.2	0.1	0.4	0.11	0.03	0.10
48	0.44	-0.20	-0.28	0.4	0.2	0.3	0.05	0.01	0.02
49	-0.41	-1.10	0.83	0.4	4.4	2.9	0.04	0.28	0.16
50	-0.36	-1.08	0.59	0.3	4.3	1.5	0.03	0.31	0.09
51	-0.01	0.67	0.18	0.0	1.7	0.1	0.00	0.26	0.02
52	-0.45	-0.96	-0.04	0.5	3.4	0.0	0.08	0.37	0.00
53	3.67	-1.11	-0.64	30.7	4.5	1.7	0.86	0.08	0.03
54	-0.24	-0.24	-0.48	0.1	0.2	1.0	0.03	0.03	0.13
55	-0.36	0.07	-1.20	0.3	0.0	6.0	0.06	0.00	0.66
56	-0.19	0.38	-0.19	0.1	0.5	0.2	0.02	0.09	0.02
57	-0.36	0.07	-1.20	0.3	0.0	6.0	0.06	0.00	0.66
58	-0.11	0.44	-0.11	0.0	0.7	0.1	0.01	0.13	0.01
59	0.13	0.92	0.06	0.0	3.1	0.0	0.01	0.42	0.00
60	-0.14	0.27	-0.48	0.0	0.3	0.9	0.01	0.03	0.10
61	0.30	-0.40	0.38	0.2	0.6	0.6	0.03	0.05	0.04
62	-0.31	-0.22	-0.55	0.2	0.2	1.3	0.04	0.02	0.13
63	-0.38	-0.43	-0.27	0.3	0.7	0.3	0.11	0.14	0.06
64	-0.41	-0.59	-0.62	0.4	1.3	1.6	0.11	0.23	0.26
65	0.01	0.33	-0.08	0.0	0.4	0.0	0.00	0.07	0.00
66	-0.24	-0.06	-0.26	0.1	0.0	0.3	0.04	0.00	0.04
67	-0.48	-0.91	-0.37	0.5	3.0	0.6	0.08	0.29	0.05
68	-0.40	-0.31	-0.91	0.4	0.3	3.5	0.09	0.05	0.47
69	-0.02	0.87	-0.33	0.0	2.8	0.5	0.00	0.37	0.05
70	-0.28	-0.05	-0.24	0.2	0.0	0.2	0.06	0.00	0.05
71	-0.36	-0.14	-0.56	0.3	0.1	1.3	0.08	0.01	0.20
72	0.01	0.33	-0.08	0.0	0.4	0.0	0.00	0.07	0.00
73	-0.30	0.07	-0.88	0.2	0.0	3.3	0.05	0.00	0.45
74	0.05	0.63	0.00	0.0	1.5	0.0	0.00	0.23	0.00
75	0.72	0.47	0.32	1.2	0.8	0.4	0.10	0.04	0.02
76	3.67	-1.11	-0.64	30.7	4.5	1.7	0.86	0.08	0.03
77	-0.02	0.72	-0.51	0.0	1.9	1.1	0.00	0.14	0.07
78	-0.20	0.31	-0.55	0.1	0.4	1.3	0.02	0.05	0.15
79	-0.32	-0.36	-0.32	0.2	0.5	0.4	0.08	0.10	0.08
80	-0.27	0.23	-0.53	0.2	0.2	1.2	0.05	0.03	0.18
81	0.44	-0.35	-0.55	0.4	0.4	1.3	0.05	0.03	0.08
82	-0.41	-0.45	-0.90	0.4	0.8	3.4	0.09	0.12	0.46
83	0.64	0.55	-0.21	0.9	1.1	0.2	0.09	0.07	0.01
84	-0.25	0.18	-0.25	0.1	0.1	0.3	0.05	0.02	0.05
85	0.04	0.49	0.27	0.0	0.9	0.3	0.00	0.16	0.05
86	-0.09	0.52	-0.15	0.0	1.0	0.1	0.01	0.16	0.01

Capítol 3

Models loglineals i models gràfics

3.1 Introducció

Les relacions d'independència condicional no són fàcils de detectar ni d'interpretar. La teoria de grafs d'independència és una eina relativament recent per resumir i clarificar aquestes interaccions, i es troba en el treball de Darroch, Lauritzen i Speed [DLS80], seguit d'altres com [WL83], popularitzats per Whittaker [Whi90a] i, més recentment, Edwards [Edw95] ha publicat un llibre on inclou el programa MIM dissenyat per a l'estudi dels models gràfics d'independència. En aquest capítol, a banda d'introduir els conceptes d'independència i independència condicional, n'especificarem algunes propietats. A continuació introduirem alguns conceptes de la teoria de grafs necessaris per a definir els grafs d'independència condicional. Seguidament com a models aplicats a dades multivariants donem els models loglineals i els models gràfics amb la representació gràfica de les seves estructures d'independència. Aquests conceptes juntament amb la generació de models i la **deviància** d'un model, ens donaran part de la fonamentació i de l'estudi de les anàlisis multicondicionals.

3.1.1 Independència i independència condicional

Sigui X una variable aleatòria amb funció de densitat o de massa $f_X(x)$ on, si X és discreta, podem escriure-ho com a $\Pr(X = j)\ \forall j$.

Definició 3.1.1.1 Dues variables aleatòries X i Y es diu que són *independents*, notat com $X \perp Y$, si la seva funció de densitat conjunta factoritza com el producte de les seves

funcions de densitats marginals.

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Una definició alternativa d'independència és que la funció de densitat condicional de Y, donat $X = x$ no és funció de x, no depèn d'x, la qual cosa pot ser escrita com:

$$f_{Y|X}(y|x) = f_Y(y)$$

L'avantatge d'aquesta darrera caracterització d'independència és que prescindim de la funció de densitat de la variable X.

Tenim a més les següents propietats:

- Propietat de factorització: existeixen funcions g i h, tals que $f_{XY}(x,y) = g(x)h(y)$
- Propietat de reducció: la independència conjunta implica la independència marginal: $X \perp (Y,Z) \Rightarrow X \perp Y, X \perp Z$, el recíproc és generalment fals

Definim ara la independència condicional.

Definició 3.1.1.2 *Considerem ara tres variables aleatòries X, Y i Z, si per cada valor de Z, X i Y són independents en la distribució condicional donat $Z = z$, llavors direm que X i Y són* **condicionalment independents** *donat Z i ho representarem seguint la notació deguda a Dawid [Daw79] com $X \perp Y|Z$.*

El mateix autor va caracteritzar la independència condicional com la factorització de la densitat conjunta $f_{X,Y,Z}(x,y,z)$ en dos factors, un dels quals no depèn de x i l'altre no depèn de y, és a dir:

$$f_{X,Y,Z}(x,y,z) = l(x,z)m(y,z)$$

Com a propietats també tenim que:

- $X \perp Y|Z \iff f_{X|ZY}(x;z,y) = f_{X|Z}(x;z) \forall x,y,z$
- $X \perp Y|Z \iff f_{XYZ}(x,y,z) = \dfrac{f_{XZ}(x,z)f_{YZ}(y,z)}{f_Z(z)} \forall x,y,z$
- $X \perp (Y_1,Y_2)|Z \Rightarrow X \perp Y_1|Z, X \perp Y_2|Z$

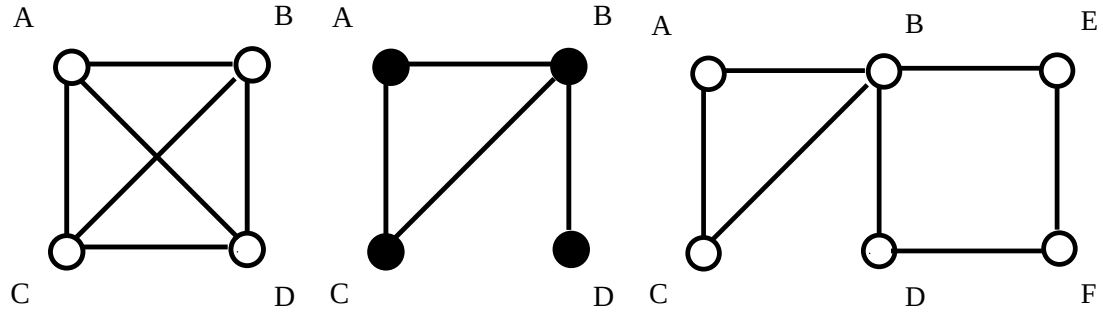
3.1.2 Teoria de grafs

Abans d'entrar en els gràfics d'independència d'un model, que veurem a la següent secció i clau de volta del modelatge gràfic, introduirem alguns conceptes de teoria de grafs que usarem posteriorment.

Un **graf** $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ és una estructura consistent en un conjunt finit $\mathcal{V}$ de *vèrtexs* i un conjunt finit $\mathcal{A}$ de línies que els uneixen anomenades *arestes* o *arcs*. Parlarem d'arestes i *grafs simètrics*, si les línies no estan orientades, i arcs i *grafs dirigits* si estan orientades. Els vèrtexs els designarem usualment amb lletres majúscules, que representaran posteriorment les nostres variables, $A, B, \ldots$ i les arestes com [AB] o [BA] equivalentment.

Considerarem a partir d'ara sempre grafs simètrics i *simples*, és a dir, on les arestes no tenen sentit i on cada parell de vèrtexs pot ser unit o no per una única aresta i una aresta no pot començar i acabar en el mateix vèrtex.

Representarem els grafs en diagrames com els següents:



Graf 1 Graf 2 Graf 3

Els vèrtexs els representarem per punts o cercles; punts per representar variables discretes i cercles per a les variables contínues, seguint la notació que empren diversos autors i d'entre ells Edwards [Edw95].

Direm que dos vèrtexs són **adjacents** si hi ha una aresta entre ells. Ho representarem per $X \sim Y$. En el Graf 1 el vèrtexs A i D són adjacents, en el Graf 2 A i D no són adjacents.

Direm que un graf és **complet** si hi ha una aresta entre cada parell de vèrtexs. El Graf 1 és complet, els Grafs 2 i 3 no són complets.

Qualsevol subconjunt de vèrtexs $u \subset \mathcal{V}$ ens indueix un **subgraf** $\mathcal{G}_u = (u, \mathcal{F})$ format pel conjunt de vèrtexs u i on $\mathcal{F}$ són totes les arestes d'$\mathcal{A}$ que tenen com a vèrtex inicial i final vèrtexs de u. Un subconjunt $u \subset \mathcal{V}$ és anomenat *complet* si indueix un subgraf complet. Per exemple el subconjunt A,B,C en el Graf 2 és complet.

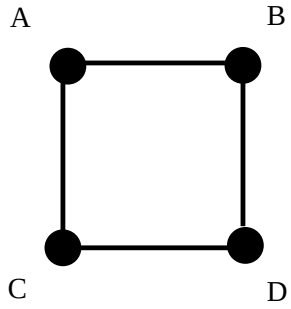
Un subconjunt $u \subset \mathcal{V}$ és anomenat una **clica** si és maximalment complet, és a dir,

u és complet i si $u \subset w$ llavors w no és complet. Les cliques en el graf 2 són {A,B,C} i {B,D}. El concepte de clica serà important en els models gràfics.

Una seqüència de vèrtexs $X_0, X_1, \ldots, X_n$ tals que $X_i \sim X_{i+1}$ per $i = 0, \ldots, n-1$ s'anomena un **camí** entre X_0 i X_n de longitud n. Per exemple, en el Graf 3, A, C, B, D, F és una camí de longitud 4 entre A i F. Un graf es diu **connex** si hi ha un camí entre qualsevol parell de vèrtexs.

Un camí amb inici i final en el mateix vèrtex, $X_1, X_2, \ldots, X_n, X_1$ és anomenat un **cicle de longitud n** o **n-cicle**. Si els n vèrtexs $X_1, X_2, \ldots, X_n$ d'un n-cicle $X_1, X_2, \ldots, X_n, X_1$ són diferents i cada vèrtex $X_i \sim X_k$ si i només si $k = i+1$ o $k = i-1$ o bé $i = 1$ i $k = n$, llavors s'anomena **cicle sense cordes**.

Com a exemple, en el Graf 4 tenim un 4-cicle sense cordes:



Graf 4

Direm que un graf **triangula** si no té cicles de longitud més gran o igual a quatre. La propietat de triangulació té relació amb l'existència d'estimacions màxim versemblants, que seran importants en la determinació de la deviància.

Per a qualssevol tres subconjunts a, b i c de $\mathcal{V}$, direm que c **separa** a i b si tots els camins d'a a b intersecccionen c, és a dir com a mínim tenen un vèrtex de c. Per exemple, en el Graf 2, {B} separa {A,C} i {D}.

Finalment definirem la **frontera** d'un subconjunt $u \subset \mathcal{V}$ que és definida com el conjunt de tots els vèrtexs de $\mathcal{V} - u$ que són adjacents als vèrtexs de u.

3.1.3 Independència condicional i graf d'independència condicional

Sigui $X = X_1, \ldots, X_n$ un conjunt de variables aleatòries, $K = \{1, \ldots, n\}$ llavors el **graf d'independència condicional de X** és el graf $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ on per tot $i, j \in K$ diferents tenim:

$$(i,j) \notin \mathcal{A} \iff X_i \perp X_j | X_{K-\{i,j\}}$$

és a dir, si dues variables A i B són condicionalment independents donades tota la resta de variables, A i B no seran vèrtexs adjacents en el graf $\mathcal{G}$.

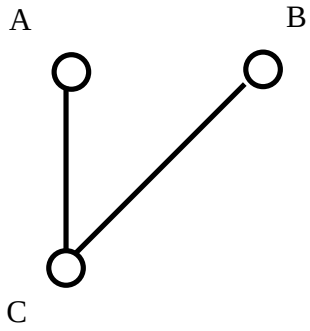
3.1.4 Exemples

Anem a introduir a continuació alguns exemples sense distingir les variables contínues de les discretes:

Si $K = 2$ i si $X_1 = A$ i $X_2 = B$ són independents $f(x) = f_1(x_1)f_2(x_2)$, llavors $A \perp B$ i el graf d'independència condicional és:

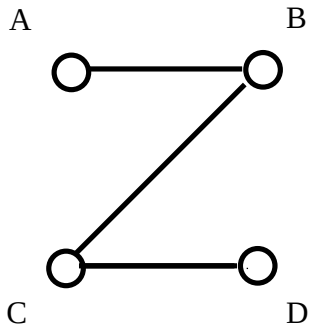
Graf 5

Si $K = 3$, $X_1 = A$, $X_2 = B$ i $X_3 = C$, i si A i B són independents condicionalment a C, és a dir, $f(x) = f_{13}(x_1, x_3)f_{23}(x_2, x_3)/f_3(x_3)$, $A \perp B|C$, llavors el graf d'independència condicional és:



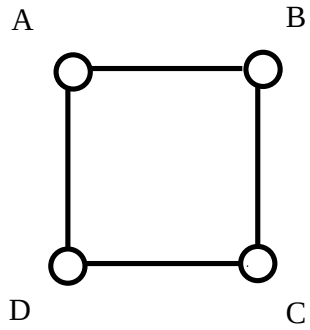
Graf 6

Si $K = 4$ i si $A \perp C|(B, D)$, $A \perp D|(B, C)$ i $B \perp D|(A, C)$ llavors el graf d'independència condicional és:



Graf 7

Si $K = 4$ i si $A \perp C|(B, D)$ i $B \perp D|(A, C)$ llavors el graf d'independència condicional és:



Graf 8

3.1.5 Teorema de separació i propietats de Markov

Sigui $X = X_1, \ldots, X_n$ un conjunt de variables aleatòries, $K = \{1, \ldots, n\}$. Siguin a, b i c subconjunts disjunts dels vèrtexs. Llavors si a separa b i c en el graf d'independència condicional, aleshores $X_b \perp X_c|X_a$:

La demostració d'aquest teorema es basa en les propietats de reducció i independència per blocs, podent trobar la demostració completa a [Whi90a].

Les propietats de Markov ens donen l'equivalència entre la definició de graf d'independència condicional i altres dues propietats equivalents (la demostració de l'equivalència d'aquestes tres propietats la podem trobar a l'article de Duran-Rúbies [DR99]).

- Propietat **dos a dos**: $\forall i, j$ no adjacents, $X_i \perp X_j|X_a$ on $a = K - \{i, j\}$
- Propietat **global**: $\forall a, b, c$ subconjunts disjunts de vèrtexs i b i c separats per a, llavors $X_b \perp X_c|X_a$
- Propietat **local**: $\forall i$ i a =frontera(i) i b els restants vèrtexs, tenim que $X_i \perp X_b|X_a$

Vegem en un exemple aquestes equivalències.

Sigui el graf d'independència condicional de la Figura 3.1. Per la propietat dos a dos tenim:

- $X_a \perp X_d|(X_b, X_c, X_e)$
- $X_a \perp X_e|(X_b, X_c, X_d)$
- $X_b \perp X_d|(X_a, X_c, X_e)$
- $X_b \perp X_e|(X_a, X_c, X_d)$

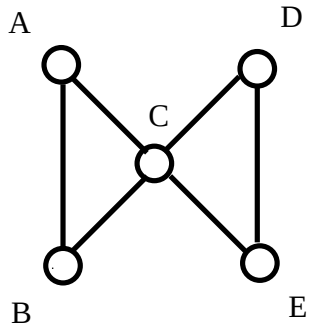


Figura 3.1: Graf d'independència condicional de les propietats de Markow

Per la propietat global tenim:

- $X_a \perp X_d | X_c$
- $X_a \perp X_e | X_c$
- $X_b \perp X_d | X_c$
- $X_b \perp X_e | X_c$

és a dir $(X_a, X_b) \perp (X_d, X_e) | X_c$

Per la propietat local tenim:

- $X_a \perp (X_d, X_e) | (X_b, X_c)$
- $X_b \perp (X_d, X_e) | (X_a, X_c)$
- $X_d \perp (X_a, X_b) | (X_c, X_e)$
- $X_e \perp (X_a, X_b) | (X_c, X_d)$

La qual cosa ens permet concloure que per la propietat global tenim dos grups de variables (a, b, c) per un costat i (c, d, e) per l'altre i així per al coneixement de la variable a, només ens caldria el coneixement de les variables b i c, utilitzant les propietats locals.

Com a exemple d'aplicació ho podem aplicar a les dades extretes de Mardia, Kent i Bibby [MKB79] on les dades són qualificacions de 88 estudiants en diferents matèries: mecànica(V), vectors(Y), àlgebra(X), anàlisi(Y) i estadística(Z). Totes són mesurades en la mateixa escala (0-100).

La Figura 3.2 ens mostra que les qualificacions per anàlisi i estadística són condicionalment independents de la mecànica i vectors, donada l'àlgebra. Una implicació del model serà que per predir la qualificació d'estadística, les qualificacions de l'àlgebra i l'anàlisi són suficients.

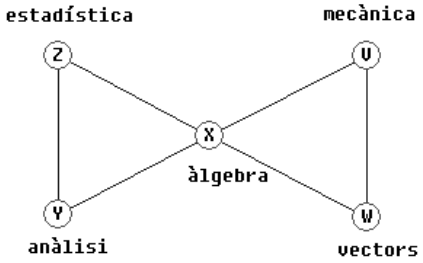


Figura 3.2: Gràfic de qualificacions en 5 matèries

3.2 Models loglineals i grafs

En aquesta secció donarem entrada als models loglineals com a eina útil per a la modelització de taules de contingència. I estendrem el concepte per a més de dues variables categòriques. Un cop obtinguts aquests models els emprarem per a generar models amb les relacions volgudes i sobre aquests aplicar els conceptes d'independència condicional i de modelatge gràfic.

3.2.1 Models loglineals

Considerem una taula de contingència de dimensions 2x2, en la qual denotem per x_{ij} la freqüència observada en la cel·la (i,j) i per x_{i+} i x_{+j} els totals per la i-èsima filera i per la i-èsima columna respectivament, i finalment per N el gran total de la taula.

En una taula de contingència, els valors esperats per a cadascuna de les cel·les, sota la hipòtesi d'independència vénen donats per l'expressió:

$$\hat{m}_{ij} = \frac{x_{i+}x_{+j}}{N}, \quad i = 1, 2, \quad j = 1, 2$$

Prenent logaritmes a ambdós costats de l'expressió,

$$\log \hat{m}_{ij} = \log x_{i+} + \log x_{+j} - \log N$$

i pensant en termes d'una taula AxB amb $\sharp A$ fileres i $\sharp B$ columnes, s'observa una semblança en relació a la notació de l'anàlisi de la variància. Llavors, la forma additiva suggereix que el paràmetre m_{ij} pot ser expressat de la forma:

$$\log m_{ij} = u + u_i^A + u_j^B,$$

o anàlogament:

$$m_{ij} = e^{u+u_i^A+u_j^B}$$

on u és la mitjana total dels logaritmes de les freqüències esperades,

$$u = \frac{1}{\sharp A \sharp B} \sum_{i=1}^{\sharp A} \sum_{j=1}^{\sharp B} \log m_{ij},$$

$u + u_i^A$ és la mitjana dels logaritmes de les freqüències esperades en les cel·les B al nivell i de la primera variable,

$$u + u_i^A = \frac{1}{\sharp B} \sum_{j=1}^{\sharp B} \log m_{ij},$$

i similarment,

$$u + u_j^B = \frac{1}{\sharp A} \sum_{i=1}^{\sharp A} \log m_{ij},$$

Com u_i^A i u_j^B representen desviacions respecte la mitjana total u

$$\sum_i u_i^A = \sum_j u_j^B = 0 \tag{3.1}$$

Si nosaltres penséssim en la possibilitat d'existència d'interaccions, podríem afegir un terme d'interacció al model d'independència, obtenint el conegut com a model *saturat*

$$\log m_{ij} = u + u_i^A + u_j^B + u_{ij}^{AB},$$

on tindríem en addició a (3.1)

$$\sum_{i=1}^{\sharp A} u_{ij}^{AB} = \sum_{j=1}^{\sharp B} u_{ij}^{AB} = 0 \tag{3.2}$$

3.2.2 Models per a taules tridimensionals

Considerem ara una taula de dimensions 2 x 2 x 2. Per a poder modelar aquesta taula introduïm la següent notació: sigui x_{ijk} la observació en la i-èsima filera, j-èsima columna i k-èsima capa de la taula i sigui m_{ijk} el corresponent valor esperat sota un determinat model. Tindrem, en referència a la notació de les marginals,

$$x_{ij+} = x_{ij1} + x_{ij2} \quad i, j = 1, 2$$

$$x_{i++} = x_{i11} + x_{i12} + x_{i21} + x_{i22} = x_{i1+} + x_{i2+} \quad i = 1, 2$$

i

$$x_{+++} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} x_{ijk} = \sum_{i=1}^{2} x_{i++}$$

I similarment, $m_{ij+} = m_{ij1} + m_{ij2}$, etc.

Si les tres variables corresponents a la nostra taula són independents llavors, i per analogia amb el model d'independència amb dues variables, les estimacions dels valors esperats per a cadascuna de les cel·les, vénen donats per l'expressió:

$$\hat{m}_{ijk} = \left(\frac{x_{i++}}{N}\right)\left(\frac{x_{+j+}}{N}\right)\left(\frac{x_{++k}}{N}\right)N$$

Prenent logaritmes

$$\log \hat{m}_{ijk} = \log x_{i++} + \log x_{+j+} + \log x_{++k} - 2\log N$$

Aquesta forma additiva dels logaritmes de l'estimació dels valors esperats és una altra vegada relacionable amb la notació inherent a l'anàlisi de la variància. Si anomenem en relació al paràmetre m_{ijk} ,

$$u = \frac{1}{2^N}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\log m_{ijk}$$

i

$$u_i^A = \frac{1}{2^{N-1}}\sum_{j=1}^{2}\sum_{k=1}^{2}\log m_{ijk} - u$$

etc, llavors podem escriure $\log m_{ijk}$ segons la notació ANOVA

$$\log m_{ijk} = u + u_i^A + u_j^B + u_k^C$$

on

$$\sum_{i=1}^{2} u_i^A = \sum_{j=1}^{2} u_j^B = \sum_{k=1}^{2} u_k^C = 0$$

ja que els termes representen desviacions respecte la gran mitjana u. Si suposem que les tres variables no són independents, sinó que tenim qualsevol altra tipologia de relació, independència condicional, independència dos a dos, totes aquestes relacions es poden modelar segons el *model loglineal generalitzat*

$$\log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC} \tag{3.3}$$

on

$$\sum_i u_i^A = \sum_j u_j^B = \sum_k u_k^C = 0$$

$$\sum_i u_{ij}^{AB} = \sum_j u_{ij}^{AB} = \sum_i u_{ik}^{AC} = \sum_k u_{ik}^{AC} = \tag{3.4}$$
$$= \sum_j u_{jk}^{BC} = \sum_k u_{jk}^{BC} = 0$$
$$\sum_i u_{ijk}^{ABC} = \sum_j u_{ijk}^{ABC} = \sum_k u_{ijk}^{ABC} = 0$$

Aquest model general no imposa cap restricció en els m_{ijk} i no queda restringit a taules 2 x 2 x 2, sinó a qualsevol taula $\sharp A$ x $\sharp B$ x $\sharp C$.

Ja hem vist que imposant

$$u_{ij}^{AB} = u_{ik}^{AC} = u_{jk}^{BC} = u_{ijk}^{ABC} = 0$$

per cada i, j, k en l'expressió (3.3) ens condueix al model d'independència completa de les tres variables.

Si ara prenem el model loglineal generalitzat (3.3) i imposem que

$$u_{ij}^{AB} = u_{ijk}^{ABC} = 0 \tag{3.5}$$

per cada i, j, k llavors

$$m_{+jk} = e^{u+u_j^B+u_k^C+u_{jk}^{BC}} \sum_i e^{u_i^A+u_{ik}^{AC}} \tag{3.6}$$

$$m_{i+k} = e^{u+u_i^A+u_k^C+u_{ik}^{AC}} \sum_j e^{u_j^B+u_{jk}^{BC}} \tag{3.7}$$

$$m_{++k} = e^{u+u_k^C} \sum_{i,j} e^{u_i^A+u_j^B+u_{ik}^{AC}+u_{jk}^{BC}} \tag{3.8}$$

Dividint el producte de les expressions (3.6) i (3.7) per (3.8) ens porta a:

$$m_{ijk} = \frac{m_{i+k}m_{+jk}}{m_{++k}} \tag{3.9}$$

on obtenim que aquest model implica que les variables 1 i 2 són independents fixat cadascun dels valors de la variable 3, és a dir, per cada valor fixat de k, tenim independència en la corresponent subtaula A x B.

Si imposem per a qualsevol i, j, k

$$u_{ik}^{AC} = u_{jk}^{BC} = u_{ijk}^{ABC} = 0 \tag{3.10}$$

ens porta al model

$$m_{ijk} = \frac{m_{ij+}m_{++k}}{m_{+++}} \tag{3.11}$$

On aquest model implica que les variables 1 i 2 preses juntament són independents de la variable 3.

Finalment un altre model, referit usualment com el *model sense interaccions de segon ordre*, el podríem obtenir imposant per a tot i, j, k:

$$u_{ijk}^{ABC} = 0 \tag{3.12}$$

No es consideraran tots els models possibles, sinó que només es consideraran **models jeràrquics**, on termes d'ordre superior només hi seran inclosos si, al mateix temps, hi són inclosos els termes d'ordre inferior.

És a dir, u_{123} només podrà ser inclòs en el model si també hi són inclosos u_{12}, u_{13} i u_{23}. La principal raó és d'ordre interpretatiu, ja que els corresponents termes d'ordre superior mesuren sempre desviacions de termes d'ordre inferior.

Com a conclusió, per a obtenir un model taula tridimensional amb un determinat model partirem del model loglineal generalitzat

$$\log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC} \tag{3.13}$$

amb les relacions entre els paràmetres $u_{(---)}$ corresponents (3.4) i sobre el qual imposarem les restriccions adients per a obtenir un model o altre (exemples (3.5), (3.10) i (3.12)).

3.2.3 Models loglineals d'ordre superior

Com a generalització dels casos anteriors de taules bi i tridimensionals, podrem construir un model loglineal generalitzat per a qualsevol ordre segons la fórmula:

$$\begin{aligned}\log m_{ijk\ldots n} = u + u_i^A + u_j^B + u_k^C + \ldots + u_n^Z + u_{ij}^{AB} + u_{ik}^{AC} + \ldots + u_{in}^{AZ} + \\ + u_{jk}^{BC} + \ldots + u_{jn}^{BZ} + \ldots + u_{ijk}^{ABC} + \ldots + u_{ijn}^{ABZ} + \ldots\end{aligned} \tag{3.14}$$

on s'hauran d'acomplir les relacions:

$$\begin{aligned}\sum_i u_i^A = \sum_j u_j^B = \sum_k u_k^C = \ldots = \sum_n u_n^Z = 0 \\ \sum_i u_{ij}^{AB} = \sum_j u_{ij}^{AB} = \sum_i u_{ik}^{AC} = \sum_k u_{ik}^{AC} = \\ = \sum_j u_{jk}^{BC} = \sum_k u_{jk}^{BC} = \ldots = \sum_k u_{kn}^{CZ} = \sum_n u_{kn}^{CZ} = \ldots = 0 \\ \sum_i u_{ijk}^{ABC} = \sum_j u_{ijk}^{ABC} = \sum_k u_{ijk}^{ABC} = \ldots = \sum_n u_{jkn}^{BCZ} = \ldots = 0\end{aligned} \tag{3.15}$$

3.2.4 Tipus de restriccions. Models amb restricció corner zero.

Hem vist en la secció anterior que imposàvem que els paràmetres del model tinguessin suma igual a zero (3.15). Hi ha, però, un segon tipus de modelització que no imposa la restricció sobre la suma dels paràmetres, sinó que imposa que el darrer dels paràmetres de cada filera o columna sigui zero (**còrner zero**), per perdre així també un grau de llibertat.

Així, per exemple, el model amb dues variables i dues categories amb paràmetres:

$$\log m_{ij} = u + u_i^A + u_j^B + u_{ij}^{AB}$$

amb restriccions:

$$\sum_i^2 u_i^A = \sum_j^2 u_j^B = 0 \qquad \sum_{i=1}^2 u_{ij}^{AB} = \sum_{j=1}^2 u_{ij}^{AB} = 0$$

es podria reescriure com:

$$\log m_{ij} = k + \mu_i^A + \eta_j^B + \delta_{ij}^{AB}$$

amb paràmetres:

$$\mu_i^A = \begin{pmatrix} \mu_1^A \\ 0 \end{pmatrix} \qquad \eta_j^B = \begin{pmatrix} \eta_1^B \\ 0 \end{pmatrix} \qquad \delta_{ij}^{AB} = \begin{pmatrix} \delta_{11}^{AB} & 0 \\ 0 & 0 \end{pmatrix}$$

Es pot trobar, en aquest cas, el pas d'una parametrització a l'altra igualant les dues expressions dels logaritmes de les freqüències:

$$\begin{pmatrix} u + u_1^A + u_1^B + u_{11}^{AB} & u + u_1^A + u_2^B + u_{12}^{AB} \\ u + u_2^A + u_1^B + u_{21}^{AB} & u + u_2^A + u_2^B + u_{22}^{AB} \end{pmatrix} = \begin{pmatrix} k + \mu_1^A + \eta_1^B + \delta_{11}^{AB} & k + \mu_1^A + \eta_2^B + \delta_{12}^{AB} \\ k + \mu_2^A + \eta_1^B + \delta_{21}^{AB} & k + \mu_2^A + \eta_2^B + \delta_{22}^{AB} \end{pmatrix}$$

si imposem les restriccions en els dos conjunts de paràmetres, obtenim la igualtat:

$$\begin{pmatrix} u + u_1^A + u_1^B + u_{11}^{AB} \\ u + u_1^A - u_1^B - u_{11}^{AB} \\ u - u_1^A + u_1^B + u_{11}^{AB} \\ u - u_1^A - u_1^B - u_{11}^{AB} \end{pmatrix} = \begin{pmatrix} k + \mu_1^A + \eta_1^B + \delta_{11}^{AB} \\ k + \mu_1^A \\ k + \mu_2^A \\ k \end{pmatrix}$$

que pot ser reescrita com:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} u \\ u_1^A \\ u_1^B \\ u_{11}^{AB} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} k \\ \mu_1^A \\ \eta_1^B \\ \delta_{11}^{AB} \end{pmatrix}$$

i d'aquí podem obtenir les matrius de canvi de paràmetres:

$$\begin{pmatrix} u \\ u_1^A \\ u_1^B \\ u_{11}^{AB} \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.25 \\ 1 & 0.5 & 0 & 0.25 \\ 1 & 0 & 0.5 & 0.25 \\ 1 & 0 & 0 & 0.25 \end{pmatrix} \begin{pmatrix} k \\ \mu_1^A \\ \eta_1^B \\ \delta_{11}^{AB} \end{pmatrix}$$

i anàlogament

$$\begin{pmatrix} k \\ \mu_1^A \\ \eta_1^B \\ \delta_{11}^{AB} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 0 & 2 & 0 & -2 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} u \\ u_1^A \\ u_1^B \\ u_{11}^{AB} \end{pmatrix}$$

Si tenim un model amb dues variables amb tres categories cadascuna:

$$\log m_{ij} = u + u_i^A + u_j^B + u_{ij}^{AB}$$

amb restriccions:

$$\sum_i^3 u_i^A = \sum_j^3 u_j^B = 0 \qquad \sum_{i=1}^3 u_{ij}^{AB} = \sum_{j=1}^3 u_{ij}^{AB} = 0$$

es podria reescriure com:

$$\log m_{ij} = k + \mu_i^A + \eta_j^B + \delta_{ij}^{AB}$$

amb paràmetres:

$$\mu_i^A = \begin{pmatrix} \mu_1^A \\ \mu_2^A \\ 0 \end{pmatrix} \qquad \eta_j^B = \begin{pmatrix} \eta_1^B \\ \eta_2^B \\ 0 \end{pmatrix} \qquad \delta_{ij}^{AB} = \begin{pmatrix} \delta_{11}^{AB} & \delta_{12}^{AB} & 0 \\ \delta_{21}^{AB} & \delta_{22}^{AB} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Podem obtenir la igualtat d'expressions dels logaritmes de les freqüències:

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} u \\ u_1^A \\ u_2^A \\ u_1^B \\ u_2^B \\ u_{11}^{AB} \\ u_{12}^{AB} \\ u_{21}^{AB} \\ u_{22}^{AB} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} k \\ \mu_1^A \\ \mu_2^A \\ \eta_1^B \\ \eta_2^B \\ \delta_{11}^{AB} \\ \delta_{12}^{AB} \\ \delta_{21}^{AB} \\ \delta_{22}^{AB} \end{pmatrix}$$

I d'aquí trobar les matrius de canvi de paràmetres:

$$\begin{pmatrix} k \\ \mu_1^A \\ \mu_2^A \\ \eta_1^B \\ \eta_2^B \\ \delta_{11}^{AB} \\ \delta_{12}^{AB} \\ \delta_{21}^{AB} \\ \delta_{22}^{AB} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 1 & 0 & 0 & -2 & -2 & -1 & -1 \\ 0 & 1 & 2 & 0 & 0 & -1 & -1 & -2 & -2 \\ 0 & 0 & 0 & 2 & 1 & -2 & -1 & -2 & -1 \\ 0 & 0 & 0 & 1 & 2 & -1 & -2 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 4 & 2 & 2 & 1 \\ 0 & 0 & 0 & 1 & 0 & 2 & 4 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 1 & 4 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 4 \end{pmatrix} \begin{pmatrix} u \\ u_1^A \\ u_2^A \\ u_1^B \\ u_2^B \\ u_{11}^{AB} \\ u_{12}^{AB} \\ u_{21}^{AB} \\ u_{22}^{AB} \end{pmatrix}$$

i anàlogament podem obtenir la seva inversa.

Aquestes dues parametritzacions dels models loglineals es poden fer per a qualsevol dimensió de variables i de categories, són equivalents, i es poden obtenir les matrius de canvi de paràmetres per als dos models sigui quina sigui la dimensió, amb la problemàtica de l'increment de la dimensió dels paràmetres del model.

3.2.5 Exemples de generació de models

En aquesta secció realitzarem uns quants exemples de generació de models loglineals. Per a això, només haurem de definir un conjunt de paràmetres adequat a l'ordre del model que vulguem.

Per exemple, per tal de generar un model per a dues variables A i B, amb tres i dues modalitats respectivament prendrem l'equació del model

$$\log m_{ij} = u + u_i^A + u_j^B + u_{ij}^{AB} \tag{3.16}$$

per a $i = 1, \cdots, 3$ i $j = 1, 2$ i fixarem els termes d'interacció u segons les relacions volgudes. Així si volem que no hi hagi relació entre A i B, independència, tots els termes u_{ij}^{AB} hauran de romandre iguals a 0, i tenint en compte sempre les restriccions imposades a (3.1) podrem imposar els valors dels altres termes u.

Un exemple de paràmetres podrien ser:

$u_1^A = 0.3$, $u_2^A = 0.6$, $u_3^A = -0.9$, $u_1^B = 0.4$, $u_2^B = -0.4$ i $u = 2$

on, per exemple, la casella (1,1) valdria $\log m_{1,1} = u + u_1^A + u_1^B + u_{1,1}^{AB} = 2 + 0.3 + 0.4 = 2.7$ i per tant l'efectiu seria

$m_{ij} = e^{2.7} = 14.9$

això ens donarà, un cop aproximats els efectius a l'enter més proper, la següent taula de contingència, on és fàcilment comprovable la hipòtesi d'independència de les dues variables, tal com ha estat imposat pel model:

15	20	5
7	9	2

Per exemple, per tal de generar un model per a quatre variables, amb tres modalitats cadascuna, prendrem l'equació del model

$$\log m_{ijkl} = u + u_i^A + u_j^B + u_k^C + u_l^D + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{AD} + \ldots \tag{3.17}$$

i fixarem els termes d'interacció u segons les relacions volgudes. Així si volem que no hi hagi relació entre un parella de variables X i Y, tots els termes u_{xy}^{XY} hauran de romandre iguals a 0, i tenint en compte sempre les restriccions imposades a (3.15) podrem imposar els valors dels altres termes u.

Per exemple, si només volem que hi hagi relació entre els parells de variables AB, BC i BD segons el model:

$$\log m_{ijkl} = u + u_i^A + u_j^B + u_k^C + u_l^D + u_{ij}^{AB} + u_{ik}^{BC} + u_{jk}^{BD} \tag{3.18}$$

donarem valors 0 a tots els altres termes i a les interaccions d'ordre superior. I prenent com a terme independent $u = 2$, termes per a les variables idèntics i iguals a $(0.4, 0.4, -0.8)$ i termes d'interacció els valors $(0.3, 0.3, -0.6, -0.2, -0.2, 0.4, -0.1, -0.1, 0.2)$, obtenim 1244 valors que es detallen a la següent taula de Burt:

648	0	0	405	214	29	287	287	74	287	287	74
0	453	0	249	130	74	199	199	55	199	199	55
0	0	143	80	46	17	63	63	17	63	63	17
405	249	80	734	0	0	346	346	42	346	346	42
214	130	46	0	390	0	153	153	84	153	153	84
29	74	17	0	0	120	50	50	20	50	50	20
287	199	63	346	153	50	549	0	0	244	244	61
287	199	63	346	153	50	0	549	0	244	244	61
74	55	17	42	84	20	0	0	146	61	61	24
287	199	63	346	153	50	244	244	61	549	0	0
287	199	63	346	153	50	244	244	61	0	549	0
74	55	17	42	84	20	61	61	24	0	0	146

Per a veure l'equivalència de les dues modelitzacions presentades anteriorment, anem a generar dos models loglineals diferents i amb el paquet estadístic SPSS 11.5.1, que treballa amb les restriccions de corner zero, veurem els paràmetres estimats i la seva equivalència amb els paràmetres de suma zero usats.

Usem el model per a dues variables amb dues categories cadascuna $\log m_{ij} = u + u_i^A + u_j^B$, sense interaccions, amb paràmetre constant $u = 5$ i efectes principals iguals a $(1, -1)$ i $(-1, 1)$ respectivament.

Això ens dóna la taula:

148	1097
20	148

Els paràmetres equivalents en el model de corner zero són:

$$\begin{pmatrix} k = 5 \\ \mu_1^A = 2 \\ \eta_1^B = -2 \\ \delta_{11}^{AB} = 0 \end{pmatrix}$$

Els paràmetres obtinguts per estimació per l'SPSS són:

$$\begin{pmatrix} k = 4.9972 \\ \mu_1^A = 2.0031 \\ \eta_1^B = -2.0015 \\ \delta_{11}^{AB} = -0.0016 \end{pmatrix}$$

podem veure que no coincideixen plenament, això és degut a l'arrodoniment que es fa de les freqüències per tal d'obtenir nombres enters.

En un model més complicat, dues variables amb tres categories cadascuna, donat pel model d'efectes sumats nuls, on obviem els paràmetres redundants, i de paràmetres:

$$\begin{pmatrix} u = 4 \\ u_1^A = 0.4 \\ u_2^A = 0.4 \\ u_1^B = -0.1 \\ u_2^B = -0.2 \\ u_{11}^{AB} = 0.3 \\ u_{12}^{AB} = 0.0 \\ u_{21}^{AB} = -0.2 \\ u_{22}^{AB} = 0.3 \end{pmatrix}$$

ens dóna la taula:

99	67	81
60	90	99
20	15	49

Els paràmetres obtinguts per estimació amb SPSS:

$$\begin{pmatrix} k = 3.8918 \\ \mu_1^A = 0.5026 \\ \mu_2^A = 0.7033 \\ \eta_1^B = -0.8961 \\ \eta_2^B = -1.1838 \\ \delta_{11}^{AB} = 1.0968 \\ \delta_{12}^{AB} = 0.9940 \\ \delta_{21}^{AB} = 0.3953 \\ \delta_{22}^{AB} = 1.0885 \end{pmatrix}$$

Si, a partir d'aquests paràmetres estimats, obtenim els paràmetres del model de suma igual a zero aquests serien:

$$\begin{pmatrix} u = 3.9976 \\ u_1^A = 0.4004 \\ u_2^A = 0.3988 \\ u_1^B = -0.1026 \\ u_2^B = -0.1935 \\ u_{11}^{AB} = 0.2997 \\ u_{12}^{AB} = 0.0000 \\ u_{21}^{AB} = -0.1995 \\ u_{22}^{AB} = 0.2969 \end{pmatrix}$$

que són aproximadament els mateixos emprats per la generació del model sota l'efecte de l'arrodoniment.

3.2.6 Influència dels paràmetres en la generació de models

Quan generem models el que fem és obtenir freqüències de la taula de contingència mitjançant el model loglineal donat. Aquestes freqüències no són res més que la suma del terme constant, més els efectes principals, més les interaccions definides en el nostre model.

$$\log m_{ijk\ldots n} = u + u_i^A + u_j^B + u_k^C + \ldots + u_n^Z + u_{ij}^{AB} + u_{ik}^{AC} + \ldots$$

La qüestió que ens plantegem és quina influència tenen els termes d'interacció sobre els coeficients del terme constant i els efectes principals.

Si partim del model més senzill, amb dues variables i dues categories cadascuna, amb els paràmetres $(u, u_1^A, u_1^B, u_{11}^{AB})$, obtenim la següent taula (Taula 3.1) on donem les taules de contingència, l'estadístic de χ^2, amb 1 grau de llibertat, i el $p - valor$ associat.

Com que l'efecte del factor interacció és un efecte multiplicador, el fet de tenir una interacció igual a 0.1 ens comporta estar multiplicant per $e^{0.1} = 1.10517$, la qual cosa fa que per a termes constants grans el canvi sigui suficient per no poder suposar independència, és a dir l'existència d'un factor d'interacció apreciable. Aquest canvi no és tan gran per a termes constants petits, però veiem que una interacció de 0.2, equivalent al fet d'estar multiplicant per 1.22140 l'exponencial del terme constant, és suficient per rebutjar la independència amb un terme constant igual a 3.

Taula 3.1: Taules generades i paràmetres en els models loglineals

20 | 20 20 | 20 $(3,0,0,0)$ $\chi^2 = 0$ $p = 1$	55 | 55 55 | 55 $(4,0,0,0)$ $\chi^2 = 0$ $p = 1$	2981 | 2981 2981 | 2981 $(8,0,0,0)$ $\chi^2 = 0$ $p = 1$
22 | 18 18 | 22 $(3,0,0,0.1)$ $\chi^2 = 0.8$ $p = 0.371$	60 | 49 49 | 60 $(4,0,0,0.1)$ $\chi^2 = 2.22$ $p = 0.136$	3294 | 2697 2697 | 3294 $(8,0,0,0.1)$ $\chi^2 = 118.981$ $p = 0.000$
25 | 16 16 | 25 $(3,0,0,0.2)$ $\chi^2 = 3.951$ $p = 0.047$	67 | 45 45 | 67 $(4,0,0,0.2)$ $\chi^2 = 8.643$ $p = 0.003$	3641 | 2441 2441 | 3641 $(8,0,0,0.2)$ $\chi^2 = 473.528$ $p = 0.000$

Anem ara a fixar-nos en la importància dels valors dels efectes principals i, d'aquests i de les interaccions, per a un valor constant fix. Per això prendrem el valor fix $u = 3$ i variarem els valors del efectes principals i de les interaccions.

Podem veure a la Taula 3.2 que, com és de suposar, els efectes principals d'A i de B afecten a les fileres i columnes respectivament, però tal com veiem en els exemples de la primera filera, la introducció d'efectes principals no afecta a la independència de la taula, excepte petits desajustos produïts per l'arrodoniment de les freqüències.

Amb els exemples de la darrera filera podem veure que el factor signe només importa en relació al signe global dels factors, és a dir, si hi ha una permutació dels signes els resultats són els mateixos amb una permutació de caselles en la taula de freqüències. No importa on és el signe, sinó podríem dir la paritat del signes.

Una altra observació és que si els valors d'un efecte principal el parteixo entre els dos efectes principals, els efectius de la diagonal principal són els mateixos, variant els de la

Taula 3.2: Taules generades i paràmetres en els models loglineals (II)

20 | 20 20 | 20 $(3, 0, 0, 0)$ $\chi^2 = 0$ $p = 1$	33 | 33 12 | 12 $(3, 0.5, 0, 0)$ $\chi^2 = 0$ $p = 1$	33 | 12 33 | 12 $(3, 0, 0.5, 0)$ $\chi^2 = 0$ $p = 1$	55 | 20 20 | 7 $(3, 0.5, 0.5, 0)$ $\chi^2 = 0.006$ $p = 0.940$
22 | 18 18 | 22 $(3, 0, 0, 0.1)$ $\chi^2 = 0.800$ $p = 0.371$	37 | 30 11 | 13 $(3, 0.5, 0, 0.1)$ $\chi^2 = 0.625$ $p = 0.429$	37 | 11 30 | 13 $(3, 0, 0.5, 0.1)$ $\chi^2 = 0.625$ $p = 0.429$	60 | 18 18 | 8 $(3, 0.5, 0.5, 0.1)$ $\chi^2 = 0.615$ $p = 0.433$
25 | 16 16 | 25 $(3, 0, 0, 0.2)$ $\chi^2 = 3.951$ $p = 0.047$	40 | 27 10 | 15 $(3, 0.5, 0, 0.2)$ $\chi^2 = 2.848$ $p = 0.091$	40 | 10 27 | 15 $(3, 0, 0.5, 0.2)$ $\chi^2 = 2.848$ $p = 0.091$	67 | 16 16 | 9 $(3, 0.5, 0.5, 0.2)$ $\chi^2 = 3.020$ $p = 0.082$
110 | 27 10 | 5 $(3, 1, 0.5, 0.2)$ $\chi^2 = 1.510$ $p = 0.219$	181 | 16 16 | 3 $(3, 1, 1, 0.2)$ $\chi^2 = 1.270$ $p = 0.260$	37 | 18 18 | 13 $(3, 0.25, 0.25, 0.1)$ $\chi^2 = 0.729$ $p = 0.323$	40 | 16 16 | 15 $(3, 0.25, 0.25, 0.2)$ $\chi^2 = 3.416$ $p = 0.065$
25 | 45 6 | 25 $(3, 0.5, -0.5, 0.2)$ $\chi^2 = 2.703$ $p = 0.100$	16 | 67 9 | 16 $(3, 0.5, -0.5, -0.2)$ $\chi^2 = 3.020$ $p = 0.082$	45 | 25 25 | 6 $(3, 0.5, 0.5, -0.2)$ $\chi^2 = 2.703$ $p = 0.100$	6 | 25 25 | 45 (3,-0.5,-0.5,-0.2) $\chi^2 = 2.703$ $p = 0.100$

diagonal secundària.

A més podem tornar a observar que la presència d'efectes principals poden fer augmentar la influència de les interaccions en freqüències grans, ja que els efectius diagonals són influïts pels efectes principals.

Finalment, també hem de tenir present que aquests resultats poden ser afectats per l'arrodoniment a enters i això sol ser més important quan tenim efectius petits.

3.2.7 Models loglineals d'interacció i graf d'interacció

Donat un model loglineal generalitzat de qualsevol ordre, si en els termes de desenvolupament del model

$$\log m_{ijk\ldots} = u + u_i^A + u_j^B + u_k^C + \ldots + u_{ij}^{AB} + u_{ik}^{AC} + \ldots$$

hi ha $u_a = 0$ per alguns subconjunts d'índexs de K, llavors diem que és un **model loglineal d'interacció**. Als termes u se'ls anomena *interaccions*.

El **graf d'interaccions** d'un model loglineal d'interaccions és el graf no orientat $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, on per tot parell de vèrtexs i j existeix un terme en el desenvolupament del model tal que conté i i j.

Per exemple, el graf d'interacció del model donat a la Fórmula 3.18 seria:

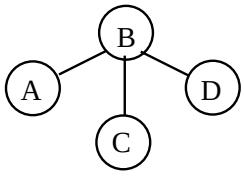


Figura 3.3: Graf d'interacció del model generat amb 4 variables

3.2.8 Models jeràrquics, gràfics i descomposables

Un model loglineal d'interacció és **jeràrquic** si per a qualssevol subconjunts a i b de K es té:

$$u_a = 0 \Rightarrow u_{a \cup b} = 0$$

És a dir, per poder-hi haver un ordre d'interacció superior hi ha d'haver el terme de grau inferior. Només considerarem sempre model jeràrquics.

Una **classe generatriu** d'un model jeràrquic és un conjunt C de subconjunts dels vèrtexs K on, dos a dos, un no és inclòs en l'altre i són interpretats com els conjunts maximals d'interaccions preses. És a dir, una classe generatriu caracteritza un model jeràrquic.

Un **model gràfic** és un model loglineal d'interacció on $u_a = 0$ és present en la descomposició en u-termes si i només si a és complet. Un model gràfic és doncs un model jeràrquic on la classe generatriu és el conjunt de les cliques del gràfic d'interaccions.

Un model és **descomposable** si és un model gràfic on el graf triangula, és a dir no tenim 4-cicles sense cordes.

Tenim la següent seqüència d'inclusions en els models:
{Model Descomposable} $\subset$ {Mod. Gràfic } $\subset$ {Mod. Jeràrquic} $\subset$ { Mod. Loglineal}

3.2.9 Exemples

Siguin els següents models loglineals:

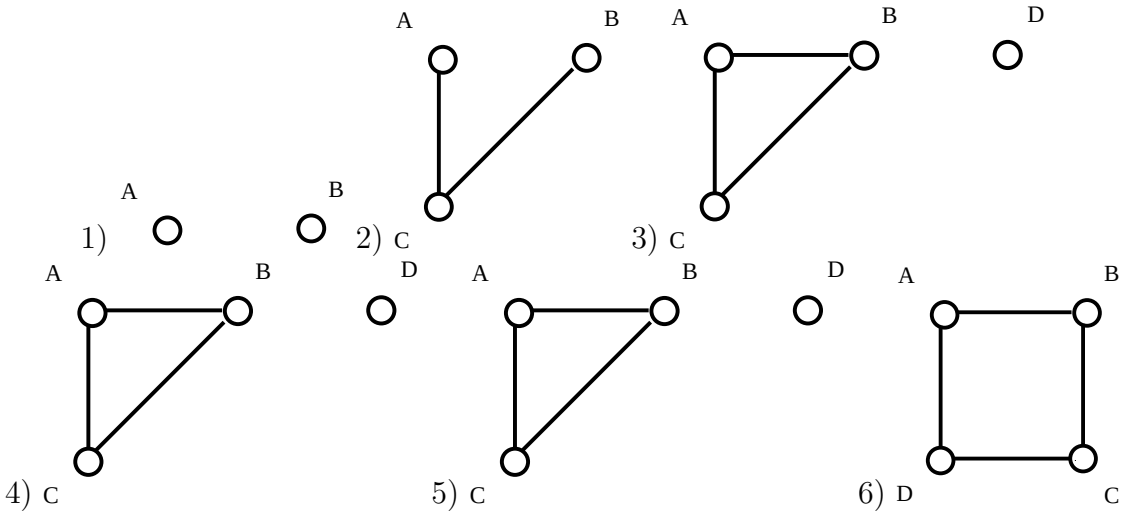
1. $\log m_{ij} = u + u_i^A + u_j^B$
2. $\log m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ik}^{AC} + u_{jk}^{BC}$
3. $\log m_{ijkl} = u + u_i^A + u_j^B + u_k^C + u_l^D + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC}$
4. $\log m_{ijkl} = u + u_i^A + u_j^B + u_k^C + u_l^D + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}$
5. $\log m_{ijkl} = u + u_l^D + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC}$
6. $\log m_{ijkl} = u + u_i^A + u_j^B + u_k^C + u_l^D + u_{ij}^{AB} + u_{il}^{AD} + u_{jk}^{BC} + u_{kl}^{CD}$

Els models loglineals d'interacció 1) 2) 3) 4) i 6) són jeràrquics, el 5) no ho és.

Les classes generatrius dels models jeràrquics són:

1) C={{A},{B}} 2) C={{A,C},{B,C}}
3) C={{A,B},{A,C},{B,C},{D}} 4) C={{A,B,C},{D}}
6) C={{A,B},{A,D},{B,C},{B,D}}

Els grafs d'interaccions d'aquest models són:



Els models jeràrquics són notats com:

1) A, B 2) AC, BC

3) AB, AC, BC, D 4) ABC, D

6) AB, AD, BC, CD

Els models jeràrquics 1) 2) 4) i 6) són models gràfics, el 3) és jeràrquic però no és gràfic, ja que {A,B,C} forma una clica del graf d'interaccions i $u^{ABC} = 0$.

Finalment, els models 1), 2) i 4) són descomposables. El model 6) no és descomposable, ja que té un 4-cicle sense corda, no triangula.

3.2.10 Independència condicional i interacció

Sigui el graf $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ i considerem el model loglineal de graf $\mathcal{G}$. Si $(i, j) \notin \mathcal{A}$, llavors tots els termes u_a tals que contenen $\{i, j\}$ són nuls. Llavors la propietat de factorització de la independència condicional permet deduir la propietat $X_i \perp X_j | X_{K-\{i,j\}}$ i com és una equivalència, el graf d'interaccions és un graf d'independència condicional.

3.3 Models discrets

En aquesta secció tractarem els models per a dades discretes. Des de la introducció dels models loglineals en els anys seixanta aquests han estat usats en una gran varietat de camps d'aplicació en anàlisi de dades discretes. Posteriorment, i deguda a Darroch, Lauritzen i Speed [DLS80] fou la connexió entre models loglineals i grafs i models gràfics. Les obres de Christensen [Chr90] i Whittaker [Whi90a] han contribuït a la seva difusió cobrint no només els models gràfics per a taules discretes sinó també per a models continus.

3.3.1 Deviància

Considerem una taula de contingència de tres variables discretes A, B i C, siguin $\sharp A$, $\sharp B$ i $\sharp C$ el nombre de modalitats de cadascuna d'aquestes variables: podem formar una taula de triple entrada creuant A, B i C, on la freqüència de cada cel·la és n_{ijk}. Podem escriure la freqüència esperada d'una cel·la com $m_{ijk} = Np_{ijk}$, on p_{ijk} és la probabilitat d'una cel·la.

Sota un model multinomial amb N observacions la versemblança d'una taula donada n_{ijk} és:

$$\mathcal{L}(\{p_{ijk}\}|\{n_{ijk}\}) = \frac{N!}{\prod_{ijk} n_{ijk}} \prod_{ijk} p_{ijk}^{n_{ijk}}$$

els valors que maximitzen aquesta expressió per un model donat són els estimadors màxim versemblants (EMVs) i els denotem per $\hat{p}_{ijk}$. Com el logaritme és una funció monòtona, els EMVs també maximitzen la funció de logversemblança:

$$\ell(\{p_{ijk}\}|\{n_{ijk}\}) = \ln\left(\frac{N!}{\prod_{ijk} n_{ijk}}\right) + \sum_{ijk} n_{ijk} \ln p_{ijk}$$

La **deviància** d'un model $\mathcal{M}_0$ és la raó de versemblança entre $\mathcal{M}_0$ i el model saturat, sense restriccions, $\mathcal{M}_f$, és a dir:

$$G^2 = 2(\hat{\ell}_f - \hat{\ell}_0)$$

on $\hat{\ell}_f$ i $\hat{\ell}_0$ són els maximitzats logversemblants sota les hipòtesis $\mathcal{M}_f$ i $\mathcal{M}_0$ respectivament.

La deviància té tres propietats que la fan ser una bona mesura del model $\mathcal{M}_0$, aquestes són:

1. Quan el model $\mathcal{M}_0 = \mathcal{M}_f$, la deviància és zero

2. Si la hipòtesi nul·la $\mathcal{M}_0$ és la hipòtesi certa, davant de la hipòtesi alternativa del model saturat $\mathcal{M}_f$ llavors la deviància es distribueix asimptòticament com una χ^2_k, on els graus de llibertat (k) són la diferència de parametres entre $\mathcal{M}_0$ i $\mathcal{M}_f$. D'aquí, $E(\text{deviància}) = k$ sota $\mathcal{M}_0$.

3. Si $\mathcal{M}_0$ no és certa, llavors la distribució de la deviància és significativament més gran que una χ^2_k, per tant, $E(\text{deviància}) > k$.

Sota $\mathcal{M}_f$, els EMVs de p_{ijk} són n_{ijk}/N per tant la deviància pot ser escrita com:

$$G^2 = 2\left\{\sum_{ijk} n_{ijk}\ln(n_{ijk}/N) - \sum_{ijk} n_{ijk}\ln(\hat{p}_{ijk})\right\} = 2\sum_{ijk} n_{ijk}\ln\left(\frac{n_{ijk}}{\hat{m}_{ijk}}\right) =$$

$$= 2\sum o\ln\frac{(o-e)^2}{e}$$

on $\hat{m}_{ijk} = N\hat{p}_{ijk}$ són les freqüències absolutes esperades sota $\mathcal{M}_0$.

Deviància, Informació i Entropia

El nombre d'informació de divergència de Kullback-Leibler és, en el cas discret:

$$I(f,g) = E\left(log\frac{f}{g}\right) = \sum f(i)log\frac{f(i)}{g(i)}$$

L'entropia de Shannon en el cas discret és:

$$-Elogf = -\sum f(i)logf(i)$$

I la relació entre aquests dos indicadors:

$$I(f,g) = Elog(\frac{f}{g})$$

Sota un determinat model $\mathcal{M}$ la Deviància és :

$$G^2 = 2\sum_i n_i ln\frac{n_i}{\hat{m}_i}$$

on $\hat{p}_i = \frac{\hat{m}_i}{N}$ és l'EMV de $p_i = \frac{n_i}{N}$ per tant :

$$G^2 = 2\sum_i n_i ln\frac{n_i}{N\hat{p}_i} = 2NI\left(\frac{n_i}{N},\hat{p}_i\right)$$

Per tant, la deviància és $2N$ vegades la Informacó de divergència de Kullback-Leibler entre els valors observats i els valors obtinguts sota el model $\mathcal{M}$

Deviància i χ^2

Sabem que sota un determinat model $\mathcal{M}$ el coeficient de χ^2 d'una taula de contingència és:

$$\chi^2 = \sum \frac{(o-e)^2}{e} = \sum_i \frac{(n_i - N\hat{p}_i)^2}{N\hat{p}_i} \tag{3.19}$$

Per altra banda la deviància és:

$$G^2 = 2NI\left(\frac{n_i}{N}, \hat{p}_i\right) = 2\sum_i n_i log\left(\frac{n_i}{N\hat{p}_i}\right) = 2\sum_i N\hat{p}_i \frac{n_i}{N\hat{p}_i} log\left(\frac{n_i}{N\hat{p}_i}\right) \tag{3.20}$$

Si considerem que quan $t \to 1$:

$$tlog(t) \simeq (t-1) + \frac{(t-1)^2}{2}$$

Llavors l'expressió (3.20) és aproximadament igual a:

$$2\sum_i \left(N\hat{p}_i(\frac{n_i}{N\hat{p}_i} - 1) + N\hat{p}_i \frac{(\frac{n_i}{N\hat{p}_i} - 1)^2}{2} \right) =$$

$$= 2\sum_i (n - N\hat{p}_i) + 2\sum_i \left(\frac{N\hat{p}_i}{2} (\frac{n_i - N\hat{p}_i}{N\hat{p}_i})^2 \right) =$$

$$= 0 + \sum_i (\frac{(n_i - N\hat{p}_i)^2}{N\hat{p}_i}) = \sum_i (\frac{(n_i - N\hat{p}_i)^2}{N\hat{p}_i})$$

Tenim doncs, que G^2 és aproximadament igual a l'expressió (3.19), i per tant la distribució de G^2 és aproximadament una χ^2 amb els mateixos graus de llibertat que a (3.19).

$$G^2 \simeq \chi^2 \tag{3.21}$$

Diferència de deviàncies

Si tenim dos models enniuats $\mathcal{M}_0 \subseteq \mathcal{M}_1$ llavors es defineix la diferència de deviàncies com:

$$d = G_0^2 - G_1^2 = 2\sum_i n_i \ln \frac{n_i}{\hat{m}_i^0} - 2\sum_i n_i \ln \frac{n_i}{\hat{m}_i^1} = 2\sum_i n_i ln \frac{\hat{m}_i^1}{\hat{m}_i^0}$$

Aquest estadístic té una distribució asimptòtica χ^2, amb graus de llibertat igual a la diferència de graus entre els dos models ([Rao73], pp. 418-420).

Goodman ([Goo69]) va notar que degut a la forma multiplicativa dels valors estimats per a models jeràrquics, l'anterior expressió és igual a:

$$2\sum_i \hat{m}_i^1 ln\frac{\hat{m}_i^1}{\hat{m}_i^0}$$

i, per tant, coincideix amb dues vegades la informació de divergència de Kullback-Leibler entre els models $\mathcal{M}_0$ i $\mathcal{M}_1$:

$$I(\hat{m}_i^1, \hat{m}_i^0) = I(\mathcal{M}_1, \mathcal{M}_0) = \frac{d}{2}$$

3.3.2 Deviància i selecció de models

Tal com hem vist la deviància ens pot servir per a seleccionar un model, però el mètode de màxima versemblança sempre dóna major suport al model amb més paràmetres. Això ha portat a buscar criteris d'informació que corregeixin aquest efecte.

- **El criteri AIC** Akaike [Aka69] va proposar un criteri alternatiu per a la selecció de models (**Aikake's Information Criterion - AIC**) on basant-se en arguments de la teoria de la informació, si el que volem són prediccions precises de $f(y)$ donat un model $f_M(y)$, arriba a que ha de minimitzar l'expressió:

  $$AIC = -2\hat{\ell}_M + 2p = G_M^2 + 2p$$

  on p és el nombre de paràmetres del model.

  El criteri és minimitzar la deviància del model, que disminuirà si introduïm més paràmetres, corregit amb el doble del nombre de paràmetres en el model.

- **El criteri BIC** Si ho plantegem des d'un enfocament bayesià, calculem les probabilitats a posteriori dels possibles models i triem el model de màxima probabilitat a posteriori (**Bayesian Information Criterion - BIC**). Aquesta maximització ens porta a una equivalència amb minimitzar l'expressió donada per Schwarz [Sch78], que pot ser escrita equivalentment com:

  $$BIC = -2\hat{\ell}_M + p\log n = G_M^2 + p\log n$$

  El criteri torna a ponderar la deviància amb l'increment del nombre de paràmetres, mirant de trobar el millor model a posteriori.

Es pot veure que ambdós criteris són expressions de la forma $G_M^2 + pg(n)$ on $g(n)$ és una funció del nombre de dades, que val 2 en el cas de l'AIC i $\log n$ en el BIC, per la qual cosa el criteri BIC triarà models amb menys paràmetres. Altres funcions $g(n)$ són possibles donant altres criteris. (per a més detalls vegeu [Peñ02], cap 11.)

Capítol 4

Estudi de les inèrcies en anàlisis de correspondències

4.1 Inèrcies en correspondències simples

Una de les finalitats de les anàlisis de correspondències és la reducció de la dimensionalitat de les dades sense una pèrdua important de la informació continguda en elles. Per això, aquesta mesura de la informació es tradueix en la mesura de la inèrcia del núvol de punts. Veurem, a continuació, el càlcul d'aquestes inèrcies, la seva relació amb el coeficient de contingència per a taules de doble entrada i les problemàtiques en les anomenades taules quadrades, fent un especial enfocament a les no simètriques..

4.1.1 Generalitats

Siguin dues variables categòriques, Q_1 i Q_2, amb I i J modalitats respectivament, i la taula de contingència producte del creuament de les dues variables. Sigui f_{ij} la freqüència relativa d'individus amb la modalitat $i \in I, j \in J$, tal com ja ha estat definida anteriorment.

Considerem la taula de perfils filera:

$$\left(\frac{f_{ij}}{f_{i.}}\right)_{j=1,\ldots,J} = \left(\frac{n_{ij}}{n_{i.}}\right)_{j=1,\ldots,J}$$

llavors la inèrcia del núvol de perfils filera serà:

$$\mathcal{I}_I = \sum(\mathcal{I}(i)) = \sum(m_i d^2(i, cdg) =$$

amb la mètrica definida a la secció (2.4.3.1) i donat que el centre de gravetat ve determinat per:

$$cdg = \sum_i f_{i.} \frac{f_{ij}}{f_{i.}} = f_{.j}$$

llavors

$$= \sum_i f_{i.} \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 = \sum_i f_{i.} \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij} - f_{.j} f_{i.}}{f_{i.}} \right)^2 =$$

$$= \sum_i \sum_j \frac{f_{i.}}{f_{.j}} \frac{(f_{ij} - f_{.j} f_{i.})^2}{f_{i.}^2} = \sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} \tag{4.1}$$

on l'expressió (4.1) ens determina la inèrcia aportada per cadascuna de les caselles de la taula de contingència. Aquesta expressió també es pot expressar com a diferència de quadrats:

$$= \sum_i \sum_j \left(\frac{f_{ij}}{\sqrt{f_{.j} f_{i.}}} - \sqrt{f_{.j} f_{i.}} \right)^2$$

desenvolupant el quadrat, com:

$$\sum_i \sum_j \left(\frac{f_{ij}^2}{f_{i.} f_{.j}} + f_{.j} f_{i.} - 2 f_{ij} \right) = \sum_i \sum_j \frac{f_{ij}^2}{f_{i.} f_{.j}} - 1$$

o bé en funció dels efectius absoluts:

$$\mathcal{I} = \sum_i \sum_j \frac{(f_{ij} - f_{.j} f_{i.})^2}{f_{i.} f_{.j}} = \sum_i \sum_j \frac{\left(\frac{n_{ij}}{n} - \frac{n_{i.}}{n} \frac{n_{.j}}{n} \right)^2}{\frac{n_{i.}}{n} \frac{n_{.j}}{n}} = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{n_{i.} n_{.j}} \tag{4.2}$$

4.1.2 Inèrcia, coeficient de contingència χ^2 i deviància

Donada una taula de contingència es pot expressar el seu coeficient de contingència, χ^2 mesura de la independència de les dues variables, com:

$$\chi^2 = \sum \frac{(o-e)^2}{e} = \sum_i \frac{(n_i - N\hat{p}_i)^2}{N\hat{p}_i} = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n_{..}} \right)^2}{\frac{n_{i.} n_{.j}}{n_{..}}} \tag{4.3}$$

i es comprova que aquesta expressió és n vegades la inèrcia total del núvol de punts, directament a partir de la fórmula (4.2):

$$\chi^2 = n\mathcal{I}$$

A més, tal com hem vist a la secció 3.3.1, la deviància sota un determinat model $\mathcal{M}$ és:

$$G^2 = 2NI\left(\frac{n_i}{N}, \hat{p}_i\right) = 2\sum_i n_i log\left(\frac{n_i}{N\hat{p}_i}\right) = 2\sum_i N\hat{p}_i\frac{n_i}{N\hat{p}_i}log\left(\frac{n_i}{N\hat{p}_i}\right) \simeq \chi^2 \qquad (4.4)$$

per tant, com que són aproximadament iguals, tenen distribucions aproximades χ^2 amb els mateixos graus de llibertat

I com que donada una taula de contingència, el seu coeficient de contingència és n vegades la inèrcia total del núvol de punts, $\chi^2 = n\mathcal{I}$, la inèrcia i la deviància segueixen aproximadament la mateixa distribució χ^2 amb els mateixos graus de llibertat i els valors són aproximadament els mateixos reescalats pel factor n.

$$G^2 \simeq n\mathcal{I} \qquad (4.5)$$

4.2 ACS de matrius quadrades

El cas especial que considerarem és l'aplicació de l'AC a taules quadrades, on les fileres i les columnes es refereixen al mateix conjunt d'objectes.

En diferents àmbits científics, alguns d'ells detallats a la Taula 4.1, ens trobem davant la presència de taules quadrades, on les fileres i les columnes es refereixen al mateix conjunt d'objectes o categories.

Taula 4.1: **Exemples de taules objecte d'estudi:**

Sociologia	mobilitat social
Economia	dades d'importació/exportació
Marketing	preferències de marques
Biblioteconomia	cites entrecreuades de revistes
Política	preferències de vot
Demografia	migracions
Classificació	freqüències de malclassificació
Psicologia	dades de confusió
Etologia	transicions entre estats de conducta
Ocupació	prioritat entre opcions de treball

Aquestes taules quadrades les podem descompondre en dos tipus: les simètriques, on

les relacions entre les fileres i les columnes són simètriques i les no simètriques, de més ampli ús, on no hi ha d'haver necessàriament una relació simètrica entre fileres i columnes.

Les principals problemàtiques per a aquests dos tipus de taules sorgeixen en les diagonals. En aquestes, habitualment els efectius poden ser o molt grans amb relació amb la resta d'efectius (diagonals carregades) o nuls per l'estructura de la taula (zeros estructurals). Aquesta darrera problemàtica, els zeros estructurals, també es pot extrendre fora de la diagonal sobretot quan les dues variables categòriques són ordinals i no té sentit el creuament d'una categoria per a categories inferiors: un exemple el podríem trobar en la taula de edat del matrimoni creuada amb edat del divorci: no hi pot haver divorci abans del matrimoni, per tant ens trobarem amb zeros estructurals.

Anem a descriure breument els dos tipus de taules -no simètriques i simètriques- i les problemàtiques ja descrites.

4.2.1 AC de matrius quadrades no simètriques: Problemàtica

Anem a presentar amb un exemple la problemàtica associada a les taules quadrades no simètriques. Com a exemple de treball, considerem la taula de mobilitat social origen-destí de treballadors presentada a la Taula 4.2 obtinguda de Kazmierczak [Kaz78].

Taula 4.2: *Moviment de treballadors a les rodalies de París*

	char	ivry	krem	gent	vitr	alfo	choi	bonn	vale	orly	rung	fres	thia	join	sucy
Charenton	6238	269	45	14	204	824	57	250	70	76	16	36	0	403	189
Ivry	270	11268	1113	1113	257	2483	530	708	166	878	166	205	281	457	174
Kremlin	34	585	11353	1001	1493	32	143	62	133	207	327	549	226	133	0
Gentilly	0	106	1389	10695	425	100	99	220	27	111	215	1037	26	152	117
Vitry	186	667	894	281	11263	1009	1577	148	123	1021	154	265	860	314	90
Alfort	713	258	134	75	632	16420	595	1675	563	250	29	0	118	507	297
Choisy	0	181	78	41	763	148	5590	24	396	964	104	38	745	25	87
Bonneuill	51	81	68	0	133	1094	109	9235	107	92	0	28	39	1831	491
Valenton	31	34	34	28	34	316	271	148	6161	628	0	0	59	83	228
Orly	14	108	492	177	353	104	528	209	568	6461	315	408	551	191	130
Rungis	0	21	160	83	81	33	23	20	64	248	1455	110	106	21	0
Fresnes	0	53	310	260	156	0	0	0	0	82	481	3889	131	0	0
Thiais	0	66	21	0	151	40	421	24	43	248	26	0	1498	25	0
Joinville	327	43	0	63	206	801	42	1362	0	40	54	90	35	17045	774
Sucy	0	0	0	26	26	20	28	159	591	102	0	0	0	403	5624

Aquestes dades, on les columnes són els orígens dels treballadors i les fileres les destinacions, ens reflexen els moviments dels treballadors entre les diferents zones de les rodalies de la zona sud de París (Figura 4.1), és un exemple de taula de mobilitat deguda al treball en l'àmbit de la sociologia.

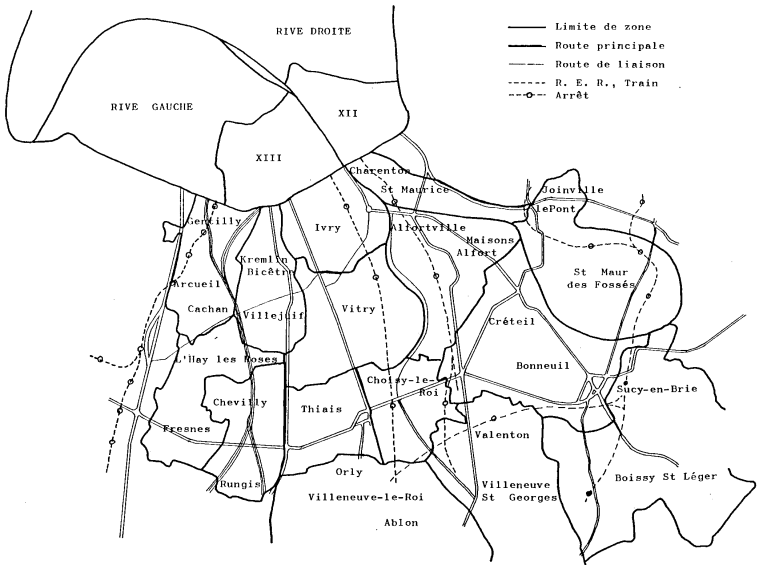


Figura 4.1: Plànol de les rodalies sud de París

L'anàlisi de correspondències simples de la Taula 4.2 dóna les representacions en els dos primers eixos factorials que podrem trobar en la Figura 4.2:

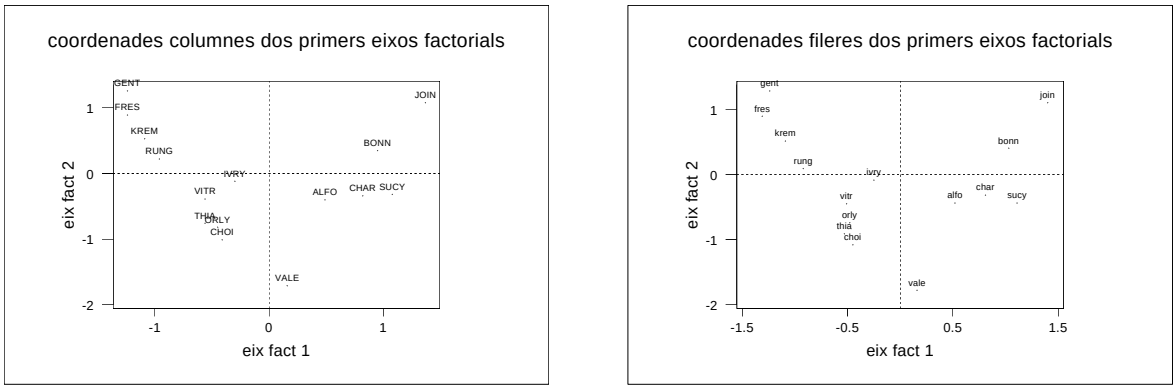


Figura 4.2: Gràfics factorials de l'AC de la Taula 4.2

La superposició dels dos gràfics de fileres i columnes dels moviments ens donarà les relacions i oposicions entre orígens i destinacions. Ho podem trobar en la Figura 4.3.

Tal i com es pot observar a la Figura 4.3, el gràfic ens dóna que hi ha una relació molt forta, com a origen i destinació, de la mateixa zona, puix que la seva representació en el pla factorial és aproximadament la mateixa.

L'explicació d'aquest fet es troba en l'anàlisi de la inèrcia i l'estudi de la descomposició de la inèrcia en aquesta taula per a cadascuna de les caselles.

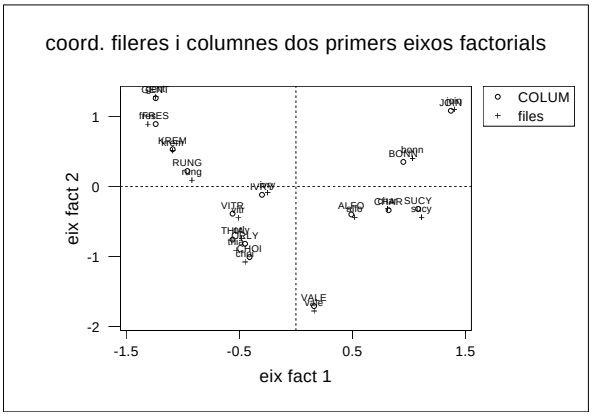


Figura 4.3: Coordenades conjuntes fileres i columnes Taula 4.2

En l'AC la *inèrcia total*, mesura de la variació de la taula P de terme general p_{ij} és:

$$I_T = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

on p_{ij} és l'efectiu de cada casella i r_i i c_j les marginals de fileres i columnes respectivament, essent r i c els vectors marginals.

La inèrcia deguda a cada casella (i,j) és:

$$\mathcal{I}_{ij} = \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

Es pot descompondre la inèrcia total de la taula en la inèrcia de la diagonal $\mathcal{I}(D)$ i en la de fora de la diagonal $\mathcal{I}(\overline{D})$, com a la suma de totes les inèrcies dels individus pertanyents a aquesta o aquella.

A la Taula 4.3 trobarem les inèrcies corresponents a cada casella i sobre la qual en podem extreure el resultat referent a les inèrcies de la diagonal i de fora de la diagonal que trobarem descrit a la Taula 4.4.

És a dir, es pot veure en aquesta darrera taula que més del 90% de la inèrcia es troba en les caselles de la diagonal, per tant la representació gràfica factorial obtinguda és bàsicament el moviment dels treballadors que tenen l'origen i el destí dins la pròpia zona. Aquest percentatge tan elevat és un tret característic de les matrius quadrades no simètriques objecte d'estudi.

El nostre objectiu serà analitzar les metodologies existents de solució principalment dins l'àmbit de les anàlisis factorials i proposar una metodologia d'anàlisi integradora que superi la influència d'aquestes caselles productores d'inèrcies estructurals per a una millor anàlisi de les relacions globals entre les modalitats.

Taula 4.3: Inèrcia deguda a cada casella de la Taula 4.2

Char	Ivry	Krem	Gent	Vitr	Alfo	Choi	Bonn	Vale	Orly	Rung	Fres	Thia	Join	Sucy
0.502	0.001	0.004	0.004	0.002	0.001	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.002	0.001
0.002	0.344	0.001	0.001	0.007	0.000	0.002	0.003	0.004	0.001	0.001	0.002	0.001	0.009	0.003
0.004	0.002	0.374	0.000	0.000	0.011	0.004	0.006	0.003	0.004	0.000	0.000	0.001	0.009	0.004
0.004	0.005	0.000	0.448	0.003	0.010	0.004	0.004	0.004	0.004	0.000	0.002	0.002	0.008	0.002
0.003	0.002	0.002	0.005	0.300	0.005	0.002	0.007	0.004	0.000	0.001	0.002	0.002	0.009	0.004
0.000	0.007	0.010	0.009	0.005	0.351	0.002	0.000	0.002	0.005	0.002	0.005	0.002	0.010	0.003
0.002	0.002	0.004	0.004	0.000	0.005	0.281	0.004	0.000	0.001	0.000	0.002	0.006	0.006	0.002
0.003	0.005	0.006	0.006	0.005	0.001	0.003	0.352	0.003	0.004	0.001	0.002	0.002	0.000	0.000
0.002	0.003	0.004	0.003	0.004	0.003	0.000	0.002	0.457	0.000	0.001	0.002	0.001	0.004	0.000
0.002	0.003	0.001	0.003	0.002	0.006	0.000	0.003	0.000	0.277	0.000	0.000	0.002	0.005	0.001
0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.000	0.000	0.245	0.000	0.000	0.001	0.001
0.001	0.002	0.000	0.000	0.001	0.004	0.002	0.002	0.002	0.001	0.008	0.382	0.000	0.004	0.001
0.001	0.001	0.001	0.001	0.000	0.001	0.003	0.001	0.003	0.000	0.000	0.001	0.171	0.001	0.001
0.002	0.008	0.010	0.008	0.008	0.008	0.006	0.000	0.006	0.007	0.002	0.003	0.003	0.469	0.000
0.002	0.003	0.004	0.003	0.003	0.005	0.002	0.002	0.001	0.002	0.001	0.001	0.001	0.001	0.492

Taula 4.4: **Descomposició de la inèrcia**

Inèrcia Total	5.99681
Inèrcia de la diagonal	5.44542
Inèrcia fora de la diagonal	0.55139

4.2.2 AC de matrius quadrades simètriques

L'anàlisi de correspondències de matrius quadrades simètriques, té per ell mateix especials propietats a més de la problemàtica ja mencionada de les diagonals. Tot i que aquesta problemàtica es veu encara més accentuada, ja que en la majoria dels estudis ens trobem amb taules de preferències on a les caselles de la diagonal ens trobem amb o bé una màxima associació o bé una relació nul·la. En aquest darrer cas, són les taules on la relació és exclusiva dels elements de fora de la diagonal, tot i que els valors diagonals no són omesos generalment en l'anàlisi.

La segona problemàtica la tenim en el fet que en la descomposició d'una matriu simètrica A pot ser escrita de la forma

$$A = UD_\lambda U^T$$

on tots els elements de U i D_λ són reals. Si hi ha cap valor propi negatiu en aquesta descomposició, llavors la DVS de la matriu A, amb les mètriques idèntiques donades per les marginals fileres i columnes, presentarà un valor propi positiu $\mu_k = -\lambda_k$ associat als vectors propis esquerre u_k i dret $-u_k$ respectivament, són els anomenats **vectors propis inversos**, on caldrà tenir present que l'existència de vectors propis inversos associats a valors propis grans esdevindrà important de cares a la mesura de l'aproximació d'A i al

seu estudi ([Ben73]).

4.2.3 AC de matrius quadrades antisimètriques

Anàlogament al cas de l'anàlisi de correspondències de matrius quadrades simètriques, l'anàlisi de matrius quadrades antisimètriques, té especials propietats a més de la problemàtica ja mencionada de les diagonals. Haurem de tenir en compte la particularitat de la descomposició de matrius antisimètriques, ja que, en general, en la descomposició via valors singulars d'una matriu antisimètrica A qualsevol podem obtenir:

$$A = UD_\lambda V' = UD_\lambda JU^T$$

on J és una matriu ortogonal diagonal en blocs (en dues dimensions):

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

on cada bloc representa una rotació en sentit antihorari de 90º i D_λ conté els valors propis en parells ordenats $\lambda_1, \lambda_1, \lambda_2, \lambda_2, \ldots$ Quan el número de fileres d'A és senar, llavors $\lambda_1 = 0$ i $J_1 = 0$.

També degut la particularitat d'aquestes matrius, el gràfics factorials de fileres i columnes són els mateixos rotats 90º. Per a la seva millor comprensió, podem aplicar els dos conceptes deguts a Gower [Gow77], Constantine i Gower [CG78]: el concepte triangle i el concepte direcció. El concepte triangle significa que per mesurar com de forta és l'antisimetria hem de mirar l'àrea del triangle format per les dues categories i l'origen. El concepte direcció ens indica si els residuals són positius o negatius en el gràfic. Asimetria positiva ens indica que i va a j, més que no pas j va a i.

4.3 Inèrcies en correspondències múltiples

En aquesta secció i en la següent, tractarem la inèrcia en les anàlisis de correspondències múltiples, seguint la via oberta en les seccions anteriors, però ara no restringint-nos només a dues variables categòriques, sinó a múltiples variables categòriques. Realitzarem la descomposició de la inèrcia total en les inèrcies de les caselles o subtaules de la diagonal de Burt, atenent a les diferents presentacions d'aquestes anàlisis, i estudiarem la problemàtica que presenten les taules de la diagonal de la matriu de Burt.

4.3.1 Generalitats

Una de les presentacions més habituals de l'ACM és mitjançant la taula de Burt. Donada una taula de Burt, on el número de total de qüestions és Q i essent $J = \sum J_q$ el número total de modalitats, es pot demostrar que la inèrcia total de la taula de Burt és la mitjana de la inèrcia de cadascuna de les subtaules individuals, on denotem la inèrcia de la subtaula (q, q') com $\phi^2_{qq'}$:

$$\mathcal{I}(B) = \frac{1}{QQ} \sum_q \sum_{q'} \phi^2_{qq'}$$

Demostració:

Donada la taula de Burt:

$$B = \begin{pmatrix} D_1 & N_{12} & \dots & N_{1Q} \\ N_{21} & D_2 & \dots & N_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ N_{Q1} & N_{Q2} & \dots & D_Q \end{pmatrix}$$

cadascuna de les subtaules D i N té les mateixes marginals fileres r_i, mateixes marginals columnes c_j i els mateixos efectius totals n. Per tant la taula de Burt tindrà com a marginals fileres Q vegades els anteriors Qr_i, el mateix per a les marginals columnes Qc_j i com a total d'individus a la taula Q^2n.

Si apliquem la fórmula (4.2) a la taula B, de terme general b_{ij} obtenim:

$$\sum_i \sum_j \frac{\left(b_{ij} - \frac{Qr_iQc_j}{Q^2n}\right)^2}{Qr_iQc_j} = \frac{1}{Q^2} \sum_i \sum_j \frac{\left(b_{ij} - \frac{r_ic_j}{n}\right)^2}{r_ic_j} = \frac{1}{Q^2} \sum_q \sum_{q'} \phi^2_{qq'}$$

4.3.2 Inèrcia de les taules B i Z

Ja hem estudiat prèviament la relació entre la descomposició de valors propis de les taules B i Z. Anem a completar aquesta relació estudiant la descomposició de les inèrcies.

Proposició 4.3.2.1 *La inèrcia total del núvol de modalitats (núvol de perfils columnes) és:* $\mathcal{I} = \frac{J}{Q} - 1$

Demostració:

anem a descompondre la inèrcia del núvol de modalitats $\mathcal{N}(\mathcal{J})$ com a la suma de les masses de les modalitats per la distància al quadrat d'aquestes al centre de gravetat.

$$\mathcal{I}(\mathcal{N}(\mathcal{J})) = \sum_{j\in J} f_{.j} d^2(j, G^I) = \sum_{j\in J} f_{.j} \left(\sum_{i\in I} \frac{1}{f_{i.}} (\frac{f_{ij}}{f_{.j}} - f_{i.})^2 \right) =$$

$$= \sum_{j\in J}\sum_{i\in I} f_{.j} \frac{1}{f_{i.}} \left(\frac{f_{ij} - f_{i.}f_{.j}}{f_{.j}} \right)^2 = \sum_{j\in J}\sum_{i\in I} f_{.j} n \left(\frac{f_{ij} - \frac{1}{n} f_{.j}}{f_{.j}} \right)^2 =$$

$$= \sum_{j\in J}\sum_{i\in I} n \frac{\left(f_{ij} - \frac{1}{n} f_{.j}\right)^2}{f_{.j}} = \sum_{j\in J}\sum_{i\in I} n \frac{f_{ij}^2 + \frac{f_{.j}^2}{n^2} - \frac{2}{n} f_{ij} f_{.j}}{f_{.j}} = \sum_{j\in J}\sum_{i\in I} \left(\frac{n f_{ij}^2}{f_{.j}} + \frac{f_{.j}}{n} - 2 f_{ij} \right) =$$

$$= \sum_{j\in J}\sum_{i\in I} \left(\frac{n f_{ij}}{z_{.j}} + \frac{1}{n}\frac{z_{.j}}{nQ} - 2\frac{z_{ij}}{nQ} \right) = \sum_{j\in J}\sum_{i\in I} n \frac{z_{ij}}{nQ}\frac{1}{z_{.j}} + \sum_{j\in J}\sum_{i\in I} \frac{z_{.j}}{n^2 Q} - \frac{2}{nQ}\sum_{j\in J}\sum_{i\in I} z_{ij} =$$

$$= \sum_{j\in J}\sum_{i\in I} n \frac{z_{ij}}{nQ}\frac{1}{z_{.j}} + \sum_{j\in J}\sum_{i\in I} \frac{z_{.j}}{n^2 Q} - \frac{2}{nQ}\sum_{j\in J}\sum_{i\in I} z_{ij} = \sum_{j\in J} \frac{1}{Q}\frac{1}{z_{.j}} \sum_{i\in I} z_{ij} + \frac{1}{n^2 Q}\sum_{j\in J}\sum_{i\in I} z_{.j} - \frac{2}{nQ} nQ =$$

$$= \sum_{j\in J} \frac{1}{Q}\frac{1}{z_{.j}} z_{.j} + \frac{1}{n^2 Q}\sum_{j\in J} n z_{.j} - 2 = \frac{J}{Q} + \frac{1}{n^2 Q} n n Q - 2 = \frac{J}{Q} + 1 - 2 = \frac{J}{Q} - 1$$

□

Així en l'exemple, que hem tractat a la Secció 2.9.8, podem veure que la inèrcia obtinguda: $I_{Tot} = 2.71429$ és igual a: $\frac{J}{Q} - 1 = \frac{26}{7} - 1 = 2.7142857$

Podem calcular, també, les inèrcies degudes a una modalitat i una qüestió.

Proposició 4.3.2.2 *Les inèrcies, respectivament, d'una modalitat i d'una qüestió són:* $\mathcal{I}(j) = \frac{1}{Q}(1 - \frac{z_{.j}}{n})$ *i* $\mathcal{I}(q) = \frac{1}{Q}(J_q - 1)$

Demostració: anem a descompondre la inèrcia d'una modalitat j fent servir que $f_{i.} = \frac{1}{n}$, $f_{.j} = \frac{z_{.j}}{nQ}$, $f_{ij} = 0$ o bé $f_{ij} = \frac{1}{nQ}$ i $\sum_i f_{ij}^2 = \frac{z_{.j}}{(nQ)^2}$:

$$\mathcal{I}(j) = f_{.j} d^2(j, G^I) = f_{.j} \left(\sum_{i\in I} \frac{1}{f_{i.}} (\frac{f_{ij}}{f_{.j}} - f_{i.})^2 \right) = \sum_{i\in I} \frac{f_{.j}}{f_{i.}} \frac{f_{ij}^2 + f_{.j}^2 f_{i.}^2 - 2 f_{ij} f_{.j} f_{i.}}{f_{.j}^2} =$$

$$= \sum_{i\in I} \frac{f_{ij}^2}{f_{i.} f_{.j}} + \sum_{i\in I} \frac{f_{.j}}{f_{i.}} - 2\sum_{i\in I} \frac{f_{ij} f_{i.}}{f_{.j}} = n\frac{nQ}{z_{.j}}\frac{z_{.j}}{(nQ)^2} + n\frac{1}{n}\frac{z_{.j}}{nQ} - 2 f_{.j} =$$

$$= \frac{1}{Q} + \frac{z_{.j}}{nQ} - 2\frac{z_{.j}}{nQ} = \frac{1}{Q}(1 - \frac{z_{.j}}{n})$$

I com la inèrcia d'una qüestió és la deguda al conjunt de modalitats de la mateixa:

$$\mathcal{I}(q) = \sum_{j\in J_q} f_{.j} d^2(j, G^I) = \sum_{j\in J_q} \frac{1}{Q}(1 - \frac{z_{.j}}{n}) = \frac{1}{Q}(J_q - 1)$$

□

Podem observar que la inèrcia d'una modalitat esdevé mínima, quan $z_{.j} = n$, és a dir, quan aquella modalitat l'han escollida tots els individus enquestats. I contràriament, la inèrcia és més gran com més petit és l'efectiu $z_{.j}$ d'individus que han escollit aquella modalitat.

Pel que fa a l'inèrcia d'una qüestió, podem veure que creix segons el nombre de modalitats de resposta, essent el mïnim igual a $\frac{1}{Q}$ quan tenim dues modalitats a la qüestió $J_q = 2$.

Finalment podem tornar a obtenir que la inèrcia total del núvol és $\frac{J}{Q} - 1$, si sumem les inèrcies de totes les qüestions:

$$\mathcal{I}(\mathcal{N}(\mathcal{J})) = \sum_{q \in Q} \mathcal{I}(q) = \sum_{q \in Q} \frac{1}{Q}(J_q - 1) = \frac{1}{Q} \sum_{q \in Q} (J_q - 1) = \frac{1}{Q}(J - Q) = \frac{J}{Q} - 1$$

4.4 Problemàtica de la inèrcia de les submatrius diagonals

L'aportació de les submatrius diagonals de la diagonal de Burt està determinada, tal com vegem en la següent proposició.

Proposició 4.4.0.3 *Donada una subtaula de la diagonal de la taula de Burt, que al seu torn és una taula quadrada diagonal, es pot demostrar que la inèrcia de la subtaula és igual al nombre de modalitats de la taula menys un, és a dir:* $\phi^2(D) = J_q - 1$

Demostració:

$$\phi^2 = \frac{1}{n_{..}} \sum_{i=1}^{J_q} \sum_{j=1}^{J_q} \frac{(o-e)^2}{e} = \frac{1}{n_{..}} \sum_{i=1}^{J_q} \sum_{j=1}^{J_q} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}})^2}{\frac{n_{i.}n_{.j}}{n_{..}}} =$$

$$= \frac{1}{n_{..}} \sum_{i=j=1}^{J_q} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}})^2}{\frac{n_{i.}n_{.j}}{n_{..}}} + \frac{1}{n_{..}} \sum_{i=1}^{J_q} \sum_{j=1, j \neq i}^{J_q} \frac{n_{i.}n_{.j}}{n_{..}} =$$

$$= \frac{1}{(n_{..})^2} \sum_{i=j=1}^{J_q} \frac{n_{ij}^2 n_{..}^2 + n_{i.}^2 n_{.j}^2 - 2n_{..}n_{ij}n_{i.}n_{.j}}{n_{i.}n_{.j}} + \frac{1}{(n_{..})^2} \sum_{i=1}^{J_q} \sum_{j=1, j \neq i}^{J_q} n_{i.}n_{.j} =$$

com $n_{i.} = n_{.j} = n_{ij}$ per a les caselles on $i = j$

$$= \frac{1}{(n_{..})^2} \sum_{i=j=1}^{J_q} \left(n_{..}^2 - 2n_{..}n_{ij}\right) + \frac{1}{(n_{..})^2} \sum_{i=1}^{J_q} \sum_{j=1}^{J_q} n_{i.}n_{.j} =$$

$$= \frac{1}{(n_{..})^2} J_q (n_{..})^2 - \frac{2n_{..}}{(n_{..})^2} \sum_{i=j=1}^{J_q} n_{ij} + \frac{1}{(x_{..})^2} \sum_{i=1}^{J_q} x_{i.} \sum_{j=1}^{J_q} x_{.j} = J_q - 2 + 1 = J_q - 1$$

□

Per la qual cosa, l'aportació de les taules de la diagonal és calculable a priori i no depèn de les relacions entre les variables. És deguda només al nombre de variables i modalitats de cadascuna d'aquestes, per tant ens trobem davant la presència d'una inèrcia estructural. La inèrcia de la diagonal de la taula de Burt és per tant:

$$\mathcal{I}(D) = \frac{J - Q}{Q^2}$$

Anem a exemplificar aquesta inèrcia estructural de la diagonal en la (Taula 4.5) donada per Abascal i Grande [AG89].

Taula 4.5: *Actitud davant els serveis oferts pel districte*

edat	renda	estudis	hàbitat	ús biblio	actitud	categ.
2	3	3	4	4	4	4
2	3	3	3	4	4	4
4	1	1	4	2	2	1
3	3	3	3	4	4	3
2	2	1	1	2	1	1
1	2	2	4	3	3	3
4	4	2	3	2	2	1
3	1	1	2	1	1	1
1	1	1	1	1	1	1
2	4	3	3	4	4	4
1	2	3	3	5	4	3
4	3	2	1	1	1	1
3	1	3	4	5	4	4
3	1	3	3	5	4	4
2	4	3	3	4	4	4
2	2	1	2	3	3	2
1	2	1	3	3	3	2
1	2	2	4	4	3	3
1	2	2	3	4	3	3
2	3	2	3	4	3	3

En l'exemple, s'analitzen les actituds davant els serveis oferts pel municipi (4 categories), l'ús de la biblioteca (5 categories) i l'edat (4 categories), els nivells de renda (4 categories), nivells d'estudis (3 categories), mida de l'hàbitat (4 categories) i la categoria sòcio-professional (4 categories). En la taula de Burt corresponent (Taula 4.7) es reflexen les relacions de les modalitats de les variables en cadascuna de les subtaules i on les sub-

taules de la diagonal són al seu torn taules diagonals que contenen els efectius de cada modalitat.

En resum tenim 7 variables i 28 modalitats sobre un total de 24 individus, per tant la inèrcia aportada per la diagonal a priori serà de:

$$\frac{28-7}{7^2} = 0.428571$$

Si en el nostre exemple descomposem la inèrcia en la part corresponent a la diagonal i a fora de la diagonal, tal com podem trobar a la Taula 4.6, tenim efectivament aquesta inèrcia estructural, deguda a les subcaselles de la diagonal, que sobre el total d'inèrcia de la taula ens dóna una aportació inercial propera al 38.5%.

Taula 4.6: *Taula d'inèrcies de l'actitud davant els serveis oferts*

Inèrcia Total	1.11356
Inèrcia subtaules diagonal	0.428571
Inèrcia no diagonal	0.684989

Podem destacar alguns treballs en l'àmbit de propostes per a una millor interpretació de l'ACM: Benzécri [Ben79] i Lebart [LMW84], proposen que els valors propis es reescalin atenent a les inèrcies produïdes artificiosament. Aquest reescalat pren en comptes dels valors propis λ_k uns nous valors propis λ'_k donats per:

$$\lambda'_k = \left(\frac{Q}{Q-1}\right)^2 * \left(\sqrt{\lambda_k} - \frac{1}{Q}\right)^2 \qquad \text{si i només si} \qquad \sqrt{\lambda_k} > \frac{1}{Q}$$

on el factor $\frac{1}{Q}$ ens corregeix l'efecte artificiositat de l'anàlisi i seria l'aportació mínima obtinguda en funció del nombre de modalitats per a una variable.

Zarraga [Zar89] en la seva anàlisi de correspondències múltiples per bandes, calcula la inèrcia no trivial de la taula de Burt en funció de les inèrcies de les bandes, un cop corregides aquestes per a eliminar-ne la part inercial deguda a les modalitats d'una mateixa qüestió.

Per la seva banda Greenacre [Gre87] [Gre89] i en treballs posteriors, dóna com a alternativa el **Joint Correspondence Analysis, JCA**, on de forma algorísmica iterativa dóna una anàlisi per a les subtaules no diagonals de la taula de Burt.

Hem vist que algunes de les tècniques incideixen en la modificació dels valors propis per a recalcular la qualitat de la interpretació, però d'altres ataquen directament la problemàtica de les inèrcies, tal com podrem veure en la següent secció.

Taula 4.7: *Taula de Burt de l'actitud davant els serveis oferts*

ed0-25	6	0	0	0	1	5	0	0	2	3	1	1	0	3	2	1	0	2	2	1	1	0	4	1	1	1	4	0
ed26-35	0	9	0	0	1	2	3	3	3	1	5	1	2	4	2	1	1	1	6	0	2	0	2	5	2	1	1	5
ed36-50	0	0	5	0	3	1	1	0	2	0	3	0	2	2	1	1	1	0	1	2	1	1	0	3	2	0	1	2
ed50-	0	0	0	4	1	1	1	1	2	2	0	1	0	2	1	1	2	0	1	0	1	2	1	0	3	0	1	0
re0-15	1	1	3	1	6	0	0	0	4	0	2	1	2	1	2	3	1	0	0	2	3	1	0	2	4	0	0	2
re15-25	5	2	1	1	0	9	0	0	5	3	1	1	2	4	2	0	2	3	3	1	1	1	6	1	2	2	5	0
re25-40	0	3	1	1	0	0	5	0	0	2	3	1	0	3	1	1	0	0	4	0	1	0	1	3	1	0	2	2
re40-	0	3	0	1	0	0	0	4	0	1	3	0	0	3	1	0	1	0	3	0	0	1	0	3	1	0	0	3
eselem	2	3	2	2	4	5	0	0	9	0	0	2	4	2	1	3	3	2	1	0	4	2	3	0	6	2	1	0
esmitg	3	1	0	2	0	3	2	1	0	6	0	1	0	3	2	1	1	1	3	0	1	1	4	0	2	0	4	0
essupe	1	5	3	0	2	1	3	3	0	0	9	0	0	6	3	0	0	0	6	3	0	0	0	9	0	0	2	7
ha0-5	1	1	0	1	1	1	1	0	2	1	0	3	0	0	0	2	1	0	0	0	3	0	0	0	3	0	0	0
ha5-20	0	2	2	0	2	2	0	0	4	0	0	0	4	0	0	2	1	1	0	0	2	1	1	0	3	1	0	0
ha20-1m	3	4	2	2	1	4	3	3	2	3	6	0	0	1	1	0	0	1	1	7	2	0	1	4	6	1	1	5	4
ha1mil-	2	2	1	1	2	2	1	1	1	2	3	0	0	0	6	0	1	1	3	1	0	1	2	3	1	0	2	3
bibmai	1	1	1	1	3	0	1	0	3	1	0	2	2	0	0	4	0	0	0	0	4	0	0	0	4	0	0	0
bibgai	0	1	1	2	1	2	0	1	3	1	0	1	1	1	1	0	4	0	0	0	1	3	0	0	4	0	0	0
bibaveg	2	1	0	0	0	3	0	0	2	1	0	0	1	1	1	0	0	3	0	0	0	0	3	0	0	2	1	0
bibsov	2	6	1	1	0	3	4	3	1	3	6	0	0	7	3	0	0	0	1	0	0	0	0	4	6	0	0	5	5
bibfrq	1	0	2	0	2	1	0	0	0	0	3	0	0	2	1	0	0	0	0	3	0	0	0	3	0	0	1	2
serdes	1	2	1	1	3	1	1	0	4	1	0	3	2	0	0	4	1	0	0	0	5	0	0	0	5	0	0	0
sernoin	0	0	1	2	1	1	0	1	2	1	0	0	1	1	1	0	3	0	0	0	0	3	0	0	3	0	0	0
serinpo	4	2	0	1	0	6	1	0	3	4	0	0	1	4	2	0	0	3	4	0	0	0	7	0	0	2	5	0
seinmo	1	5	3	0	2	1	3	3	0	0	9	0	0	6	3	0	0	0	6	3	0	0	0	9	0	0	2	7
catmcas	1	2	2	3	4	2	1	1	6	2	0	3	3	1	1	4	4	0	0	0	5	3	0	0	8	0	0	0
catnoqu	1	1	0	0	0	2	0	0	2	0	0	0	1	1	0	0	0	2	0	0	0	0	2	0	0	2	0	0
catqual	4	1	1	1	0	5	2	0	1	4	2	0	0	5	2	0	0	1	5	1	0	0	5	2	0	0	7	0
catmoqu	0	5	2	0	2	0	2	3	0	0	7	0	0	4	3	0	0	0	5	2	0	0	0	7	0	0	0	7

4.5 Metodologies existents de solució

La gran influència de la diagonal, en el sentit propi de diagonals en l'ACS o de les taules diagonals de l'ACM, en aquestes taules objecte d'estudi ha portat a la necessitat d'estudiar propostes de modificació de les anàlisis en el sentit de treure la influència d'aquestes caselles. Aquestes propostes de modificació es basen principalment en reconstituir els elements de les diagonals, amb metodologies diferents, atenent a diferents propostes de reconstitució, però també s'ha abordat el problema des d'una metodologia d'intentar descompondre l'anàlisi en diferents parts. A les següents seccions tractarem les propostes que diferents autors han proposat, les compararem i exemplificarem sobre l'exemple dels moviments de París i, finalment, proposarem la nostra metodologia de resolució.

4.5.1 Reconstitucions proposades

Algunes de les propostes fetes, i que són objecte de comparació, han estat:

1. Anul·lació dels elements de les diagonals carregades
2. Reemplaçament dels elements de la diagonal sota la hipòtesi d'independència
3. Tractament *missing data* dels elements de la diagonal
4. Reemplaçament dels elements de la diagonal mitjançant la fórmula de reconstitució
5. Descomposició de la matriu com l'addició d'una matriu simètrica i d'una d'anti-simètrica

Anem a descriure i a estudiar aquestes propostes.

4.5.2 Anul·lació dels elements de les diagonals carregades

Una primera proposta de solució dels problemes causats per les caselles amb inèrcies estructurals està basada en l'eliminació dels efectius de les caselles. Per la qual cosa es redefineix tot element de la diagonal o caixa de la diagonal de la matriu objecte d'estudi com:

$$\hat{p}_{ij} = 0$$

amb la qual cosa es pretén evitar que aquests influeixin en l'anàlisi, però aquesta redefinició comporta que la inèrcia de la nova diagonal no sigui nul·la, la qual cosa era el que es

pretenia. Es pot demostrar que la inèrcia de la diagonal després de la redefinició no és nul·la, en relació als marges originals, sinó que és igual a :

$$\mathcal{I}(D) = \sum_{ij} r_i c_j$$

I a més, aquesta redefinició comporta una alteració dels marges, és a dir una modificació de les mètriques de l'anàlisi.

La proposta basada en l'anul·lació dels efectius de les caselles de la diagonal, ens dóna una qualitat de representació en el primer pla factorial (gràfic 4.4) del 41.528% i la següent descomposició de la inèrcia:

Inèrcia total	1.32642
Inèrcia diagonal	0.0739544
Inèrcia no diagonal	1.25247

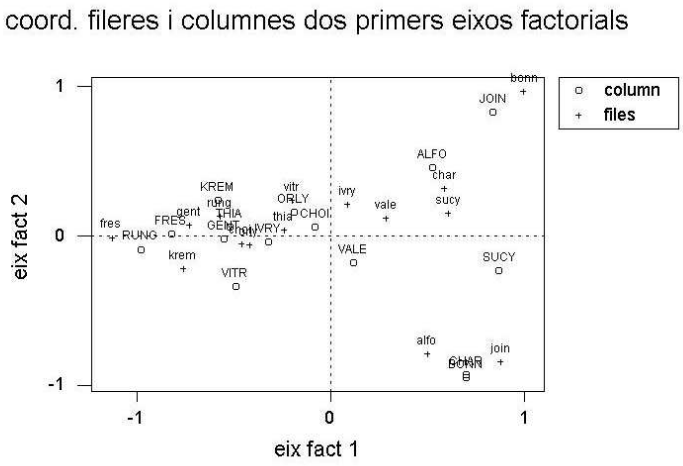


Figura 4.4: Gràfic factorial de l'anul·lació d'efectius

S'obté una considerable reducció de la inèrcia total i la de les subtaules de la diagonal, però no la seva anul·lació.

4.5.3 Reemplaçament dels elements de la diagonal sota la hipòtesi d'independència

Aquesta proposta, així com les següents, està sustentada en base als *mètodes d'anàlisi factorial en referència a un model* proposats per Escofier[Esc84], ja que la taula objecte

d'estudi és la taula original menys el model d'independència imposat als elements diagonals.

La component inercial de la redefinició proposada anteriorment és deguda a que la nullitat de la inèrcia no s'obté substituint els elements per zeros sinó que s'obté substituint els elements pel producte de les marginals. De Leeuw i van der Heijden [LH88] suggereixen redefinir els elements de la diagonal com a producte dels marges, és a dir, sota un model d'independència:

$$\hat{p}_{ij} = r_i c_j$$

Aquesta reconstitució de la matriu de dades sota el supòsit d'independència comportaria que les caselles reconstituïdes tinguessin una inèrcia nul·la $\mathcal{I}(D) = 0$, però això no es així ja que no es conserven els marges de les taules i, per tant no s'aconsegueix el propòsit. És més, la reconstitució està influenciada pels valors de les caselles influents, ja que aquestes tenen un pes molt important en la definició de les marginals.

L'anàlisi de correspondències de la taula reconstituïda sota la hipòtesi d'independència ens dóna les següents representacions en els dos primers eixos factorials (Figura 4.5), en les quals podem veure que encara es conserva la gran relació d'un mateix lloc com a origen i com a destinació. Aquesta relació és deguda al fort paper que té la casella de la diagonal en les marginals corresponents, ja que la marginal en un terme mitjà és deguda en un 70% als valors de les caselles de la diagonal.

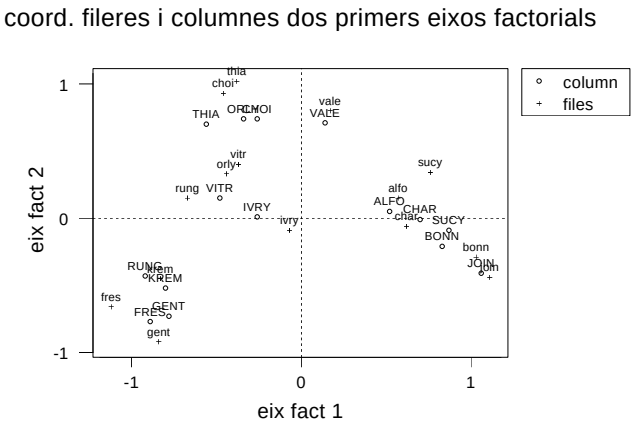


Figura 4.5: AC de la reconstitució sota hipòtesi d'independència

La qualitat de representació de la representació 4.5 és d'un 63.3% amb la següent descomposició de la inèrcia:

Inèrcia total	1.09097
Inèrcia diagonal	0.255799
Inèrcia no diagonal	0.835175

Tal com podem veure, la inèrcia de la diagonal no és 0, sinó que en haver canviat les marginals, els resultats de substituir per la hipòtesi d'independència no es corresponen a les noves diagonals.

4.5.4 Tractament *missing data* de les caselles diagonals

La tècnica usada en aquesta secció s'inclou en les tècniques per al tractament de taules incompletes proposades per Nora [Nor74], també tractades per Benzécri [Ben73] així com per Greenacre [Gre84] i de Leeuw i van der Heijden [LH88] entre d'altres, essent ara aplicada al tractament de dades influents.

La influència de les caselles de la diagonal en les marginals porta a una proposta de solució basada en el tractament de les dades de la diagonal com a dades **missing**, és a dir, considerarem a tots els efectes que aquestes dades són mancants i per tant la seva reconstitució a partir de les marginals no serà influïda per aquestes. Es defineix cada element de la diagonal o caixa diagonal com:

$$\hat{p}_{ij} = \frac{r_i^0 c_j^0}{1^0 - (r_i^0 + c_j^0)}$$

on el superíndex zero significa que l'element de la diagonal ha estat exclòs dels respectius marginals. En la Figura 4.6 podem trobar-ne la representació gràfica resultant:

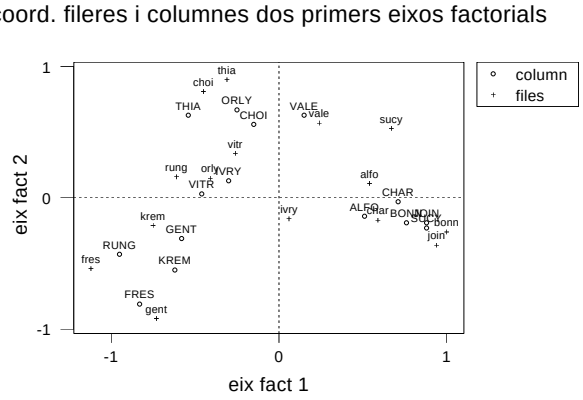


Figura 4.6: AC de la reconstitució tipus *missing*

Podem veure que en aquest darrer gràfic ja s'ha perdut l'estructura deguda a la presència de grans quantitats en les caselles de la diagonal i només resten les quantitats degudes a les caselles de fora de la diagonal.

Aquesta reconstitució de la matriu per la imputació d'efectius com *missing data* comporta que les cel·les reconstituïdes tenen una inèrcia gairebé nul·la, no es conserven les marginals de la taula però la reconstitució ara no està influïda pels valors diagonals influents*.

La qualitat de representació en la Figura 4.6 és d'un 46.8% amb la següent descomposició de la inèrcia:

Inèrcia total	1.16711
Inèrcia diagonal	0.0178457
Inèrcia no diagonal	1.14926

En aquest cas ja veiem que el fet de reconstituir eliminant la influència de les dades de la diagonal de les marginals ens comporta una reducció substancial de la inèrcia deguda a la diagonal.

4.5.5 Reemplaçament dels elements de la diagonal mitjançant la fórmula de reconstitució

Basada en el tractament tipus missing aplicat a la diagonal i proposat per Mutombo [Mut73] i Nora [Nor74], per a dades *missing*, hi ha el tractament k-EM. Aquest es basa en un algorisme fonamentat en l'algorisme EM de Dempster i altres [DLD76], però amb la introducció d'un nou paràmetre, l'ordre k de reconstitució inherent a les anàlisis factorials.

Aquest algorisme està basat en fer una primera resconstitució tipus missing, és a dir, definim:

$$\hat{p}_{ij} = \frac{r_i^0 c_j^0}{1^0 - (r_i^0 + c_j^0)}$$

i aquest serà el valor inicial de l'algorisme.

Les etapes de l'algorisme són:

I) Es comença el procés assignant els valors inicials de $\hat{p}_{ij}$

II) Es realitza una AC de la matriu amb els valors reconstituïts tipus *missing* (Etapa M, o de maximització, de l'algorisme EM), per a trobar els eixos i les coordenades factorials.

*La mateixa tècnica pot ser usada per al tractament de zeros estructurals fora de la diagonal o de les subtaules diagonals en ACM

III) S'utilitza la fórmula de reconstitució de les anàlisis factorials, d'ordre $k = 1$ del valor $\hat{p}_{ij}$ (Etapa E, o d'estimació, de l'algorisme EM)

IV) Es reiteren els passos (II) i (III) fins a la convergència dels valors reconstituïts.

Així:

$$\hat{p}_{ij}^{(1,k)} = r_i^{(1,k-1)} c_j^{(1,k-1)} \left\{ 1 + \sum_{\alpha=1}^{1} \sqrt{\lambda_\alpha^{(1,k-1)}} \psi_{\alpha i}^{(1,k-1)} \phi_{\alpha i}^{(1,k-1)} \right\}$$

on els superíndexs (1,k) representen l'ordre de reconstitució i l'ordre d'interacció respectivament.

Una vegada obtinguda aquesta convergència, s'incrementa l'ordre de reconstitució, essent ara $k = 2$, prenent com a valor inicial l'obtingut en l'etapa (IV) del procés anterior:

$$\hat{p}_{ij}^{(2,0)} = r_i^{(1,m)} c_j^{(1,m)} \left\{ 1 + \sum_{\alpha=1}^{1} \sqrt{\lambda_\alpha^{(1,m)}} \psi_{\alpha i}^{(1,m)} \phi_{\alpha i}^{(1,m)} \right\}$$

i es reiteren les etapes (II) a (IV) fins a assolir de nou la convergència, amb ordre de reconstitució $k = 2$.

Es finalitza el procés augmentant l'ordre de reconstitució d'acord, per exemple, amb els valors propis significatius en l'anàlisi de la taula final indicador habitual de la qualitat de representació de cada eix, utilitzant a l'inici de cada nova etapa com a valor inicial el valor obtingut al final del procés d'iteració d'ordre inferior.

$$\hat{p}_{ij}^{(l,k)} = r_i^{(l,k-1)} c_j^{(l,k-1)} \left\{ 1 + \sum_{\alpha=1}^{l} \sqrt{\lambda_\alpha^{(1,k-1)}} \psi_{\alpha i}^{(1,k-1)} \phi_{\alpha i}^{(1,k-1)} \right\}$$

En aquesta metodologia es conserven les marginals havent-ne exclòs d'elles els elements de la diagonal.

En la Figura 4.7 trobem la representació gràfica d'aquesta anàlisi amb una reconstitució d'ordre 1.

En aquest cas ens trobem amb una qualitat de representació del primer pla factorial del 52.3% amb la següent descomposició de la inèrcia:

Inèrcia total	1.07918
Inèrcia diagonal	0
Inèrcia no diagonal	1.07918

Tal com podem veure, en aquest cas la inèrcia de la diagonal és 0, ja que una reconstitució 1-EM equival a una reconstitució iterada sota el supòsit d'independència.

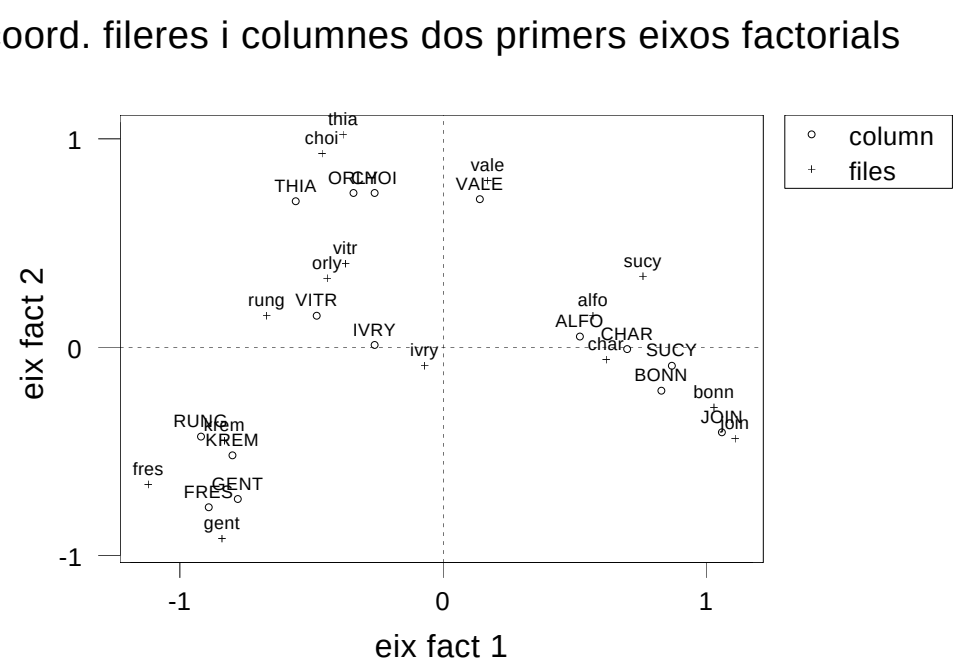


Figura 4.7: AC de la reconstitució 1-EM

Altres propostes en aquesta mateixa línia, d'imposar un model sobre les dades, han estat proposades, vegi's per exemple [Fou85], però també hi ha metodologies alternatives com la de la següent secció.

4.5.6 Descomposició de la matriu com l'addició d'una matriu simètrica i d'una d'antisimètrica

En un altre vessant totalment diferent trobem la proposta de Constantine i Gower [CG78], revisada per Greenacre [Gre00], de solució basada en la descomposició de la matriu com a suma d'una matriu simètrica i d'una altra d'antisimètrica. La matriu objecte d'estudi $\mathbf{P\text{-}rc}^T$ pot ésser descomposada en $\mathbf{Q{+}R}$ on $\mathbf{Q}$ és una matriu simètrica i $\mathbf{R}$ una matriu antisimètrica. Si anomenem $\mathbf{A{=}P\text{-}rc}^T$, llavors podem trobar $\mathbf{Q}$ **i** $\mathbf{R}$ de la següent manera:

$$\mathbf{Q} = \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{A} + \mathbf{A^T}) \qquad i \qquad \mathbf{R} = \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{A} - \mathbf{A^T})$$

Utilitzant les propietats de les matrius $\mathbf{Q}$, de marginals $\mathbf{w}$, i $\mathbf{R}$ i com que:

$$\mathbf{P} - \mathbf{ww^T} = \mathbf{Q} - \mathbf{ww^T} + \mathbf{R}$$

Això ens permet fer una descomposició de la inèrcia en la part deguda a la simetria i a la no-simetria. L'anàlisi de $\mathbf{Q\text{-}ww}^T$ reflexa la inèrcia de la simetria (valors diagonals i

moviments recíprocs) i l'anàlisi de $\mathbf{R}$ reflexa la no-simetria (els balanços dels moviments recíprocs).

Llavors en comptes de l'anàlisi estàndard, la descomposició en valors singulars generalitzada (GSVD) de la tripleta $(\mathbf{P} - \mathbf{rc^T}, \mathbf{D}_r^{-1}, \mathbf{D}_c^{-1})$

les anàlisis proposades són les GSVD:

$$(\mathbf{Q} - \mathbf{ww^T}, \mathbf{D_w^{-1}}, \mathbf{D_w^{-1}})$$

$$(\mathbf{R}, \mathbf{D_w^{-1}}, \mathbf{D_w^{-1}})$$

És a dir se centra en referència a les marginals de $\mathbf{Q}$ i es prenen també les mètriques de la part simètrica.

Els gràfics factorials d'ambdues parts els podem trobar en la Figura 4.8, on podem observar una perfecta relació entre fileres i columnes en el primer gràfic i unes relacions particulars entre fileres i columnes en el segon gràfic.

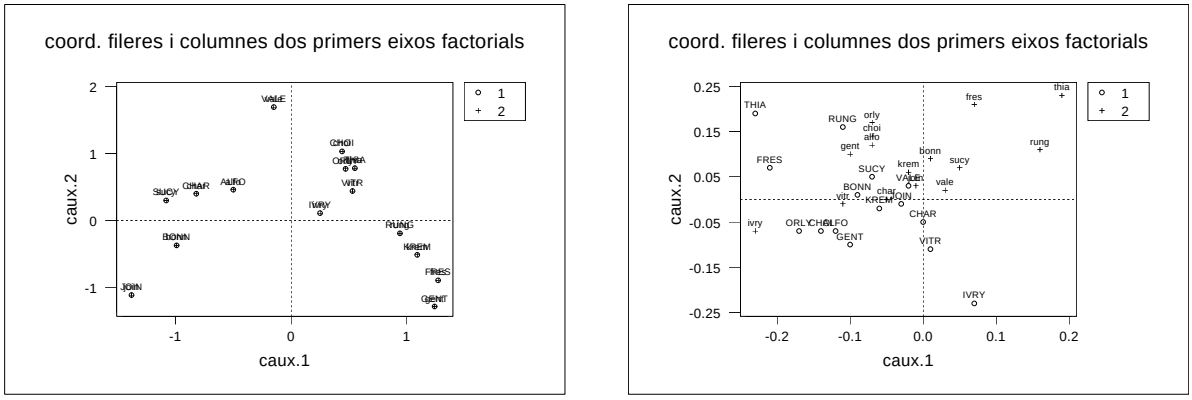


Figura 4.8: AC de la part simètrica i de la part no simètrica

4.6 Proposta d'anàlisi per a matrius quadrades no simètriques

Atenent a les diferents propostes realitzades pels diferents investigadors, l'anàlisi que suggerim està dividida en tres etapes. Aquestes, per la seva relació de similitud amb els moviments migratoris, seran anomenades:

1. Anàlisi Migració Intra

2. Anàlisi Migració Entre i Recíproca

3. Anàlisi Saldos Migratoris

Passarem ara a descriure cadascuna d'aquestes tres parts i posteriorment en farem un exemple de la seva aplicació.

4.6.1 Anàlisi Migració Intra

La primera etapa és l'anomenada **anàlisi de la migració intra** o l'anàlisi dels moviments dins de la mateix zona, és a dir l'estudi de l'importància dels valors de la diagonal en referència als seus respectius marges, això ens donarà quins percentatges d'individus romanen dins la mateixa categoria en la taula. Aquesta primera part ens permetrà veure en quines categories els efectius romanen en la mateixa categoria.

Per a aquesta primera etapa el que farem serà calcular els percentatges del moviment intern de la categoria en referència a l'efectiu global vers la mateixa categoria. L'anàlisi d'aquests percentatges ens proporcionarà quines són les categories en les quals bàsicament el moviment és intern.

4.6.2 Anàlisi Migració Entre i Recíproca

La segona etapa és l'anomenada **anàlisi de la migració entre i recíproca** o l'anàlisi dels moviments simètrics entre cadascuna de les parelles de zones, és a dir, l'estudi de la component simètrica, definida ja en apartats anteriors, però a la qual li traurem la influència dels valors de la diagonal.

Per a treure la influència de la diagonal, en l'estudi de la part simètrica el que farem és aplicar la reconstitució tipus **k-EM** a la matriu $\mathbf{Q\text{-}ww}^T$, matriu de la component simètrica. Aquesta reconstitució ens permetrà treure la càrrega inercial de la diagonal i quedar-nos amb el que són exclusivament les relacions fora de la diagonal, però només els simètrics. Això ens permetrà determinar quines relacions de simetria tenim entre les diferents modalitats de la taula.

4.6.3 Anàlisi Saldos Migratoris

Finalment la darrera etapa és l'anomenada **anàlisi dels saldos migratoris** o l'anàlisi de la diferència de moviments no recíprocs entre subjectes, és a dir, l'estudi de la component no simètrica dels moviments entre subjectes. Aquesta part ens permetrà detectar les

asimetries en els intercanvis entre categories. Analitzat conjuntament amb els moviments simètrics ens permetran conformar l'anàlisi de les relacions no diagonals.

Per a realitzar aquesta darrera part realitzarem l'estudi ja presentat de la matriu $\mathbf{R}$, matriu corresponent a la no-simetria, on realitzarem sobre aquestes dades la descomposició en valors singulars generalitzada de la tripleta:

$$(\mathbf{R}, \mathbf{D_w^{-1}}, \mathbf{D_w^{-1}})$$

És a dir centrem en referència a les marginals i prenem també les mètriques de la part simètrica $\mathbf{Q}$.

Haurem de tenir en compte la particularitat de la descomposició de matrius anti-simètriques ja mencionat a la secció (4.2.3).

4.7 Exemples d'aplicació

En aquesta secció anem a aplicar la proposta d'anàlisi efectuada a dos exemples. El primer d'ells correspon als moviments migratoris obligats pel treball, entre les comarques de Catalunya corresponents al cens de 1996, matriu quadrada no simètrica, el qual el tractarem en base als tres ítems presentats:

- Anàlisi Migració Intra
- Anàlisi Migració Entre i Recíproca
- Anàlisi Saldos Migratoris

Trobem a la taula 4.8 les dades dels moviments entre les 41 comarques catalanes deguts al treball, corresponents a l'any 1996 [Ide01]. Clarament ens trobem amb una matriu quadrada no simètrica, on la importància de la inèrcia de la diagonal és del 97.77% sobre una inèrcia total de 28.6486.

En el segon exemple, un exemple de correspondències múltiples, les dades corresponen a una enquesta socioeconòmica referida a *Les condicions de vida i aspiracions dels francesos* [LHvE80], els resultats de la qual són emprats per M. Greenacre i J. Blasius [GB94] per il·lustrar el Joint Correspondence Analysis (JCA). A aquestes dades farem una reconstitució de les caselles de la diagonal mitjançant la fórmula de reconstitució. Podrem comprovar que el resultat obtingut és gairebé equivalent al que els autors esmentats presenten com a JCA.

Taula 4.8: *Moviment de treballadors a les 41 comarques catalanes*

A. Camp	10119	4	28	1	0	7	6	417	9	1	33	95	225	0	0	289	10	2	0	2	1	5	1	2	0	0	1	0	6	3	0	13	7	1	0	991	1	1	2	24	6
A. Emporda	0	32542	7	0	0	16	9	2	2	453	67	4	753	4	3	0	3	0	127	1244	38	1	1	28	0	1	7	139	0	2	17	7	11	107	0	9	0	0	5	84	49
A. Penedes	17		623368	1	0	262	31	15	3	10	983	487	1879	42	1	4	343	5	5	12	21	5	2	7	0	1	0	14	0	4	0	2	5	7	4	71	2	1	1	279	70
A. Urgell	0	3		25508	3	3	13	0	0	2	21	1	162	7	70	0	3	0	2	2	3	0	22	10	9	3	2	0	0	0	1	11	72	2	42	3	2	6	2	32	10
A. Ribagorça	0	0	0	0	1029	5	1	1	0	4	6	0	91	0	0	0	1	0	0	2	0	1	2	0	23	0	2	0	0	0	0	1	69	0	1	3	1	2	47	6	3
Anoia	17	19	411	7	0	27531	180	5	1	14	1115	13	1773	12	1	100	36	0	12	6	54	0	5	13	1	0	11	1	1	2	2	331	34	6	9	14	0	20	2	300	86
Bages	0	15	54	16	0		25248169	15	3	7	804	15	2130	436	20	4	17	0	8	31	56	1	4	272	1	5	3	7	1	3	10	40	28	19	159	35	1	3	1	1185	272
B. Camp	579	4	41	1	1	20		1737061	134	7	130	170	1194	9	0	95	33	23	16	30	31	29	7	12	1	3	10	0	50	263	2	2	69	6	4	9449	9	8	9	119	47
B. Ebre	34	2	8	0	0	1	2		23419487	1	31	23	380	1	1	4	9	4	0	7	10	1180	0	3	4	2	2	1	4	61	4	7	23	6	0	331	34	2	0	22	22
B. Emporda	1	524	7	3	0	7	19	2		131730	122	9	1164	2	11	0	3	0	82	2179	65	0	0	21	3	1	0	49	0	0	15	2	6	319	0	10	0	2	2	179	86
B. Llobregat	14	36	968	3	3	693	464	47	16		28142346	131	74642	19	9	10	646	1	17	100	486	13	5	126	5	5	6	10	0	12	6	20	56	81	3	210	0	16	6	9424	1912
B. Penedes	137	1	889	1	0	29	18	89	12	3		43012050	2087	3	0	16	469	2	2	9	13	50	0	9	0	0	6	7	6	4	2	0	14	7	5	640	1	2	1	232	45
Barcelones	61	226	1080	32	5	586	806	206	73	186	52938		241613094	85	45	24	1040	14	69	563	6693	34	32	564	24	9	30	30	12	36	62	77	243	663	24	618	1	32	17		3855416455
Bergueda	1	4	18	9	0	15	760	5	1	7	62	4		50110802	70	0	16	0	5	10	21	0	3	278	0	3	1	2	0	1	36	4	9	5	90	6	0	0	1	90	32
Cerdanya	0	2	3	71	0	1	9	0	0	3	25	0	358		104354	0	0	0	2	32	13	0	5	10	0	0	1	1	0	0	22	3	5	20	6	1	0	0	0	63	10
Conca Barb.	395	2	14	0	1	64	1	110	8	1	37	9	237	1	0	5101	0	17	0	1	4	4	2	4	0	1	2	0	3	1	0	17	29	3	0	262	0	30	2	30	10
Garraf	17	5	855	3	2	44	33	39	5	0	1272	547	5563	6	6		1123481	0	2	13	51	3	2	6	1	0	2	1	2	7	122	0	14	9	0	127	0	1	2	328	135
Garrigues	14	0	7	5	1	10	11	20	0	4	28	8	173	0	0	71		44767	0	9	3	0	14	6	7	2	303	0	6	12	1	40	830	1	1	43	3	49	3	20	8
Garrotxa	0	148	3	1	0	1	5	3	0	53	22	0	262	1	4	0	0	0	16914	656	27	0	0	24	0	0	0	208	0	0	92	0	1	186	0	4	0	1	0	33	21
Girones	2	683	6	4	1	40	21	6	0	1065	150	6	1072	9	6	1	7	0		26544429	107	1	1	33	0	0	0	707	0	0	33	2	7	2620	0	19	0	1	2	180	101
Maresme	5	52	59	0	0	51	65	11	10	69	1577	8	25345	10	7	2	43	1	25		31981561	10	6	99	1	0	10	25	0	3	8	0	13	1908	1	61	0	4	4	2011	2505
Montsia	19	2	3	1	0	3	10	74	1468	2	24	12	252	0	0	2	4	0	1	8		415782	3	5	2	0	1	0	0	34	0	0	16	6	1	184	21	1	1	30	7
Noguera	3	3	2	40	8	39	10	9	2	5	45	5	387	3	3	7	1	16	0	8	8	0	9383	2	20	11	212	0	1	0	5	307	1265	8	22	26	0	277	14	36	19
Osona	5	18	12	3	3	26	229	5	4	22	184	1	1672	61	26	0	15	0	27	45	48	2		145904	0	1	2	12	0	2	307	8	15	98	9	18	0	0	1	256	1065
Pallars J.	1	0	4	13	7	2	4	5	0	4	23	1	268	1	4	0	0	0	1	3	14	2	24	0	3890	42	6	0	0	2	0	19	131	4	3	16	0	7	5	38	8
Pallars S.	3	1	1	10	0	1	7	4	0	1	24	3	198	1	1	0	3	0	3	5	5	0	10	3		481839	5	0	0	1	0	3	94	2	0	3	0	3	19	37	7
Pla d'Urgell	5	1	4	6	1	32	8	9	3	1	37	2	190	0	4	11	4	78	3	2	6	1	170	2	5	0	8507	1	0	1	4	208	866	3	1	18	0	401	7	27	17
Pla Estany	0	179	0	0	0	1	0	0	0	67	8	1	131	0	1	0	2	1	300	1268	11	0	0	8	0	1	0	7345	0	0	4	0	2	91	0	2	0	0	0	23	13
Priorat	7	1	3	0	0	3	3	245	11	0	9	7	120	0	0	2	3	27	1	2	5	2	0	1	0	1	0	0	2095	217	0	1	13	1	0	224	2	1	0	19	1
Ribera Ebre	7	0	7	1	0	0	1	175	102	5	10	8	163	0	2	3	7	3	0	0	4	9	4	0	0	3	0	0		876113	0	2	18	2	0	181	59	1	1	26	8
Ripolles	0	30	8	0	1	2	8	2	1	25	40	0	382	9	52	0	0	0	159	129	24	0	0	362	1	0	0	13	0	0	8980	0	2	18	0	6	0	0	0	75	59
Segarra	1	2	3	11	1	162	19	6	2	1	42	2	273	1	3	20	3	7	1	1	9	0	49	3	1	0	35	0	0	0		15895	117	0	26	3	0	198	1	45	7
Segria	11	10	21	61	18	61	38	52	12	8	155	12	1345	4	7	25	11	245	4	36	22	3	451	18	62	24	656	3	5	41	14		15755346	11	24	136	1	214	43	148	61
Selva	0	114	8	1	0	9	15	5	4	241	105	10	1393	10	3	1	5	1	167	3331	1608	0	1	87	0	0	1	62	0	1	5	4		1034022	3	11	0	0	3	284	700
Solsones	0	0	0	54	0	17	141	7	0	0	13	1	123	36	0	4	0	0	0	3	4	0	8	3	5	3	4	1	0	0	2	22	35		53871	1	0	1	4	28	8
Tarragones	1000	8	109	5	0	29	19	3918	101	11	287	688	2205	1	1	140	101	19	2	21	34	44	2	13	0	3	1	1	25	115	3	6	73	9		252114	11	9	1	210	70
Terra Alta	2	0	1	1	0	2	1	30	67	0	15	9	94	0	1	0	2	0	0	0	3	11	2	3	1	0	0	0	4	280	0	5	2	1	0		953434	2	0	8	1
Urgell	7	4	6	4	1	56	9	9	1	0	43	2	270	0	1	30	3	23	0	6	15	1	220	6	10	3	380	1	0	0	0	907	466	5	4	26	0	8858	5	41	17
Val d'Aran	0	2	0	3	7	0	1	0	0	0	9	0	77	0	1	0	2	0	0	1	2	0	3	1	6	4	1	0	0	0	0	0	53	1	0	4	0	0	2913	11	2
Valles Oc.	16	54	292	12	1	199	677	25	16	63	6364	76	37886	25	4	7	109	7	23	132	578	14	12	205	4	7	7	11	1	10	13	21	54	159	6	129	1	9		10194624	6843
Valles Or.	25	25	41	0	0	46	170	12	2	59	1417	16	15007	20	3	0	44	1	9	149	790	3	0	561	0	3	1	12	0	3	10	15	11	632	2	42	0	4	1		898982507

4.7.1 Moviments deguts al treball entre les 41 comarques catalanes

A les dades de la taula 4.8 els aplicarem l'anàlisi proposada a la secció 4.6. Per a ajudar a interpretar més bé els moviments, ho referirem al mapa adjunt de les 41 comarques catalanes (Figura 4.9).

Figura 4.9: Les 41 comarques catalanes. (Origen Web de l'Idescat. Disseny propi)

Anàlisi Migració Intra

Començarem amb l'estudi de la migració intra, és a dir la importància dels moviments de treballadors de la pròpia zona sobre els total d'efectius que arriben a la zona. En la Figura 4.10 trobem una representació gràfica d'aquests moviments.

Podem observar que, del total de moviments sobre cadascuna de les zones, trobem que com a mínim un 61.2% d'aquests moviments pertanyen a gent de la pròpia zona (Baix Llobregat) mentre que arriba a uns màxims de més d'un 90% a la Val d'Aran, Segrià, Osona, Alt Urgell, Alt Empordà i Garrotxa.

És a dir, molt més de la meitat de les migracions sobre un mateixa zona són internes. Es dóna la circumstància que les zones amb una migració Intra més altes es corresponen a zones més perifèriques i a zones nord, mentre que aquelles on la migració intra és més baixa són les que estan situades en zona d'influència de grans capitals, sobretot de Barcelona. Ho podem visualitzar globalment en la Figura 4.11.

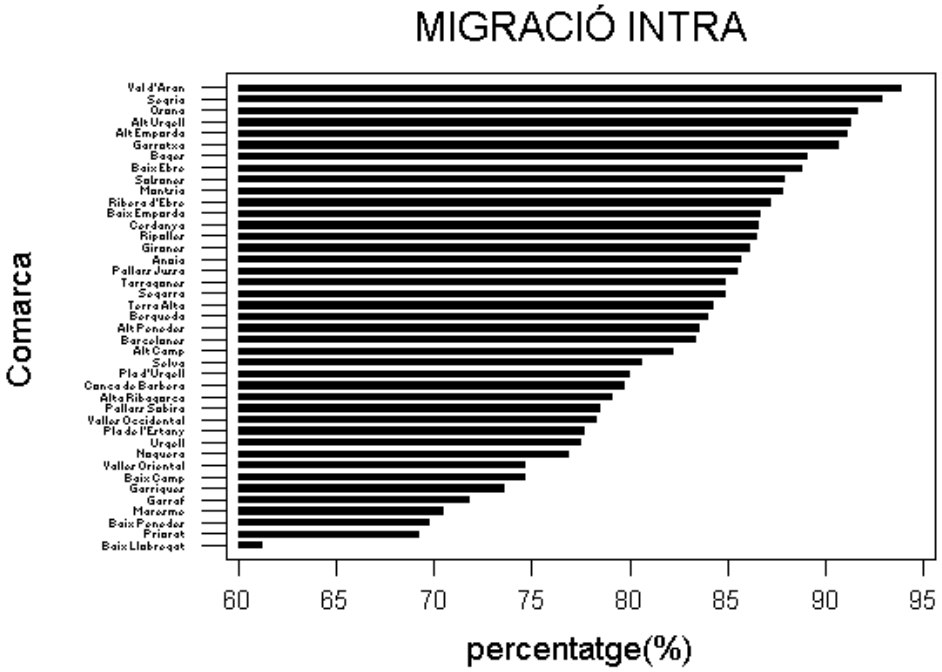


Figura 4.10: Anàlisi de la Migració Intra

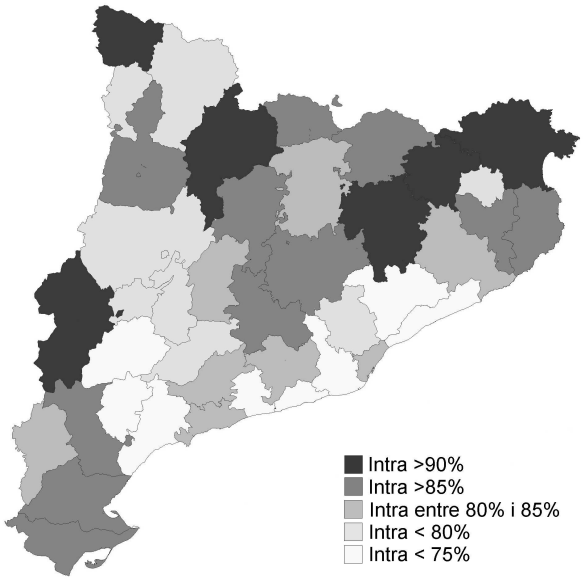


Figura 4.11: Mapa de la Migració Intra

Anàlisi Migració Inter i Recíproca

L'Anàlisi de la migració recíproca entre zones, comporta fer l'anàlisi de la matriu de moviments simètrics, però sense la influència dels valors de la diagonal, per la qual cosa es realitzarà una anàlisi tipus **3-EM** de la matriu de simetria. Podem observar en la Figura 4.12 la representació gràfica d'aquests moviments:

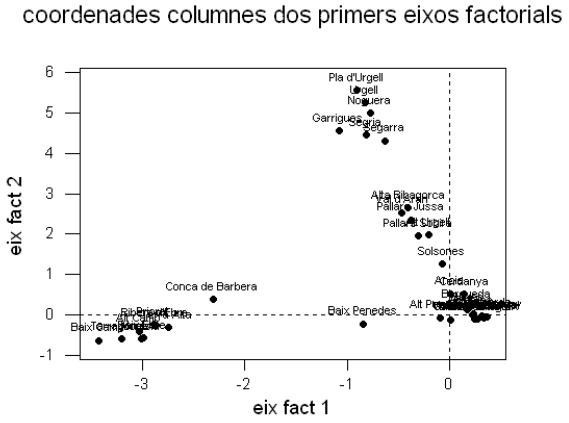


Figura 4.12: Gràfic del primer pla factorial de l'anàlisi de la Migració Inter i Recíproca

En l'anàlisi dels dos primers eixos factorials (qualitat=33.7%) podem trobar en el primer eix les zones on els intercanvis són importants situades totes elles a les comarques meridionals amb dues comarques de transició. Queden configurades a l'extrem de l'eix pel Baix Camp i el Tarragonès, diferenciades del Montsià i Baix Ebre, per una banda, i més cap a l'origen de l'eix Alt Camp, Priorat, Ribera d'Ebre i Terra Alta. Les comarques de transició són la Conca de Barberà i el Baix Penedès. En l'origen de l'eix hi ha un gran cúmul de punts.

Mentre que en el segon eix trobem un altre grup, format per les comarques de la regió occidental amb la mateixa estructura de dues zones i comarques de transició cap a l'origen. Garrigues, Noguera, Pla d'Urgell, Segarra, Segrià, i Urgell formarien la primera zona, i més propera a l'origen, una segona zona formada per Alt Urgell, Alta Ribagorça, els dos Pallars i la Val d'Aran, i de transició cap al gran centre el Solsonès i l'Anoia.

Més explícit és el gràfic amb els tres primers eixos factorials (qualitat=46.4%) (Figura 4.13) on es reprodueix l'estructura amb un tercer eix on les comarques nord-orientals ens repeteixes l'estructura zonal (una zona i comarques de transició) i es veu clarament que destaca la posició central del Barcelonès respecte els tres eixos independents de desplaçaments. L'anàlisi ens permet també elaborar un mapa de Catalunya classificat amb comarques amb gran relació entre i recíproca (Figura 4.14).

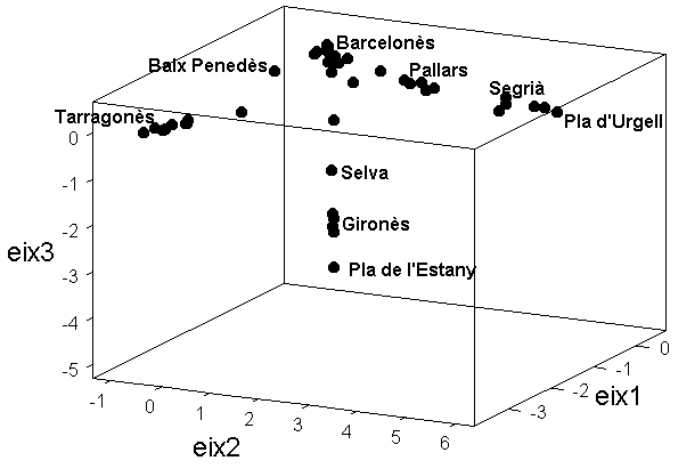


Figura 4.13: Gràfic 3-EM de l'Anàlisi de la Migració Inter i Recíproca

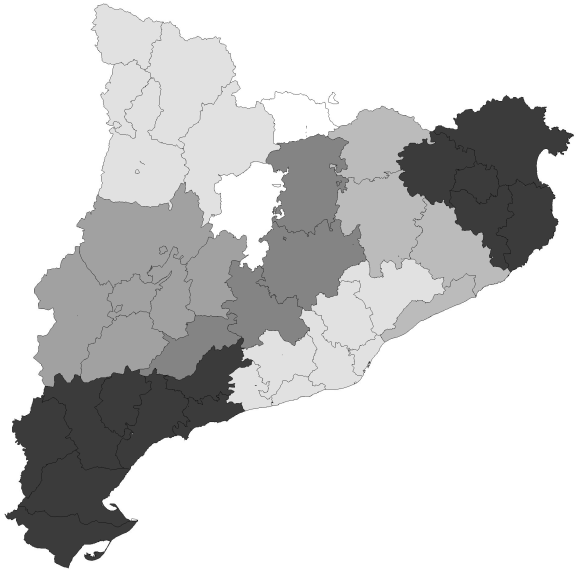


Figura 4.14: Mapa de les zones de Migració Inter i recíproca

Anàlisi dels Saldos Migratoris

Finalment l'estudi dels saldos migratoris, és a dir les diferències entre moviments rebuts d'altres zones i moviments vers altres zones. La realització d'aquest estudi comportarà l'estudi de la matriu corresponent a la no simetria.

En la Figura 4.15 podem trobar la representació gràfica d'aquests moviments on ens queden representats els orígens i les destinacions dels saldos migratoris.

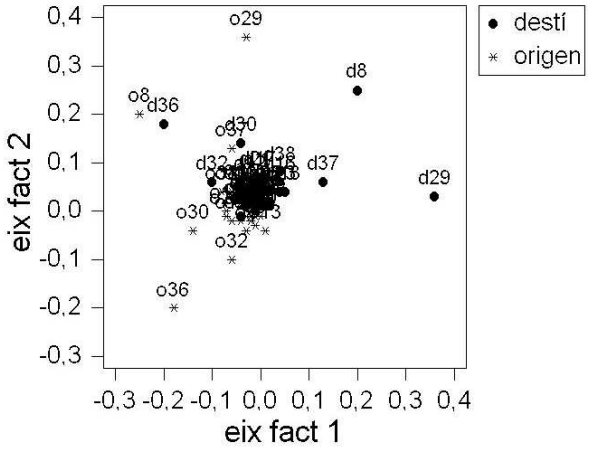


Figura 4.15: Anàlisi dels Saldos Migratoris

Podem veure en aquest darrer gràfic dels dos primers eixos factorials (Q=31.15%) les representacions de les relacions de saldos migratoris, podent comprovar primerament que donat el caràcter antisimètric de la matriu analitzada, les representacions dels individus com a productors de saldos positius o negatius, és a dir com a receptors o com a productors de moviments de treballadors, està subjecte a una rotació de 90 graus tal i com ja havíem indicat.

Per altra banda, podem veure relacions de saldos entre les comarques del Baix Camp i el Tarragonès, amb un gran saldo a favor d'aquesta darrera comarca. Uns moviments amb origen al Priorat i destins a l'Alt i Baix Camp i al Tarragonès i finalment d'origen a la Terra Alta i destí a la Ribera d'Ebre.

Es pot observar que aquestes zones amb grans diferències de saldos migratoris òbviament corresponen a zones que no tenen una gran migració intra, ja que s'ha de complir que la migració interna sigui baixa i a més no tinguin una gran relació simètrica amb una altra comarca, sinó més aviat tinguin un pol atractor local.

El tercer eix factorial ens aporta unes inter-relacions triangulars entre les Garrigues, la

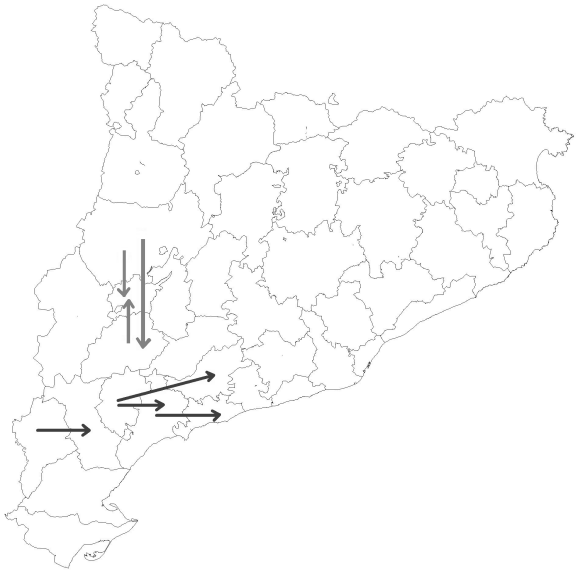


Figura 4.16: Representació dels principals Saldos Migratoris

Noguera i el Pla d'Urgell. Aquestes relacions les podem trobar representades en la Figura 4.16.

Totes aquestes relacions, que hem descrit corresponents a orígens i destinacions, es podrien trobar també construint triangles sobre el gràfic factorial. Per fer-ho, hauríem d'escollir sobre el gràfic factorial aquells triangles que ens donessin unes àrees més grans, formats escollint o bé només els orígens o bé només les destinacions. Aquests triangles els podem visualitzar fàcilment unint els punts de la Figura 4.15.

4.7.2 Condicions de vida i aspiracions dels francesos

En aquest segon exemple les dades corresponen a una enquesta socioeconòmica referida a *Les condicions de vida i aspiracions dels francesos* [LHvE80]. Nosaltres treballarem només amb les variables referides a sexe, nivell d'educació (1-cap, 2-primària, 3-secundària 4-batxillerat 5-universitat), allotjament(1-hipoteca, 2-propietari, 3-llogater 4-allotjament gratuït), propietat d'accions (1-si, 2-no), propietat de béns immobles(1-si, 2-no), edat categoritzada (1-menor25, 2-25:34, 3-35:49,4-50:64, 5-65 o més) i mida de la població (1-menys 2000, 2000:50000, 3-50000:100000, 4-100000:500000, 5- més 500000)

A aquestes dades farem una reconstitució dels blocs de caselles de la diagonal de la taula de Burt, mitjançant la fórmula de reconstitució 2-EM. Podem trobar les dades a la Taula 4.12. Si calculem les inèrcies de cadascuna de les subtaules de la taula de Burt

(Taula 4.9), podem comprovar que la descomposició de la inèrcia total en les parts degudes a la diagonal i a la no diagonal són, en aquest exemple, iguals a :

$$\mathcal{I}_{Total} = \mathcal{I}_D + \mathcal{I}_{\neg D} = 0.401911 = 0.367347 + 0.0345644$$

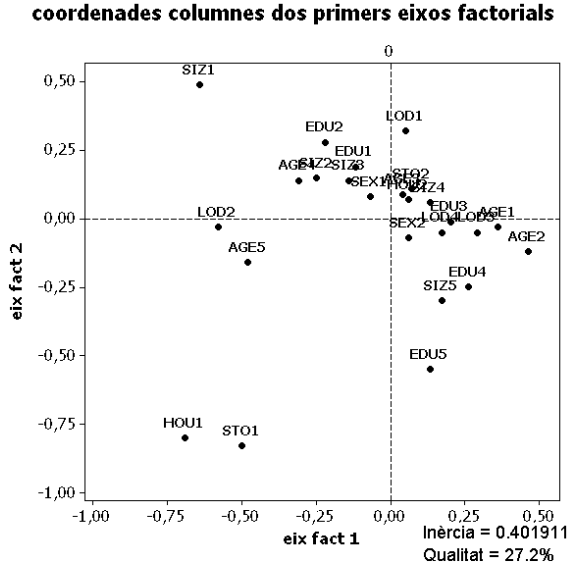
és a dir, més de un 90.4% de la inèrcia es troba en els blocs de la diagonal. Això, des del punt de vista geomètric, farà que hi hagi una tendència de portar tots els punts en els gràfics factorials cap a fora, com podrem veure comparant els gràfics factorials de les figures 4.17 i 4.18.

Taula 4.9: *Inèrcies per cada subtaula de contingència de la Taula de Burt*

	Sexe	Educació	Allotjam	Accions	Propietat	Edat	MidaMun
Sexe	1	0.0087	0.0083	0.0003	0.0006	0.0023	0.0018
Educació	0.0087	4	0.0232	0.0357	0.0091	0.1479	0.0715
Allotjam	0.0083	0.0232	3	0.0274	0.0374	0.1454	0.1426
Accions	0.0003	0.0357	0.0274	1	0.1056	0.0290	0.0081
Propietat	0.0006	0.0091	0.0374	0.1056	1	0.0103	0.0042
Edat	0.0023	0.1479	0.1454	0.0290	0.0103	4	0.0274
MidaMun	0.0018	0.0715	0.1426	0.0081	0.0042	0.0274	4

L'anàlisi de correspondències múltiples de la Taula de Burt (Taula 4.10) ens dóna el gràfic factorial de la Figura 4.17

Figura 4.17: *ACM de la taula de Burt corresponent a les dades de la taula 4.10*

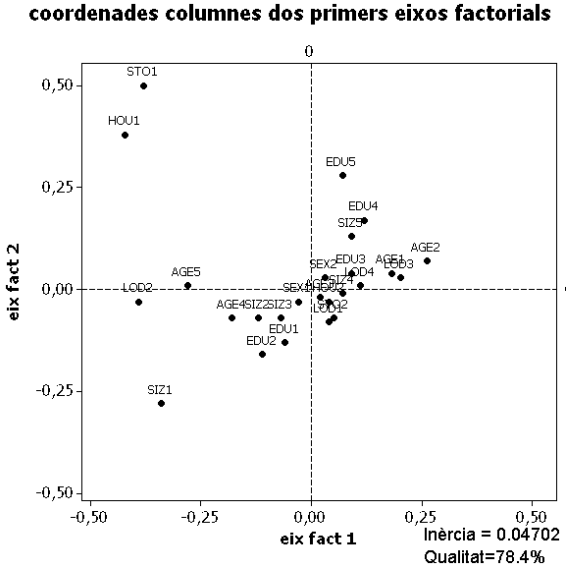


Taula 4.10: *Taula de Burt de les 7 variables de les condicions de vida i aspiracions dels francesos*

	SEXM	SEXF	ED1	ED2	ED3	ED4	ED5	LO1	LO2	LO3	LO4	ST1	ST2	HO1	HO2	AG1	AG2	AG3	AGE	AG5	SI1	SI2	SI3	SI4	SI5
SEXM	469	0	102	164	69	65	69	62	151	224	32	54	415	35	434	18	111	169	84	87	42	40	81	161	145
SEXF	0	531	94	163	91	102	81	58	142	302	29	67	464	47	484	22	136	187	104	82	41	47	94	168	181
EDU1	102	94	196	0	0	0	0	17	58	106	15	11	185	10	186	9	26	55	45	61	19	16	36	70	55
EDU2	164	163	0	327	0	0	0	45	116	151	15	27	300	21	306	0	57	124	82	64	43	35	67	109	73
EDU3	69	91	0	0	160	0	0	16	37	98	9	19	141	14	146	14	47	60	27	12	12	13	33	52	50
EDU4	65	102	0	0	0	167	0	27	40	86	14	28	139	20	147	16	62	63	17	9	8	15	21	56	67
EDU5	69	81	0	0	0	0	150	15	42	85	8	36	114	17	133	1	55	54	17	23	1	8	18	42	81
LOD1	62	58	17	45	16	27	15	120	0	0	0	11	109	7	113	3	23	68	20	6	7	20	27	48	18
LOD2	151	142	58	116	37	40	42	0	293	0	0	60	233	48	245	9	24	91	80	89	59	33	62	72	67
LOD3	224	302	106	151	98	86	85	0	0	526	0	45	481	23	503	22	180	182	79	63	11	29	80	191	215
LOD4	32	29	15	15	9	14	8	0	0	0	61	5	56	4	57	6	20	15	9	11	6	5	6	18	26
STO1	54	67	11	27	19	28	36	11	60	45	5	121	0	39	82	2	18	35	27	39	4	9	22	36	50
STO2	415	464	185	300	141	139	114	109	233	481	56	0	879	43	836	38	229	321	161	130	79	78	153	293	276
HOU1	35	47	10	21	14	20	17	7	48	23	4	39	43	82	0	3	11	27	20	21	7	12	12	25	26
HOU2	434	484	186	306	146	147	133	113	245	503	57	82	836	0	918	37	236	329	168	148	76	75	163	304	300
AGE1	18	22	9	0	14	16	1	3	9	22	6	2	38	3	37	40	0	0	0	0	2	4	8	15	11
AGE2	111	136	26	57	47	62	55	23	24	180	20	18	229	11	236	0	247	0	0	0	13	16	34	102	82
AGE3	169	187	55	124	60	63	54	68	91	182	15	35	321	27	329	0	0	356	0	0	31	34	71	100	120
AGE4	84	104	45	82	27	17	17	20	80	79	9	27	161	20	168	0	0	0	188	0	25	20	26	60	57
AGE5	87	82	61	64	12	9	23	6	89	63	11	39	130	21	148	0	0	0	0	169	12	13	36	52	56
SIZ1	42	41	19	43	12	8	1	7	59	11	6	4	79	7	76	2	13	31	25	12	83	0	0	0	0
SIZ2	40	47	16	35	13	15	8	20	33	29	5	9	78	12	75	4	16	34	20	13	0	87	0	0	0
SIZ3	81	94	36	67	33	21	18	27	62	80	6	22	153	12	163	8	34	71	26	36	0	0	175	0	0
SIZ4	161	168	70	109	52	56	42	48	72	191	18	36	293	25	304	15	102	100	60	52	0	0	0	329	0
SIZ5	145	181	55	73	50	67	81	18	67	215	26	50	276	26	300	11	82	120	57	56	0	0	0	0	326

Després de fer la reconstitució 2-EM, obtenim la taula de dades Taula 4.12 i el gràfic 4.18 és el gràfic del primer pla factorial de les dades reconstituïdes

Figura 4.18: *ACM de la taula de Burt corresponent a les dades de la taula 4.12 on s'han reconstituït els elements de les subtaules de la diagonal*



La descomposició de la inèrcia total en suma de subtaules de la diagonal i de fora de la diagonal queda, després de la reconstitució, com $\mathcal{I}_{Total} = \mathcal{I}_D + \mathcal{I}_{\neg D} = 0.0470249 = 0.0124605 + 0.0345644$ i la descomposició d'aquestes inèrcies la podem trobar a la taula 4.11.

Taula 4.11: *Inèrcies per cada subtaula de contingència de la Taula de Burt reconstituïda*

	Sexe	Educació	Allotjam	Accions	Propietat	Edat	MidaMun
Sexe	0.0002	0.0087	0.0083	0.0003	0.0006	0.0023	0.0019
Educació	0.0087	0.0834	0.0232	0.0357	0.0091	0.1480	0.0715
Allotjam	0.0083	0.0232	0.1890	0.0274	0.0374	0.1454	0.1426
Accions	0.0003	0.0357	0.0274	0.1819	0.1056	0.0290	0.0081
Propietat	0.0006	0.0091	0.0374	0.1056	0.0468	0.0103	0.0042
Edat	0.0023	0.1479	0.1454	0.0290	0.0103	0.0650	0.0274
MidaMun	0.0019	0.0715	0.1426	0.0081	0.0042	0.0274	0.0444

Taula 4.12: *Taula de Burt reconstituïda aplicant l'algorisme 2-EM als blocs diagonals*

	SEXM	SEXF	ED1	ED2	ED3	ED4	ED5	LO1	LO2	LO3	LO4	ST1	ST2	HO1	HO2	AG1	AG2	AG3	AGE	AG5	SI1	SI2	SI3	SI4	SI5
SEXM	223	246	102	164	69	65	69	62	151	224	32	54	415	35	434	18	111	169	84	87	42	40	81	161	145
SEXF	246	285	94	163	91	102	81	58	142	302	29	67	464	47	484	22	136	187	104	82	41	47	94	168	181
EDU1	102	94	45	78	29	25	19	17	58	106	15	11	185	10	186	9	26	55	45	61	19	16	36	70	55
EDU2	164	163	78	137	46	38	28	45	116	151	15	27	300	21	306	0	57	124	82	64	43	35	67	109	73
EDU3	69	91	29	46	27	30	27	16	37	98	9	19	141	14	146	14	47	60	27	12	12	13	33	52	50
EDU4	65	102	25	38	30	37	37	27	40	86	14	28	139	20	147	16	62	63	17	9	8	15	21	56	67
ED5	69	81	19	28	27	37	38	15	42	85	8	36	114	17	133	1	55	54	17	23	1	8	18	42	81
LOD1	62	58	17	45	16	27	15	15	32	65	7	11	109	7	113	3	23	68	20	6	7	20	27	48	18
LOD2	151	142	58	116	37	40	42	32	172	76	13	60	233	48	245	9	24	91	80	89	59	33	62	72	67
LOD3	224	302	106	151	98	86	85	65	76	348	37	45	481	23	503	22	180	182	79	63	11	29	80	191	215
LOD4	32	29	15	15	9	14	8	7	13	37	4	5	56	4	57	6	20	15	9	11	6	5	6	18	26
STO1	54	67	11	27	19	28	36	11	60	45	5	60	61	39	82	2	18	35	27	39	4	9	22	36	50
STO2	415	464	185	300	141	139	114	109	233	481	56	61	818	43	836	38	229	321	161	130	79	78	153	293	276
HOU1	35	47	10	21	14	20	17	7	48	23	4	39	43	23	59	3	11	27	20	21	7	12	12	25	26
HOU2	434	484	186	306	146	147	133	113	245	503	57	82	836	59	859	37	236	329	168	148	76	75	163	304	300
AGE1	18	22	9	0	14	16	1	3	9	22	6	2	38	3	37	2	13	15	6	5	2	4	8	15	11
AGE2	111	136	26	57	47	62	55	23	24	180	20	18	229	11	236	13	91	90	30	22	13	16	34	102	82
AGE3	169	187	55	124	60	63	54	68	91	182	15	35	321	27	329	15	90	128	66	58	31	34	71	100	120
AGE4	84	104	45	82	27	17	17	20	80	79	9	27	161	20	168	6	30	66	44	42	25	20	26	60	57
AGE5	87	82	61	64	12	9	23	6	89	63	11	39	130	21	148	5	22	58	42	43	12	13	36	52	56
SIZ1	42	41	19	43	12	8	1	7	59	11	6	4	79	7	76	2	13	31	25	12	17	10	19	24	13
SIZ2	40	47	16	35	13	15	8	20	33	29	5	9	78	12	75	4	16	34	20	13	10	9	17	27	24
SIZ3	81	94	36	67	33	21	18	27	62	80	6	22	153	12	163	8	34	71	26	36	19	17	33	56	50
SIZ4	161	168	70	109	52	56	42	48	72	191	18	36	293	25	304	15	102	100	60	52	24	27	56	111	110
SIZ5	145	181	55	73	50	67	81	18	67	215	26	50	276	26	300	11	82	120	57	56	13	24	50	110	129

4.8 Conclusions

En aquest capítol s'ha fet un estudi de la descomposició de la inèrcia en correspondències simples i correspondències múltiples i s'han vist les principals problemàtiques associades a la diagonal en ACS de matrius simètriques i subtaules de la diagonal en ACM, així com diverses de les propostes per afrontar aquesta problemàtica dins de l'àmbit de les anàlisis factorials.

En base a les metodologies de tractament dins de l'àmbit de les anàlisis factorials, s'ha desenvolupat una nova proposta basada en la descomposició en la part simètrica i l'antisimètrica i la reconstitució k-EM de la part simètrica, després d'analitzar la influència de les diagonals. Aquesta proposta d'anàlisi dóna una interpretació força acurada de les relacions entre les caselles de la taula de contingència objecte de l'anàlisi.

Aquesta mateixa reconstitució k-EM s'ha aplicat a la reconstitució de les taules de la diagonal, millorant-ne la representació factorial i obtenint uns resultats equivalents al que es coneix com Joint Correspondence Analysis.

Capítol 5

ACM respecte a un model i ACM condicional

5.1 Introducció

En aquest capítol introduïm la teoria i les notacions necessàries així com la formulació de les ACM respecte un model i ACM condicional, per tal de posteriorment mirar d'introduir les anàlisis multicondicionals a partir del desenvolupament de les anàlisis condicionals. Podrem veure que l'anàlisi condicional és un cas particular de l'anàlisi en referència a un model però hem cregut convenient fer les dues presentacions, ja que algunes de les metodologies estudiades per redefinir inèrcies treballaven amb anàlisis respecte a un model determinat.

5.2 ACM respecte a un model

En aquesta secció s'introdueix l'anàlisi factorial d'una taula de dades respecte a una taula model donada. Es presenta seguint la proposta de B. Escofier [Esc84] la definició dels núvols d'individus i variables, la matriu de pesos i la matriu de distàncies, que ens donen els elements de la nostra anàlisi. També es presenta l'anàlisi de correspondències ordinària com una anàlisi respecte al model d'independència.

5.2.1 Metodologia

En aquesta secció introduïm les notacions, la definició dels núvols de punts i les mètriques de l'anàlisi de correspondències múltiples en referència a un model.

Notacions

Sigui $K_{IJ} = \{K_{ij}; i \in I, j \in J\}$ una taula de nombre positius amb I fileres i J columnes. Sigui $M_{IJ} = \{M_{ij}; i \in I, j \in J\}$ una taula model de les mateixes dimensions que la taula K_{IJ}. Aquest model pot provenir del producte dels marges (supòsit d'independència), productes dels marges modificats, a partir de terceres variables,...

Siguin $P_I = \{P_i; i \in I\}$ i $Q_J = \{Q_j; j \in J\}$ dues taules de nombres positius de dimensions respectives I i J. Aquestes taules poden ser les marginals de K_{IJ}, combinacions lineals de les marginals de K_{IJ} i M_{IJ}, valors exteriors, ...

Notem respectivament amb minúscules k_{IJ}, m_{IJ}, p_I i $q_J$les taules obtingudes a partir de dividir les anteriors, K_{IJ}, M_{IJ}, P_I i Q_J, per la suma dels seus respectius totals.

Notem també k_I, k_J, m_I i m_J les marginals de k_{IJ} i m_{IJ}:

$$\begin{array}{ll} k_{i.} = \sum_j k_{ij} & k_{.j} = \sum_i k_{ij} \\ m_{i.} = \sum_j m_{ij} & m_{.j} = \sum_i m_{ij} \end{array}$$

Núvol de punts

Al conjunt de fileres de I associem un núvol de punts, notat $N(I)$, núvol dins l'espai $\mathbb{R}^J$, de la següent manera:

$$\frac{k_{iJ}}{p_i} - \frac{m_{iJ}}{p_i} = \left\{ \frac{k_{ij}}{p_i} - \frac{m_{ij}}{p_i}; j \in J \right\}$$

Si $p_I = k_I = m_I$ llavors $\frac{k_{iJ}}{p_i}$ i $\frac{m_{iJ}}{p_i}$, els perfils filera de la taula de dades i de la taula model respectivament, poden ser considerats com a mesures de probabilitat.

L'espai $\mathbb{R}^J$, està dotat de la mètrica χ^2 de centre q_J. La distància entre dos punts serà:

$$D^2(i, i') = \sum_j \left\{ \frac{k_{ij} - m_{ij}}{p_i} - \frac{k_{i'j} - m_{i'j}}{p'_i} \right\}^2 \frac{1}{q_j}$$

Dos punts i i i' seran propers si les desviacions entre la filera en la taula de dades i la filera en el model ponderades per p_i, pes associat al punt i, són semblants per a tot j.

El núvol J es defineix simètricament.

Les coordenades dels centres de gravetat dels núvols $N(I)$ i $N(J)$, notades ω_I i ω_J són:

$$\omega_I(j) = \sum_i p_i \frac{k_{ij} - m_{ij}}{p_i} = k_j - m_j$$

$$\omega_J(i) = \sum_j q_j \frac{k_{ij} - m_{ij}}{p_j} = k_i - m_i$$

Si els marges en J de la taula i el model són iguals entre ells, el núvol de línies és centrat. De la mateixa manera, si els marges en I són iguals, el núvol de columnes és centrat.

Les distàncies entre els punts del núvol seran ben representades per les seves projeccions en els primers eixos factorials si el centre de gravetat és proper a l'origen. Caldrà, doncs, que els marges del model no difereixin massa dels marges de la taula de dades.

Factors de l'anàlisi

La tripleta (X,D,M) objecte d'estudi és en aquest cas:

- X matriu de coordenades dels punts del núvol. $x_{ij} = \frac{k_{ij} - m_{ij}}{p_i}$
- D matriu diagonal dels pesos dels elements. $d_{ii} = p_i$
- M matriu diagonal de la mètrica de l'espai. $m_{jj} = \frac{1}{q_{jj}}$

I la matriu $X'DXM$ objecte d'anàlisi té per terme general en $\mathbb{R}^J$:

$$a_{jj'} = \sum_i \frac{k_{ij} - m_{ij}}{p_i} \frac{k_{ij'} - m_{ij'}}{q_j}$$

Es pot mostrar fàcilment la dualitat de les anàlisis dels núvols $N(I)$ i $N(J)$.

Cas particular: l'Anàlisi de Correspondències

Considerem el model on m_{ij} és igual a producte dels marges de k_{ij}: $m_{ij} = k_{i.}k_{.j}$, corresponent a l'hipòtesi d'independència:tenim $m_I = k_I$ i $m_J = k_J$. Si, a més imposem $p_I = k_I$ i $q_J = k_J$, llavors aquest model ens porta exactament a l'anàlisi de correspondències de la taula k_{IJ}.

En efecte, un punt vindrà representat dins $\mathbb{R}^J$ per les coordenades:

$$\left\{ \frac{k_{ij}}{k_{i.}} - k_{.j}; j \in J \right\}$$

En el cas més general, on tinguem solament $m_I = k_I = p_I$ i $m_J = k_J = q_J$, els resultats es poden obtenir per l'AFC clàssic de la taula de terme general

$$k_{ij} - m_{ij} + k_{i.}k_{.j} \tag{5.1}$$

Anàlisi que resulta ser l'anàlisi factorial de correspondències de **les dades menys el model més el producte de les marginals**, marginals comunes a les dades i el model.

5.3 ACM condicionat per una variable

En aquesta secció introduïm l'ACM condicionat com a pas previ de l'ACM multicondicional. En ella veurem com la necessitat de controlar la influència de variables externes en l'anàlisi de relacions entre variables, ens porta a definir en primer lloc les anàlisis condicionades i posteriorment les multicondicionades.

5.3.1 Introducció

El propòsit de l'ACM és d'una banda estudiar els lligams entre diverses variables qualitatives definides en una població donada i d'una altra estudiar l'estructura induïda per aquestes variables sobre la població.

En ACM s'analitza la desviació de cada perfil respecte al perfil mitjà (hipòtesi d'independència):

- A $\mathbb{R}^n \rightarrow \frac{k_{ij}}{nQ} - \frac{k_j}{nQ}$
- A $\mathbb{R}^I \rightarrow \frac{k_{ij}}{k_j} - \frac{1}{n}$

Però, a vegades, totes les variables o la majoria d'elles estan lligades a una variable qualitativa denotada per T, que pot ésser tractada com una *evolució en el temps*. Aquest cas fou àmpliament estudiat i tractat per B. Escofier [Esc87], mostrant que en aquest cas una ACM clàssica principalment mostraria aquests lligams, fins i tot si la variable T no han estat inclosa en el conjunt de les dades objecte de l'anàlisi.

B. Escofier proposà l'anomenada **anàlisi de correspondències múltiples condicional** [Esc87] d'acord a l'eliminació de la influència de la variable T en el temps, ja que estem interessats en lligams estables en el temps i no en la evolució temporal d'aquests lligams.

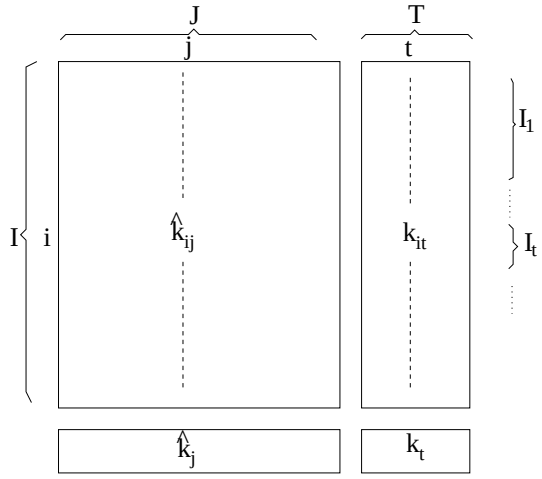
El problema és trobar una anàlisi del tipus 'anàlisi de correspondències múltiples' condicionat en relació a la variable T externa de les variables tractades, que respecti les principals de les propietats de les ACM: construcció i projecció dels dos núvols de punts, dualitat i fórmula de transició entre les projeccions dels dos núvols i l'equivalència amb l'anàlisi d'una taula de Burt.

5.3.2 Metodologia

Notacions

Notem I la població que comprèn n individus; J el conjunt de totes les modalitats de les Q variables qualitatives; i T, la variable qualitativa de les qual en volem treure'n la influència, T ens indicarà la variable pròpiament o el conjunt de les seves modalitats. La variable T defineix una partició d'I on les classes són denotades I_t on t és una modalitat de la variable T

Les dades poden ésser codificades per dues taules disjuntives completes: l'una creua I i J i l'altra creua I i T. La Figura 5.1 ens resumeix les notacions.



$\hat{k}_{ij} = 1$ or 0

$k_{it} = 1$ or 0

$Q = \sum_{j \in J} \hat{k}_{ij} \quad \forall i$

n= card I

$\hat{k}_j = \sum_{i \in I} \hat{k}_{ij}$

$k_t = \sum_{i \in I} k_{it} =$ card I_t

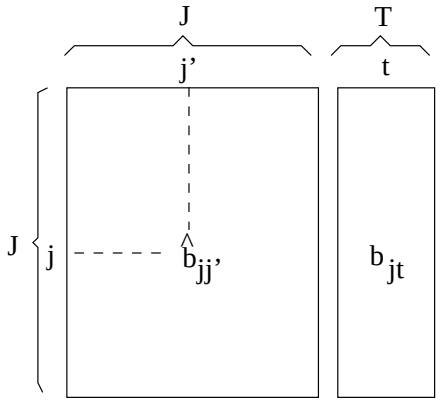
Figura 5.1: *Taula de notacions disjuntiva*

Podem trobar l'informació de la taula $I * J$ condensada en la taula de Burt associada que té per terme general $\hat{b}_{jj'}$. També podem considerar la banda que creua J i la variable de condicionament T, en la qual notem per b_{jt} el terme general i a la qual ens referirem més endavant com a *banda de Burt*. La Figura 5.2 ens resumeix les notacions de la taula de Burt.

Sigui S una mètrica diagonal d'un espai euclidià definida per la inversa del pes de cada modalitat $\hat{k}_j$ sobre la massa total.

$$s_{jj} = \frac{nQ}{\hat{k}_j}$$

Siguin m_t els pesos de les modalitats, definits per la proporció total de cada modalitat sobre la quantitat d'individus $m_t = k_t/n$.



$$\hat{b}_{jj'} = \sum_{i \in I} \hat{k}_{ij} \hat{k}_{ij'}$$

$$b_{jt} = \sum_{i \in I} \hat{k}_{ij} k_{it} = \sum_{i \in I_t} \hat{k}_{ij}$$

$$\sum_j b_{jt} = k_t Q$$

$$\sum_t b_{jt} = \hat{b}_{jj} = \hat{k}_j$$

Figura 5.2: *Taula de notacions disjuntives de Burt*

Escofier va demostrar que existeix una solució única, que satisfà les condicions requerides anteriorment, tant en l'estudi del condicionament per T del núvol dels individus, del núvol de les modalitats i de la taula de Burt.

Anem a estudiar els núvols d'individus, de modalitats i la taula de Burt.

Núvol dels individus

Tenim el núvol de punts a R^J dividit en T subnúvols:

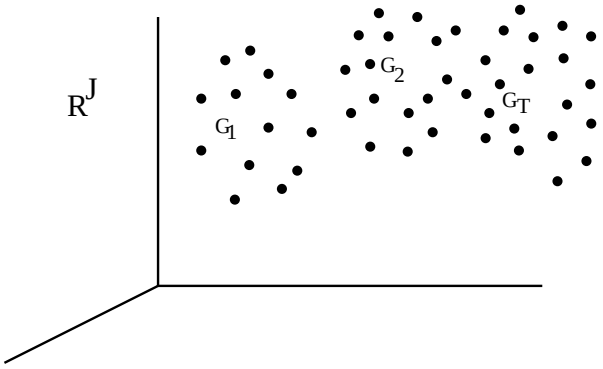


Figura 5.3: *Representació il·lustrativa de la descomposició basada en la partició*

L'anàlisi del núvol global reflexarà dues components de variabilitat:

- La dispersió *inter* períodes
- La dispersió *intra* períodes

En el núvol dels individus $N(I)$, T defineix una partició en subnúvols: a la modalitat t de T correspon el subnúvol de la població qüestionat en l'instant t. Seguint la hipòtesi inicial, si les respostes varien molt en el temps, els subnúvols I_t serien molt allunyats

els uns dels altres. Per eliminar dins del núvol $N(I)$ la dispersió deguda al temps, serà suficient recentrar cadascun dels subnúvols I_t segons el seu origen; és a dir representem un individu, no pas per la seva desviació de la mitjana general, sinó per la seva desviació de la mitjana del subgrup que ha estat qüestionat al mateix instant t que ell.

La mitjana del subnúvol $N(I_t)$ dins l'espai $\mathbb{R}^J$ té en l'eix j la coordenada :

$$\frac{1}{k_t}\sum_{i\in I_t}\frac{k_{ij}}{Q}=\frac{1}{Qk_t}\sum_{i\in I_t}k_{ij}=\frac{b_{jt}}{Qk_t}$$

Per tant l'individu $i\in I_t$, ja centrat, serà representat en $\mathbb{R}^J$ pel punt de coordenades:

$$\frac{k_{ij}}{Q}-\frac{b_{jt}}{Qk_t}=\frac{1}{Q}\left(k_{ij}-\frac{b_{jt}}{k_t}\right)$$

i dos individus seran propers si la seva diferència amb la mitjana del seu grup és anàloga.

En comptes de centrar cada subnúvol, es podria proposar projectar ortogonalment segons la direcció de l'evolució, ja sigui eliminant el primer factor lligat al temps, ja sigui considerant el subespai lligat als centre de gravetat. Això pot portar a eliminar una estructura estable en el temps, quan tots ells evolucionen en el temps segons la direcció de més gran dispersió.

Núvol de les modalitats

Ens cal ara en el les distàncies del núvol de les modalitats, $N(J)$, eliminar la part corresponent al lligam entre les qüestions i la variable T. Aquesta part apareixerà en la projecció del núvol $N(J)$ en el subespai engendrat per les modalitat de T, subespai que notarem E_T, i on els vectors indicatrius de les modalitats e_t són ortogonals.

$$e_t=\left(\ 00001111110000\dots\ \right)$$
$$\parallel e_t\parallel=k_t$$

La projecció del perfil $\left(\frac{k_{ij}}{k_j}\right)_i$ de la modalitat j sobre E_T es pot escriure en funció dels e_t ortogonals.

Si anomenem $u=\left(\frac{k_{ij}}{k_j}\right)_i$ i com $u-pr(u)$ és ortogonal a $e_i,\forall i$, llavors

$$pr(u)=\alpha_1e_1+\alpha_2e_2+\ldots+\alpha_te_t$$

$$<u-\alpha_1e_1-\ldots-\alpha_te_t,e_i>=0,\forall i\Rightarrow<u,e_i>-\alpha_i<e_i,e_i>=0,\forall i$$

d'aquí obtenim:

$$\alpha_i = \frac{< u, e_i >}{< e_i, e_i >} = \frac{\sum_i \frac{k_{ij}}{k_j} k_{it}}{\sum_i k_{it} k_{it}} = \frac{\sum_i \frac{k_{ij}}{k_j} k_{it}}{k_t} = \sum_i \frac{k_{ij}}{k_j} \frac{k_{it}}{k_t} = \frac{1}{k_j k_t} \sum k_{ij} k_{it} = \frac{b_j t}{k_j k_t}$$

Per tant:

$$pr(u) = \sum_t \frac{b_{jt}}{k_j k_t} e_t$$

La coordenada, doncs, sobre l'eix t és igual a $\frac{b_{jt}}{k_j k_t}$ on t és la classe a la qual pertany i.

La distància en E_T entre les projeccions de j i de j' val:

$$D^2(\text{proj } j, \text{proj } j') = \sum_t \sum_{i \in I_t} \left((\frac{b_{jt}}{k_j k_t}) - (\frac{b_{j't}}{k_{j'} k_t}) \right)^2 n = \sum_t \left((\frac{b_{jt}}{k_j}) - (\frac{b_{j't}}{k_{j'}}) \right)^2 \frac{n}{k_t}$$

Que és la distància χ^2 entre dos perfils de j i j' en la taula $J * T$, és a dir la distància induïda per T.

Resumint, per eliminar la part de la distància induïda per T, només cal projectar el núvol $N(J)$ sobre l'ortogonal d'E_T. La columna j serà llavors representada pel punt de coordenades:

$\frac{1}{k_j}(k_{ij} - \frac{b_{jt}}{k_t})$ si $i \in I_t$

Notem l'analogia que es té amb la definició de les correlacions parcials per a variables numèriques.

Anàlisi de correspondències i taula model

L'expressió del condicionament per la variable T ha portat a la construcció de dos nous núvols d'individus i de modalitats. Les coordenades dels punts en aquests dos núvols són:

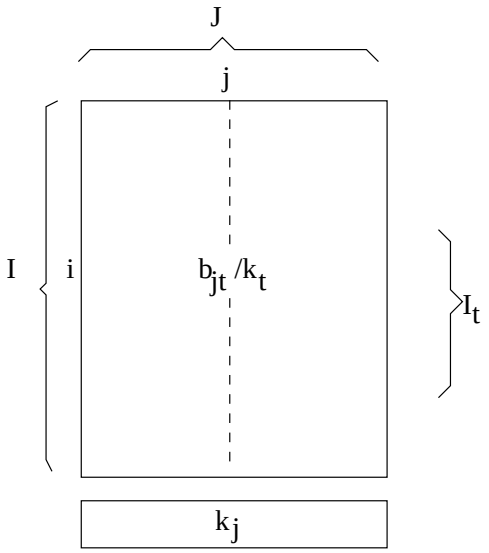
La coordenada j de $i \in I_t$ en $\mathbb{R}^J$ és:

$$\frac{1}{Q}(k_{ij} - \frac{b_{jt}}{k_t})$$

La coordenada $i \in I_t$ de j dins $\mathbb{R}^I$ és:

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

L'expressió $\frac{b_{jt}}{k_t}$ apareix en els dos núvols, considerem per tant la taula de dimensions $I * J$ on la filera i conté el perfil $\frac{b_{jt}}{k_t}$ de la classe I_t de l'element i. Els marges d'aquesta taula són els mateixos que els de la taula disjuntiva completa.



$$\sum_j \frac{b_{jt}}{k_t} = Q\frac{k_t}{k_t} = Q$$

$$\sum_i \frac{b_{jt}}{k_t} = \sum_t \sum_{i\in I_t} \frac{b_{jt}}{k_t} =$$

$$= \sum_t b_{jt} = \sum_{i\in I} k_{ij} = k_j$$

El perfil de la filera $i \in I_t$ d'aquesta taula és $\frac{b_{jt}}{Qk_t}$ així com el de la columna j és $\frac{b_{jt}}{k_j k_t}$. Les coordenades dels punts dels núvols de les fileres i de les columnes obtinguts no són altres que les diferències dels perfils de les fileres (columnes respectivament) de la taula disjuntiva completa i les d'aquesta taula.

Podem considerar dins l'anàlisi condicional aquesta taula, que ens serveix de referència, com a una taula model. L'anàlisi condicionada és l'anàlisi d'aquesta taula en referència a aquest model i podem aplicar la teoria de les AC respecte un model [Esc84], puix que tenen ambdues taules els mateixos marges. Haurem de realitzar una AFC la taula formada per *dades - model + producte de marges*

$$k^*_{ij} = k_{ij} - \frac{b_{jt}}{k_t} + \frac{k_j}{n} \text{ si } i \in I_t$$

On la taula k^*_{ij} té els mateixos marges que la taula k_{ij},i on les mètriques són les mateixes.

Aquesta taula, que no és pas una taula disjuntiva completa, tradueix la situació on el lligam entre I i J és degut només a l'evolució temporal. Acumulant les fileres (idèntiques) de la mateixa classe hom obté la banda $J * T$.

Obtenim així una anàlisi **inter** en el sentit que en el núvol d'individus només es té en compte la inèrcia inter, la inèrcia **intra** dins les classes I_t ha quedat suprimida.

Operador Projecció

L'operació de recentrat respecte al centroide de cada període equival a la projecció sobre els subespais ortogonals als subespais generats per les variables indicatrius e_t de T en $\mathbb{R}^I$.

Anem a calcular l'operador projecció sobre l'espai generat per les indicatrius del e_t.

L'operador projecció és per definició:

$$P = T(T'D_I^{-1}T)^{-1}T'D_I^{-1}$$

en el nostre cas com $D_I = I$

$$P = T(T'T)^{-1}T'$$

Llavors essent la matriu T d'indicatrius de les e_t

$$T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

L'operador projecció $P = T(T'T)^{-1}T'$ és:

$$P = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{K_1^t} & 0 & \cdots & 0 \\ 0 & \frac{1}{K_2^t} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{K_T^t} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{1}{K_1^t} & \cdots & \frac{1}{K_1^t} & 0 & \cdots & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{K_1^t} & \cdots & \frac{1}{K_1^t} & 0 & \cdots & 0 & \cdots \\ 0 & \cdots & 0 & \frac{1}{K_2^t} & \cdots & \frac{1}{K_2^t} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \frac{1}{K_2^t} & \cdots & \frac{1}{K_2^t} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

on K_T^t indica el cardinal de la modalitat.

5.4 Inèrcies en ACM condicional

L'objectiu que pretenem és determinar la dependència d'un conjunt de variables qualitatives respecte a un model, generat per una variable condicionadora. Per al seu estudi, desenvolupem el càlcul efectiu de les inèrcies Inter i Intra, en les quals es pot descomposar la inèrcia total del núvol de punts, i la distribució que segueixen aquestes inèrcies. Això ens permetrà, conegudes les distribucions, determinar quan aquest condicionament és o no significant.

5.4.1 Descomposició de la inèrcia

Tal com hem vist en l'anàlisi de correspondències múltiples, la inèrcia total del núvol de punts pot ser calculada com:

$$\mathcal{I}_T = \frac{J}{Q} - 1$$

Donada una partició del núvol de punts en T subnúvols, tal com mostra la Figura 5.3, l'anàlisi de la inèrcia - dispersió del núvol global - reflexarà dues components de variabilitat:

- La dispersió *inter* períodes *(Inèrcia Intergrups o Between-groups)*: I_B
- La dispersió *intra* períodes *(Inèrcia Intragrups o Within-groups)*: I_W

Segons el teorema de Huyghens [Gre84] donat $y_1, \ldots, y_n$ un núvol de punts amb masses $m_1, \ldots, m_n$ en un espai euclidi multidimensional amb una mètrica definida per una matriu simètrica semidefinida positiva S, on $\bar{y}$ és el centroide dels punts $\bar{y} = \sum_i m_i y_i / \sum_i m_i$,

llavors la inèrcia del núvol respecte a qualsevol punt y és igual a la inèrcia respecte al centroide més la distància al quadrat entre y i $\bar{y}$ ponderada per la massa total del núvol.

$$\sum_i m_i \parallel y_i - y \parallel_S^2 = \sum_i m_i \parallel y_i - \bar{y} \parallel_S^2 + \left(\sum_i m_i\right) \parallel y - \bar{y} \parallel_S^2$$

on la norma respecte a la mètrica S és $\forall a, b \in \mathbb{R}^J$:

$$\parallel a - b \parallel_S^2 \equiv (a-b)'S(a-b)$$

Si tenim la presència de T subnúvols de punts $I_1, \ldots, I_T$, tindrem un centroide global $\bar{y}$ i t subcentroides $\bar{y}_t$, i associat a cadascun d'ells la massa del subnúvol m_t. Llavors la inèrcia total del núvols de punts és igual a la suma de les inèrcies dels T centroides dels grups (és a dir, inèrcies intergrups) més la suma de les inèrcies de cada grup respecte el seu centroide (és a dir inèrcies intragrup), on y_i^t és l'individu i pertanyent al grup t, i m_i^t la seva massa.

$$\mathcal{I}_T = \sum_t \sum_{i \in I_T} m_i^t \parallel y_i^t - \bar{y} \parallel_S^2 = \sum_t m_t \parallel \bar{y}_t - \bar{y} \parallel_S^2 + \sum_t m_i \sum_{i \in I_T} m_i^t \parallel y_i^t - \bar{y}_t \parallel_S^2 =$$

$$= \mathcal{I}_B + \mathcal{I}_W$$

On la demostració la tenim com a aplicació directa del teorema de Huyghens a cada subnúvol de punts, prenent com a punts d'aplicació els diferents centroides dels subnúvols.

En ACM Condicional, podem considerar la inèrcia intergrups com a la inèrcia deguda als diferents centres de gravetat dels subnúvols definits per les modalitats condicionadores.

El centre de gravetat del núvol és:

$$\bar{y} = \left(\frac{\hat{k}_j}{nQ}\right)_{j=1,\ldots,J}$$

i els centres de gravetat dels subnúvols de punts són:

$$\bar{y}_t = \left(\frac{b_{jt}}{Qk_t}\right)_{j=1,\ldots,J}$$

per tant la inèrcia inter vindrà donada per les distàncies entre els subcentres i el centre total, és a dir:

$$\mathcal{I}_B = \sum_{t=1}^{T} m_t(\bar{y}_t - \bar{y})'S(\bar{y}_t - \bar{y})$$

on la mètrica S ve donada per l'invers de les coordenades del centre de gravetat i les masses m_t per l'efectiu de cada modalitat sobre el total d'individus $m_t = k_t/n$.

Per tant la inèrcia inter és:

$$\mathcal{I}_B = \sum_t \frac{k_t}{n} \sum_j \frac{nQ}{k_j} \left(\frac{b_{jt}}{Qk_t} - \frac{k_j}{nQ} \right)^2 = Q \sum_t \sum_j \frac{k_t}{k_j} \left(\frac{b_{jt}}{Qk_t} - \frac{k_j}{nQ} \right)^2 =$$

$$= \frac{1}{Q} \sum_t \sum_j \frac{k_t}{k_j} \left(\frac{b_{jt}}{k_t} - \frac{k_j}{n} \right)^2 = \frac{1}{Q} \sum_t \sum_j \frac{k_t}{k_j} \left(\frac{b_{jt}^2}{k_t^2} + \frac{k_j^2}{n^2} - 2\frac{b_{jt}}{k_t}\frac{k_j}{n} \right) =$$

$$= \frac{1}{Q} \sum_t \sum_j \left(\frac{b_{jt}^2}{k_t k_j} + \frac{k_j k_t}{n^2} - \frac{2}{n} b_{jt} \right) =$$

$$= \frac{1}{Q} \left(\sum_t \sum_j \frac{b_{jt}^2}{k_t k_j} + \sum_t \sum_j \frac{k_j k_t}{n^2} - \sum_t \sum_j \frac{2}{n} b_{jt} \right) =$$

$$= \frac{1}{Q} \sum_t \sum_j \frac{b_{jt}^2}{k_t k_j} + \frac{1}{n^2 Q} \sum_t k_t \sum_j k_j - \frac{2}{nQ} \sum_t \sum_j b_{jt} =$$

$$= \frac{1}{Q} \sum_t \sum_j \frac{b_{jt}^2}{k_t k_j} + \frac{1}{n^2 Q} nnQ - \frac{2}{nQ} nQ =$$

$$= \frac{1}{Q} \sum_t \sum_j \frac{b_{jt}^2}{k_t k_j} + 1 - 2 = \frac{1}{Q} \sum_t \sum_j \frac{b_{jt}^2}{k_t k_j} - 1 \tag{5.2}$$

I per tant, coneguda la inèrcia total de la taula de correspondències i la relació entre la inèrcia total, la inèrcia inter i la inèrcia intra, la inèrcia intra és:

$$\mathcal{I}_W = \mathcal{I}_T - \mathcal{I}_B = \frac{J}{Q} - 1 - \frac{1}{Q} \sum_t \sum_j \frac{b_{jt}^2}{k_t k_j} - 1 =$$

$$\mathcal{I}_W = \frac{J - \sum_t \sum_j \frac{b_{jt}^2}{k_t k_j}}{Q} - 2 \tag{5.3}$$

Quan la variable condicionadora defineixi una partició a l'atzar, llavors la inèrcia $\mathcal{I}_B$ s'aproximarà a zero i $\mathcal{I}_W \cong \mathcal{I}_{TOT}$.

5.4.2 Distribució de les inèrcies Inter i Intra

Considerem la banda de Burt ja definida anteriorment, de dimensió JxT, creuament de les modalitats de les variables a condicionar i de les modalitats condicionadores:

b_{11}	b_{12}	$\dots$	b_{1T}
b_{21}	b_{22}	$\dots$	b_{2T}
b_{j1}	b_{j2}	$\dots$	b_{jT}
b_{J1}	b_{J2}	$\dots$	b_{JT}

aquesta pot ser considerada com una taula de contingència on les marginals són:

$$\left(\hat{k}_j\right)_{j=1,\dots,J} \quad , \quad (Qk_t)_{t=1,\dots,T} \quad \text{i gran total} \quad nQ$$

Per tant el coeficient χ^2 d'aquesta banda, vindrà donat per:

$$\chi^2_{banda} = \sum_t \sum_j \frac{\left(b_{jt} - \frac{k_j Q k_t}{nQ}\right)^2}{\frac{k_j Q k_t}{nQ}} = \sum_t \sum_j \frac{\left(b_{jt} - \frac{k_j k_t}{n}\right)^2}{\frac{k_j k_t}{n}} =$$

$$= \sum_t \sum_j \frac{b^2_{jt} + \frac{k_j^2 k_t^2}{n^2} - 2b_{jt}\frac{k_j k_t}{n}}{\frac{k_j Q k_t}{nQ}} = \sum_t \sum_j \left(n\frac{b^2_{jt}}{k_j k_t} + \frac{k_j k_t}{n} - 2b_{jt}\right) =$$

$$= n\sum_t \sum_j \frac{b^2_{jt}}{k_j k_t} + \frac{1}{n}\sum_t \sum_j k_j k_t - 2\sum_t \sum_j b_{jt} =$$

$$= n\sum_t \sum_j \frac{b^2_{jt}}{k_j k_t} + \frac{1}{n} nnQ - 2nQ = n\sum_t \sum_j \frac{b^2_{jt}}{k_j k_t} + nQ - 2nQ =$$

$$\chi^2_{banda} = n\sum_t \sum_j \frac{b^2_{jt}}{k_j k_t} - nQ \tag{5.4}$$

Donats els resultats anteriors (5.2) i (5.4), podem observar que el resultat d'aquest coeficient χ^2_{banda} és exactament nQ vegades el valor de la inèrcia inter.

$$\mathcal{I}_B = \frac{1}{nQ}\chi^2_{banda}$$

Per tant, conegut que el coeficient χ^2_{banda} segueix asimptòticament una distribució de χ^2, sota la hipòtesi d'independència, la distribució de la inèrcia inter $\mathcal{I}_B$ segueix una distribució χ^2 escalada*:

$$\mathcal{I}_B \sim \frac{\chi^2}{nQ}$$

I on els graus de llibertat de la distribució χ^2 obtinguda són $(T-1)(J-1)$.

*Un resultat anàleg en el context d'anàlisi de correspondències de juxtaposició de taules de contingència fou provat per Leclerc [Lec75], on no tenia present la independència de les variables

Si les variables d'estudi fossin independents, els graus de llibertat de la distribució χ^2 obtinguda són $(T-1)(J-Q)$, ja que les modalitats de cadascuna de les Q variables condicionades estan lligades, això ens fa tenir un nombre de graus de llibertat en la component de la dimensió de variables a condicionar igual a $J-Q$. Aquest no és un cas interessant, ja que en els casos usuals, amb associacions entre variables desconegudes, podem assumir els graus de llibertat assumits anteriorment els quals ens donaran uns intervals de confiança més grans i, per tant, uns tests més conservadors.

$$\mathcal{I}_B \sim \frac{\chi^2_{(T-1)(J-1)}}{nQ}$$

I coneguda la distribució, coneixem els moments:

$$E(\mathcal{I}_B) = \frac{(T-1)(J-1)}{nQ}$$
$$Var(\mathcal{I}_B) = \frac{2(T-1)(J-1)}{(nQ)^2}$$

I per tant, per 5.3, la inèrcia intra $\mathcal{I}_W$ té la distribució:

$$\mathcal{I}_W \sim \frac{J}{Q} - 1 - \frac{\chi^2_{(T-1)(J-1)}}{nQ} = \frac{J-Q}{Q} - \frac{\chi^2_{(T-1)(J-1)}}{nQ}$$

amb moments:

$$E(\mathcal{I}_W) = \frac{J-Q}{Q} - \frac{(T-1)(J-1)}{nQ} = \frac{(J-Q)(n-T+1)}{nQ}$$
$$Var(\mathcal{I}_W) = \frac{2(T-1)(J-1)}{(nQ)^2}$$

De totes maneres, per a obtenir un test més fiable, sempre podem ortogonalitzar el conjunts de les Q variables qualitatives seqüencialment entre elles abans de calcular la variables condicionadora.

5.4.3 ACM condicional i Inferència

En l'ACM Condicional i per tal d'eliminar la influència del la variable qualitativa T, la metodologia imposa recentrar cada subnúvol $\mathcal{I}_T$ al seu centroide. Això comporta que després de la realització d'una ACM condicional la inèrcia total del núvol recentrat sigui igual a la inèrcia intragrups original, puix el recentrat elimina la inèrcia intergrups.

$$\mathcal{I}_{Condicional} = \mathcal{I}_W$$

Per altra banda, donat un condicionament, si el condicionament és insignificant hi haurà una $\mathcal{I}_B$ despreciable, ja que els centroides de cada subnúvol seran molt propers al centroide global, o el que és el mateix, la $\mathcal{I}_W$ serà gairebé equivalent a la $\mathcal{I}_T$.

Aquests dos fets comporten que la influència d'un condicionament pugui ser determinada per la importància de la $\mathcal{I}_W$ en referència a la inèrcia total $\mathcal{I}_T$, o alternativament, per la presència de una $\mathcal{I}_B$ gairebé despreciable. És fàcilment demostrable que aquest fet és independent de la distribució dels efectius k_t.

Per tant, coneguda la distribució de les inèrcies Inter i Intra, sota la hipòtesi d'una partició aleatòria, això ens permet establir intervals de confiança i realitzar contrastos per als valors de les inèrcies obtinguts en el condicionament.

Un interval de confiança per a la inèrcia resultant del condicionament $\mathcal{I}_B$, a un nivell α, vindrà donat per:

$$\left(0, \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ}\right)$$

Serà equivalent al càlcul de l'interval de confiança per a la $\mathcal{I}_W$, a un nivell α, que vindrà donat per:

$$\left(\frac{J-Q}{Q} - \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ}, \frac{J-Q}{Q}\right)$$

La longitud d'aquests intervals depèn dels paràmetres que defineixen la distribució de la inèrcia, n, Q, J i T. Llavors, per exemple, és fàcil veure que la longitud dels intervals de confiança de la inèrcia inter $\mathcal{I}_B$ decreix amb la mida de la mostra n. Calculant els intervals de confiança per a un nombre d'individus cada vegada incrementat, amb $\alpha = 0.05$, obtenim la següent taula:

n	Q	J	T	Interval
100	4	20	5	(0 , 0.209188)
200	4	20	5	(0 , 0.104594)
400	4	20	5	(0 , 0.052297)
500	4	20	5	(0 , 0.041838)
1000	4	20	5	(0 , 0.020919)
2000	4	20	5	(0 , 0.010459)

Com que els diversos sub-centres són millors estimadors del centre global, en particions aleatòries, la inèrcia inter haurà de decréixer.

Per provar els resultats previs, considerem, com un exemple, la taula de correspondències múltiples Z i la taula de la variable condicionadora T:

a_1	a_2	b_1	b_2	b_3
1	0	1	0	0
0	1	1	0	0
1	0	1	0	0
1	0	0	1	0
1	0	0	0	1
0	1	0	1	0
0	1	0	0	1
0	1	0	0	1
0	1	1	0	0

t_1	t_2
1	0
1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

on $n = 9$, $J = 5$, $Q = 2$ i $T = 2$. llavors la taula de Burt i la banda de Burt són les següents:

a_1	a_2	b_1	b_2	b_3
4	0	2	1	1
0	5	2	1	2
2	2	4	0	0
1	1	0	2	0
1	2	0	0	3

t_1	t_2
4	0
1	4
3	1
1	1
1	2

el centre global és:

$$C_0 = \left(\frac{4}{18}, \frac{5}{18}, \frac{4}{18}, \frac{2}{18}, \frac{3}{18}\right)$$

i els subcentres són:

$$C_1 = \left(\frac{4}{10}, \frac{1}{10}, \frac{3}{10}, \frac{1}{10}, \frac{1}{10}\right) \qquad \text{and} \qquad C_2 = \left(\frac{0}{8}, \frac{4}{8}, \frac{1}{8}, \frac{1}{8}, \frac{2}{8}\right)$$

Per tant, podem calcular la inèrcia total $\mathcal{I}_{TOT} = 1.5$, la inèrcia inter $\mathcal{I}_B = 0.388750$ i la inèrcia intra $\mathcal{I}_W = 1.111250$.

Per altra banda, el coeficient χ^2 de la banda de Burt és $\chi^2_{band} = 6.99750$, i per tant,

$$\frac{\chi^2_{band}}{nQ} = \frac{6,99750}{9 * 2} = 0.388750$$

I com

$$\mathcal{I}_B \sim \frac{\chi^2_{(T-1)(J-1)}}{nQ} = \frac{\chi^2_4}{18}$$

l'interval de confiança amb $\alpha = 0.05$ és $\mathcal{I}_B(\alpha) = (0, 0.43417)$, i per això el condicionament no és significant, no hi ha dependència entre la taula Z i les modalitats condicionadores.

5.4.4 Simulació i test de bondat d'ajust

Per tal de comprovar els resultats teòrics obtinguts, hem realitzat, primer, diverses simulacions per tal de determinar el comportament de $\mathcal{I}_B$, en funció dels paràmetres, n, Q, J i T i després, hem realitzat el test de bondat d'ajust de Kolmogorov-Smirnov, per determinar la bondat d'ajust de la funció de distribució χ^2.

Simulació

Primer hem simulat diferents particions d'acord amb els diferents nombres de modalitats de la variable condicionadora ($T = 2, 3, 4, 5, 6$) sobre una taula de tres variables amb dues categories cadascuna i 12 individus. Hem realitzat 20 simulacions aleatòries per a cada valor de T, i d'elles hem calculat la seva $\mathcal{I}_B$.

La Figura 5.4 mostra que $\mathcal{I}_B$ és positivament relacionat amb el nombre de classes de T, la qual cosa està relacionada amb tenir més graus de llibertat la distribució χ^2 escalada.

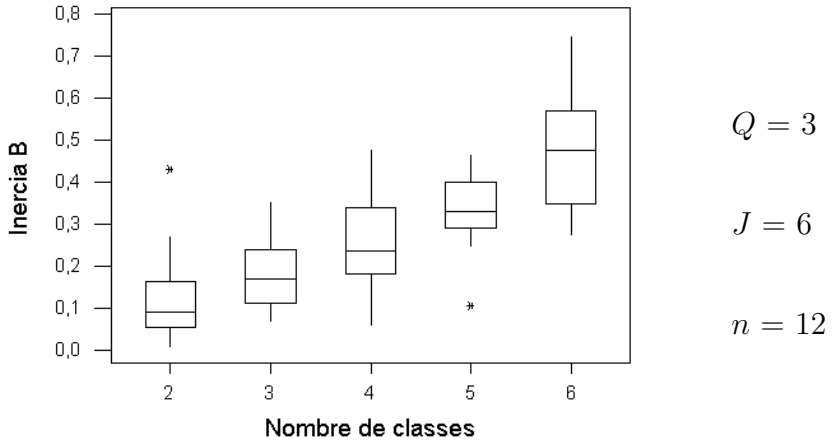


Figura 5.4: *Increment de la inèrcia between segons el nombre de modalitats de T*

A més, hem realitzat 1000 simulacions aleatòries amb diferents valors dels paràmetres Q, J, n i T: el nombre de variables Q pot prendre els valors 2, 4 i 8, amb J_q modalitats cadascuna, on J_q=3, 5 i 7; el nombre de modalitats condicionadores T=3, 5 i 7, i finalment el total d'individus n=100, 500 i 2000.

Per a cadascuna de les combinacions, hem obtingut l'interval de confiança teòric al 95% per la $\mathcal{I}_B$ i el percentatge de valors simulats dins d'aquest interval. Els resultats són mostrats en la Taula 5.1. Podem veure que el percentatge no és molt distant del 95% teòric en la majoria de les simulacions.

Taula 5.1: *Simulació de valors de $\mathcal{I}_B$ per a diferents valors de Q, J, T i n*

			$n = 100$		$n = 500$		$n = 2000$	
Q	J	T	95% CI	% In	95% CI	% In	95% CI	% In
2	6	3	(0,0.077537)	94.3	(0,0.0155073)	93.9	(0,0.00387683)	95.0
2	6	5	(0,0.131481)	97.1	(0,0.0262962)	95.4	(0,0.00657406)	93.8
2	6	7	(0,0.182075)	94.9	(0,0.0364150)	95.6	(0,0.00910376)	95.0
2	10	3	(0,0.131481)	94.4	(0,0.0262962)	94.8	(0,0.00657406)	95.6
2	10	5	(0,0.230971)	94.8	(0,0.0461943)	94.8	(0,0.0115486)	96.0
2	10	7	(0,0.325854)	94.0	(0,0.0651708)	95.5	(0,0.0162927)	94.7
2	14	3	(0,0.182057)	96.3	(0,0.0364150)	94.3	(0,0.00910376)	95.4
2	14	5	(0,0.325854)	94.6	(0,0.0651708)	95.5	(0,0.0162927)	94.8
2	14	7	(0,0.464041)	96.2	(0,0.0928083)	94.6	(0,0.0232021)	93.6
4	12	3	(0,0.065741)	94.4	(0,0.0131481)	95.5	(0,0.00328703)	94.3
4	12	5	(0,0.115486)	95.0	(0,0.0230971)	95.1	(0,0.00577428)	94.2
4	12	7	(0,0.162927)	95.7	(0,0.0325854)	95.6	(0,0.00814635)	95.7
4	20	3	(0,0.115486)	94.3	(0,0.0230971)	95.9	(0,0.00577428)	94.6
4	20	5	(0,0.209188)	94.7	(0,0.0418376)	95.2	(0,0.0104594)	95.6
4	20	7	(0,0.299677)	95.1	(0,0.0599355)	94.4	(0,0.0149839)	95.7
4	28	3	(0,0.162927)	94.3	(0,0.0325854)	95.0	(0,0.00814635)	95.8
4	28	5	(0,0.299677)	94.7	(0,0.0599355)	95.0	(0,0.0149839)	93.8
4	28	7	(0,0.432510)	95.0	(0,0.0865020)	95.6	(0,0.0216255)	95.8
8	24	3	(0,0.057743)	94.8	(0,0.0115486)	92.9	(0,0.00288714)	94.7
8	24	5	(0,0.104594)	94.9	(0,0.0209188)	93.7	(0,0.00522970)	94.3
8	24	7	(0,0.149839)	94.3	(0,0.0299677)	95.6	(0,0.00749193)	94.0
8	40	3	(0,0.104594)	95.3	(0,0.0209188)	95.5	(0,0.00522970)	94.2
8	40	5	(0,0.194256)	94.3	(0,0.0388512)	94.8	(0,0.00971280)	95.6
8	40	7	(0,0.281661)	95.4	(0,0.0388512)	94.9	(0,0.0140830)	94.3
8	56	3	(0,0.149839)	95.0	(0,0.0299677)	94.9	(0,0.00749193)	95.2
8	56	5	(0,0.281661)	95.2	(0,0.0563322)	95.5	(0,0.0140830)	94.1
8	56	7	(0,0.410726)	95.5	(0,0.0821451)	94.2	(0,0.0205363)	93.8

Test de bondat d'ajust

Per tal de determinar si:

$$nQ\mathcal{I}_B \sim \chi^2_{(T-1)(J-1)}$$

hem usat l'estadístic de Kolmogorov-Smirnov [Kol33] on la hipòtesi nul·la és que les dades escalades segueixen asimptòticament una distribució χ^2 amb $(T-1)(J-1)$ graus de llibertat.

Generant particions aleatòries per a diferents valors de n, J, Q i T hem obtingut la corresponents distribució empírica de la inèrcia inter, $\mathcal{I}_B$, i els seus moments i l'estadístic de bondat d'ajust de Kolmogorov-Smirnov (ks) amb el corresponent p-valor.

Realitzant simulacions de mida 100, per a cada combinació hem obtingut els resultats mostrats a la Taula 5.2.

Taula 5.2: *Test de ks per a diferents valors de Q, J, T i n*

n	Q	J	T	mitjana($nQ\mathcal{I}_B$)	var($nQ\mathcal{I}_B$)	ks	p-valor
400	2	8	4	19.37131	39.57659	0.1108	0.1896
1000	2	8	4	17.52190	34.56822	0.0961	0.3147
1000	2	8	5	24.69165	42.85737	0.0754	0.6202
1000	3	15	5	47.8183	107.2000	0.0910	0.3793
2000	3	15	5	45.84285	78.14225	0.0780	0.5762
200	4	16	5	49.2355	135.4396	0.1215	0.1044
400	4	16	4	36.06192	77.27660	0.0565	0.9074
1000	4	16	5	47.8575	113.4476	0.0709	0.6965
2000	4	16	5	46.4625	106.0357	0.0885	0.4142

En totes les generacions provades hem obtingut p-valors prou grans, la qual cosa corrobora que la distribució χ^2 amb els corresponents graus de llibertat és un model teòric per contrastar la inèrcia inter sota la hipòtesi nul·la de partició aleatòria.

5.4.5 Exemple d'aplicació

Finalment, presentem un exemple d'aplicació.

Considerem un conjunt de $n = 376$ individus i quatre variables, amb quatre modalitats cadascuna. La relació entre les variables està basada en un model gràfic [MN89] [FJ93].

El model aquí considerat es mostra a la Figura 5.5, generat mitjançant un model loglineal amb interaccions de primer ordre:

$$\log m_{ijkl} = u + u_{a(i)} + u_{b(j)} + u_{c(k)} + u_{d(l)} + u_{ab(ij)} + u_{ad(il)} + u_{bd(jl)} + \epsilon$$

amb iguals paràmetres dels efectes principals $u_{a(i)} = u_{b(j)} = u_{c(k)} = u_{d(l)}$

$$\begin{pmatrix} 0.4 & 0.4 & -0.4 & -0.4 \end{pmatrix}$$

i igual paràmetres d'interacció $u_{ab(ij)} = u_{ad(il)} = u_{bd(jl)}$

$$\begin{pmatrix} 0.3 & 0.3 & -0.2 & -0.4 \\ -0.2 & -0.1 & 0.1 & 0.2 \\ -0.2 & -0.3 & 0.3 & 0.2 \\ 0.1 & 0.1 & -0.2 & 0 \end{pmatrix}$$

on ϵ és una fluctuació aleatòria sota la hipòtesi de distribució normal.

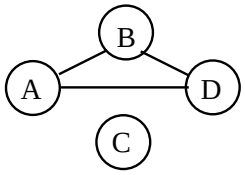


Figura 5.5: *Model gràfic d'interacció*

Considerem la taula disjuntiva amb només les variables A i D, que notem Z_{AD}, i les dues taules condicionades per les variables B i C. En ambdós condicionaments tenim $J = 8, Q = 2, T = 4, n = 376$.

La inèrcia d'aquestes tres taules és:

Inèrcia total Z_{AD}	Z_{AD} condicionada per B	Z_{AD} condicionada per C
3	0.05113	0.01249

Sota la hipòtesi de partició a l'atzar, un interval de confiança de nivell $\alpha = 0.05$ per a $\mathcal{I}_B$ és (0 , 0.04344). Llavors, podem veure que la inèrcia resultant pel condicionament per B queda fora de l'interval, mostrant la dependència de les variables A i D en relació a B, mentre que la inèrcia del condicionament per C és dins l'interval de confiança, mostrant la independència entre les variables A i D i la variable C.

Aquests resultats ens donen la dependència de dues de les variables A, D respecte una tercera variable, confirmant les relacions establertes entre les variables donades pel model prèviament definit.

Hem realitzat aquest tests sobre tots els models gràfics de 4 i 5 variables, confirmant, els resultats obtinguts, les relacions establertes pel models gràfics. Alguns dels resultats sobre 4 variables, amb $J = 8$, $Q = 2$ i $T = 4$, amb els corresponents intervals de confiança es mostren en la Taula 5.3.

Les inèrcies fora de l'interval de confiança es mostren amb negreta, indicant-nos la dependència de les variables A i D respecte la tercera variable. Podem veure que només les dues variables A i D estan lligades amb la variable condicionadora, és a dir I_B fora de l'interval, quan el model gràfic d'interacció ens dóna aquest lligam.

Per tal de veure la importàcia dels paràmetres d'interacció, amb el mateix model de relació donat per la Figura 5.5, fixem un nivell base (0) dels paràmetres d'interacció

$$\begin{pmatrix} 0.3 & 0.3 & -0.3 & -0.3 \\ 0.3 & 0.3 & -0.3 & -0.3 \\ -0.3 & -0.3 & 0.3 & 0.3 \\ -0.3 & -0.3 & 0.3 & 0.3 \end{pmatrix}$$

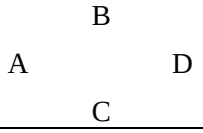
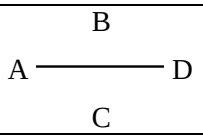
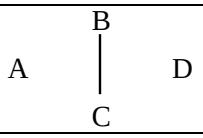
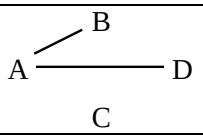
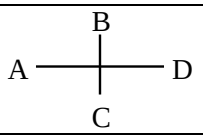
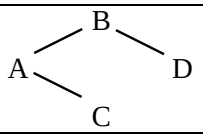
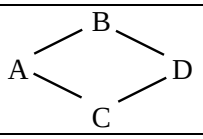
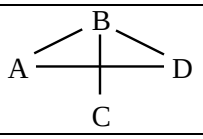
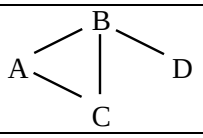
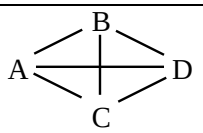
i definim el nivell superior (1) com el doble del nivell base i el nivell inferior (-1) com la meitat del nivell base.

Després d'això, repetim l'anàlisi de les inèrcies Between, en el disseny que creua el nivell de la relació AD i el nivell de tota la resta de relacions diferents de zero (és a dir les relacions AB i BD), amb taules amb 1000 individus aproximadament. Els resultats es mostren a la Taula 5.4

Les inèrcies fora de l'interval de confiança estan en negreta, mostrant-nos la dependència de les variables A i D en relació a la variable B en tots els casos, sigui quin sigui el nivell, i mai en referència a la variable C. Podem observar també que com més gran és el nivell de relació, més gran és la inèrcia Between.

Com a conclusió, l'estudi de la inèrcia Between ens pot ajudar a comprendre més profundament les relacions directes o indirectes entre variables.

Taula 5.3: **Inèrcies Between dels models sota condicionament**

Model	mida	I_B cond. per B	I_B cond. per C	Interval Conf.
	n=304	0.00735	0.00735	(0,0.053735)
	n=348	0.00369	0.00369	(0,0.046941)
	n=348	0.00612	0.00127	(0,0.046941)
	n=354	0.02749	0.01129	(0,0.046145)
	n=603	0.00298	0.00104	(0,0.027090)
	n=374	**0.05203**	0.02747	(0,0.043677)
	n=400	**0.0493**	**0.0493**	(0,0.040838)
	n=396	**0.06351**	0.00283	(0,0.041251)
	n=396	**0.06002**	0.03193	(0,0.041251)
	n=452	**0.06295**	**0.05978**	(0,0.036140)

Taula 5.4: **Inèrcies Between dels models condicionats segons el nivell**

nivell AD	nivells AB,BD	n	Cond. per B	Cond. per C	Int. Conf.
1	0	n=1000	**0.08046**	0.00024	(0,0.01634)
0	0	n=996	**0.07822**	0.00035	(0,0.01640)
-1	0	n=1000	**0.05917**	0.00024	(0,0.01634)
1	1	n=1000	**0.29836**	0.00060	(0,0.01634)
0	1	n=1000	**0.26491**	0.00037	(0,0.01634)
-1	1	n=1004	**0.24659**	0.00020	(0,0.01627)
1	-1	n=1012	**0.03246**	0.00001	(0,0.01614)
0	-1	n=1012	**0.02004**	0.00037	(0,0.01614)
-1	-1	n=1000	**0.02032**	0.00017	(0,0.01634)

Capítol 6

ACM multicondicional

6.1 Introducció

En aquesta secció entrarem en les anàlisis multicondicionals, on tractarem d'estendre el condicionament vist en la secció anterior pel condicionament per diverses variables. Veurem les dificultats que comporta en alguns casos i les alternatives o propostes de condicionament múltiple que desenvoluparem per a tractar-ho.

6.2 ACM condicionat per més d'una variable

6.2.1 Introducció

S'ha vist en la secció anterior l'ACM condicional per una variable, però, a vegades, totes les variables o la majoria d'elles estan lligades no a una sinó a un conjunt de variables qualitatives denotades per T_1,...,T_k, en aquest cas una ACM clàssica principalment mostraria aquests lligams, fins i tot si el conjunt de variables T_1,...,T_k no han estat introduïdes en el conjunt de dades.

Si a nosaltres ens interessa, no pas l'estudi de les relacions estables dins de cada grup de variables T_1,...,T_k, sinó l'estudi de les relacions estables sense l'influència 'temporal' d'aquests grups, és necessari eliminar la influència del conjunt de variables T_1,...,T_k, el que nosaltres anomenarem **condicionament múltiple**.

El nostre problema serà ara trobar una anàlisi del tipus *anàlisi de correspondències múltiples* condicionat en relació a un conjunt de variables exteriors a les variables tractades, que respecti les principals de les propietats de les ACM: construcció i projecció dels dos núvols de punts, dualitat i fórmula de transició entre les projeccions dels dos núvols

i l'equivalència amb l'anàlisi d'una taula de Burt

6.2.2 Notacions

Notem I la població que comprèn n individus; J el conjunt de totes les modalitats de les Q variables qualitatives, T_1,...,T_k el conjunt de les variables qualitatives de les quals en volem treure'n la influència. Cadascuna de les variables T_1,...,T_k defineixen una partició d'I on les classes són denotades I_{t_i} on t_i és una modalitat de la variable T_i

6.2.3 Problemàtica de la no ortogonalitat

Una de les temptatives per fer el condicionament respecte més d'una variable, ens porta al problema que el condicionament requereix que els vectors directors de les modalitats siguin ortogonals, cosa que es compleix amb una sola variable condicionadora, però quan introduïm una segona o tercera variable condicionadora, això no es compleix. Per la qual cosa es prova d'ortogonalitzar les modalitats de la segona qüestió condicionadora a les primeres, completant una base, però aquí ens trobem amb un problema

Per realitzar la ortogonalització de Gram-Schmidt es requereix la inversa, per tal de fer el càlcul dels coeficients no nuls en la igualtat que ens diu per quin vector jo puc substituir. Però una solució de la no existència de la inversa el trobem en la inversa generalitzada.

6.2.4 Matriu inversa generalitzada

La matriu $(A^tA)^{-1}$ que permet solucionar el problema dels mínims quadrats existirà sempre que A^tA no sigui singular. Si A^tA és singular la solució al problema es basa en la existència de la matriu pseudoinversa.

Donada una matriu $A \in M_{(n,m)}$ i un vector $y \in \mathbb{R}^n$ definim com a **matriu pseudoinversa d'**A, representada per A^-, la matriu que aplicada al vector y ens proporciona la solució x_0 que fa mínim $\parallel Ax_0 - y \parallel$, és a dir que ens dóna

$$A^-y = x_0$$

Estudiem l'existència d'aquesta matriu:

- Si A és invertible llavors $A^- = A^{-1}$

- Si A és una matriu del tipus

$$A = \begin{pmatrix} \mu_1 & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \mu_r & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in M_{(n,m)} \qquad \text{amb } \mu_i \neq 0$$

  llavors la pseudoinversa és:

$$A^- = \begin{pmatrix} \frac{1}{\mu_1} & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \frac{1}{\mu_r} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in M_{(m,n)}$$

- La generalització per a qualsevol altra matriu $A \in M_{(n,m)}$ ve donada per la existència de matrius $Q_1 \in M_{(n,n)}$, $Q_2 \in M_{(m,m)}$ ortogonals i

$$S = \begin{pmatrix} \mu_1 & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \mu_r & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in M_{(n,m)} \qquad \text{amb } \mu_i \neq 0$$

  tal que $A = Q_1 S Q_2^t$ - matrius de vectors propis per la dreta, valors propis i vectors propis per l'esquerra respectivament - verificant-se que la pseudoinversa d'A és:

$$A^- = Q_2 S^- Q_1^t$$

  La demostració d'això es troba en:

  Si $A = Q_1 S Q_2^t$ llavors AA^t és una matriu simètrica d'on puc obtenir fàcilment:

$$AA^t = Q_1 S Q_2^t Q_2 S^t Q_1^t = Q_1 S S^t Q_1^t$$

  D'on puc obtenir fàcilment els vectors propis Q_1 i els valors propis $S_1 = SS^t$.

  Anàlogament:

$$A^t A = Q_2 S^t Q_1^t Q_1 S Q_2^t = Q_2 S^t S Q_2^t$$

D'on puc obtenir fàcilment els vectors propis Q_2 i els valors propis $S_2 = S^t S$.

Coneixent Q_1 i Q_2, això ens permet aïllar de l'expressió $A = Q_1 S Q_2^t$ els valors propis S,

$$S = Q_1^t A Q_2$$

Finalment coneguts Q_1, Q_2 i S es pot demostrar fàcilment que :

$$A^- = Q_2 S^- Q_1^t$$

6.3 Altres condicionaments

Donada la complexitat del problema de la no ortogonalitat i la seva difícil solució en els problemes reals, s'han estudiat tota una altra sèrie de propostes de condicionaments, en base a diferents possibilitats de condicionament en referència a dues variables condicionadores.

6.3.1 Introducció

En aquesta secció desenvolupem, sobre una sèrie de models loglineals generats, diferents propostes de condicionament en referència a dues variables condicionadores **B** i **C**. Les propostes que s'estudien són:

1. condicionament simultani per B i C
2. condicionament pel encreuament de B i C
3. condicionament per la juxtaposició de B i C
4. condicionament successiu per B i per C
5. condicionament successiu per C i per B

Passem a descriure breument cadascuna d'aquestes propostes: la primera d'elles, el condicionament simultani per B i per C, consisteix en per tal de solucionar el problema de la no independència entre les variables condicionadores construir una matriu de dades que contingui els dos condicionaments de forma independent. Per a fer-ho si anomenem M_B a la matriu de dades condicionades per B i M_C anàlogament per C, llavors la nostra matriu objecte d'estudi és:

$$\begin{pmatrix} M_B & 0 \\ 0 & M_C \end{pmatrix}$$

Aquest artifici ens permet una ortogonalitat dels dos condicionaments, ara bé, la complicació sorgeix en el moment de la seva interpretació, ja que tenim una doble projecció.

L'encreuament de B i C ens dóna una nova variable $B * C$ que té un nombre de modalitats igual al producte de modalitats de B per C. Aquesta seria la millor opció des del punt de vista teòric, però té dos inconvenients. El primer d'ells és que habitualment no es tenen els resultats de la resta de variables en relació a l'encreuament de dues altres, i depenent de la manera que ens hagin estat facilitat les dades, fins i tot pot ser impossible d'obtenir-ho. El segon problema que ens hem trobat és que a mesura que augmenta el nombre de modalitats de la variable condicionadora, augmenta la inèrcia I_B, ja que tal com hem vist, hi ha més subcentres i per tant més allunyament de la inèrcia total.

La tercera proposta es basa en la juxtaposició de taules de contingència. Analitzem la matriu:

$$\begin{pmatrix} M_B \\ M_C \end{pmatrix}$$

on la seva inèrcia, en aquest cas, serà la mitjana de les dues inèrcies corresponents a les matrius M_B i M_C. L'interès i la complicació el trobarem en l'anàlisi de la doble representació de les modalitats.

Finalment les dues darreres propostes són els condicionaments successius. Primerament s'ha realitzat el condicionament successiu per B i després per C i en segon terme el condicionament successiu primer per C i després per B. Com que no hi ha necessàriament ortogonalitat, aquestes dues darreres propostes de condicionament no tenen perquè coincidir en el seu resultat.

6.3.2 Exemple d'aplicació a models

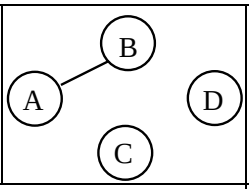
Anem a aplicar aquestes propostes a diferents models loglineals basat en 4 variables categòriques A, B, C, D amb 4 modalitats cadascuna. Per a cadascun dels models donarem les inèrcies de les taules de Burt i Z.

Per a cada model disposarem d'una taula on hi trobarem tant la inèrcia total, com les de les inèrcies dels condicionaments per B i per C, com les de les cinc propostes esmentades anteriorment.

Taula 6.1: **Resultats model loglineal 48**

Model	B A D C	Mida	304		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50367	2,99265	1,49701	2,99265	1,49701
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,53551	3,99402	2,97714	1,48330	2,99265	1,49701
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,98399	1,48929	2,98399	1,48929		

Taula 6.2: **Resultats model loglineal 54**

Model		Mida	348		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50306	2,97520	1,47875	2,98943	1,49335
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,51386	3,97209	2,95844	1,46313	2,98231	1,48590
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,96449	1,46895	2,96448	1,46894		

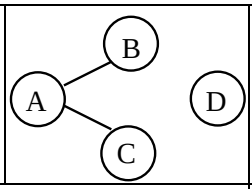
Taula 6.3: **Resultats model loglineal 50**

Model	B A ——— D C	Mida	348		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1.52416	2,99631	1,52046	2,99631	1,52046
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,59894	4,04093	2,97478	1,49859	2,99631	1,52046
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,99165	1,51584	2,99165	1,51584		

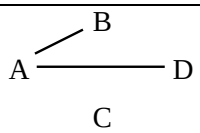
Taula 6.4: **Resultats model loglineal 46**

Model	B A | D C	Mida	348		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50751	2,99388	1,50218	2,99873	1,50640
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,56002	4,00858	2,98493	1,49452	2,99631	1,50429
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,99258	1,50105	2,99265	1,50112		

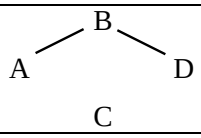
Taula 6.5: **Resultats model loglineal 55**

Model		Mida	354		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,51030	2,97489	1,48620	2,97489	1,48620
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,52555	3,97239	2,94704	1,46040	2,97489	1,48620
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,95311	1,46628	2,95311	1,456628		

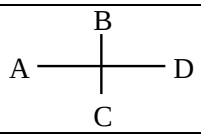
Taula 6.6: **Resultats model loglineal 44**

Model		Mida	354		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52412	2,97251	1,49403	2,98871	1,51335
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,57711	4,00738	2,94439	1,46694	2,98061	1,50348
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,96094	1,48306	2,96096	1,48309		

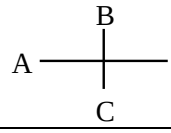
Taula 6.7: **Resultats model loglineal 45**

Model		Mida	352		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50108	2,96032	1,46211	2,99719	1,49831
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,49315	3,96042	2,94166	1,44385	2,97875	1,48002
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,95714	1,45903	2,95718	1,45906		

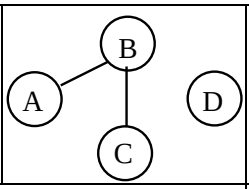
Taula 6.8: **Resultats model loglineal 43**

Model		Mida	350		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52561	2,99371	1,51925	2,99701	1,52252
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,60259	4,04177	2,97543	1,50091	2,99536	1,52088
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,99085	1,51637	2,99080	1,51632		

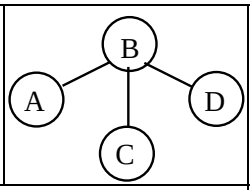
Taula 6.9: **Resultats model loglineal 43(2)**

Model		Mida	603		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52443	2,99702	1,52137	2,99896	1,52327
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,60366	4,04464	2,99238	1,51660	2,99799	1,52232
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,99598	1,52031	2,99597	1,52031		

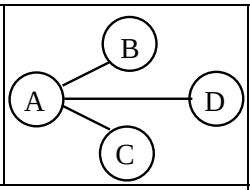
Taula 6.10: **Resultats model loglineal 56**

Model		Mida	352		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50217	2,97482	1,47769	2,99828	1,50048
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,51253	3,97817	2,96247	1,46590	2,98655	1,48896
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,972923	1,47588	2,97297	1,47592		

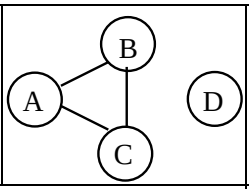
Taula 6.11: **Resultats model loglineal 57**

Model		Mida	376		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50129	2,95059	1,45306	2,99838	1,49968
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,48663	3,95274	2,93956	1,44251	2,97449	1,47609
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,94731	1,44997	2,94836	1,45097		

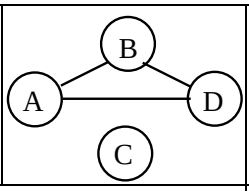
Taula 6.12: **Resultats model loglineal 58**

Model		Mida	398		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52314	2,97669	1,49999	2,97669	1,49999
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,57158	3,999998	2,94976	1,47431	2,97669	1,49999
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,95373	1,47823	2,95373	1,47823		

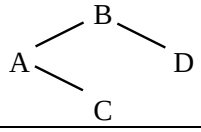
Taula 6.13: **Resultats model loglineal 53**

Model		Mida	376		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,51024	2,96641	1,47718	2,96906	1,47940
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,50953	3,95658	2,92582	1,43795	2,96774	1,47828
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,93998	1,45208	2,93930	1,45145		

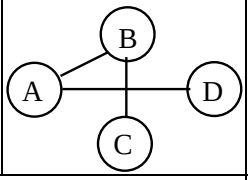
Taula 6.14: **Resultats model loglineal 49**

Model		Mida	376		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52869	2.94887	1,47448	2.98751	1,51564
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,56723	3,99012	2,92428	1,44961	2,96819	1,49470
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,93697	1,46227	2,93701	1,46233		

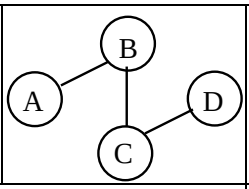
Taula 6.15: **Resultats model loglineal 42**

Model		Mida	374		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50319	2,94797	1,45292	2,97253	1,47649
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,47193	3,92941	2,90899	1,41650	2,96025	1,46455
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,92122	1,42827	2,92120	1,42825		

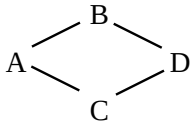
Taula 6.16: **Resultats model loglineal 59**

Model		Mida	374		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52613	2,96488	1,49124	2,99546	1,52172
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,58623	4,01296	2,94923	1,47597	2,98017	1,50623
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,95982	1,48627	2,95981	1,48625		

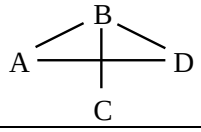
Taula 6.17: **Resultats model loglineal 60**

Model		Mida	371		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50446	2,97120	1,47626	2,97100	1,47598
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,49969	3,95224	2,92536	1,43120	2,97110	1,47581
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,94230	1,44792	2,94224	1,44786		

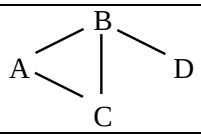
Taula 6.18: **Resultats model loglineal 41**

Model		Mida	400		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50416	2,95070	1,45616	2,95070	1,45616
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,44994	3,91232	2,88958	1,39922	2,95070	1,45616
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,90658	1,41513	2,90658	1,41513		

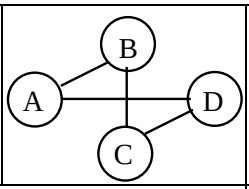
Taula 6.19: **Resultats model loglineal 47**

Model		Mida	396		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,53155	2,93649	1,46175	2,99717	1,52845
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,56513	3,99020	2,91592	1,44094	2,96683	1,49459
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,93409	1,45937	2,93524	1,46066		

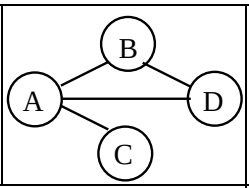
Taula 6.20: **Resultats model loglineal 52**

Model		Mida	396		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50245	2,93998	1,44340	2,96807	1,47113
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,45283	3,91452	2,90127	1,40722	2,95402	1,45709
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,91584	1,42108	2,91602	1,42132		

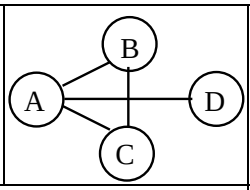
Taula 6.21: **Resultats model loglineal 61**

Model		Mida	396		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52931	2,97083	1,49978	2,96709	1,49344
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,56094	3,99322	2,92395	1,45063	2,96896	1,49631
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,93792	1,46432	2,93798	1,46438		

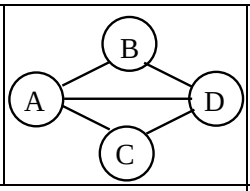
Taula 6.22: **Resultats model loglineal 62**

Model		Mida	406		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52638	2,95390	1,47593	2,97175	1,49732
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,54677	3,97325	2,91832	1,44079	2,96283	1,48647
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,92873	1,45139	2,92877	1,45144		

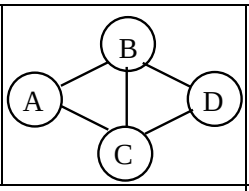
Taula 6.23: **Resultats model loglineal 63**

Model		Mida	406		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52759	2,97052	1,49587	2,97296	1,49977
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,56323	3,99563	2,93267	1,45851	2,97174	1,49780
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,94898	1,47478	2,94849	1,47434		

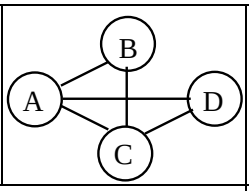
Taula 6.24: **Resultats model loglineal 64**

Model		Mida	428		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,53205	2,95161	1,47996	2,95161	1,47996
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,523582	3,95992	2,89421	1,42175	2,95161	1,47996
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,90622	1,43359	2,90622	1,43359		

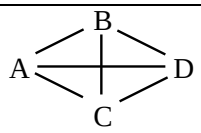
Taula 6.25: **Resultats model loglineal 65**

Model		Mida	421		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,50270	2,94590	1,44927	2,94814	1,45171
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,43935	3,90098	2,88824	1,39485	2,947802	1,45046
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,90514	1,41099	2,90520	1,41106		

Taula 6.26: **Resultats model loglineal 66**

Model		Mida	428		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,53012	2,97693	1,50686	2,94520	1,47118
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,55180	3,97804	2,90689	1,43419	2,96106	1,48880
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,92742	1,45437	2,92606	1,45298		

Taula 6.27: **Resultats model loglineal 51**

Model		Mida	452		
Inèrcia Total		Condicionat per B		Condicionat per C	
In Z	In B	In Z	In B	In Z	In B
3	1,52881	2,93705	1,45934	2,94022	1,46377
Condicionat per B i C		Condicionat per B*C		Juxtaposició B C	
In Z	In B	In Z	In B	In Z	In B
6,50025	3,92312	2,87483	1,39738	2,93864	1,46153
Condicionat succ B C		Condicionat succ C B			
In Z	In B	In Z	In B		
2,89267	1,41498	2,89283	1,41520		

6.3.3 Interpretació

Totes aquestes inèrcies productes del condicionament ens porten als següents resultats: en primer lloc el condicionament simultani per B i C no és fàcilment interpretable, des del punt de vista de les inèrcies, ja que aquesta sobrepassa àmpliament la inèrcia total i no es pot relacionar amb ella. Per aquests motius i les complicacions que ja tenia de forma intrínseca descartarem aquest condicionament.

El condicionament per l'encreuament, no és tan fàcil d'interpretar atenent al valor de la diferència entre la inèrcia total i la inèrcia condicionada, com el condicionament simple. El motiu és el gran increment que sofreix la I_B, ja que si ens fixem en el model de la Taula 6.1, sense cap relació entre les variables, les inèrcies degudes al condicionament per $B = 2,99265$ i per $C = 2,99265$ són valors molt propers a la inèrcia total 3, mentre que la inèrcia pel condicionament $B * C = 2,97714$ és un valor molt allunyat degut només a l'increment de modalitats.

Els valors inercials per la juxtaposició de B i C, només fan que confirmar-nos que són la mitjana dels valors de les inèrcies de les dues inèrcies per separat. La problemàtica de la seva interpretació també ens portarà a descartar aquesta proposta.

Finalment, els condicionaments successius, ens porten a observar que, primer de tot, no hi ha massa diferència en termes absoluts de l'ordre del condicionament. L'error en tots els models estudiats no supera les dues mil·lèsimes. En segon lloc es pot observar,

que tot i que els valors finals són aproximadament iguals, les diferències intermèdies són diferents, atenent a la presència de les variables condicionadores en el model.

Així, per exemple, la variació de les inèrcies del model 57 (Taula 6.11) on la variable B està fortament lligada a la resta de variables, el resultat de les inèrcies és significativament gran quan B és la variable condicionadora, ja sigui com a primera o com a segona condicionadora. En la taula 6.28 queden resumides les inèrcies d'aquests condicionaments.

Taula 6.28: Taula inèrcies model 6.11

Inèrcia total	1r condicionament	2n condicionament
3	(per B) 2,95059	(per C) 2,94731
3	(per C) 2,99838	(per B) 2,94836

Aquest resultat es dóna en tots els models estudiats i es pot interpretar anàlogament com es fa amb els models loglineals i la diferència de deviàncies.

6.4 Proposta d'ACM multicondicional

Per tal de realitzar una ACM multicondicional, la situació ideal tal com hem vist seria tenir o bé variables condicionadores independents o el que és equivalent ortogonalitzar-les. Ara bé, tant la primera situació com la segona són dificultoses.

La primera perquè mai les variables són independents totalment, cosa gairebé improbable, o perquè no podem construir una variable com a creuament de totes les modalitats de les variables condicionadores, ja que moltes vegades no ens porporcionen tota la informació en forma disjuntiva, sinó ja agregada. La segona situació, l'ortogonalització, si bé és sempre factible, la dificultat d'interpretació dels resultats pot fer-ho gairebé inservible com a metodologia.

Per tant, la proposta seria fer, en aquells casos que no sigui possible construir la variable condicionadora resultat de creuar totes les modalitats, un condicionament respecte a totes les variables en la línia dels presentats en la secció anterior, i analitzar-los des d'aquesta perspectiva.

El que farem, doncs, és un procés d'analitzar l'efecte condicionador de cadascuna de les variables sobre els parells restants. Un cop fet això i donat un nivell de significació, guardarem en el nostre model, les arestes corresponents a inèrcies Between significants, donat un nivell α.

A l'hora de confeccionar el model, tindrem en compte també les relacions que ens dóna la taula de contingència productes del creuament de tots els parells de variables.

Exposem a continuació un resultat genèric obtingut en l'anàlisi de les inèrcies d'una taula de 4 variables i els mètodes que anomenarem **mètode directe** i **mètode invers** d'obtenció

6.4.1 Resultats genèrics i descripció del mètode

Hem elaborat una rutina que ens dóna per un costat l'anàlisi de totes les parelles possibles de taules de contingència i per altra les inèrcies between de cada parell en referència a una tercera variable. Els resultats els podem trobar en les Taules 6.29 i 6.30.

Taula 6.29: Resultats de contingència per a cada parell

variables	df-s	χ^2	inèrcia	p-valors
AB	9	32.3691	0.0879595	0.000172
AC	9	2.2435	0.0060965	0.987006
AD	9	32.3691	0.0879595	0.000172
BC	9	1.9216	0.0052217	0.992648
BD	9	7.4403	0.0202183	0.591374
CD	9	1.9216	0.0052217	0.992648

D'aquesta primera taula ja en podríem trobar un primer model que ens recull els lligams de les variables, dues a dues, sense tenir en compte els efectes de terceres variables. Podem veure, en aquest cas genèric uns p-valors molt petits corresponents als parells AB i AD. Això ens podria portar al següent model obtingut mitjançant contingència:

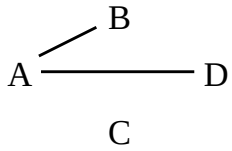


Figura 6.1: Graf del model de contingència

La rutina ens dóna a continuació els resultats que trobem a la següent taula (Taula 6.30), on a més de la inèrcia between, donem el p-valor del contrast associat i l'extrem de la regió d'acceptació.

Taula 6.30: Inèrcies between per a cada parell en referència a una tercera variable

variables	var. condi.	df2	χ^2	Iner. between	p-valor	extrem sup.
AB	C	21	4.1651	0.0056591	0.999972	0.0443894
AB	D	21	39.8094	0.0540889	0.007844	0.0443894
AC	B	21	34.2907	0.0465906	0.033725	0.0443894
AC	D	21	34.2907	0.0465906	0.033725	0.0443894
AD	B	21	39.8094	0.0540889	0.007844	0.0443894
AD	C	21	4.1651	0.0056591	0.999972	0.0443894
BC	A	21	34.6126	0.0470280	0.031124	0.0443894
BC	D	21	9.3619	0.0127200	0.986069	0.0443894
BD	A	21	64.7382	0.0879595	0.000002	0.0443894
BD	C	21	3.8431	0.0052217	0.999986	0.0443894
CD	A	21	34.6126	0.0470280	0.031124	0.0443894
CD	B	21	9.3619	0.0127200	0.986069	0.0443894

D'aquesta taula en podem treure els condicionaments significants i els no significants donat un nivell de significació, que podem prendre, per exemple igual a $\alpha = 0.05$. Per exemple vegem clarament que el parell AB condicionat per D ens dóna un valor de la inèrcia condicional per sobre de l'extrem superior de la regió d'acceptació amb un p-valor molt inferior a α. Per tant això ens marcaria una relació a tenir en compte.

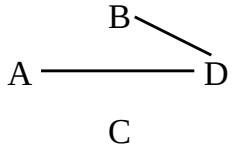
Per altra banda podem veure de la mateixa manera que el parell AB condicionat per C ens dóna un valor de la inèrcia condicional molt per sota de l'extrem amb un p-valor proper a 1. Aquesta relació ens implicaria la no dependencia condicional.

D'aquí han sorgit les dues maneres d'obtenir el model que explicitem en la següent secció.

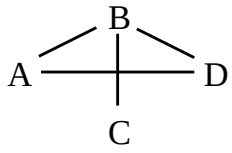
6.4.2 Mètodes d'obtenció del model

Inicialment vàrem obtenir el model pel **mètode directe**, és a dir, un cop obtingudes les relacions de condicionament significatives, ens caldrà unir cadascuna de les variables condicionades amb la seva condicionadora i representar aquestes arestes en un graf.

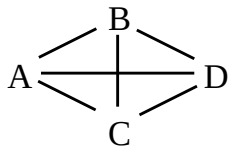
Per exemple la relació significativa AB condicionat per D, això ens afegiria les arestes AD i BD al nostre graf:



a continuació afegiríem la propera relació significativa, la relació AC amb B



i així amb totes les relacions significatives, arribant al model directe final següent:

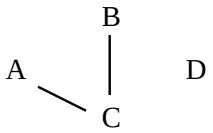


on queden confoses les relacions entre els vèrtexs i els múltiples condicionaments. Això ens va portar al mètode invers.

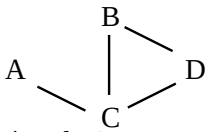
L'obtenció del model pel **mètode invers**, es basa en un cop obtingudes les relacions de condicionament no significants, unir cadascuna de les variables condicionades amb la seva condicionadora, representar aquestes arestes no significants en un graf i obtenir el complementari d'aquest graf, és a dir aquell on les arestes són les que no es troben en el graf d'inèrcies no significants.

Aquesta metodologia ens assegurarà que no sortirà le relació AB a menys que no sigui significant en tots els condicionaments respecte una tercera variable. Aquest **mètode invers** precisarà la majoria de vegades ser complementat amb el graf obtingut del model de contingència. Aquesta complementació la prendrem en el sentit d'afegir aquelles arestes que no hi siguin directament ni tampoc hi siguin per via indirecta de relació a través d'una altra variable.

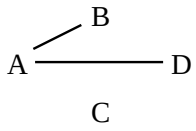
L'obtenció del mètode invers començaria afegint al graf les arestes AC i BC, corresponents al condicionament de AB per C



i, a continuació, afegir-hi les arestes AC i DC corresponents al condicionament AD per C, BD i CD corresponents al condicionament BC per D i així totes les no significatives obtenint el graf:

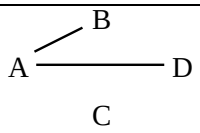
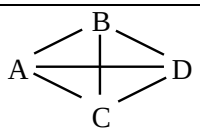
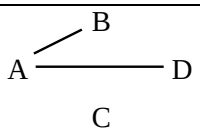
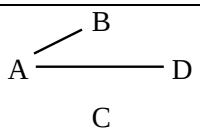


Finalment el graf obtingut pel mètode invers seria el graf complementari a aquest:



Així doncs obtindríem els grafs que es detallen en la següent Taula:

Taula 6.31: **Models obtinguts amb els diferents mètodes**

Contingència	Mètode Directe	Mètode Invers	Mètode invers + cont.
			

Com podem veure, el mètode directe no ens dóna el mateix resultat que els altres tres mètodes, hi ha moltes més arestes. Això ens ha portat a realitzar un estudi, amb models coneguts per tal de veure l'eficàcia del mètodes directe, invers i invers més contingència.

6.4.3 Modelització de dades generades

Presentem a continuació tres taules que recullen la primera models generats per a tres variables i la resposta dels models directe, invers i invers+contingència, comparat amb el model obtingut amb el paquet MIM i les dues següents models per a 4 variables i les respostes dels models.

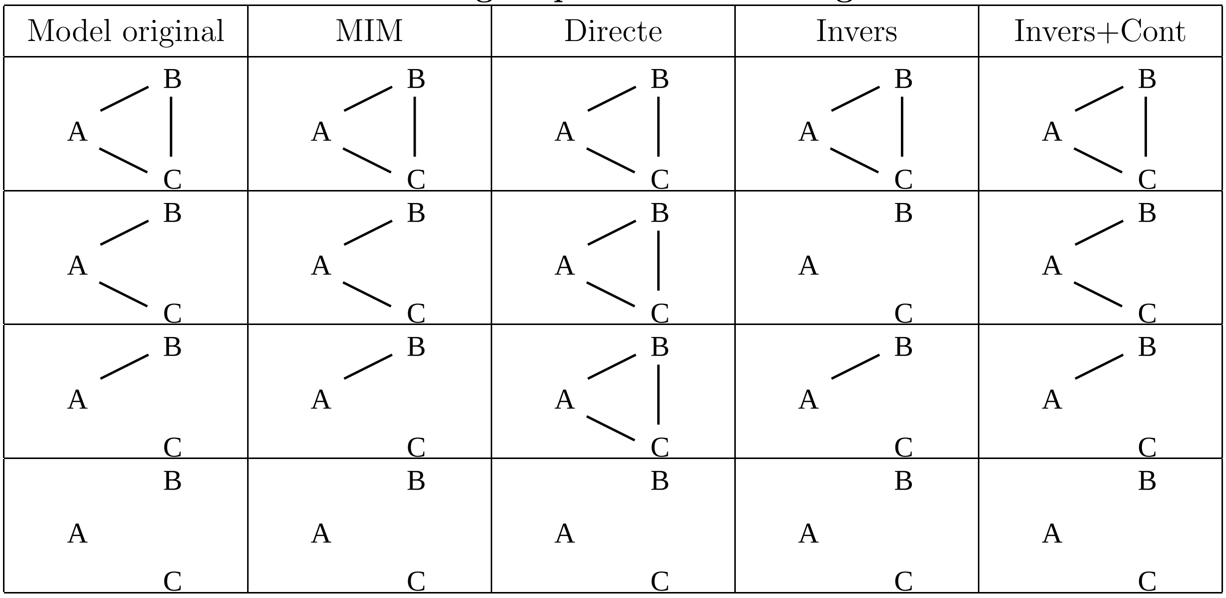
De l'anàlisi d'aquestes taules en podem treure les següents conclusions:

- Per als 4 models amb tres variables, el MIM obté en tots els casos el model original, el mètode invers complementat també, mentre que l'invers sense complementar amb contingència coincideix amb 3 de 4 i el mètode directe amb 2 de 4.

- Per a la primera taula amb quatre variables, amb 10 models amb interaccions iguals, el MIM n'aconsegueix trobar 5 de 10, el mètode invers complementat 7 de 10 i el mètode directe 2 de 10.

- Per a la segona taula amb quatre variables, amb 10 models amb interaccions diferents, el MIM n'aconsegueix 2 de 10, el mètode invers complementat 6 de 10 i el mètode directe 8 de 10. En aquest cas al mètode invers complementat només se

Taula 6.32: **Models obtinguts per a 3 variables iguals interaccions**

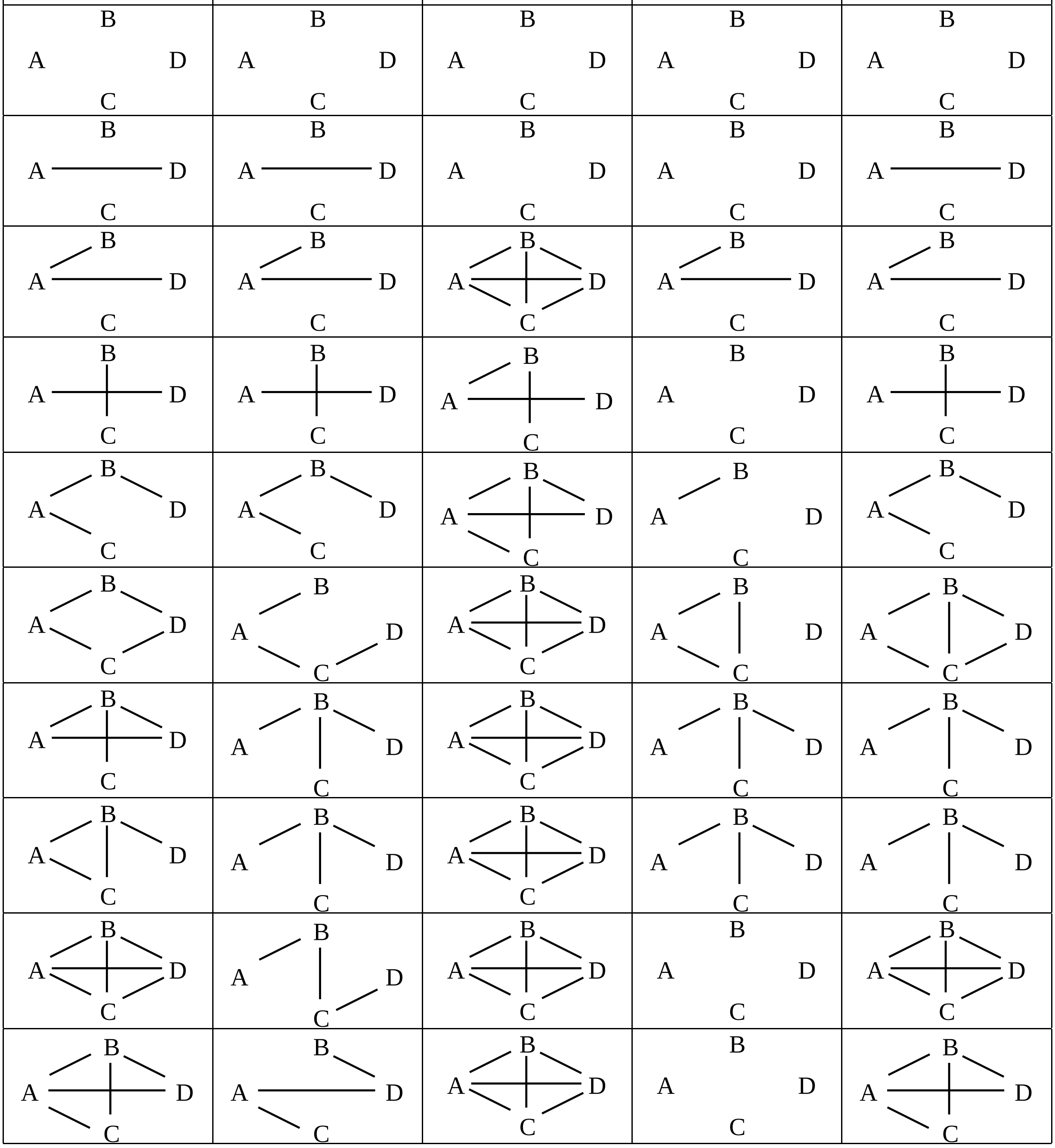


li han afegit les arestes mínimes per a tal de garantir totes les relacions sorgides d'independència via indirecte. Si afegíssim totes les relacions directes, assoliríem un 9 de 10.

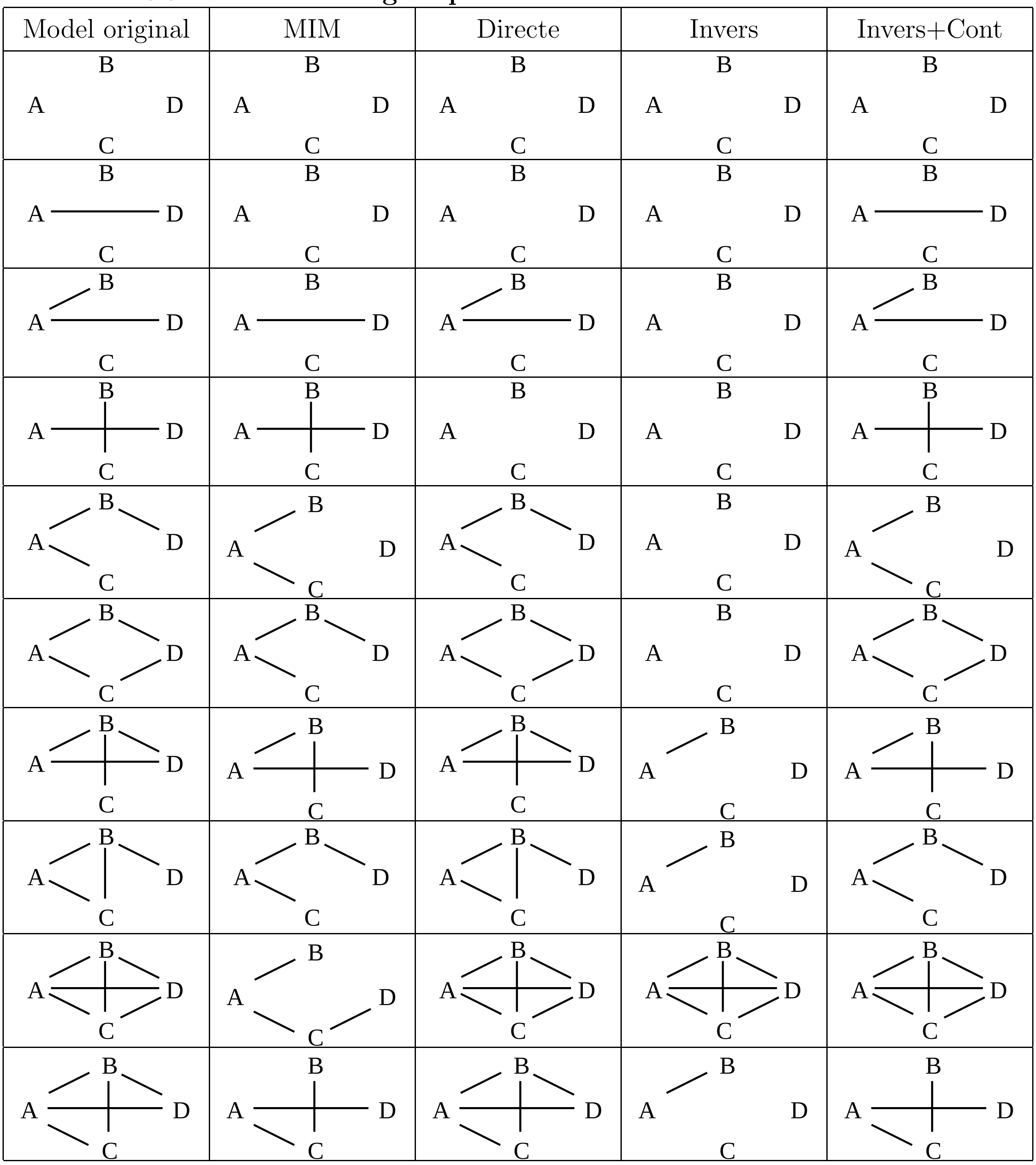
- El model directe troba usualment més relacions de les existents, ja que en no tenir en compte els condicionaments successius, troba relacions indirectes.
- El model invers, troba menys relacions de les existents, ja que la seva visió és molt més restrictiva. Ara bé complementat amb contingència dóna resultats molt aproximats al model real.
- El model directe presenta molts més errors quan hi ha interaccions amb els mateixos efectes i, en canvi, dóna resultats molt més aproximats amb els models amb diferents interaccions.
- El fet que el MIM no trobi molts de models pot ser degut al seu algorisme de selecció de models ja que vol trobar models gràfics i en els models generats no hi ha interaccions de tercer ordre.
- Finalment, hem de tenir en compte que els models obtinguts a partir d'inèrcies no contemplen interaccions superiors a ordre 2 encara que els gràfics d'independència ens ho puguin fer semblar.

Taula 6.33: **Models obtinguts per a 4 variables iguals interaccions**

Model original	MIM	Directe	Invers	Invers+Cont



Taula 6.34: **Models obtinguts per a 4 variables diferents interaccions**



Aquests resultats, en no fer condicionaments successius i en no haver-hi interaccions d'ordre superior, poden donar-nos resultats esbiaixats pel que fa al MIM, però els considerem prou bons pel que fa sobretot al mètode invers complementat i en menor mesura al mètode directe.

Per això, el que farem a continuació és aplicar aquesta metodologia a unes dades reals amb models desconeguts, un exemple amb 3 variables i un altres amb 7 variables i comparar els models obtinguts amb el MIM amb la modelització directa i la modelització inversa complementada amb contingència.

6.4.4 Modelització de dades reals

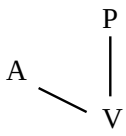
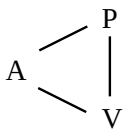
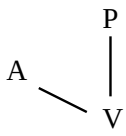
En aquesta secció presentem dos exemples aplicat a dades reals. Les primeres dades corresponen a l'exemple donat per M.L. Radelet i G.L. Pierce [RP91], extret d'[Agr02], en relació a un estudi sobre els efectes de les característiques racials en persones acusades d'homicidi. El segon exemple és el ja tractat en el Capítol 4, en la secció 4.7.2 d'aquesta tesi. Les dades corresponen 1000 registres referents a 7 variables socioeconòmiques sobre condicions de vida.

Exemple 1: Característiques racials i sentència

En aquest exemple, els 674 acusats estan classificats d'acord amb 3 variables A=raça de l'acusat, V=raça de la víctima, ambdues amb les categories -blanc, negre- i la variable P=veredicte de pena de mort amb les categories -sí, no-. (Aquestes dades ens proporcionen un cas més de la coneguda paradoxa de Simpson)

Podem trobar a continuació els resultats de la nostra modelització comparats amb els obtinguts amb el MIM.

Taula 6.35: **Models per al veredicte segons les races**

MIM	Directe	Invers+Cont
		

Tal com es pot observar els resultats obtinguts amb el modelat per MIM i el mètode invers complementat ens donen el mateix model, mentre que en el mètode directe apareix també la relació entre la raça de l'acusat i el veredicte de pena de mort.

Exemple 2: Condicions de vida i aspiracions

El segon exemple que modelitzarem són les dades corresponents a 1000 registres referents a 7 variables socioeconòmiques sobre les condicions de vida i aspiracions dels francesos que tenim a la Taula 4.10. Aquestes variables han estat etiquetades d'acord amb la següent taula:

etiqueta	A	B	C	D	E	F	G
variable	Sexe	Educació	Allotjam	Accions	Propietat	Edat	MidaMun

Els resultats que es presenten a continuació són els obtinguts amb el paquet MIM i els obtinguts amb el mètode invers complementat. Hi ha diversos resultats obtinguts amb el MIM ja que s'han experimentat diferents models que s'explicaran després dels resultats obtinguts.

En la Taula 6.36 podrem trobar els models, les deviàncies d'aquests, els graus de llibertat i finalment els criteris AIC i BIC de cada model. S'indiquen els dos millors segons els criteris d'informació en negreta.

Taula 6.36: **Models per a les condicions de vida i aspiracions del francesos**

Model	Fórmula	Dev	df	AIC	BIC
SW1	ABG,CFG,CDG,CDE,BFG,BDG	1191,81	3723	-6254,19	-24525,76
SW2	DFG,DE,CFG,CE,BE,BFG,ABG	1227,84	3741	-6254,16	-24614,08
SW3	DF,DE,CF,CE,BD,ABDG	1387,59	3770	-6152,41	-24654,65
SW4	DE,CFG,CE,BFG,BDF,A	1254,50	3769	-6283,50	-24780,83
SW5	BFG,BDF,CDE,BCD,ABCG	1197,00	3652	-6107,00	-24030,12
SW6	CFG,CDG,CDE,BFG,BDG,A	1238,66	3751	-6263,34	-24672,33
UN1	FG,A,B,C,D,E	2116,35	3965	-5813,65	-25272,90
SF1	DE,CG,CF,CE,BF,AC	1577,50	3934	-6290,50	**-25597,60**
SFU	DE,CFG,CE,BFG,BDF,AC	1233,02	3766	-6298,98	-24781,58
AIC	CDG,CDF,CDE,BDF,AC	1426,08	3876	**-6325,92**	-25348,37
BIC	G,F,DE,CE,BD,A	2003,74	3973	-5942,26	-25440,78
IN	AC,CB,CD,CE,CF,CG,BD,BF,BG,DE,DF	1406,61	3895	**-6383,39**	**-25499,10**

- Els models **SW** han partit d'una primera etapa *stepwise* i després s'ha eliminat alguna de les arestes, que no eren eliminades automàticament per respectar el fet que fossin models gràfics i descomposables, continuant el procés *stepwise* a continuació. S'han obtingut 6 models d'aquest tipus.
- El model **UN** s'ha realitzat sense posar restriccions de tipus descomposable o gràfic.

- Els models **SF** s'han realitzat amb la metodologia *forward* el **SF1** amb restriccions de model gràfic i descomposable i el **SFu** sense restriccions.
- Els models **AIC** i **BIC** s'han elaborat amb la metodologia *stepwise* on el criteri de selecció estava basat respectivament en els criteris AIC i BIC. No són els models seleccionats entre tots els models directament pel mínim AIC o BIC.
- Finalment, el model **IN** ha estat obtingut mitjançant la metodologia inercial, amb mètode invers complementat.

Podem veure que aquest darrer, en el nostre procés de selecció, és el millor pel que fa referència al criteri d'informació d'Akaike i és el segon millor segons el criteri d'informació Bayesià, per tant el model seleccionat és prou correcte i millor que força dels obtinguts per procediments de selecció automàtica.

6.5 Conclusions

En aquest capítol s'ha abordat la problemàtica de l'ACM multicondicional, en el qual s'han estudiat diverses propostes, defugint l'otogonalització per la seva complexitat a l'hora d'interpretar els resultats. En aquestes s'ha vist que el condicionament successiu era l'opció més factible, sempre i quan no tinguem l'opció de creuar les variables i els resultats d'aquests creuaments. A més el condicionament successiu és interpretable via les inèrcies inter i intra dels condicionaments resultants. En el camp dels condicionaments successius hi pot haver més treball a fer per aprofundir i no restringir-ho a només dues variables condicionadores, si no a més, tot i que els creuaments múltiples ens poden portar a categories creuades amb nuls o molt pocs individus que satisfacin el creuament.

Una proposta alternativa és la que hem presentat basada en l'efecte del condicionament de cadascuna de les variables sobre la resta de variables. Això en ha portat a diverses solucions, d'entre les quals destaquem pels bons resultats obtinguts el mètode invers complementat. Tot i això, hi pot haver petites diversificacions a l'hora de complementar el mètode invers amb el model de contingència.

Els resultats obtinguts pel mètode invers complementat són equiparables als obtinguts mitjançant el paquet MIM.

Capítol 7

Aplicació a un exemple

En aquest darrer capítol i com a síntesi del treball de recerca el que s'ha fet és, per tal de no analitzar la totalitat de la matriu de Burt d'un conjunt de variables categòriques, aplicar primer la metodologia de selecció de models basada en les inèrcies condicional per a trobar un model per a les dades i un cop obtingut aquest model, fer-ne la reconstitució basada en la metodologia $k - EM$ segons el model obtingut anteriorment.

Les dades escollides són les corresponents a l'enquesta sobre Condicions laborals i aspiracions del francesos, donades en la secció 4.7.2 i ja tractades com a exemple de modelització en el capítol anterior, on s'ha modelitzat seguint diferents metodologies.

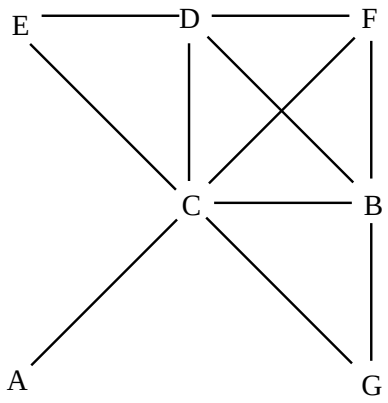
Per a aquestes variables conservem la codificació que ja hem fet servir en el capítol anterior i que és la que trobem en la següent Taula:

etiqueta	A	B	C	D	E	F	G
variable	Sexe	Educació	Allotjam	Accions	Propietat	Edat	MidaMun

En primer lloc, utilitzarem el model trobat mitjançant bel mètode invers complementat corresponent al model **IN** de la Taula 6.36 del capítol anterior, i que podem veure en la Figura 7.1, per a fer la reconstitució de les caselles corresponents a les subtaules de Burt que no estan associades a cap aresta del gràfic. A més també reconstituirem les subtaules diagonals de la taula de Burt. Podem veure en aquest gràfic el paper central que ocupa la variable C corresponent a l'Allotjament, contraposat a la variable A, -Sexe- que només està lligada a la variable allotjament i a cap de les altres.

Podem trobar a la Taula 7.1 un esquema de la reconstitució, on es troben marcades amb l'índex corresponent les subtaules de Burt que reconstituirem.

Figura 7.1: **Graf d'interaccions del model invers complementat**



Taula 7.1: **Taula de reconstitucions de les subtaules de Burt**

variables	A	B	C	D	E	F	G
A	11	12		14	15	16	17
B	21	22			25		
C			33				
D	41			44			47
E	51	52			55	56	57
F	61				65	66	67
G	71			74	75	76	77

A continuació farem una reconstitució $2 - EM$ segons les caselles de la Taula 7.1 i obtindrem la taula de dades reconstituïdes que podem trobar a la Taula 7.2. En aquesta reconstitució hem tingut un petit problema ja que ens sortien algunes caselles amb valors negatius en no fer una reconstitució completa i hem truncat aquests efectius a zero.

Taula 7.2: *Taula de Burt reconstituïda de les 7 variables de les condicions de vida i aspiracions dels francesos*

	SEXM	SEXF	ED1	ED2	ED3	ED4	ED5	LO1	LO2	LO3	LO4	ST1	ST2	HO1	HO2	AG1	AG2	AG3	AGE	AG5	SI1	SI2	SI3	SI4	SI5
SEX1	228	244	95	161	72	74	67	62	151	224	32	64	405	54	419	17	103	164	99	97	53	45	89	149	137
SEX2	244	288	100	166	88	94	83	58	142	302	29	56	475	35	501	24	147	193	91	75	35	43	88	180	189
EDU1	95	100	40	67	30	30	27	17	58	106	15	11	185	24	172	9	26	55	45	61	19	16	36	70	55
EDU2	161	166	67	114	49	49	45	45	116	151	15	27	300	43	283	0	57	124	82	64	43	35	67	109	73
EDU3	72	88	30	49	27	29	26	16	37	98	9	19	141	9	153	14	47	60	27	12	12	13	33	52	50
EDU4	74	94	30	49	29	32	28	27	40	86	14	28	139	5	164	16	62	63	17	9	8	15	21	56	67
EDU5	67	83	27	45	26	28	24	15	42	85	8	36	114	7	145	1	55	54	17	23	1	8	18	42	81
LOD1	62	58	17	45	16	27	15	14	33	64	7	11	109	7	113	3	23	68	20	6	7	20	27	48	18
LOD2	151	142	58	116	37	40	42	33	121	117	16	60	233	48	245	9	24	91	80	89	59	33	62	72	67
LOD3	224	302	106	151	98	86	85	64	117	311	34	45	481	23	503	22	180	182	79	63	11	29	80	191	215
LOD4	32	29	15	15	9	14	8	7	16	34	4	5	56	4	57	6	20	15	9	11	6	5	6	18	26
STO1	64	56	11	27	19	28	36	11	60	45	5	24	95	39	82	2	18	35	27	39	25	15	28	33	20
STO2	405	475	185	300	141	139	114	109	233	481	56	95	782	43	836	38	229	321	161	130	62	72	147	296	307
HOU1	54	35	24	43	9	5	7	7	48	23	4	39	43	31	55	0	0	24	36	46	31	15	26	19	0
HOU2	419	501	172	283	153	164	145	113	245	503	57	82	836	55	872	42	260	335	153	123	55	72	150	314	333
AGE1	17	24	9	0	14	16	1	3	9	22	6	2	38	0	42	2	15	16	4	1	0	2	5	15	19
AGE2	103	147	26	57	47	62	55	23	24	180	20	18	229	0	260	15	96	98	25	6	0	13	31	95	118
AGE3	164	193	55	124	60	63	54	68	91	182	15	35	321	24	335	16	98	129	62	52	24	29	59	120	125
AGE4	99	91	45	82	27	17	17	20	80	79	9	27	161	36	153	4	25	62	51	57	35	23	42	54	37
AGE5	97	75	61	64	12	9	23	6	89	63	11	39	130	46	123	1	6	52	57	70	45	25	44	43	16
SIZ1	53	35	19	43	12	8	1	7	59	11	6	25	62	31	55	0	0	24	35	45	30	15	26	19	0
SIZ2	45	43	16	35	13	15	8	20	33	29	5	15	72	15	72	2	13	29	23	25	15	10	19	26	19
SIZ3	89	88	36	67	33	21	18	27	62	80	6	28	147	26	150	5	31	59	42	44	26	19	36	53	43
SIZ4	149	180	70	109	52	56	42	48	72	191	18	33	296	19	314	15	95	120	54	43	19	26	53	113	121
SIZ5	137	189	55	73	50	67	81	18	67	215	26	20	307	0	333	19	118	125	37	16	0	19	43	121	145

Podem trobar a la Taula 7.3 la descripció de les inèrcies. En els dos primers eixos factorials tenim resumida el 85.68% d'inèrcia de la taula, què és $I_{Total} = 0.0644673$.

Taula 7.3: Valors propis, percentatges d'inèrcia i percentatges acumulats

Valor propi	% Inèrcia	% Iner.Acum.
0,0522153	80,9951	80,995
0,0030183	4,6820	85,677
0,0023843	3,6985	89,375
0,0018739	2,9068	92,282
0,0013742	2,1316	94,414
0,0009827	1,5243	95,938
0,0009108	1,4128	97,351
0,0005066	0,7859	98,137
0,0003729	0,5784	98,715
0,0002382	0,3694	99,085
0,0002114	0,3280	99,413
0,0001690	0,2621	99,675
0,0001067	0,1655	99,840
0,0000454	0,0704	99,911
0,0000403	0,0625	99,973
0,0000070	0,0109	99,984
0,0000052	0,0081	99,992
0,0000023	0,0036	99,996
0,0000014	0,0022	99,998
0,0000011	0,0016	99,999
0,0000002	0,0003	100,000
0,0000001	0,0002	100,000
0,0000000	0,0000	100,000
0,0000000	0,0000	100,000

La descomposició de la inèrcia total com a aportació d'inèrcia de cadascuna de les inèrcies de les subtaules de la matriu de Burt, després de la reconstitució, la podem trobar a la Taula 7.4. Podem veure en aquesta taula que les inèrcies corresponents a les subtaules no reconstituïdes ens donen una inèrcia de 0.030842 sobre el total, mentre que la resta correspon a la inèrcia de les subtaules reconstituïdes, tant les pertanyents a la diagonal com aquelles donades pel model inercial.

Taula 7.4: *Inèrcies per cada subtaula de contingència de la Taula de Burt reconstituïda segons el model inercial*

	Sexe	Educació	Allotjam	Accions	Propietat	Edat	MidaMun
Sexe	0,0000121	0,000042	0,0001692	0,0000459	0,0001514	0,0002464	0,0002262
Educació	0,000042	0,0001251	0,0004752	0,0007306	0,000474	0,0029798	0,001433
Allotjam	0,0001692	0,0004752	0,000766	0,0005627	0,0007188	0,0029358	0,0028168
Accions	0,0000459	0,0007306	0,0005627	0,0001741	0,0020128	0,0005863	0,0008029
Propietat	0,0001514	0,000474	0,0007188	0,0020128	0,0017436	0,0025661	0,0026129
Edat	0,0002464	0,0029798	0,0029358	0,0005863	0,0025661	0,004576	0,0040041
MidaMun	0,0002262	0,001433	0,0028168	0,0008029	0,0026129	0,0040041	0,0038825

Per a finalitzar l'exemple presentem els gràfics factorials d'aquestes dades, com que la taula és simètrica només en presentem un:

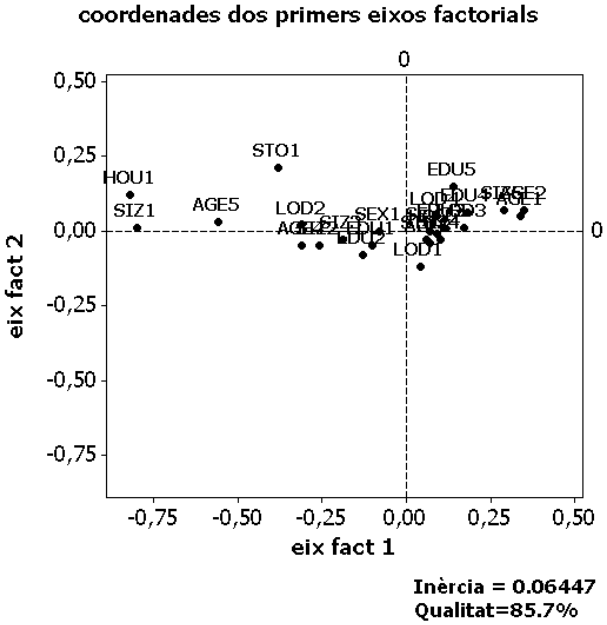


Figura 7.2: Gràfic factorial de les dades de la Taula 7.2

I per a millorar la interpretació, proporcionem les coordenades i ajudes de la interpretació de les fileres i columnes, en una única taula, ja que són les mateixes per simetria.

Aquests resultats ens permeten comparar-los amb les taules i informacions de les dades originals (Taula 4.10) i les taules i informacions després de reconstituir la diagonal (Taula 4.12).

Taula 7.5: Coordenades, contribucions i cosinus quadrats

modal	cor.1	cor.2	cor.3	cnt.1	cnt.2	cnt.3	cos.1	cos.2	cos.3
SEX1	-0,08	0,00	0,00	0,8	0,0	0,0	0,93	0,00	0,00
SEX2	0,07	0,00	0,00	0,8	0,0	0,0	0,93	0,00	0,00
EDU1	-0,10	-0,05	-0,07	0,6	1,9	5,1	0,26	0,05	0,11
EDU2	-0,13	-0,08	-0,06	1,6	9,4	7,1	0,56	0,19	0,11
EDU3	0,11	0,01	-0,01	0,5	0,0	0,1	0,45	0,00	0,00
EDU4	0,18	0,06	0,08	1,5	3,0	6,0	0,55	0,06	0,10
EDU5	0,14	0,15	0,14	0,8	17,0	17,8	0,24	0,31	0,25
LOD1	0,04	-0,12	-0,09	0,0	8,8	6,4	0,02	0,21	0,12
LOD2	-0,31	0,02	0,02	7,8	0,6	0,4	0,94	0,00	0,00
LOD3	0,17	0,01	0,02	4,2	0,1	0,8	0,89	0,00	0,01
LOD4	0,09	0,05	0,02	0,1	0,6	0,2	0,17	0,05	0,01
STO1	-0,38	0,21	-0,20	4,7	25,5	27,7	0,59	0,18	0,16
STO2	0,06	-0,03	0,03	0,8	4,4	4,6	0,59	0,18	0,15
HOU1	-0,82	0,12	0,11	16,5	6,4	7,0	0,94	0,02	0,02
HOU2	0,09	-0,01	-0,01	2,0	0,3	0,3	0,96	0,01	0,01
AGE1	0,34	0,05	-0,03	1,2	0,5	0,3	0,46	0,01	0,00
AGE2	0,35	0,07	-0,04	8,5	5,5	2,9	0,91	0,03	0,01
AGE3	0,07	-0,04	0,02	0,4	2,2	0,7	0,32	0,10	0,03
AGE4	-0,31	-0,05	0,00	5,0	2,5	0,0	0,94	0,03	0,00
AGE5	-0,56	0,03	-0,02	14,6	0,7	0,5	0,93	0,00	0,00
SIZ1	-0,80	0,01	0,02	15,6	0,1	0,1	0,98	0,00	0,00
SIZ2	-0,26	-0,05	0,04	1,7	1,1	0,8	0,76	0,03	0,02
SIZ3	-0,19	-0,03	0,03	1,7	1,0	1,2	0,86	0,03	0,03
SIZ4	0,10	-0,03	0,03	0,9	1,6	2,1	0,75	0,08	0,08
SIZ5	0,29	0,07	-0,06	7,6	6,7	7,7	0,86	0,04	0,04

Podem veure, en referència a la qualitat de representació del primer pla factorial, que hem passat d'un 27.2% en les dades inicials, a un 78.4% en la reconstitució de la diagonal i ara a un 85.7%. Pot semblar poca aquesta darrera diferència, però si en fixem en només el primer eix en la reconstitució de la diagonal ens aporta un 49.8%, mentre que en la reconstitució sorgida del model inercial el primer eix té un 81.0%.

Aquesta diferència queda també manifestada en els gràfics factorials, ja que la disposició dels punts en el gràfic de la Figura 7.2 és molt més propera al primer eix, amb una dispersió molt petita en el segon eix, com passava en els gràfics factorials de les dades originals i de la reconstitució de la diagonal.

Es pot fer una anàlisi més completa amb la informació que podem extreure de les contribucions i cosinus quadrats, així com aquelles informacions que poguéssim obtenir d'experts en l'àrea de coneixement.

Capítol 8

Epíleg

En aquest capítol presentem les conclusions relacionades amb els objectius marcats a l'inici d'aquesta tesi doctoral i exposem un llistat de problemes, idees i altres aportacions que han anat sorgint a mesura que s'avançava en la investigació. És del tot evident que el tema d'investigació no l'hem tancat, ans al contrari, a mesura que hem anat aprofundint, hem anat trobant nous problemes i noves vies de desenvolupament. Som conscients que l'estudi amb profunditat de cadascuna de les qüestions obertes és, en realitat, una línia de recerca a estudiar en el futur. A més reiterar que l'objectiu no ha estat trobar un model per a les dades sinó simplificar l'anàlisi descriptiva de les mateixes.

8.1 Conclusions

- El nostre objectiu principal era no analitzar les totalitats de les taules de correspondències simples i múltiples, sinó només aquelles parts d'interès. El que hem fet en el Capítol 2 ha estat fer una presentació dels mètodes d'anàlisis factorials de correspondències, però no una presentació des del punt de vista clàssic de l'anàlisi de perfils, sinó com a casos particulars de l'anàlisi canònica i l'anàlisi canònica generalitzada. Aquesta presentació ha comportat que el capítol hagi quedat molt extens ja que s'ha volgut presentar-ho veient la seva equivalència amb la presentació més clàssica.

- En el Capítol 3 s'han presentat les bases de la modelització loglineal amb dos motius. El primer de tot ha estat perquè s'utilitzaran models generats per a estudiar el condicionament en les anàlisis condicionals i per a fer servir les metodologies de modelització com a guia dels resultats dels models inercials que hem obtingut.

- En aquest Capítol 3 s'ha realitzat també un estudi de la influència dels paràmetres en la generació de models loglineals, així com s'han trobat les formulacions de canvi de parametrització de dos dels tipus de restriccions més habituals amb les que es formulen els models loglineals.

- En el Capítol 4 tenim els resultats del treball que ens havíem proposat en primer terme: estudiar la problemàtica de la influència de les diagonals i les propostes de solució dins els marc de les anàlisis factorials. Un cop fet aquest estudi, hem desenvolupat una nova proposta en base a les metodologies de tractament dins de l'àmbit de les anàlisis factorials, basada en la descomposició en la part simètrica i l'antisimètrica i la reconstitució k-EM de la part simètrica, després d'analitzar la influència de les diagonals. Aquesta proposta d'anàlisi dóna una interpretació força acurada de les relacions entre les caselles de la taula de contingència objecte de l'anàlisi.

- També en el Capítol 4 s'ha afrontat la problemàtica de les inèrcies de les subtaules de la diagonal de la matriu de Burt i de les propostes de reformulació. També dins del camp de les anàlisis factorials, s'ha estudiat l'aplicació de la metodologia emprada en ACS per tal d'aplicar-la a les subtaules de la diagonal de la taula de Burt. La mateixa reconstitució k-EM s'ha aplicat a la reconstitució de les taules de la diagonal, obtenint una millora de la representació factorial i obtenint uns resultats equivalents al que es coneix com Joint Correspondence Analysis (JCA).

- En el Capítol 5 presentem els ACM respecte a un model i ACM condicional, es presenta la seva formulació i l'expressió del darrer com a un cas particular del primer.

- A continuació en el Capítol 5, realitzem l'estudi de les inèrcies en ACM condicional, descomposant aquesta inèrcia en inèrcia Inter i Intra. L'estudi d'aquestes inèrcies, amb la seva formulació i l'estudi de les seves distribucions, ens permet veure quan un condicionament és o no significatiu. Es realitzen simulacions i tests i es presenten exemple d'aplicació. Això ens permet fer un primer pas cap a la simplificació de l'estudi de les relacions: veure quan aquestes són o no significatives.

- En el Capítol 6, es mira d'estendre l'anàlisi condicional a l'anàlisi multicondiconal, veient les dificultats que comporta per la no ortogonalitat de les modalitats de les diferents variables. Es desenvolupen diferents apropaments defugint l'ortogonalitza-

ció per la seva dificultat d'interpretació, veient que la més útil pot ser l'anàlisi via els condicionaments successius. Es realitza un estudi en base a models generats dels diferents apropaments.

- En aquest Capítol 6 realitzem una proposta de simplificació de les relacions basada en una anàlisi multicondicional. Aquesta anàlisi, però, no és successiva, sinó estudiant l'efecte condicionador de cada variable sobre les altres amb les eines desenvolupades en el Capítol 5.

- Es presenta una metodologia desenvolupada per a buscar les relacions bàsiques entre les nostres variables, basada en la inèrcia condicional. Aquesta metodologia es desenvolupa sobre models generats, per a comprovar el seu bon funcionament i s'usa la modelització loglineal per a contrastar els resultats obtinguts. Es comprova que la metodologia inercial funciona i s'aplica a exemples generats i a exemples de dades reals.

- Aquesta metodologia queda limitada a les interaccions d'ordre 2, ja que no s'ha aprofundit en ordres superiors, tot i que en ordres superiors podríem tenir altres problemàtiques com podria ser l'excessiu nombre de caselles buides en taules de contingència per a més de 2 dimensions. A més, hem de tenir present que habitualment no té sentit buscar associacions d'ordre elevat amb dades provinents d'enquestes, que són una de les fonts de dades més freqüents.

- S'ha aplicat en el Capítol 7 la metodologia desenvolupada per anul·lar la influència de subtaules lligades a relacions espúries, obtingudes pel condicionament entre les variables. Aquesta metodologia s'ha comparat amb els resultats de les taules original i amb la diagonal reconstituïda.

8.2 Línies d'investigació futures

- Una de les línies d'investigació futures seria no restringir-nos a les metodologies en el camp de les anàlisis factorials, sinó expandir-les a altres dominis, com ja hi ha apropaments a la problemàtica de la reconstitució de caselles de les taules via modelització.

- Un altre punt que ja hem destacat seria aprofundir en la metodologia de condicionaments successius d'ordre superior a dos i les interaccions d'ordre superior.

- Estudiar la problemàtica de l'excessiva dispersió en els creuaments de múltiples categoritzacions, quan es realitzin els condicionaments successius.

- Procurar un procés d'automatització mitjançant una aplicació informàtica que realitzi les anàlisis condicionals per a eliminar les relacions no significatives en les anàlisis.

- Continuar la mateixa metodologia d'eliminació de relacions no significatives en les anàlisis de correlacions canòniques generalitzades amb variables categòriques i a traves d'aquí a les tècniques de relacions entre diferents conjunts de variables.

Bibliografia

[AG89] E. Abascal and I. Grande. *Métodos multivariantes para la investigación comercial:Teoría, problemas y programación comercial*. Ariel Economía, 1989.

[Agr02] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics, 2002.

[Aka69] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.

[AMR86] T. Aluja and M. Martí-Recober. Complementary between log-linear models and correspondence analysis. Technical report, Dept. EIO-UPC, 1986.

[BBM80] J.-P. Benzécri, C. Bougarit, and J.-C. Madre. Ajustement d'un tableau a ses marges d'après la formule de reconstitution. *Les Cahiers de l'Analyse des Données*, 5(2):163–172, 1980.

[Ben64] J.-P. Benzécri. Cours de linguistique mathematique. *Publication Mimeo. Faculté des Scienties - Rennes*, 1964.

[Ben69] J.-P. Benzécri. Statistical analysis as a tool to make patterns emerge from data. *Methodologies of Pattern Recognition. Academic-Press*, pages 35–74, 1969.

[Ben73] J.-P. Benzécri. *L'Analyse des Données*. Dunod, Paris, 1973. vols. I,II.

[Ben79] J.-P. Benzécri. Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données*, 4:377–378, 1979.

[Ben92] J.-P. Benzécri. *Correspondence Analysis Handbook*. Marcel Dekker, Inc. New York., 1992.

[Car68] J.D. Carroll. Generalisation of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th annual convention of the American Psychological Association*, 3:227–228, 1968.

[Car85] A. Carlier. Analyse des évolutions sur tables de contingences, quelques aspects opérationels. *Data Analysis and Informatics*, pages 421–428, 1985.

[Cau92a] H. Caussinus. Projections révélatrices. *Modèles pour l'analyse des donneés multidimensionelles*, 1992.

[Cau92b] H. Caussinus. The use of probabilistic models to produce optimal graphical dispays of high-dimensional data sets. *Journal of the Italian Statistical Society*, 1(1):51–66, 1992.

[CdF87] H. Caussinus and A. de Falguerolles. Tableaux carres: Modelisation et mèthodes factorielles. *Revue de Statistique Appliquée*, 35:35–52, 1987.

[CG78] A.G. Constantine and J.C. Gower. Graphical representation of asymmetry. *Applied Statistics*, 27:297–304, 1978.

[Chr90] R. Christensen. *Log-linear Models*. Springer-Verlag, New York, 1990.

[CK96] A. Carlier and P.M. Kroonenberg. Decompositions and biplots in three-way correspondence analysis. *Psychometrika*, 61(2):355–373, 1996.

[Cua91] C.M. Cuadras. *Métodos de Análisis Multivariante*. P.P.U., Barcelona, 1991.

[Daw79] A.P. Dawid. Conditional independence in statistical theory (with discussion). *Journal Royal Statistical Society series B*, 41:1–31, 1979.

[DEAB97] J. Daunis-Estadella and T. Aluja-Banet. Análisis de correspondencias de matrices cuadradas no simétricas. problemática y tratamientos. *Actas XXIII Congreso Nacional de Estadística e Investigación Operativa*, pages 40.3–40.4, 1997. D.L. CS - 94 - 1997.

[DEABTH97] J. Daunis-Estadella, T. Aluja-Banet, and S. Thió-Henestrosa. Reconstitución de datos influyentes en análisis de correspondencias. *NGUS'97 IV International Meeting of Multidimensional Data Analysis*, pages 108–111, 1997. D.L. BI-1514-97.

[DLD76] A.P. Dempster, N.M. Laird, and Rubin D.R. Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society series B*, 93:1–38, 1976.

[DLS80] J.N. Darroch, S.L. Lauritzen, and T.P Speed. Markov fields and loglinear interaction models for contingency tables. *Ann. Stat.*, 8:522–539, 1980.

[DR99] J.M. Duran-Rúbies. Models gràfics d'independència. *Qüesttió*, 23(1), 1999.

[EC65] B. Escofier-Cordier. *L'Analyse Factorielle des correspondances*. PhD thesis, Université de Rennes, 1965. Publicada a Cahiers du Bureau Universitaire de Recherche Opérationelle, núm 13, (1969) pàg.25-39.

[Edw95] D. Edwards. *Introduction to Graphical Modelling*. Springer Verlag, 1995.

[EP90] B. Escofier and J. Pagès. *Analyses Factorielles Simples et Multiples*. Dunod, Paris, 1990.

[Esc84] B. Escofier. Analyse factorielle en référence à un modèle. application à l'analyse de tableaux d'échanges. *Revue de Statistique Appliquée*, 32(4), 1984.

[Esc87] B. Escofier. Analyse des correspondances multiples conditionelle. Technical report, INRIA, 2 Versailles, 1987.

[EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1, 1936.

[Fie94] S.E. Fienberg. *The Analysis of Cross-Classified Categorical Data*. MIT Press, Massachusetts, 1994.

[Fis40] R.A. Fisher. The precision of discriminant functions. *Annals Eugen.*, 10:422–429, 1940.

[FJ93] A. de Falguerolles and S. Jmel. Un modèle graphique pour la selection de variables qualitatives. *Revue de Statistique Appliquée*, 41:23–41, 1993.

[Fou85] T. Foucart. Tableaux symmétriques et tableaux d'échanges. *Revue de Statistique Appliquée*, 33(2):37–54, 1985.

[Gab71] K.R. Gabriel. The biplot -graphical display of matrices with application to principal component analysis. *Biometrika*, 58:453–467, 1971.

[Gab02] K.R. Gabriel. Goodness of fit of biplots and correspondence analysis. *Biometrika*, 89(2):423–436, 2002.

[GB94] M.J. Greenacre and J. Blasius. *Correspondence Analysis in the Social Sciences*. Academic Press, London., 1994.

[Gif91] A. Gifi. *Nonlinear Multivariate Analysis*. John Wiley & Sons. Inc. West Sussex., 1991.

[Goo69] L.A. Goodman. On partitioning χ^2 an detecting partial association in three-way contingency tables. *J. Roy. Statistical Society. Ser. B*, 31:486–498, 1969.

[Goo86] Goodmann. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables (with discussion). *Internat. Statistical Review*, 54(3):243–309, 1986.

[Gow77] John C. Gower. The analysis of asymmetry and orthogonality. *Recent Developments in Statistics*, 1977.

[Gre84] M.J. Greenacre. *Theory and applications of Correspondence Analysis*. Academic Press, London, 1984.

[Gre87] M.J. Greenacre. Correspondence analysis of multivariate categorical data by weighted least squares. *Biometrika*, 75:457–467, 1987.

[Gre89] M.J. Greenacre. Measuring total variation and its components in multiple correspondence analysis. *Suplement to the Proceedings of the Conference: Journées Internationales de l'Analyse des Données*, 1989.

[Gre90] M.J. Greenacre. Some limitations of multiple correspondence analysis. *Computational Statistics Quarterly 3, Physica-Verlag*, pages 249–256, 1990.

[Gre91] M.J. Greenacre. Interpreting multiple correspondence analysis. *Applied Stochastic Models and Data Analysis*, 7:195–210, 1991.

[Gre93a] M.J. Greenacre. *Correspondence Analysis in Practice*. Academic Press, London., 1993.

[Gre93b] M.J. Greenacre. Multivariate generalisations of correspondence analysis. *Multivariate analysis: future directions*, 2:249–256, 1993.

[Gre00] M.J. Greenacre. Correspondence analysis of square asymmetric tables. *Applied Statistics*, 49:297–310, 2000.

[Gut35] L. Guttman. The quantification of a class of attributes: a theory and a method of scale construction. *The prediction of personal adjustement, SSRC New-York*, 1935.

[Hay52] C. Hayashi. On prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Inst. of Stat. Math - Tokyo*, 3(2), 1952.

[HFL89] P.G.M. van der Heijden, A. de Falguerolles, and J. de Leeuw. A combined approach to contingence tables analysys using correspondence analysis and loglinear analysis. *Applied Statistics*, 38(2):249–292, 1989.

[Hir35] H.O. Hirschfeld. A connection between correlation and contingency. *Cambridge Philosofical Society Proceedings*, 31:520–524, 1935.

[HL85] P.G.M. van der Heijden and J. de Leeuw. Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, 50:429–448, 1985.

[Hor35] P. Horst. Measuring complex attitudes. *J. Social Psychology*, 6:369–374, 1935.

[Hot36] H. Hotelling. Relation between two sets of variates. *Biometrika*, 28:321–377, 1936.

[Ide01] Institut d'Estadística de Catalunya Idescat. Estadística de població 1996. *Fluxos de mobilitat obligada pel treball o estudi. Dades comarcals i municipals*, 12:21–25, 2001.

[IK94] M. Ishii-Kuntz. *Ordinal Log-Linear Models*. Sage Publications, 1994.

[Kaz78] J.B. Kazmierczak. Migrations interurbaines dans la banlieue sud de paris. *Cahiers de l'Analyse des Données*, 3(2):203–218, 1978.

[KB80] D. Knoke and P. J. Burke. *Log-Linear Models*. Sage Publications, 1980.

[Kol33] A.N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giorn. Dell'Ist. Ital. degli Att.*, 4:83–91, 1933.

[LB94] C. Lauro and S. Balbi. From exloratory to confirmatory analysis of structured qualitative data. Technical report, Dipartamento di Matematica e Statistica, Università Federico II, Napoli, Italia., 1994(?).

[LD84] C. Lauro and L. D'Ambra. L'analyse non-symétrique des correspondances. *Data Analysis and Informatics III*, pages 433–446, 1984.

[Lec75] A. Leclerc. L'analyse des correspondances sur la juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, 23(3):5–16, 1975.

[LH88] J. de Leeuw and P.G.M. van der Heijden. Correspondence analysis of incomplete contingence tables. *Psychometrika*, 53(2):223–233, 1988.

[LHvE80] L. Lebart and Y. Houzel van Effentere. Le système d'enquête sur les conditions de vie et aspirations des français. *Consommation*, 1, 1980.

[LM85] L. Lebart and A. Morineau. *SPAD Système portable pour l'analyse des données*. CESIA, Paris, 1985.

[LMF85] L. Lebart, A. Morineau, and J.-P. Fénélon. *Tratamiento Estadístico de datos*. Marcombo, S.A., Barcelona, 1985.

[LMW84] L. Lebart, A. Morineau, and K.M. Warwick. *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons, Inc., New York., 1984.

[MKB79] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[MN89] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall. London, 1989. 2nd edition.

[Mut73] F.K. Mutombo. *Traitement des données manquantes et rationalisation d'un réseau de stations de mesures*. PhD thesis, Université Paul et Marie Curie, Paris VI, 1973.

[Nis80] S. Nishisato. *Analysis of Categorical Data: Dual Scaling and its Applications*. Univ. Toronto Press, 1980.

[Non92] R. Nonell. *El Condicionament en les Anàlisis Factorials Descriptives: l'Anàlisi Parcial Interna i Simultània*. PhD thesis, Univ. Politècnica de Catalunya, Barcelona, 1992.

[Nor74] C. Nora. *Une méthode de reconstitution et d'analyse de données incomplétes*. PhD thesis, Université Paul et Marie Curie, Paris VI, 1974.

[NTA00] R. Nonell, S. Thió, and T. Aluja. Some alternatives for conditional principal component analysis. *Applied Stochastic Models in Business and Industry*, 16(2):147–158, 2000.

[Peñ02] Daniel Peña. *Análisis de datos multivariantes*. Mc Graw-Hill, 2002.

[Rao64] C.R. Rao. The use and interpretation of principal components analysis in applied research. *Sankhya, A*, 26:329–357, 1964.

[Rao73] C.R. Rao. *Linear Statistical Inference and Its Applications. Secon Edition*. John Wiley & Sons, New York., 1973.

[RP91] M.L. Radelet and G.L. Pierce. Choosing those who will die: Race and the death penalty in florida. *Florida Law Review*, 43:1–34, 1991.

[Sap90] G. Saporta. *Probabilités, Analyse des Données et Statistique*. Éditions Technip, Paris., 1990.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461 –464, 1978.

[Tuk62] J.W. Tukey. The future of data analysis. *AMS*, 33:1–67, 1962.

[Vol93] M. Volle. *Analyse des Données*. Ed. Economica, Paris., 1993.

[WdF94] J. Whittaker and A. de Falguerolles. Graphical models and optimal scoring in multi-way contingency tables. Technical report, Publications de le Laboratoire de Statistique et Probabilités. Univ. Paul Sabatier, Toulose., 1994.

[Whi90a] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Ed. Wiley, 1990.

[Whi90b] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, 1990.

[WL83] N. Wermuth and S.L. Lauritzen. Graphical and recursive models for contingence tables. *Biometrika*, 70(3):537–552, 1983.

[Zar89] A. Zarraga. *Análisis de Correspondencias Múltiples por Bandas*. PhD thesis, Universidad del País Vasco. Fac. Ciencias Económicas y Empresariales, 1989.

Apèndix A

Macros de Minitab

En aquest capítol proporcionem algunes de les macros generades amb el Minitab, que ens han permès implementar informàticament les nostres propostes d'anàlisi. No les proporcionem totes, ja que algunes són variacions de les altres modificant els paràmetres d'entrada i les sortides volgudes. Les que trobareu a continuació cobreixen les parts corresponents a:

- la reconstitució de la matriu amb la metodologia k-EM: la macro **RECOBLOC** i les 4 macros que ella executa **GLOBAL, MODCORSI, RECCORS2, INCORSI**
- trobar la significació de la inèrcia dels condicionaments múltiples: la macro **TEST** que és la que ho realitza i les 4 macros a ella associades **TESTXI2, TESTINER, OPERPROJ, CORMUTRIS**
- realització i presentació dels resultats i eines d'ajuda d'anàlisis de correspondències **CORSI2**
- obtenció de la descomposició de les inèrcies **INCORSI2**

A.1 Macro RECOBLOC

```
macro recobloc colmod blocmodi col.1-col.ncol nfil ordrerec

#aquesta macro fa la reconstitucio k-em dels blocs que li diem a blocmodi
#a colmod hi ha el numero de modalitats de cada variable
#a blocmodi els blocs a modificar (13 es el bloc 1a fila 3a columna)
#a partir dels blocs construim dues columnes filmodi i colmodi on guardem les caselles

mconstant a b c d i j k l x y  numvar totmodi nfil ordrerec
mcolumn colmod blocmodi colfi colini filmodi colmodi
mcolumn col.1-col.ncol

let numvar=count(colmod)
let colfi(1)=colmod(1)
do i=2:numvar
  let colfi(i)=colmod(i)+colfi(i-1)
enddo

let colini=colfi-colmod+1
let totmodi=count(blocmodi)
let k=1
do l=1:totmodi
 let x=floor(blocmodi(l)/10)
 let y=blocmodi(l)-x*10
 let a=colini(x)
 let b=colfi(x)
 let c=colini(y)
 let d=colfi(y)

 do i=a:b
   do j=c:d
     let filmodi(k)=i
     let colmodi(k)=j
     let k=k+1
   enddo
 enddo
enddo
%A:GLOBAL Col.1-Col.NCOL NFIL COLmodi FILmodi ORDREREC

endmacro
```

A.2 Macro GLOBAL

```
MACRO GLOBAL C.1-C.NCOL NFIL COL FIL ORDREREC

#
MCONSTANT K1 K2 K3 k4 k5 TOTCANV NFIL ORDRE COMPT index compt2 ORDREREC MINCONT
MCOLUMN C.1-C.NCOL caux.1-caux.ncol Y.1-Y.NCOL YMIN
MCOLUMN COL FIL CORIG CRECO CDIF2 cordre cdist cmaxim cdifabs cdistabs cminim
MMATRIX mrecons miner

LET TOTCANV=COUNT(COL)
# anem a modificar totes les dades de les posicions indicades
# a les columna i filera per la substitucio tipus missing
DO K1=1:TOTCANV
   LET K2=COL(k1)
   let k3=FIL(k1)
   %a:modcorsi c.1-c.ncol k2 k3
ENDDO

copy c.1-c.ncol caux.1-caux.ncol
# anem a fer ara una reconstruccio d'ordre ORDRE a partir
# de les correspondencies simples.Creem tres columnes
# corig on hi ha les dades originals
# creco dades reconstruides(amb un valor iniciador)
# cdif2 on hi ha les diferencies entre corig i creco al quadrat
# la suma dels quadrats d'aquestes diferencies ens servira per aturar el proces

let index=0
DO ORDRE=1:ORDREREC
   DO K1=1:TOTCANV
    LET K2=COL(k1)
    let k3=FIL(k1)
    let CORIG(K1)=c.k2(k3)
    LET CRECO(K1)=C.K2(K3)-1
    LET CDIF2(K1)=(CORIG(K1)-CRECO(K1))**2
    let cdifabs(k1)=abs(corig(k1)-creco(k1))
    ENDDO
    LET COMPT=SUM(CDIF2)
    let compt2=sqrt(sum(cdifabs))
# anem a fer un bucle perque mentre la suma dels quadrat de les
# diferencies no sigui mes petit que el nombre total de canvis
# es a dir a cada canvi un error 1 de mitjana maxim
# vagi reconstruint i canviant les dades velles per les noves
    WHILE COMPT ge 0.5
    copy c.1-c.ncol y.1-y.ncol
    %a:reccors2 c.1-c.ncol nfil ordre
    Rmin c.1-c.ncol ymin
    let mincont=min(YMIN)
    DO K1=1:TOTCANV
       LET K2=COL(k1)
       let k3=FIL(k1)
       let corig(k1)=creco(k1)
       LET CRECO(K1)=C.K2(K3)
```

```
       LET CDIF2(K1)=(CORIG(K1)-CRECO(K1))**2
    ENDDO
# hem guardat a creco les noves dades i ara tornarem a les dades
# originals guardades a CAUX canviant les reconstituides
    COPY CAUX.1-CAUX.NCOL C.1-C.NCOL
    DO K1=1:TOTCANV
       LET K2=COL(k1)
       let k3=FIL(k1)
       LET C.K2(K3)=CRECO(K1)
    ENDDO
    LET CDIF2(K1)=(CORIG(K1)-CRECO(K1))**2
    let cdifabs(k1)=abs(corig(k1)-creco(k1))
    LET COMPT=sqrt(SUM(CDIF2))
    let compt2=sum(CDIFABS)
    let index=index+1
    let cdist(index)=compt
    let cordre(index)=ordre
    let cdistabs(index)=compt2
    let cmaxim(index)=max(cdif2)
    LET cminim(INDEX)=MIN(CDIF2)
    %a:incorsi c.1-c.ncol nfil miner
    ENDWHILE
ENDDO
print cordre cdist cdistabs cmaxim cminim
tsplot cdist;
Symb cordre.
tsplot cdistabs;
symb cordre.

ENDMACRO
```

A.3 Macro RECCORS2

```
MACRO RECCORS2 C.1-C.NCOL NFIL ORDRE
#
# RECONS REALITZA PER A LES COLUMNES DADES UNA RECONSTITUCIO
# DE LES COLUMNES DONADES VIA CORRESPONDENCIES SIMPLES
# DE L'ORDRE DEMANAT

MCONSTANT NFIL K1 K2 K3 NUMVP NCOL INSUMTOT ORDRE control i j t
MCOLUMN C.1-C.NCOL COMPRIF.1-COMPRIF.NCOL CORFIL.1-CORFIL.NCOL  x.1-x.ncol xmin
MCOLUMN COMPRIG.1-COMPRIG.NCOL CORCOL.1-CORCOL.NFIL
MCOLUMN SUMFIL SUMCOL VALPRO SQVALPRO INARPECO INARPEFI
MMATRIX MDADES MPESFIL MPESCOL MDADESTR MINPESFI MINPESCO MRECONS
MMATRIX MADIAG MVEPROF MVEPROG MVEPROGT MSQVAPRO MINARPEC MINARPEF
MMATRIX MCOORFIL MCOORCOL MAUX1 MAUX2 MAUX3 MDADESIN MSQVAPR

# VERIFICO QUE NO SIGUI UNA RECONSTITUCIO IMPOSSIBLE

IF ORDRE GT NCOL
    NOTE NO ES POSSIBLE UNA RECONSTITUCIO D'AQUEST ORDRE
    EXIT
ENDIF

# VERIFICO QUE TOTES LES COLUMNES TINGUIN LA MATEIXA LONGITUD

DO K1=1:NCOL
   IF COUNT(C.K1) NE NFIL
      NOTE
      NOTE LES COLUMNES NO TENEN EL MATEIX NOMBRE D'INDIVIDUS
      NOTE AIXO ES NECESSARI, PER TANT S'INTERROMP L'EXECUCIO
      EXIT
   ENDIF
ENDDO

# ************************* COMENCEM *************************
COPY C.1-C.NCOL MDADESIN
# MDADESIN ES LA MATRIU DE DADES INICIALS
RSUM C.1-C.NCOL SUMFIL
LET INSUMTOT=1/SUM(SUMFIL)
MULT INSUMTOT MDADESIN MDADES
#CALCULEM LA MATRIU DE DADES DE FREQUENCIES RELATIVES
LET SUMFIL=INSUMTOT*SUMFIL
DO K1=1:NCOL
    LET SUMCOL(K1)=SUM(C.K1)*INSUMTOT
ENDDO
#PRINT SUMFIL SUMCOL INSUMTOT
DIAG SUMFIL MPESFIL
DIAG SUMCOL MPESCOL
#TENIM A MPESFIL LA MATRIU DIAGONAL DE MARGINALS PER FILERES
#TENIM A MPESCOL LA MATRIU DIAGONAL DE MARGINALS PER COLUMNES

TRANS MDADES MDADESTR
INVE MPESFIL MINPESFI
```

```
INVE MPESCOL MINPESCO
LET INARPECO=1/(SQRT(SUMCOL))
DIAG INARPECO MINARPEC
#INARPECO ES L'INVERS DE L'ARREL DEL PES DE LES COLUMNES
#ANEM A CALCULAR LA MATRIU A DIAGONALITZAR A R-NCOL

MULT MINARPEC MDADESTR MADIAG
MULT MADIAG MINPESFI MADIAG
MULT MADIAG MDADES MADIAG
MULT MADIAG MINARPEC MADIAG
EIGEN MADIAG VALPRO MVEPROF
#TENIM A VALPRO TOTS ELS VALORS PROPIS INCLOS EL TRIVIAL
#TENIM A MVEPROF LES COMPONENTS PRINCIPALS F

MULT MINPESFI MDADES MCOORFIL
MULT MCOORFIL MINARPEC MCOORFIL
MULT MCOORFIL MVEPROF MCOORFIL
COPY MCOORFIL CORFIL.1-CORFIL.NCOL

#ANEM A CALCULAR LES COORDENADES DE LES COLUMNES
# posem valor absolut als valors propis per evitar els -0.000000
let valpro=abs(valpro)
LET SQVALPRO=SQRT(VALPRO)
DIAG SQVALPRO MSQVAPR
MULT MINARPEC MVEPROF MCOORCOL
MULT MCOORCOL MSQVAPR MCOORCOL

COPY MCOORCOL CORCOL.1-CORCOL.NCOL
#PRINT CORCOL.1-CORCOL.NCOL # #

LET SQVALPRO=1/SQRT(VALPRO)
DIAG SQVALPRO MSQVAPRO
DEFINE 0 NFIL NCOL MRECONS

DO K1=1:ORDRE
    COPY CORCOL.K1 MAUX1
    COPY CORFIL.K1 MAUX2
    LET K2=SQVALPRO(K1)
    MULT K2 MAUX1 MAUX1
    TRANS MAUX1 MAUX1
    MULT MAUX2 MAUX1 MAUX3
    MULT MAUX3 MPESCOL MAUX3
    MULT MPESFIL MAUX3 MAUX3
    LET K3=1/INSUMTOT
    MULT K3 MAUX3 MAUX3
    #PRINT k1 k2 k3 MAUX3
    ADD MAUX3 MRECONS MRECONS
# tall canviat per solucionars negatius
    copy mrecons x.1-x.ncol
    rmin x.1-x.ncol xmin
    let control=min(xmin)
    if control<0
    #note   TENIM RECONSTITUCIO NEGATIVA
```

```
    #note quan l'ordre es k1 l'element canviat es l'(i,j) que valia t
    do i=1:ncol
        do j=1:nfil
        if x.i(j)<0
            let t=x.i(j)
            #prin k1, i,,j, t
            let x.i(j)=0
        endif
        enddo
    enddo
    copy  x.1-x.ncol mrecons
    endif
# fi tall
 ENDDO
#PRINT MRECONS
#if control=1
#   copy mrecons3 mrecons
#endif copy mrecons c.1-c.ncol

ENDMACRO
```

A.4 Macro INCORSI

```
MACRO INCORSI C.1-C.NCOL NFIL MINERCIA
# INCORSI CALCULA PER A LES COLUMNES QUE LI DONEM LES INERCIES
# PER SEPARAT DE LA DIAGONAL I DE FORA DE LA DIAGONAL D'UNA
# MATRIU DE CORRESPONDENCIES SIMPLES

MCONSTANT NFIL K1 K2 NIND SUMTOT SUMDIAG SUMNDIAG SUMSUPDI SUMINFDI
MCOLUMN C.1-C.NCOL CMOD.1-CMOD.NCOL MITJAFIL CIND.1-CIND.NCOL
MCOLUMN MITJACOL
MMATRIX MDADES MMITJFIL MMITJCOL MINDEP MFINAL MINERCIA


# VERIFIQUEM QUE TOTES LES COLUMNES TINGUIN LA MATEIXA LONGITUD
DO K1=1:NCOL
    IF COUNT(C.K1) NE NFIL
        NOTE
        NOTE NO TOTES LES COLUMNES TENEN EL MATEIX NOMBRE D'INDIVIDUS
        NOTE ES NECESSARI I PER TANT S'INTERROMP L'EXECUCIO
        EXIT
    ENDIF
ENDDO

COPY C.1-C.NCOL MDADES
# MDADES ES LA MATRIU DE DADES INICIALS
# ANEM A CALCULAR EL NOMBRE D'INDIVIDUS TOTALS
LET NIND=0
DO K1=1:NCOL
    LET NIND=NIND+SUM(C.K1)
ENDDO
#COPIEM LES COLUMNES C A CMOD PER A TREBALLAR AMB AQUESTES I NO AMB LES ORIGINALS
COPY C.1-C.NCOL CMOD.1-CMOD.NCOL
#ANEM A DIVIDIR TOTES LES COLUMNES PEL NOMBRE D'INDIVIDUS TOTAL
DO K1=1:NCOL
    LET CMOD.K1=(CMOD.K1)/NIND
ENDDO

RSUM CMOD.1-CMOD.NCOL MITJAFIL
DO K1=1:NCOL
    LET MITJACOL(K1)=SUM(CMOD.K1)
ENDDO

#ANEM A POSAR A CIND.1-CIND.NCOL LES DADES AMB EL SUPOSIT D'INDEPENDENCIA
COPY MITJAFIL MMITJFIL
COPY MITJACOL MMITJCOL

TRANS MMITJCOL MMITJCOL
MULT MMITJFIL MMITJCOL MINDEP
COPY MINDEP CIND.1-CIND.NCOL
DO K1=1:NCOL
    LET CIND.K1=(CMOD.K1-CIND.K1)*(CMOD.K1-CIND.K1)/CIND.K1
ENDDO
```

```
#NOTE MATRIU D'INERCIES A CIND.1-CIND.NCOL

COPY CIND.1-CIND.NCOL MINERCIA

# INICIALITZEM ELS COMPTADORS PER A LA INERCIA TOTAL I DE LA DIAGONAL
#I CALCULEM AQUESTES

LET SUMTOT=0
LET SUMDIAG=0
LET SUMSUPDI=0
DO K1=1:NCOL
    LET SUMTOT=SUMTOT+SUM(CIND.K1)
    LET SUMDIAG=SUMDIAG+(CIND.K1(K1))
    DO K2=1:K1
        LET SUMSUPDI=SUMSUPDI+(CIND.K1(K2))
    ENDDO
ENDDO

LET SUMSUPDI=SUMSUPDI-SUMDIAG
LET SUMINFDI=SUMTOT-SUMDIAG-SUMSUPDI
LET SUMNDIAG=SUMTOT-SUMDIAG

#canvio aquesta sortida sense res
#
# PRINT SUMTOT SUMDIAG SUMNDIAG SUMSUPDI SUMINFDI

ENDMACRO
```

A.5 Macro TEST

```
macro
test colmodal dades

#
# a partir de les dades en forma disjuntiva completa les fraccionem en altres matrius segons
# la columna de mides i els les passem la macro testiner
#
#esta limitat a un maxim de 64 modalitats i 10 variables



mconstant n Q J T totmodal totalvar i ii cons1 cons2 cons3 cons4 cons5 cons6
MCONSTANT u v w graus pvalue compta inercia XI2 INER chi2 inertia extrsup
mcolumn colmodal colinici colfi colum.1-colum.64 colinfo.1-colinfo.8 variab1 variab2 variab3 variabls
mcolumn colinf.1-colinf.6 varia1 varia2 varia3 varias COLXI2
mmatrix dades mat1 mat2 mat3 mat4

let Totmodal=sum(colmodal)
let totalvar=count(colmodal)

if totmodal>64
   NOTE Error massa modalitats
   EXIT
endif
if totalvar>10
   NOTE Error massa variables
   EXIT
endif


copy dades colum.1-colum.totmodal

let colinici(1)=1
do i=2:totalvar
   let colinici(i)=colinici(i-1)+colmodal(i-1)
enddo
let colfi(1)=colmodal(1)
do i=2:totalvar
   let colfi(i)=colfi(i-1)+colmodal(i)
enddo


#note
#note ------------------------------------------
#note les variables u i v estan condicionades per w
#note ------------------------------------------
#note
let n=count(colum.1)
let q=2
```

```
let compta=1
do u=1:totalvar
   let cons1=colinici(u)
   let cons2=colfi(u)
   do v=u:totalvar
     if u<>v
    let cons3=colinici(v)
    let cons4=colfi(v)
    let i=colfi(u)-colinici(u)+1
    let j=colfi(v)-colinici(v)+1

%A:TESTXI2 colum.cons1-colum.cons2 colum.cons3-colum.cons4 n i j CHI2 inertia GRAUS PVALUE
        let colinf.1(compta)=u
        let colinf.2(compta)=v
        let colinf.3(compta)=CHI2
        let colinf.4(compta)=graus
        let colinf.5(compta)=pvalue
        let colinf.6(compta)=inertia
        let compta=compta+1

       endif
   enddo
enddo


Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinf.1 varia1
Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinf.2 varia2
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinf.1 varia1
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinf.2 varia2


Concatenate varia1 varia2 varias
name varias 'varis'
name colinf.3 'CHI2'
name colinf.4 'df-s'
name colinf.5 'p-valors'
name colinf.6 'inertia'
print varias colinf.4 colinf.3 colinf.6 colinf.5



let compta=1
do u=1:totalvar
   let cons1=colinici(u)
   let cons2=colfi(u)
   do v=u:totalvar
     if u<>v
    let cons3=colinici(v)
    let cons4=colfi(v)
    copy (colum.cons1-colum.cons2 colum.cons3-colum.cons4) mat1
    let j=cons4-cons3+1+cons2-cons1+1
    do w=1:totalvar
        if (w<>v) and (w<>u)
```

```
        let cons5=colinici(w)
        let cons6=colfi(w)
        let t=cons6-cons5+1
        copy colum.cons5-colum.cons6 mat2

        %A:testiner n j q t mat1 mat2 XI2 INER graus pvalue extrsup
#print u v w
        let colinfo.1(compta)=u
        let colinfo.2(compta)=v
        let colinfo.3(compta)=w
        let colinfo.4(compta)=graus
        let colinfo.5(compta)=pvalue
        let colinfo.6(compta)=xi2
        LET COLINFO.7(COMPTA)=INER
        let colinfo.8(compta)=extrsup
        let compta=compta+1
            endif
        enddo
      endif
   enddo
enddo

Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.1 variab1
Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.2 variab2
Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.3 variab3
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.1 variab1
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.2 variab2
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.3 variab3

Concatenate variab1 variab2 variabls

name variabls 'vars'
name variab3 'condi'
name colinfo.4 'df2'
name colinfo.5 'p-valor'
name colinfo.6 'val-xi2'
NAME COLINFO.7 'INER-B'
name colinfo.8 'extr-sup'

print variabls variab3 colinfo.4 colinfo.6 COLINFO.7 colinfo.5 colinfo.8


endmacro

macro
test colmodal dades

#
# a partir de les dades en forma disjuntiva completa les fraccionem en altres matrius segons
# la columna de mides i els les passem la macro testiner
#
#esta limitat a un maxim de 64 modalitats i 10 variables
```

```
mconstant n Q J T totmodal totalvar i ii cons1 cons2 cons3 cons4 cons5 cons6
MCONSTANT u v w graus pvalue compta inercia XI2 INER chi2 inertia extrsup
mcolumn colmodal colinici colfi colum.1-colum.64 colinfo.1-colinfo.8 variab1 variab2 variab3
mcolumn variabls colinf.1-colinf.6 varia1 varia2 varia3 varias COLXI2
mmatrix dades mat1 mat2 mat3 mat4

let Totmodal=sum(colmodal)
let totalvar=count(colmodal)

if totmodal>64
   NOTE Error massa modalitats
   EXIT
endif
if totalvar>10
   NOTE Error massa variables
   EXIT
endif


copy dades colum.1-colum.totmodal

let colinici(1)=1
do i=2:totalvar
   let colinici(i)=colinici(i-1)+colmodal(i-1)
enddo
let colfi(1)=colmodal(1)
do i=2:totalvar
   let colfi(i)=colfi(i-1)+colmodal(i)
enddo


#note
#note _____________________________________________
#note les variables u i v estan condicionades per w
#note ---------------------------------------------
#note
let n=count(colum.1)
let q=2

let compta=1
do u=1:totalvar
   let cons1=colinici(u)
   let cons2=colfi(u)
   do v=u:totalvar
     if u<>v
    let cons3=colinici(v)
    let cons4=colfi(v)
    let i=colfi(u)-colinici(u)+1
    let j=colfi(v)-colinici(v)+1

%A:TESTXI2 colum.cons1-colum.cons2 colum.cons3-colum.cons4 n i j CHI2 inertia GRAUS PVALUE
```

```
        let colinf.1(compta)=u
        let colinf.2(compta)=v
        let colinf.3(compta)=CHI2
        let colinf.4(compta)=graus
        let colinf.5(compta)=pvalue
        let colinf.6(compta)=inertia
        let compta=compta+1

       endif
   enddo
enddo


Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinf.1 varia1
Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinf.2 varia2
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinf.1 varia1
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinf.2 varia2


Concatenate varia1 varia2 varias
name varias 'varis'
name colinf.3 'CHI2'
name colinf.4 'df-s'
name colinf.5 'p-valors'
name colinf.6 'inertia'
print varias colinf.4 colinf.3 colinf.6 colinf.5



let compta=1
do u=1:totalvar
   let cons1=colinici(u)
   let cons2=colfi(u)
   do v=u:totalvar
     if u<>v
    let cons3=colinici(v)
    let cons4=colfi(v)
    copy (colum.cons1-colum.cons2 colum.cons3-colum.cons4) mat1
    let j=cons4-cons3+1+cons2-cons1+1
    do w=1:totalvar
        if (w<>v) and (w<>u)
        let cons5=colinici(w)
        let cons6=colfi(w)
        let t=cons6-cons5+1
        copy colum.cons5-colum.cons6 mat2

        %A:testiner n j q t mat1 mat2 XI2 INER graus pvalue extrsup
#print u v w
        let colinfo.1(compta)=u
        let colinfo.2(compta)=v
        let colinfo.3(compta)=w
        let colinfo.4(compta)=graus
```

```
        let colinfo.5(compta)=pvalue
        let colinfo.6(compta)=xi2
        LET COLINFO.7(COMPTA)=INER
        let colinfo.8(compta)=extrsup
        let compta=compta+1
            endif
        enddo
      endif
   enddo
enddo

Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.1 variab1
Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.2 variab2
Code (1) "A" (2) "B" (3) "C" (4) "D" (5) "E" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.3 variab3
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.1 variab1
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.2 variab2
#Code (1) "B" (2) "C" (3) "E" (4) "G" (5) "D" (6) "F" (7) "G" (8) "H" (9) "I" colinfo.3 variab3

Concatenate variab1 variab2 variabls

name variabls 'vars'
name variab3 'condi'
name colinfo.4 'df2'
name colinfo.5 'p-valor'
name colinfo.6 'val-xi2'
NAME COLINFO.7 'INER-B'
name colinfo.8 'extr-sup'

print variabls variab3 colinfo.4 colinfo.6 COLINFO.7 colinfo.5 colinfo.8


endmacro
```

A.6 Macro TESTXI2

```
MACRO
TESTXI2 COL.1-COL.NCOL NFIL MOD1 MOD2 CHIQUAD inercia grauslli PVAL
#
# INCORSI CALCULA PER A LES COLUMNES QUE LI DONEM LES INERCIES
# PER SEPARAT DE LA DIAGONAL I DE FORA DE LA DIAGONAL D'UNA
# MATRIU DE CORRESPONDENCIES SIMPLES
#

MCONSTANT NFIL K1 K2 NIND CHIQUAD SUMDIAG SUMNDIAG SUMSUPDI SUMINFDI CONS1 CONS2
MCONSTANT CONS3 CONS4 MOD1 MOD2 INERCIA grauslli PVAL margtot
MCOLUMN COL.1-COL.NCOL CMOD.1-CMOD.NCOL MITJAFIL CIND.1-CIND.NCOL
MCOLUMN MITJACOL
MMATRIX MDADES MMITJFIL MMITJCOL MINDEP MFINAL MATR1 MATR2 MATR3


COPY COL.1-COL.MOD1 MATR1
LET CONS3=MOD1+1
LET CONS4=MOD1+MOD2
COPY COL.CONS3-COL.CONS4 MATR2
TRANS MATR2 MATR2
MULT MATR2 MATR1 MATR3
#PRINT MATR3

COPY MATR3 MDADES
#COPY COL.1-COL.NCOL MDADES
# MDADES ES LA MATRIU DE DADES INICIALS

#
# NOMBRE D'INDIVIDUS TOTALS
#
LET NIND=nfil

#COPIEM LES COLUMNES C A CMOD PER A TREBALLAR AMB AQUESTES I NO AMB LES ORIGINALS
COPY MATR3 CMOD.1-CMOD.MOD1


RSUM CMOD.1-CMOD.MOD1 MITJAFIL
DO K1=1:MOD1
    LET MITJACOL(K1)=SUM(CMOD.K1)
ENDDO

let margtot=sum(mitjacol)
#prin margtot

#ANEM A POSAR A CIND.1-CIND.NCOL LES DADES AMB EL SUPOSIT D'INDEPENDENCIA
#PRINT MITJAFIL MITJACOL
COPY MITJAFIL MMITJFIL
COPY MITJACOL MMITJCOL

TRANS MMITJCOL MMITJCOL
```

```
MULT MMITJFIL MMITJCOL MINDEP
LET CONS1=1/NIND
MULT CONS1 MINDEP MINDEP
COPY MINDEP CIND.1-CIND.MOD1

#PRIN MINDEP

DO K1=1:MOD1
    LET CIND.K1=(CMOD.K1-CIND.K1)*(CMOD.K1-CIND.K1)/CIND.K1
ENDDO

#NOTE MATRIU D'INERCIES
#PRINT CIND.1-CIND.MOD1

LET CHIQUAD=0
DO K1=1:MOD1
    LET CHIQUAD=CHIQUAD+SUM(CIND.K1)
ENDDO


############################
#prin matr3, cmod.1-cmod.mod1
#chiSquare cmod.1-cmod.mod1
#print chiquad


let inercia=chiquad/margtot

LET grauslli=(MOD1-1)*(MOD2-1)

CDF CHIQUAD pval;
ChiSquare grauslli.
let pval=1-pval

#print CHIQUAD grauslli pval
ENDMACRO
```

A.7 Macro TESTINER

```
macro
testiner n J Q T mcondi mcondi2 DIFERCOR DIFERVAL graus pvalor extrsup

#aquesta macro ens calcula per al nombre d'individus n, J modalitats, Q questions
#T modalitats condicionadores dels valors disjuntius complerts que es toben a
#mcondi els que han de ser condicionats i mcondi 2 les modalitats condicionadores
#si el condicionament es o no significatiu
#per a fer-ho projecta sobre les modalitats condicionadores
#i fa servir la macro cormutris per calcular el totla de la inercia no projectada
#depres calcula el limit de l'interval  i fa la diferencia per veure si es o no
#significatiu al 95%. Posteriorment dona el p-valor i els graus de llibertat
#
mconstant n Q J T extrsup graus confi xialfa
mconstant cons1 cons2 cons3 sumaval diferval pvalor difercor
mcolumn col.1-col.J colaux1
mmatrix mcondi mcondi2 mprojec maux1 maux2 maux3 maux4 maux5 maux6 maux7 maux8 maux9

%A:operproj mcondi mcondi2 mprojec
define 1 n 1 maux1
define 1 J 1 maux2
mult mcondi maux2 maux3
let cons1=1/n
mult cons1 maux3 maux3
trans mcondi maux7
mult maux7 maux1 maux4
let cons2=1/2
mult cons2 maux4 maux4
trans maux4 maux5
mult maux3 maux5 maux6
add maux6 mprojec maux8
copy maux8 col.1-col.J

%a:cormutris col.1-col.J Q n sumaval
let diferval=J/Q-1-sumaval
let graus=(T-1)*(J-1)
let confi=0.95
InvCDF confi xialfa;
ChiSquare graus.
let difercor=diferval*(n*Q)
CDF difercor pvalor;
ChiSquare graus.
let extrsup=xialfa/(n*Q)
let pvalor=1-pvalor
brief 2

endmacro
```

A.8 Macro OPERPROJ

```
macro
projeccio m_acondi m_condi mfi

mmatrix k t ttr ktr m_condi m_acondi proj  maux1 mfi

#
#m_acondi es la matriu disjunctiva a condicionar
#m_condi es la matriu condicionadora
# mfi es la matriu projeccio sobre l'ortogonal
#entrem m_acondi a la matriu k
#entrem m_condi a la matriu t
#

copy m_acondi k
copy m_condi t

trans k ktr
trans t ttr
mult  ttr t maux1
inve maux1 maux1
mult t maux1 maux1
mult maux1 ttr maux1
# a maux1 tenim l'operador projeccio sobre T
mult maux1 k proj
# a proj tenim l'operador projeccio sobre T operat a K
#note projeccio
subt proj k mfi
#note projeccio sobre ortogonal
#print  mfi

endmacro
```

A.9 Macro CORMUTRIS

```
MACRO
CORMUtris C.1-C.N NQUES NIND sumaval

#
# Aquesta macro {\'e}s un fragment de cormubis on nom{\'e}s m'interessa el valor de la
# suma de valors propis
#
# CORMU REALITZA PER A LES COLUMNES QUE LI DONEM UNA ANALISI
# DE CORRESPONDENCIES MULTIPLES
# LES DADES PODEN ESTAN EN FORMA DISJUNTIVA COMPLETA O CONDENSADA
# I LI INDIQUEM LES COLUMNES I EL NOMBRE TOTAL DE QUESTIONS
#

MCONSTANT NQUES NIND K1 K2 K3 K4 K5 NTOTCAT INNQUES SUMVALPR NUMVAPRO NCOL NCOL2
mconstant grafics sumaval
MCOLUMN C.1-C.N D.1-D.60 SQVALPRO VEPR.1-VEPR.60
MCOLUMN CATEGOR categor2 DIAGBUR SQDIABU VALPRO CVALPRO CPESFIL CINPESFI NOM
MCOLUMN ACUMULAT INERCIA CPESCOL CINPESCO CINARPEC SCOSQUA DIS
MCOLUMN CO2.1-CO2.3 COR.1-COR.3 CT.1-CT.3 CORFIL.1-CORFIL.60 COO.1-COO.3
MCOLUMN CORCOL.1-CORCOL.60 COO2.1-COO2.3 CAUX ETIQUET etiquet2
mcolumn x.1-x.n xsum
MMATRIX MDISJ MDISJTR MBURT MANALISI MVEPRO MINPESFI MINPESCO MINARPEC MVEPRO2
MMATRIX MDIAGBU MDIAGBUI MSQDIABU MSQDIBUI MDADES MCOORFIL MCOORCOL MSQVAPR
MMATRIX MVEPRO3 MAUX mburtfi

#
# VERIFIQUEM QUE TOTES LES COLUMNES TINGUIN LA MATEIXA LONGITUD
#
DO K1=1:N
        IF COUNT(C.K1) NE NIND
                NOTE
                NOTE NO TOTES LES COLUMNES TENEN EL MATEIX NOMBRE D'INDIVIDUS
                NOTE ES NECESSARI I PER TANT S'INTERROMP L'EXECUCIO
                EXIT
        ENDIF
ENDDO

iF NQUES EQ N
    # SI SON IGUALS VOL DIR QUE LES DADES NO ESTAN EN FORMA DISJUNTIVA COMPLETA I
    # ENS INTERESSA QUE AIXI SIGUI PER FACILITAR EL CALCUL DE LA MATRIU DE BURT
    DO K2=1:N
        LET CATEGOR(K2)=MAX(C.K2)
        let categor2(k2)=Min(C.k2)
    ENDDO
    do k2=1:N
    if categor2(k2)>1
           Note hi ha alguna categoria sense efectiu, arregla-ho i torna-ho a executar.
       Note Gracies
       Exit
    endif
        enddo
```

```
    LET NTOTCAT=SUM(CATEGOR)
    LET K1=1
    LET K2=0
    DO K3=1:NQUES
        LET K2=MAX(C.K3)+K2
        INDIC C.K3 D.K1-D.K2
        LET K1=K2+1
    ENDDO

ELSEIF NQUES NE N
    LET K1=1
    LET NTOTCAT=N
    COPY C.1-C.N D.K1-D.NTOTCAT

ENDIF

RSUM D.1-D.NTOTCAT CPESFIL
LET K1=SUM(CPESFIL)
#PRINT CPESFIL k1
LET CPESFIL=CPESFIL/K1
LET CINPESFI=1/CPESFIL
DIAG CINPESFI MINPESFI

DO K1=1:NTOTCAT
    LET CPESCOL(K1)=SUM(D.K1)
ENDDO
LET K1=SUM(CPESCOL)
LET CPESCOL=CPESCOL/K1
#PRINT CPESCOL k1
LET CINPESCO=1/CPESCOL
DIAG CINPESCO MINPESCO
LET CINARPEC=SQRT(CINPESCO)
DIAG CINARPEC MINARPEC

COPY C.1-C.N MDADES
COPY D.1-D.NTOTCAT MDISJ

#
#TENIM A MDISJ LA MATRIU EN FORMA DISJUNTIVA COMPLETA
#
TRANS MDISJ MDISJTR
MULT MDISJTR MDISJ MBURT

#
#TENIM A MBURT LA MATRIU DE BURT
#
#PRINT MBURT
copy mburt mburtfi
# ANEM A FER ELS CALCULS PER TROBAR ELS VALORS
# I VECTORS PROPIS DE LA MATIRU D'ANALISI
#

DIAG MBURT DIAGBUR
```

```
copy mburt x.1-x.n
rsum x.1-x.n xsum
let xsum=xsum/nques
#print diagbur xsum
#pause

#la diferencia amb cormu {\'e}s que aquest no conte la diagonal de Burt sino
#la diagonal donada per les marginals
# ho fem copiant-ho sobre la diagonal de burt
copy xsum diagbur

DIAG DIAGBUR MDIAGBU
INVE MDIAGBU MDIAGBUI

LET INNQUES=1/NQUES
LET SQDIABU=SQRT(DIAGBUR)
DIAG SQDIABU MSQDIABU
INVE MSQDIABU MSQDIBUI
# msqdibui es la matriu inversa de l'arrel de la de diagonal de burt
MULT MBURT MSQDIBUI MANALISI
MULT MSQDIBUI MANALISI MANALISI
MULT INNQUES MANALISI MANALISI
EIGEN MANALISI CVALPRO MVEPRO
#print cvalpro
#print mvepro
#
# ANEM A ELIMINAR LES VALORS PROPIS ZERO ESTRUCTURALS I
# EL VALOR PROPI 1 GUARDANT-HO A VALPRO
#

LET VALPRO=CVALPRO
LET K3=NTOTCAT-NQUES+2
LET K4=K3-1
#print k3 ntotcat k4 valpro
if k3<>ntotcat
    DELE K3:NTOTCAT VALPRO
else
    let k4=k3
endif
DELE 1 VALPRO
COPY MVEPRO VEPR.1-VEPR.NTOTCAT
COPY VEPR.2-VEPR.K4 MVEPRO2



LET SUMVALPR=SUM(VALPRO)
let sumaval=sumvalpr

ENDMACRO
```

A.10 Macro CORSI2

```
MACRO
CORSI2 X.1-X.NCOL NFIL nomc nomf eixos
#
# CORSI2 REALITZA PER A LES COLUMNES DONADES UNA AN{\cdot}LISI
# DE CORRESPONDENCIES SIMPLES i dona els grafics per 1 i 2 eixos
#
#
# x.1-x.ncol    columnes
# nfil      nombre de fileres
# nomc nomf nom de les columnes i nom de les fileres
# eixos     nombre d'eixos pels quals volem informacio
#
MCONSTANT i j NFIL CONS1 CONS2 CONS3 NUMVAPRO NCOL INSUMTOT NCOL2 sumvapro eixos graf1
MCOLUMN X.1-X.NCOL COMPRIF.1-COMPRIF.NCOL CORFIL.1-CORFIL.NCOL CORCOL.1-CORCOL.nfil
MCOLUMN CNT.1-CNT.5 COS.1-COS.5 COR.1-COR.5  ETIQUET nomc nomf NOMAUX
MCOLUMN caux.1-caux.5 columaux columau2 colaux.1-colaux.5
MCOLUMN SUMFIL SUMCOL INARPECO VALPRO INERCIA ACUMULAT DIST SCOSQUA SQVALPRO nomaux2
MMATRIX MDADESIN MDADES MPESFIL MPESCOL MDADESTR MINPESFI MINPESCO
MMATRIX MINARPEC MADIAG MVEPRO MVEPRO2 MCOORFIL MSQVAPR MCOORCOL


#
# VERIFIQUEM QUE TOTES LES COLUMNES TINGUIN LA MATEIXA LONGITUD
# i que no volguem informacio per m{\'e}s de 5 eixos
if eixos gt 5
    NOTE
    NOTE Nom{\'e}s esta previst per fins a 5 eixos
    NOTE ES NECESSARI I PER TANT S'INTERROMP L'EXECUCIO
    EXIT
ENDIF
DO CONS1=1:NCOL
    IF COUNT(X.CONS1) NE NFIL
        NOTE
        NOTE NO TOTES LES COLUMNES TENEN EL MATEIX NOMBRE D'INDIVIDUS
        NOTE ES NECESSARI I PER TANT S'INTERROMP L'EXECUCIO
        EXIT
    ENDIF
ENDDO

COPY X.1-X.NCOL MDADESIN
# MDADESIN ES LA MATRIU DE DADES INICIALS

RSUM X.1-X.NCOL SUMFIL
LET INSUMTOT=1/SUM(SUMFIL)
MULT INSUMTOT MDADESIN MDADES
LET SUMFIL=INSUMTOT*SUMFIL
DO CONS1=1:NCOL
    LET SUMCOL(CONS1)=SUM(X.CONS1)*INSUMTOT
ENDDO

DIAG SUMFIL MPESFIL
```

```
DIAG SUMCOL MPESCOL

TRANS MDADES MDADESTR

INVE MPESFIL MINPESFI
INVE MPESCOL MINPESCO
# INARPECO ES L'INVERS DE L'ARREL DEL PES DE LES COLUMNES
LET INARPECO=1/(SQRT(SUMCOL))
DIAG INARPECO MINARPEC

#ANEM A CALCULARLA MATRIU A DIAGONALITZAR:
MULT MINARPEC MDADESTR MADIAG
MULT MADIAG MINPESFI MADIAG
MULT MADIAG MDADES MADIAG
MULT MADIAG MINARPEC MADIAG
EIGEN MADIAG VALPRO MVEPRO
DELE 1 VALPRO

NOTE VOLS Vectors propis?(y/n)
YESNO GRAF1

If graf1=1
   print mvepro
endif

# ANEM A CALCULAR ELS PERCENTATGES D'INERCIA I ELS PERCENTATGES ACUMULATS
LET INERCIA=100*VALPRO/(SUM(VALPRO))
LET  NUMVAPRO=COUNT(VALPRO)

DO  CONS1=1:NUMVAPRO
    LET ACUMULAT(CONS1)=0
    DO CONS3=1:CONS1
        LET ACUMULAT(CONS1)=ACUMULAT(CONS1)+INERCIA(CONS3)
    ENDDO
ENDDO

COPY MVEPRO COMPRIF.1-COMPRIF.NCOL

COPY COMPRIF.2-COMPRIF.NCOL MVEPRO2

NOTE
NOTE       ************************************************************
NOTE       VALORS PROPIS, PERCENTATGES D'INERCIA I PERCENTATGES ACUMULATS
NOTE       ************************************************************
let sumvapro=sum(valpro)

PRINT SUMVAPRO,VALPRO,INERCIA,ACUMULAT
#PRINT COMPRIF.1-COMPRIF.NCOL

# ANEM A CALCULAR LES COORDENADES DE LES FILERES

MULT MINPESFI MDADES MCOORFIL
MULT MCOORFIL MINARPEC MCOORFIL
```

```
MULT MCOORFIL MVEPRO MCOORFIL

COPY MCOORFIL CORFIL.1-CORFIL.NCOL

#PRINT CORFIL.2-CORFIL.NCOL

#ANEM A CALCULAR LES COORDENADES DE LES COLUMNES
#
#                 NOTA !!!!!!!
#
# posem valor absolut als valors propis per evitar els -0.000000
#
#
let valpro=abs(valpro)
LET SQVALPRO=SQRT(VALPRO)
DIAG SQVALPRO MSQVAPR
MULT MINARPEC MVEPRO2 MCOORCOL
MULT MCOORCOL MSQVAPR MCOORCOL

LET NCOL2=NCOL-1
COPY MCOORCOL CORCOL.1-CORCOL.NCOL2
#PRINT CORCOL.1-CORCOL.NCOL2
#ANEM A FER LA PRESENTACIO DE DADES

LET SCOSQUA=0
DO CONS1=1:NCOL2
    LET SCOSQUA=SCOSQUA+CORCOL.CONS1*CORCOL.CONS1
ENDDO

LET DIST=ROUND(SCOSQUA*100)/100
#prin sumcol sumfil
DO CONS1=1:eixos
   LET COR.CONS1=ROUND(CORCOL.CONS1*100)/100
   LET CNT.CONS1=ROUND(CORCOL.CONS1*CORCOL.CONS1*SUMCOL/VALPRO(CONS1)*1000)/10
   LET COS.CONS1=ROUND(100*CORCOL.CONS1*CORCOL.CONS1/SCOSQUA)/100
ENDDO
NOTE
NOTE      ************************************************************
NOTE      COORDENADES, CONTRIBUCIONS I COSINUS QUADRATS DE LES COLUMNES
NOTE      ************************************************************
    PRINT nomc COR.1-COR.eixos, CNT.1-CNT.eixos, COS.1-COS.eixos

copy cor.1-cor.eixos caux.1-caux.eixos
copy nomc nomaux

LET SCOSQUA=0
DO CONS1=2:NCOL
    LET SCOSQUA=SCOSQUA+CORFIL.CONS1*CORFIL.CONS1
ENDDO
LET DIST=ROUND(SCOSQUA*100)/100

DO CONS1=1:3
   LET CONS3=CONS1+1
```

```
   LET CORFIL.CONS1=CORFIL.CONS3
   LET COR.CONS1=ROUND(CORFIL.CONS1*100)/100
   LET CNT.CONS1=ROUND(CORFIL.CONS1*CORFIL.CONS1*SUMFIL/VALPRO(CONS1)*1000)/10
   LET COS.CONS1=ROUND(100*CORFIL.CONS1*CORFIL.CONS1/SCOSQUA)/100
ENDDO
NOTE
NOTE      ************************************************************
NOTE      COORDENADES, CONTRIBUCIONS I COSINUS QUADRATS DE LES FILERES
NOTE      ************************************************************
    PRINT nomf COR.1-COR.eixos, CNT.1-CNT.eixos, COS.1-COS.eixos

#tenim a caux les coordenades de les columnes
#tenim a cor les coordenades de les fileres

NOTE VOLS Grafics?(y/n)
YESNO GRAF1

If graf1=1

# bucle per fer els grafics dels tres primers eixos
#do i=1:2
#   do j=2:3
#if i<>j
# si no volem el bucle desactivem-lo aqui i al final
# i activem els dos lets que segueixen
let j=2
let i=1

#per canviar el signe
#let caux.i=-caux.i
plot caux.j*caux.i;
symbol;
size 0.6;
tsize 0.7;
size 0.7;
label nomc;
title "coordenades columnes dos primers eixos factorials";
Axis 1;
Label "eix fact 1";
Axis 2;
Label "eix fact 2";
reference 1 0;
type 3;
reference 2 0;
type 3.

#per canviar el signe
#let cor.2=-cor.2
plot cor.j*cor.i;
symbol;
type 2;
label nomf;
tsize 0.7;
```

```
size 0.7;
title "coordenades fileres dos primers eixos factorials";
Axis 1;
Label "eix fact 1";
Axis 2;
Label "eix fact 2";
reference 1 0;
type 3;
reference 2 0;
type 3.

stack caux.j cor.j colaux.j.
stack caux.i cor.i colaux.i.
stack nomc nomf nomaux;
subs nomaux2.

set columaux;
format (a6).
column
files
end

set columau2
1
2
end

convert columau2 columaux nomaux2 nomaux2

copy nomaux etiquet

#per canviar el signe
#let caux.j=-caux.j

plot colaux.j*colaux.i;
symbol nomaux2;
type 1 2;
label etiquet;
tsize 0.7;
size 0.7;
title "coord. fileres i columnes dos primers eixos factorials";
Axis 1;
Label "eix fact 1";
Axis 2;
Label "eix fact 2";
reference 1 0;
type 3;
reference 2 0;
type 3;
.
endif
ENDMACRO
```

A.11 Macro INCORSI2

```
MACRO
INCORSI2 C.1-C.NCOL NFIL SUMTOT SUMDIAG SUMNDIAG SUMSUPDI SUMINFDI
#
# INCORSI2 CALCULA PER A LES COLUMNES QUE LI DONEM LES INERCIES
# PER SEPARAT DE LA DIAGONAL I DE FORA DE LA DIAGONAL D'UNA
# MATRIU DE CORRESPONDENCIES SIMPLES

MCONSTANT NFIL K1 K2 NIND SUMTOT SUMDIAG SUMNDIAG SUMSUPDI SUMINFDI
MCOLUMN C.1-C.NCOL CMOD.1-CMOD.NCOL MITJAFIL CIND.1-CIND.NCOL
MCOLUMN MITJACOL
MMATRIX MDADES MMITJFIL MMITJCOL MINDEP

# VERIFIQUEM QUE TOTES LES COLUMNES TINGUIN LA MATEIXA LONGITUD
DO K1=1:NCOL
    IF COUNT(C.K1) NE NFIL
        NOTE
        NOTE NO TOTES LES COLUMNES TENEN EL MATEIX NOMBRE D'INDIVIDUS
        NOTE ES NECESSARI I PER TANT S'INTERROMP L'EXECUCIO
        EXIT
    ENDIF
ENDDO

COPY C.1-C.NCOL MDADES
# MDADES ES LA MATRIU DE DADES INICIALS

# ANEM A CALCULAR EL NOMBRE D'INDIVIDUS TOTALS
LET NIND=0
DO K1=1:NCOL
    LET NIND=NIND+SUM(C.K1)
ENDDO
#COPIEM LES COLUMNES C A CMOD PER A TREBALLAR AMB AQUESTES I NO AMB LES ORIGINALS
COPY C.1-C.NCOL CMOD.1-CMOD.NCOL

#ANEM A DIVIDIR TOTES LES COLUMNES PEL NOMBRE D'INDIVIDUS TOTAL
DO K1=1:NCOL
    LET CMOD.K1=(CMOD.K1)/NIND
ENDDO

RSUM CMOD.1-CMOD.NCOL MITJAFIL
DO K1=1:NCOL
    LET MITJACOL(K1)=SUM(CMOD.K1)
ENDDO

#ANEM A POSAR A CIND.1-CIND.NCOL LES DADES AMB EL SUPOSIT D'INDEPENDENCIA
COPY MITJAFIL MMITJFIL
COPY MITJACOL MMITJCOL

TRANS MMITJCOL MMITJCOL
MULT MMITJFIL MMITJCOL MINDEP

COPY MINDEP CIND.1-CIND.NCOL
```

```
DO K1=1:NCOL
    LET CIND.K1=(CMOD.K1-CIND.K1)*(CMOD.K1-CIND.K1)/CIND.K1
ENDDO

NOTE MATRIU D'INERCIES
PRINT CIND.1-CIND.NCOL

#
#INICIALITZEM ELS COMPTADORS PER A LA INERCIA TOTAL I DE LA DIAGONAL
#I CALCULEM AQUESTES
#
LET SUMTOT=0
LET SUMDIAG=0
LET SUMSUPDI=0
DO K1=1:NCOL
    LET SUMTOT=SUMTOT+SUM(CIND.K1)
    LET SUMDIAG=SUMDIAG+(CIND.K1(K1))
    DO K2=1:K1
        LET SUMSUPDI=SUMSUPDI+(CIND.K1(K2))
    ENDDO
ENDDO

LET SUMSUPDI=SUMSUPDI-SUMDIAG
LET SUMINFDI=SUMTOT-SUMDIAG-SUMSUPDI
LET SUMNDIAG=SUMTOT-SUMDIAG
#PRINT SUMTOT SUMDIAG SUMNDIAG SUMSUPDI SUMINFDI
ENDMACRO
```