CAPÍTOL 7

ANÀLISI CLUSTER EN UNA POBLACIÓ MULTINOMIAL

Índex Capítol

CAPÍTOL 7

ANÀLISI CLUSTER EN UNA POBLACIÓ MULTINOMIAL

Encara que es detecti la presència d'una frontera d'estil, sempre és possible que alguns dels capítols, o blocs, quedin mal classificats. L'objectiu d'aquest capítol és presentar les tècniques d'agrupació, tant jeràrquiques com no jeràrquiques, basades en l'anàlisi cluster clàssica. A més a més, es proposen dues tècniques d'agrupació dels blocs i/o capítols en dos grups el més homogenis possible, a partir de les unitats d'estadística textual resumides en una taula de contingència. Les dues tècniques estan basades en l'ajust de models per dades politòmiques. La primera tècnica minimitza la Deviança, mentre que la segona la X^2 de Pearson.

Entenem per mètodes de *Classificació* aquells en els que coneixem a priori el nombre de grups i disposem d'observacions de les quals sabem a quin grup pertanyen, i que formen la mostra d'entrenament. Tenen per objectiu principal assignar noves observacions a un d'aquests grups. A aquestes tècniques també se les coneix com a "supervised classification". L'anàlisi *Cluster*, en canvi, és una tècnica més primitiva en la que no es fa cap tipus d'assumpció pel que fa al nombre de grups o a la seva estructura, ni es disposa de mostra d'entrenament. També se les coneix com a "unsupervised classification".

Les tècniques Cluster clàssiques han estat desenvolupades per a resoldre el problema d'agrupar n objectes en m grups, a partir de mesures en l variables, $\boldsymbol{y_i}=(y_{1i}, y_{2i}, ..., y_{li})$ per cada objecte i. Els grups resultants han de ser tals que els objectes que pertanyen a un mateix grup siguin "similars" entre sí i "diferents" als dels altres grups. Els mètodes d'agrupació han de ser completament numèrics i, a més a més, el nombre de grups, m, en general, no està determinat *a priori*. L'agrupament dels objectes en clusters es fa en base a mesures de similaritat o de dissimilaritat calculables a partir de les dades.

Atès que és gairebé impossible examinar totes les possibles agrupacions d'objectes per determinar la millor, han aparegut una sèrie de tècniques que troben agrupacions

raonablement bones sense haver d'examinar-les totes. A la secció 7.1 es revisaran les mesures de dissimilaritat. A continuació es repassaran els dos grups de tècniques clàssiques per a l'anàlisi cluster: a la secció 7.2 les tècniques no jeràrquiques, basades en particions i a la secció 7.3 les tècniques jeràrquiques. A la secció 7.4 es proposaran dues tècniques no jeràrquiques per a agrupar les files d'una taula de contingència basades en l'ajust de models.

A la tesi es fan servir les tècniques clàssiques basades en particions per agrupar capítols o blocs en funció dels índexs de diversitat, mentre que per a l'ús de paraules i per a la llargada de paraula i de frase s'aplicaran les noves tècniques que es proposen per a dades en forma de taula de contingència. No es faran servir les tècniques jeràrquiques, atès que per a la nostra base de dades formada per uns 400 objectes a agrupar, els resultats que s'obtenen són de difícil interpretació.

7.1 Mesures de dissimilaritat entre objectes o variables

Alguns dels mètodes d'anàlisi per identificar una estructura en grups a partir d'una base de dades complexa es basen en la definició de mesures de similaritat, s_{ie}, o dissimilaritat, d_{ie}, entre els objectes *i-èssim* i *e-èssim*, per a tots els parells d'objectes *(i,e)*. En la majoria dels casos les mesures de similaritat poden ser transformades en mesures de dissimilaritat fent servir, per exemple, $d_{ie}=-a\cdot s_{ie}$, per a alguna constant a. Les mesures de dissimilaritat entre parells d'objectes, d_{ie}, es defineixen com:

i) $d_{ie} \geq 0,$

ii) $d_{ii} = 0, \quad i$

iii) $d_{ie} = d_{ei}$,

per a tots els $i,e=1,...,n$. Alguns mètodes, a més a més, assumeixen que els objectes poden ser representats en un espai Euclidià. Una matriu de dissimilaritats $\boldsymbol{D} = \{d_{ie}\}$ es diu que és mètrica si satisfà la desigualtat triangular:

$$d_{ie} \leq d_{ih} + d_{he} ,$$

per a qualsevol *(i,e,h)*. Per un repàs sobre mesures de dissimilaritat per a altres tipus de variables veure, per exemple, Gordon (1999).

Sigui y_{ji} el valor que pren la variable j-èssima per l'objecte i-èssim. Algunes de les mesures de dissimilaritat que s'han fet servir més en anàlisi cluster per variables quantitatives són les mètriques de Minkowski:

$$d_{ie} = (\Sigma_{j=1}^{l} \left| y_{ji} - y_{je} \right|^{\lambda})^{1/\lambda} \qquad (\lambda \geq 1) .$$

Quan $\lambda=1$, la distància de Minkowski és la distància de Manhattan o City-Block, mentre que per $\lambda=2$ tenim la distància Euclídea. En general, variant el valor de λ es canvia el pes donat a les dissimilaritats més grans.

Freqüentment s'estandaritzen les variables abans de calcular les distàncies, per tal que les l variables siguin igualment importants en determinar les distàncies. Una manera de fer-ho és ponderant-les, de manera que tinguin la mateixa variança. En aquest cas, els

pesos assignats a les variables són del tipus: $\omega_j=(\sigma_j)^{-1}$, on σ_j és la desviació tipus per la variable j-èssima, per $j=1,...,l$. La distància de Minkowski queda com:

$$d_{ie} = \left(\Sigma_{j=1}^{l} \omega_j^{\lambda} \left| y_{ji} - y_{je} \right|^{\lambda}\right)^{1/\lambda} .$$

En algunes aplicacions interessa agrupar variables en lloc d'objectes. Les mesures de similaritat per variables quantitatives sovint prenen la forma de coeficients de correlació mostrals, i en algunes aplicacions es substitueixen les correlacions negatives per els seus valors absoluts. En la tesi no agruparem variables, pel que no farem servir aquestes mesures de similaritat.

7.2 Tècniques Cluster basades en particions

Les tècniques no jeràrquiques, que impliquen particions, estan dissenyades per agrupar objectes, i no permeten les agrupacions entre variables. Les tècniques que farem servir al llarg de la tesi per agrupar capítols o blocs són d'aquest tipus.

El nombre de clusters a obtenir, m, pot ser especificat per avançat o be pot ser determinat com a part del procediment. A diferència dels mètodes jeràrquics, els objectes es poden moure entre els grups en diferents moments de l'anàlisi, i com que no s'han de calcular les matrius de dissimilaritats entre objectes, aquestes tècniques poden ser aplicades a conjunts de dades molt més grans que els mètodes jeràrquics.

En tots els mètodes no jeràrquics l'algorisme comença, be amb una assignació arbitrària dels objectes a un dels grup i calculant el centre d'aquests grups, o bé amb la tria d'uns centres de grup feta, també, de forma arbitrària. A continuació es repassarà l'algorisme *k-means*, una de les tècniques cluster clàssiques basades en particions i que en la tesi s'ha utilitzat en l'anàlisi dels índexs de diversitat.

L'algorisme k-means

Es defineix el centre del grup *k-èssim* com el vector de mitjanes dels objectes que estan assignats al grup:

$$Centre_k = \bar{y}_k = (\bar{y}_{k1},...,\bar{y}_{kl}),$$
$$\bar{y}_{kj} = \frac{1}{n_k}\sum_{i\in k} y_{ji},$$

on n_k és el nombre d'objectes en el grup k-èssim. Es calculen les distàncies des de cada objecte als centres dels m clusters. Normalment aquesta distància és la distància euclídea, i les variables tant poden estar estandaritzades com no.

Comença un procés iteratiu en el que, a cada iteració s'avança al llarg de tots els n objectes, assignant cada objecte al cluster que té el centre més proper. En acabar la iteració es recalculen els centres dels grups. S'itera fins que s'arriba a un punt en el que no es poden fer més canvis d'assignació d'objectes a grups.

En una segona versió de l'algorisme, a cada iteració canvia d'assignació només l'objecte que té una distància més curta al centre d'un cluster diferent al que està assignat en aquell moment.

L'assignació final d'objectes a clusters depèn, en alguna mesura, de la partició inicial. Si es té informació *a priori* de pertinences d'objectes a grups es aconsellable utilitzar-la. Altrament una bona estratègia consisteix en generar moltes particions inicials i, per cadascuna d'elles, veure quin és el resultat final, per quedar-nos amb aquell que sigui millor segons algun criteri, com per exemple el de mínima variança dintre dels clusters o màxima distància entre centres de clusters.

A la secció 7.4 es descriuen dos algorismes cluster, basats en el *k-means*, per agrupar les files d'una taula de contingència.

7.3 Tècniques Cluster Jeràrquiques

A la tesi no farem servir les tècniques cluster jeràrquiques. En agrupar al voltant de 400 objectes, capítols o blocs, els resultats que s'obtenen són de difícil interpretació. Tot i així, aquestes tècniques poden ser molt útils en estudis estilomètrics, i per aquest motiu incloem una breu descripció.

Les tècniques jeràrquiques, a partir de la matriu de distàncies o dissimilaritats $\boldsymbol{D}=\{d_{ie}\}$, procedeixen a través d'una sèrie d'aglomeracions, o de divisions, d'objectes i el resultat es pot representar en un gràfic conegut com a *dendograma*. Els mètodes jeràrquics proposats serveixen tant per agrupar objectes com variables.

En els *mètodes jeràrquics aglomeratius* es comença amb que cada objecte constitueix un grup diferent, per tant hi ha *n* grups. En la primera etapa s'ajunten en un grup els dos objectes més similars, desapareixent els dos objectes originals i quedant *n-1* grups, i es recalculen les distàncies des del nou grup a tots els altres. En cada etapa successiva s'ajunten els dos grups més similars (menys dissimilars) en un de nou, i es recalcula la matriu de dissimilaritats eliminant les files i columnes corresponents als dos grups que s'han unit en un de sol i afegint la fila i columna corresponent a les distàncies entre el nou cluster i els altres clusters. Es van unint clusters de dos en dos fins a que, en l'última etapa, es fusionen els dos últims grups en un de sol.

Els *mètodes jeràrquics divisius* funcionen de manera oposada: es comença amb tots els objectes en un únic cluster, i s'acaba amb tants clusters com objectes. En la primera etapa l'únic cluster es trenca en els dos grups més dissimilars possibles. En cada etapa successiva es va dividint un dels grups en els dos més dissimilars, fins a tenir tants grups com objectes. A diferència del que passa en els algorismes no jeràrquics, un cop dos objectes s'han unit en un mateix grup, ja no es separen en les etapes successives.

Representació gràfica: Dendograma

El resultat dels mètodes cluster jeràrquics pot ser representat en forma de gràfic en dues dimensions, conegut com a *dendograma*. Tal i com es pot veure en la figura 7.1, el *dendograma* mostra les agrupacions (o divisions, segons el mètode) que s'han donat en els passos successius, així com el nivell de similaritat al que s'han unit (o separat) els grups. El gràfic de la figura 7.1 s'ha realitzat, fent servir el mètode aglomeratiu, per

l'estudi de la longitud de paraula en blocs de 1000 paraules consecutives, essent els objectes els blocs, i les variables el nombre de paraules de *k* lletres, per *k=1,2,...,10+*. Els resultats mostren com les variables més similars, i per tant les primeres en unir-se en un sol clustre, són les llargades 9 i 10 i superior. A continuació aquest cluster s'uneix amb el format per la llargada 8. Els grups es van unint, fins que la darrera unió es té entre els clusters (2,3,4) i (1,5,6,7,8,9,10+).

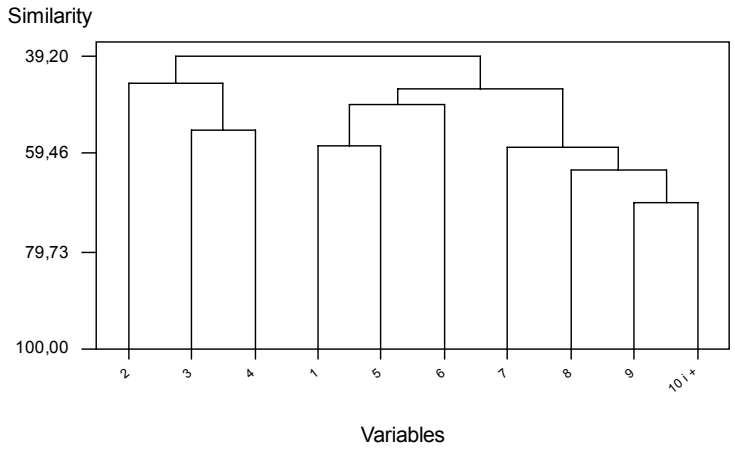

Figura 7.1: Dendograma del procés jeràrquic aglomeratiu per a obtenir clusters basats en les llargades de paraula. Els objectes són els blocs de 1000 paraules.

7.4 Cluster no jeràrquic per files d'una taula de contingència

En aquest capítol, proposem dos mètode per agrupar objectes en clusters basats en particions, semblant al *k-means* presentat a l'apartat 7.2. En tots dos s'ajusten models de regressió politòmica, en el primer es minimitza la Deviança i en el segon es minimitza la distància X^2 de Pearson, mètode a substituir en l'algorisme *k-means* la distància euclídea per la distància χ^2.

Les tècniques cluster clàssiques acostumen a partir de la distància euclídea o generalitzacions de la distància euclídea com la City Block o la de Minkowski. Per dades en forma de taula de contingència aquestes distàncies perden bona part del seu sentit. Per exemple, en el cas dels capítols de llargada variable, la distància euclídea i les seves generalitzacions mesurarien diferències en la llargada del capítol més que diferències en proporcions en cada categoria.

La taula de contingència sobre la que s'aplicaran els mètodes està formada per tantes files (*n*) com objectes estem considerant i tantes columnes (*l*) com categories de la variable haguem triat, i el valor contingut en cada cel·la y_{ji} correspon al nombre d'ocurrències en la categoria j-èssima comptades en el bloc o capítol i-èssim.

En alguns problemes es poden tractar files i columnes de forma simètrica. Greenacre (1988,1993) planteja realitzar una anàlisi cluster de les files i de les columnes d'una taula de contingència, emprant com a mesura de dissimilaritat entre objectes (siguin

files o columnes de la taula) la distància χ^2, fent servir tècniques jeràrquiques. El fet de tenir un nombre molt elevat de capítols (425 amb més de 200 paraules) o de blocs (396) fa que l'ús de tècniques jeràrquiques sigui poc informatiu, atès que tant el dendograma resultant són difícilment interpretables. En l'estudi del *Tirant*, l'agrupació de columnes té molt menys interès que la de les files. Agrupar files vol dir trobar capítols (o blocs) homogenis entre sí i diferents dels que pertanyen a un altre cluster, mentre que fer-ho amb les columnes vol dir trobar categories amb comportaments semblants.

Presentem una versió simplificada dels algorismes, en la que el nombre de clusters en els que es volen agrupar els capítols (o blocs) és dos, però l'algorisme pot generalitzar-se al cas de *m* clusters sense dificultat. Anomenarem als dos clusters: Cluster *0* i Cluster *1*.

7.4.1 Algorisme basat en la Deviança

El primer mètode que es proposa per a l'obtenció d'agrupacions de files d'una taula de contingència mitjançant l'ajust de models està molt relacionat amb el mètode d'estimació del punt de canvi en una seqüència multinomial, tractats en el capítol 6 de la tesi. Per estimar el punt de canvi per respostes politòmiques $\boldsymbol{y_i}=(y_{1i},y_{2i},...,y_{li})$, tal que:

$$\boldsymbol{y_i} \sim Multinomial(N_i, \boldsymbol{\pi_i}),$$

on $\boldsymbol{\pi_i}=(\pi_{1i},\pi_{2i},...,\pi_{li})$ és el vector de probabilitats per la mostra i-èssima, podem modelar $\boldsymbol{y_i}$ prenent com a referència la categoria *1* i per a una *r* donada i ajustant, pel mètode de la màxima versemblança, el model (6.4):

$$g(\pi_i) = (g_2(\pi_i),....,g_l(\pi_i)),$$

on:

$$g_j(\pi_i) = \log\left(\frac{\pi_{ji}}{\pi_{1i}}\right) = \beta_{0j}^{(r)} + \beta_{1j}^{(r)} Ind_{1i}^{(r)},$$

per $j=2,3,...,l$. Recordem que $Ind_{1i}^{(r)}$ és una variable indicadora que pren valor *0* per a les mostres $1,2,...,r$ i valor *1* per les mostres $r+1,...,n$.

Per a agrupar les dades en dos clusters ajustem el mateix tipus de models amb l'única diferència que la variable indicadora, $Ind_{1i}^{(c)}$, prendrà valor *0* per a les files que pertanyen *Cluster 0*, i prendrà valor *1* per a les files que pertanyen *Cluster 1*:

$$g_j(\pi_i) = \log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \beta_{j0}^{(c)} + \beta_{j1}^{(c)} Ind_{1i}^{(c)},$$

per $j=2,3,...,l$.

L'algorisme comença amb l'assignació de cada fila a un dels dos clusters, s'ajusta el model i es calcula la versemblança. A continuació, un procés iteratiu, anàleg al descrit per l'algorisme *k-means*, calcula les versemblances dels *n* models en els que la fila i-èssima ha canviat d'assignació de cluster i les altres *n-1* tenen l'assignació que tenien en acabar la iteració anterior, per $i=1,...,n$. Dels *n* possibles canvis d'assignació, es consolida aquell pel qual el model té el màxim guany en la versemblança i es comença una nova iteració. El procés iteratiu acaba quan cap canvi d'assignació provoca un augment en la versemblança del model.

L'assignació inicial de les files a un dels clusters pot ser feta, com en el cas de l'algorisme *k-means*, tant a partir de coneixements a priori com de forma arbitrària.

Exemple 1: es parteix de la taula de contingència formada pels *n=396* blocs de 1000 paraules com a files i les *l=10* categories de llargada de paraula. Degut al cost computacional que suposa el mètode, s'ha fet una sola execució, on l'assignació inicial està basada en l'estimació del punt de canvi obtinguda en el capítol 8 de la tesi, les primeres 327 files assignades al *Cluster 0*, les files de la 328 a la 396 al *Cluster 1*.

En la figura 7.2 es pot observar l'augment en el logaritme de la versemblança en funció del número de iteració.

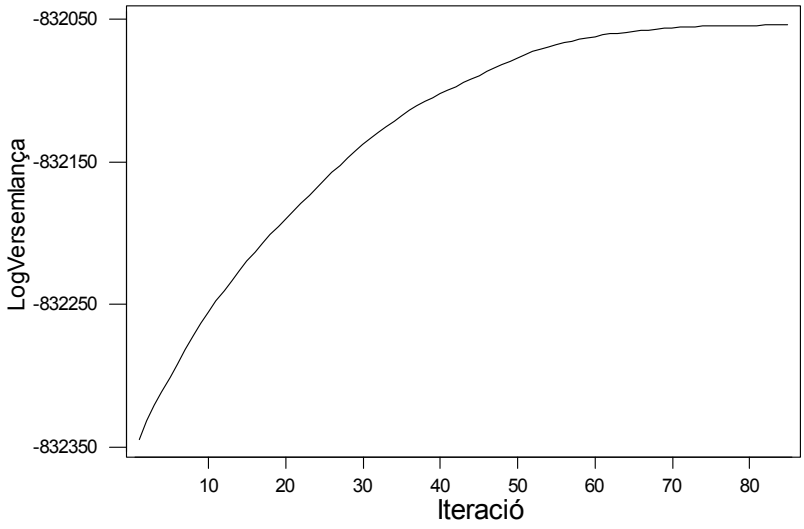

Figura 7.2: Evolució del logaritme de la versemblança amb el nombre de iteracions, per a les dades de l'exemple 1. Les files corresponen als blocs de 1000 paraules i les columnes a les 10 categories en que s'ha subdividit la llargada de paraula, i que s'analitzaran en el capítol 8 de la tesi.

7.4.2 Algorisme basat en la distància $\chi 2$

El mètode és idèntic en tot al proposat en l'apartat 7.4.1, excepte en que com a criteri d'ajust dels models ara es fa servir la distància X^2 de Pearson en comptes de la Deviança (o versemblança).

Sigui $\boldsymbol{y_i}=(y_{1i},y_{2i},....,y_{li})$ l'i-èssim vector d'observacions, amb N_i la suma dels elements de la fila, i y_{ji} el nombre d'ocurrències en la categoria j-èssima per a l'objecte i-èssim. L'algorisme comença amb l'assignació arbitrària de cada fila a un dels dos clusters, s'ajusta el model pel criteri de distància X^2 de Pearson mínima:

$$g(\pi_i) = (g_2(\pi_i),....,g_l(\pi_i)),$$

on:

$$g_j(\pi_i) = \log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \beta_{j0}^{(c)} + \beta_{j1}^{(c)} Ind_{1i}^{(c)},$$

per $j=2,3,...,l$. Recordem que $Ind_{1i}^{(c)}$ és una variable indicadora que pren valor *0* per a les files que pertanyen *Cluster 0*, i prendrà valor *1* per a les files que pertanyen *Cluster 1*.

La distància X^2 de Pearson es calcula com:

$$X^2 = \sum_{i=1}^{n} \sum_{j=1}^{l} \frac{\left(y_{ji} - \hat{y}_{ji}\right)^2}{\hat{y}_{ji}},$$

on $E\left(y_{ij}\right) = \hat{y}_{ji}$.

L'assignació inicial de les files a un dels clusters pot ser feta tant a partir de coneixements a priori com de forma arbitrària.

Es continua amb un procés iteratiu, anàleg al descrit per l'algorisme cluster basat en la deviança, que calcula les distàncies X^2 pels n models en els que la fila i-èssima ha canviat d'assignació de cluster i les altres $n-1$ tenen l'assignació que tenien en acabar la iteració anterior, per $i=1,...,n$. Dels n possibles canvis d'assignació, es consolida aquell pel qual distància X^2 és mínima i es comença una nova iteració. El procés iteratiu acaba quan cap canvi d'assignació provoca un augment en la versemblança del model.

Aquest mètode és equivalent a buscar les dues agrupacions de manera que la distància χ^2 entre els dos clusters sigui mínima. Siguin n_0 i n_1 el nombre de files assignades en el pas inicial al *Cluster 0* i *1*, respectivament, de forma que $n_0+n_1=n$. S'obtenen dues taules de contingència de dimensions $n_0 \times l$ i $n_1 \times l$. Es calculen els estadístics χ^2, definits a (5.4), per cadascuna d'aquestes dues taules que anomenarem, respectivament χ^2_0 i χ^2_1, com a suma de les distàncies entre el valor observat per a cada cel·la del cluster i el seu valor esperat, sota la hipòtesi de independència entre files i columnes:

$$\chi_k^{\,2} = \sum_{i \in k} \sum_{j=1}^{l} \frac{(y_{ji} - \hat{y}_{ji})^2}{\hat{y}_{ji}}; \qquad (7.1)$$

per $k=0,1$, on el valor esperat per la cel·la (i,j) ve donat per:

$$\hat{y}_{ji} = \frac{\sum_{h \in k} y_{jh} \sum_{m=1}^{l} y_{mi}}{N_k},$$

on:

$$N_k = \sum_{i \in k} \sum_{j=1}^{l} y_{ji},$$

és el nombre total d'observacions en el cluster *k-èssim*. A continuació calculem l'estadístic χ^2 per tota la taula, fent servir (7.1) suposant que totes les files pertanyen al mateix cluster.

Diem que la suma $\chi^2_W = \chi^2_0 + \chi^2_1$ és el valor de l'estadístic *dins dels clusters,* i és una mesura del grau de homogeneïtat de les files dins dels clusters. Com més petit és χ^2_W més homogenis són els clusters.

Sumem les cel·les d'una mateixa columna per totes les files que pertanyen a un cluster, de forma que ens quedem amb una taula de contingència de $2 \times l$. Anomenem $\hat{y}_{jk}^{(a)}$ a:

$$y_{jk}^{(a)} = \sum_{i \in k} y_{ji},$$

per $k=0,1$; i calculem l'estadístic χ^2 per aquesta taula:

$$\chi^2_B = \sum_{k=0}^{1}\sum_{j=1}^{l} \frac{(y_{jk}^{(a)} - \hat{y}_{jk}^{(a)})^2}{\hat{y}_{jk}^{(a)}},$$

on:

$$\hat{y}_{jk}^{(a)} = \frac{\left(\sum_{i=1}^{n} y_{ji}\right)\left(\sum_{m=1}^{l} y_{mk}^{(a)}\right)}{N},$$

amb:

$$N = \sum_{i=1}^{n}\sum_{j=1}^{l} y_{ji} = \sum_{k=0}^{1}\sum_{j=1}^{l} y_{jk}^{(a)}$$

el total d'observacions. Anomenem aquest estadístic χ^2_B o *entre clusters*. Com més gran és el valor de χ^2_B, més diferents són els dos clusters entre ells.

A continuació s'inicia el procés iteratiu en el que, a cada iteració només una fila canvia d'assignació. Entre tots els possibles canvis d'assignació de cluster es fa aquell en el que l'increment de l'estadístic χ^2_B de la taula resultant del canvi d'assignació és maximitzat. Aquest canvi fa que el cluster que perd un element guanyi en homogeneïtat, i per tant el valor de χ^2_i dins d'aquest cluster disminueix. Pel cluster que guanya un element, l'estadístic χ^2_i dins del cluster pot augmentar, en augmentar el nombre d'elements, però el valor de χ^2_W sempre disminueix amb aquests canvis. D'aquesta manera els dos clusters van essent cada vegada més diferents entre sí i, alhora, més homogenis. El procés continua fins que no hi ha cap canvi d'assignació que augmenti l'estadístic χ^2_B.

El nombre de iteracions necessàries per a obtenir una divisió en clusters, tal que cap canvi d'assignació resulti amb un increment en la χ^2_B, depèn de l'assignació inicial de les files a un dels dos clusters. La composició final dels clusters també pot variar amb assignacions inicials diferents.

Una mesura de la bondat de l'agrupació de les files en dos clusters és la relació entre la distància χ^2_B entre clusters i el valor de χ^2 per a la taula completa. Com més gran sigui aquesta relació χ^2_B/χ^2, major serà la proporció de no-homogeneïtat de la taula completa explicada per la divisió en dos clusters. L'algorisme d'obtenció de clusters a partir de la taula de contingència mitjançant el mètode basat en la distància χ^2 ha sigut desenvolupat en FORTRAN.

Exemple 2: les dades corresponen a la llargada de paraula, la taula de contingència està composada per *n=425* capítols de més de 200 paraules de llargada i *l=10* categories, que corresponen a les paraules de *j* lletres, per *j=1,2,...,10+*.

En la primera execució de l'algorisme, els capítols han sigut assignats inicialment de forma alternativa a un dels dos clusters, és a dir, els capítols primer, tercer, cinquè, setè i tots els senars han sigut assignats al Cluster *0*, el segon, quart, sisè i endavant al Cluster *1*. Per arribar a una solució estable calen 228 iteracions o, dit d'una altra manera, 228 canvis d'assignació.

En una segona execució, basant-nos en punt de canvi estimat en l'anàlisi de la llargada de paraula, és a dir, en coneixement a priori, els capítols del *1* al *345* han sigut assignats inicialment al Cluster *0* mentre que els capítols del *346* fins al final s'han assignat al Cluster *1*. Ara, amb 97 iteracions s'arriba a la solució estable.

Els gràfics de la figures 7.3 mostren l'evolució de χ^2_B i χ^2_W per a les dues assignacions inicials. En elles es pot observar tant el creixement de χ^2_B com el decreixement de χ^2_B en funció del número de iteració, així com el nombre de iteracions necessàries per arribar a l'estabilitat.

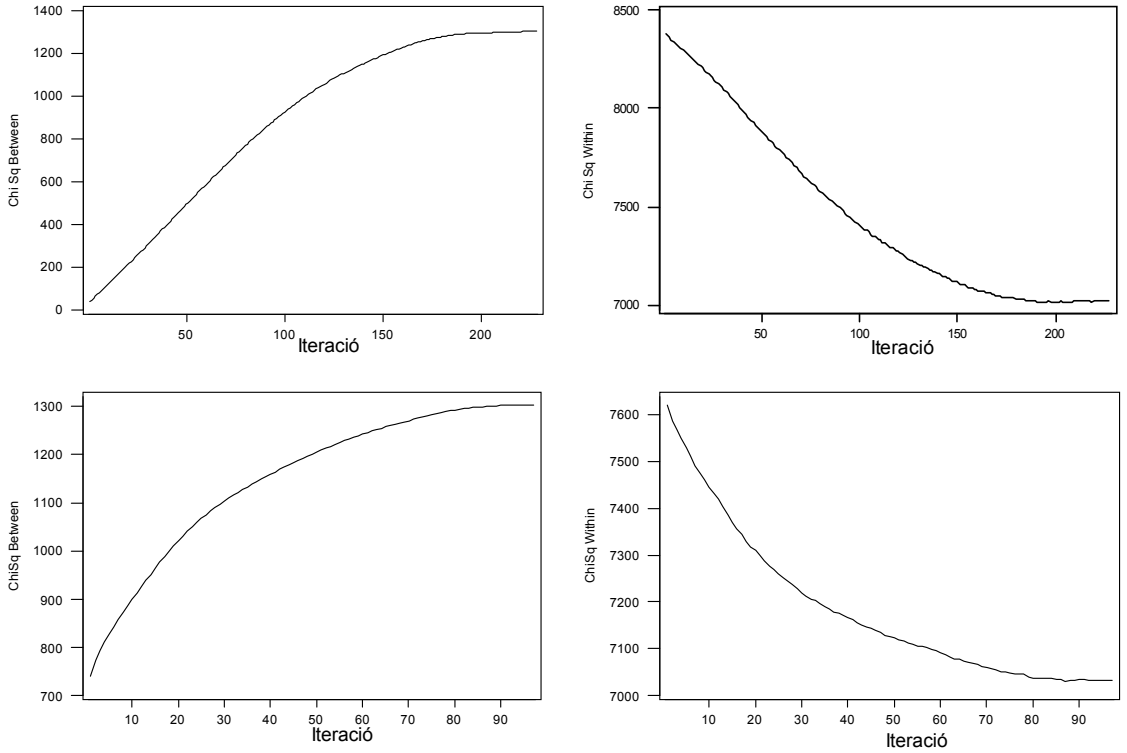

Figura 7.3: Evolució de χ^2_B i de χ^2_W en funció del nombre d'iteracions de l'algorisme per a l'exemple 2: en els gràfics de dalt l'assignació inicial de les files de la taula de contingència de llargades de paraula s'ha fet de forma alternada (capítols senars al cluster *0*, capítols parells al *1*), mentre que en els gràfics de baix, s'han assignat els capítols *1-345* al cluster *0* i els capítols *346-487* al cluster *1*.

La solució que s'obté no és única i depèn de l'assignació inicial. Una possible estratègia consisteix en generar varies assignacions inicials i triar com a resultat final aquell per al que χ^2_B és màxima. S'han realitzat més de 1000 execucions de l'algorisme amb les dades de l'exemple 2 amb assignacions inicials aleatòries. Els resultats no convergeixen cap a una única solució, obtenint-se algunes agrupacions en clusters diferents, tot i que la major part (prop del 80%) convergeix a una mateixa classificació que, a més a més, és la que presenta un valor de χ^2_B més elevat.

Exemple 3: la taula de contingència aquí és la mateixa que per l'exemple 1: les 396 files corresponen als blocs de 1000 paraules i les i *l=10* categories són les paraules de *j* lletres, per *j=1,2,...,10+*.

Aquí, els resultats que s'assoleixen són exactament els mateixos sigui quina sigui la assignació inicial que es faci. La divisió en clusters obtinguda explica el *17,1%* ($\chi^2_B/\chi^2=0.171$) de la "no-homogeneïtat" entre tots els capítols en funció de la llargada de paraula. Els valors obtinguts són: $\chi^2=8408.3$ i $\chi^2_B=1304.9$. Els resultats, tot i no ser molt diferents, no són els mateixos que pel mètode basat en la deviança.

7.4.3 Relació dels algorismes amb l'anàlisi de correspondències i discussió de resultats

Els clusters obtinguts de l'aplicació de l'algorisme separen els capítols del *Tirant* en funció, bàsicament, de la primera component de l'Anàlisi de Correspondències, com es pot veure per als resultats de l'exemple 2 en la figura 7.4. De l'agrupació obtinguda es pot veure com els dos clusters estan separats per una línia, i a l'interior de cada cluster hi ha una gran heterogeneïtat, havent-hi capítols molt diferents entre sí en el mateix cluster. En el capítol 8 de la tesi, dedicat a l'estudi de l'homogeneïtat d'estil a través de la llargada de paraula es veuran més detalls de l'aplicació.

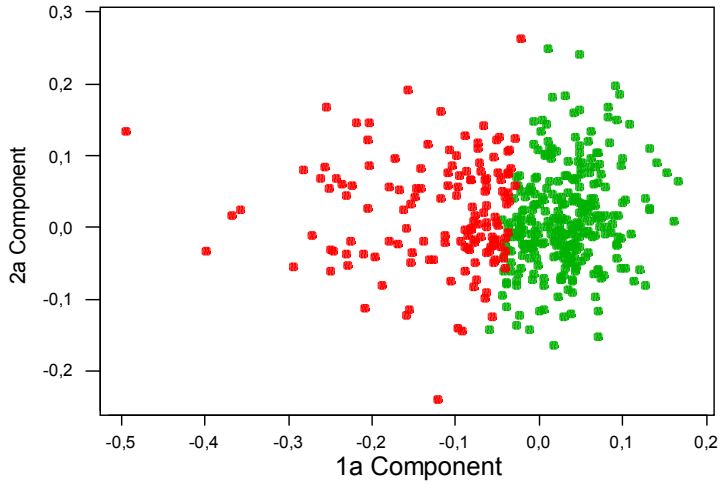

Figura 7.4: Gràfic de les projeccions de les files en les dues primeres components de l'Anàlisi de Correspondències per a les dades del exemple 1. El color verd correspon als blocs assignats al Cluster *0* mentre que els verds són aquells que pertanyen al Cluster *1*.

En anàlisi de correspondències, es defineix inèrcia t al quocient entre la distància χ^2 i N, la suma de totes les observacions:

$$t = \frac{\chi^2}{N} = \frac{\chi^2}{\sum_i \sum_j y_{ij}}$$

Greenacre (1988) comprova com el màxim de χ^2_B/χ^2 que podem obtenir serà sempre menor que la proporció de la inèrcia total que explica la primera component de l'Anàlisi de correspondències. Per l'exemple 2 el quocient queda limitat al 27%, i per l'exemple 3 al 28%.

Si es comparen els resultats de l'obtenció d'agrupacions de files d'una taula de contingència mitjançant els dos mètodes proposats, basats en l'ajust de models per dades politòmiques, en els que canvia el criteri de qualitat de l'ajust, la deviança i distància X^2, es pot observar com els resultats són molt semblants. Els clusters obtinguts tot i que no són idèntics, tant els valors de la versemblança del model com el de les distàncies χ^2_B i χ^2_W són molt propers en els dos casos. Per la taula de contingència dels exemples 1 i 3, hi ha diferències en l'assignació de 16 blocs. Segons el mètode basat en la deviança tots ells estan assignats al cluster que conté majoritàriament els capítols del final del llibre, mentre que segons el mètode basat en la distància χ^2 estan assignats a l'altre cluster. La taula 7.1 mostra les diferències en els estadístics que ens han servit per obtenir els clusters a partir dels dos mètodes proposats. En ella es pot observar com les

diferències en els resultats obtinguts fent servir els dos mètodes són petites. En la tesi es faran servir tots dos mètodes per a l'anàlisi de la llargada de paraula i de l'ús de paraules, mentre que per la llargada de frase només s'ha aplicat el mètode basat en la distància χ^2.

Estadístic	**Mètode basat en Deviança**	**Mètode basat en χ^2**
Log versemblança	-832053,386	-832055,217
χ^2_B	1333,792	1340,164
χ^2_W	6473,366	6473,522
χ^2_0	4180,047	4455,544
χ^2_1	2293,319	2017,978

Taula 7.1: Comparació dels resultats obtinguts en l'obtenció de 2 clusters a partir de la taula de contingència de blocs i categories de llargada de paraula, segons els dos mètodes proposats.