CAPÍTOL 5

EINES D'ANÀLISI EXPLORATÒRIA DE DADES

Índex Capítol

5.1 Gràfics de Control 35
5.2 Gràfics de Control per dades univariants 36
5.2.1 Gràfics de Control per variables contínues.....37
5.2.2 Gràfics de Control per dades categòriques 38
5.2.3 Gràfics Cusum 40
5.3 Gràfics de Control per dades Multivariants 41
5.3.1 Gràfics de control T^2 per a dades Normals.....41
5.3.2 Gràfics de Control per dades Multinomials.....42
5.4 Anàlisi de Correspondències.....43

CAPÍTOL 5

EINES D'ANÀLISI EXPLORATÒRIA DE DADES

En aquest capítol es repassen dues tècniques d'anàlisi exploratòria de dades. Els gràfics de control són una eina molt utilitzada en la monitorització de processos amb l'objectiu de mantenir-los en estat de control. L'anàlisi de correspondències és una tècnica gràfica per a representar en dues dimensions una taula de contingència. Els gràfics de control ens serviran per detectar gràficament punts de canvi, que l'anàlisi de correspondències ens ajudarà a interpretar de forma qualitativa.

5.1 Gràfics de Control

En qualsevol procés existeix un cert grau de variabilitat inherent o natural. Aquesta variabilitat natural, o soroll de fons, és deguda a la suma dels efectes de moltes petites causes, essencialment no controlables i inevitables. Quan el soroll de fons d'un procés és relativament petit i es manté estable, considerem que és acceptable i a les causes que l'ocasionen les anomenem causes comuns o aleatòries. En aquest cas diem que el procés està en estat de control i, tot i que no podem predir quin serà el valor exacte de la propera observació, sovint tenim informació sobre la seva distribució de probabilitat.

De forma ocasional poden aparèixer en el sistema altres tipus de causes de variabilitat. Acostumen a tenir efectes més forts que els provocats per les causes comuns i normalment provoquen un nivell de variabilitat inacceptable. El fet d'aparèixer de forma esporàdica i que els efectes siguin més grans permet identificar aquestes causes i, o bé eliminar-les, o bé posar els mitjans per tal que no es tornin a donar. Les anomenem causes assignables. Un procés en el que hi ha causes assignables presents es diu que està fora de control, perquè les observacions que genera no són independents ni estan idènticament distribuïdes.

Els gràfics de control són una tècnica emprada en el control de processos amb l'objectiu de detectar l'ocurrència de causes assignables que signifiquin canvis en el procés, per tal que es puguin investigar i prendre les accions correctives que corresponguin.

Un gràfic de control d'una variable en estudi és una representació gràfica d'una seqüència d'observacions, preses equiespaiadament en el temps, en funció del número de mostra o del temps. En el gràfic hi apareix una línia horitzontal que representa el valor mig de la variable controlada, i dues línies horitzontals més, anomenades límit superior i inferior de control, respectivament. Aquests límits estan triats de manera que, si el procés està en estat de control, els punts han d'aparèixer de forma aleatòria i gairebé la totalitat d'ells cauen a l'interior dels dos límits.

Hi ha moltes mesures de qualitat del comportament d'un gràfic. Una d'elles, la probabilitat de falsa alarma, és la probabilitat de que el procés doni un senyal de fora de control quan en realitat és estable. La majoria d'autors han adoptat la longitud mitjana de recorregut, ARL (Average Run Length), com a mètode per resumir les característiques i propietats dels gràfics de control. L'ARL és el promig del nombre de punts mostrals que cal representar en el gràfic abans de tenir-ne un que indiqui una condició de fora de control quan, en realitat, el procés està en estat de control. Pels gràfics univariants clàssics l'ARL val $1/p$, on p és la probabilitat de que un punt surti fora dels límits de control mentre el procés està en control.

En l'estudi de l'homogeneïtat de texts literaris, els gràfics de control poden ser usats per detectar quan hi ha un canvi de comportament en la variable analitzada. Aquest canvi pot indicar l'existència d'una frontera estilística, no atribuïble a la variabilitat inherent, tot i que també pot indicar, per exemple, un canvi en l'argument, .

En funció del tipus de variable a monitoritzar es pot fer una primera classificació dels gràfics de control en univariants i multivariants. Quan estudiem, per exemple, l'evolució de la llargada mitjana de paraula o la proporció de paraules de 6 lletres fem servir gràfics univariants, mentre que quan estudiem simultàniament les proporcions de paraules en cadascuna de les 10 categories en que s'ha subdividit la llargada de paraula fem servir gràfics multivariants.

5.2 Gràfics de Control per dades univariants

Els gràfics de control per a dades univariants estudien l'estabilitat d'un procés o sistema analitzant l'evolució d'una variable real, que pot ser contínua (gràfics de control per variables) o categòrica (gràfic de control per atributs).

En aquesta secció es farà un ràpid repàs als gràfics clàssics per variables tant contínues com categòriques, així com una introducció als gràfics de sumes acumulades, Cusum. Els gràfics per variables contínues es faran servir per estudiar l'evolució de les següents variables:

- llargada mitjana de paraula i les dues primeres components de l'anàlisi de correspondències per la llargada de paraula, tant per capítols com per blocs, en el capítol 8,
- llargada mitjana de frase i la primera component de l'anàlisi de correspondències per la llargada de frase en el capítol 9,

- llargada de capítol i logaritme de la llargada de capítol en el capítol 10,
- les dues primeres components de l'anàlisi de correspondències per l'ús de paraules, realitzat tant per les *12* com per les *25* paraules més freqüents, en el capítol 11,
- els 10 índexs de diversitat, tant per capítols com per blocs, i la primera component de l'anàlisi de components principals només per blocs, en el capítol 12.

Pel que fa als gràfics per variables categòriques es faran servir per estudiar l'evolució de les següents variables:

- nombre de paraules en les 10 categories en que s'ha dividit la llargada de paraula, en el capítol 8,
- nombre de frases en les 10 categories en que s'ha dividit la llargada de frase, en el capítol 9,
- nombre de vegades que apareixen les 25 paraules més abundants en el *Tirant*, un cop eliminades dues paraules sensibles al context, en el capítol 11,

5.2.1 Gràfics de Control per variables contínues

Quan es tracta de controlar una resposta mesurada de forma contínua, s'acostuma a monitoritzar simultàniament el nivell i la variabilitat de la resposta. La variabilitat intrínseca del procés es pot estimar a partir de rèpliques, observacions preses en una mateixa mostra en condicions homogènies, o, quan això no és possible perquè només es pot obtenir una observació en cada mostra, s'estima a partir d'observacions consecutives en el temps.

Quan per cada punt de mostreig disposem de mostres de grandària més gran que 1, el nivell s'acostuma a representar a través de la mitjana, i el diagrama resultant s'anomena gràfic de *mitjanes* o gràfic $\overline{Y}$. Es possible controlar la variabilitat o dispersió del procés mitjançant el gràfic de desviacions tipus o mitjançant el gràfic de *rangs*, anomenat gràfic *R*. Sigui y_{ik} l'observació k-èssima en la mostra i-èssima de grandària N. La mitjana $\overline{y}$ es calcula de la forma habitual:

$$\overline{y}_i = \frac{1}{N}\sum_{k=1}^{N} y_{ik} \ ,$$

i el rang es calcula ordenant les dades de la mostra i restant la més gran de la més petita:

$$R_i = \max_k \{y_{ik}\} - \min_k \{y_{ik}\},$$

El gràfic de rangs s'utilitza de forma molt més habitual que el de desviacions tipus en ser més senzill de calcular. Els dos gràfics que s'obtenen, el de mitjanes i el de rangs, s'analitzen tant de forma independent com conjunta.

Per al gràfic de mitjanes els límits de control superior i inferior s'obtenen de l'expressió:

$$LC = \overline{\overline{y}} \pm 3s_{\overline{y}} \ , \qquad (5.1)$$

on $\bar{\bar{y}}$ indica la mitjana de les mitjanes mostrals o mitjana global, i $s_{\bar{y}}$ la desviació tipus de les mitjanes mostrals.

Gràcies al Teorema Central del Límit sabem que, si el procés està en estat de control i, per tant, si les observacions són independents i idènticament distribuides, pel gràfic de mitjanes, amb els límits de control a tres desviacions tipus de la mitjana, *p=0,0027* és la probabilitat de que un punt caigui fora dels límits. L'ARL pel gràfic de mitjanes és:

$$ARL = \frac{1}{0{,}0027} = 370\,.$$

Hi ha molts casos en els que només es pot prendre una observació per mostra, això fa que no es pugui estimar la variabilitat intrínseca, com es fa quan hi ha rèpliques. En control de qualitat això es dóna sovint quan la mesura és molt costosa, com en el cas dels assaigs destructius, quan el nivell de producció és baix, quan les mesures successives no difereixen més que l'error de mesura o d'anàlisi i quan es vol controlar els paràmetres del procés com la temperatura, la pressió o la velocitat. En aquests casos no es poden utilitzar els gràfics de *mitjanes-rangs* i cal recórrer al diagrama I d'*observacions individuals*.

En l'estudi del *Tirant* mai tindrem rèpliques, perquè per a cada tros de text analitzat s'obté una sola mesura per a cada unitat analitzada. Per tant, utilitzarem el gràfic *I* per observacions individuals sempre que tinguem variables contínues, és a dir quan treballem amb la llargada mitjana de paraula o frase o amb un índex de diversitat, tot per un tros de text donat.

Aquest gràfic consisteix en la representació de la variable a controlar en funció del temps o de la seva posició en la seqüència. La línia central és la mitjana de totes les observacions, i els límits de control, que són una mesura de la variabilitat intrínseca del procés, es calculen:

$$LC = \bar{y} \pm 3s_y\,,$$

on, a diferencia de l'expressió (5.1), ara la s_y s'obté a partir dels rangs mòbils, les diferències entre dues observacions consecutives:

$$RM_i = y_i - y_{i-1}\,.$$

5.2.2 Gràfics de Control per dades categòriques

Sovint no podem representar una característica de forma numèrica, només podem observar si posseeix un determinat atribut o si un succés ha passat o no. Per exemple, només sabrem si una peça és correcta o no ho és, si una màquina ha tingut una avaria en un interval de temps o si una paraula del text és un article o no ho és.

Els gràfics de control que monitoritzen la proporció de unitats que posseeixen un determinat atribut o els que controlen el nombre de successos ocorreguts per unitat de temps o longitud s'anomenen gràfics per atributs.

Per variables discretes els gràfics a utilitzar depenen del tipus de distribució que segueix la variable, i de si la grandària N_i de la mostra és constant o no ho és. Si la variable és el nombre, y_i de unitats que posseeixen un determinat atribut respecte al total, N_i, de la mostra, s'acostuma a aproximar la seva distribució per una binomial. Si, en canvi, és el

nombre, z_i, de successos que es donen per unitat de temps o de longitud, la distribució s'aproxima per una Poisson.

Així tindrem que:

$$y_i \sim binomial\ (N_i, \pi_i)$$

$$z_i \sim Poisson\ (\lambda_i N_i)$$

Per mostres de grandària constant $N_i=N$ i per variables binomials el gràfic que fem servir és l'anomenat *NP*, mentre que per variables Poisson usem el gràfic *C*, quan la grandària de mostra és variable emprem els gràfics *P* i *U* per variables binomials i Poisson, respectivament.

En l'estudi de l'homogeneïtat del *Tirant* usem gràfics per variables binomials quan estudiem la proporció de paraules o frases d'una determinada llargada o el nombre d'ocurrències de les 25 paraules més freqüents la freqüència, i usem gràfics per variables Poisson quan analitzem el nombre de paraules diferents o el nombre de paraules que apareixen un sol cop en un tros de text. Considerem mostres de grandària variable els capítols, mentre que els blocs de 1000 paraules són de llargada constant.

Per calcular els límits de control s'aprofita l'aproximació normal. Per variable binomial s'obtenen com:

$$LC = \bar{y} \pm 3s_y ,$$

mentre que per variable amb distribució de Poisson ho fem mitjançant:

$$LC = \bar{z} \pm 3s_Z ,$$

on $\bar{y}$ i $\bar{z}$ son les mitjanes respectives. Si es disposa d'una seqüència de *n* mostres consecutives de grandària fixa $N_i=N$. La desviació tipus per una variable binomial s'estima com:

$$s_y = \sqrt{\frac{\bar{y}}{N}\left(1 - \frac{\bar{y}}{N}\right)},$$

mentre que per dades Poisson es fa mitjançant

$$s_z = \sqrt{\bar{z}} .$$

Pel cas de variable binomial i grandària de la mostra no constant, treballem amb la proporció de unitats en la mostra que posseeixen l'atribut:

$$\hat{\pi}_i = \frac{y_i}{N_i},$$

mentre que per variables Poisson usarem:

$$u_i = \frac{z_i}{N_i},$$

que mesura el nombre de successos per unitat de temps o longitud. En ambdós casos, en fer les estimacions, a diferència de tot el que s'havia vist fins ara, com que la desviació tipus depèn de la grandària de la mostra, els límits de control no són constants, essent més amples per mostres més petites i més estrets per mostres més grans.

Com en el cas dels gràfics de mitjanes, pels gràfics per variables categòriques es situen els límits de control de forma que la probabilitat de que, estant el procés en estat de control, surti un punt fora de límits és de *p=0,0027*, i per tant l'*ARL=370*.

5.2.3 Gràfics Cusum

Un inconvenient del gràfic de mitjanes-rangs, i en general de tots els gràfics de control clàssics és que cada punt representa la informació continguda en una sola mostra, i es perd tota la informació que proporciona la seqüència de punts anteriors. Això fa que per als gràfics de control clàssics sigui difícil detectar petits canvis en la mitjana, de l'ordre de $1{,}5\sigma_y$ o inferiors.

Page (1954) va proposar els gràfics de sumes acumulades com a una alternativa al gràfic de mitjanes-rangs. Incorpora directament tota la informació de la seqüència, representant gràficament les sumes acumulades de les desviacions dels valors mostrals respecte al valor objectiu de la mitjana del procés. Si la grandària de la mostra és més gran que *1*, $\bar{y}_i$ és la mitjana de la mostra i-èssima i μ_0 el seu valor objectiu o valor nominal, els punts per al gràfic de sumes acumulades són:

$$S_m = \sum_{i=1}^{m} (\bar{y}_i - \mu_0) \ , \tag{5.2}$$

per *m=1,2,...,n*, on S_m rep el nom de Suma acumulada fins la mostra *m*. Tot i que habitualment s'han emprat els gràfics Cusum per controlar la mitjana d'un procés, també s'han desenvolupat procediments de sumes acumulades per altres variables, com ara les binomials o les de Poisson.

En el camp dels estudis sobre l'homogeneïtat d'estil podem representar en un gràfic CUSUM qualsevol tipus de variable, tant discretes com contínues. Hilton i Holmes (1993) proven els gràfics CUSUM amb les paraules curtes (dues i tres lletres), per a comprovar si dos texts poden ser atribuïts al mateix autor. En la tesi s'han fet servir per estudiar l'evolució al llarg del *Tirant* de les llargades mitjanes de paraula i frase, la llargada de capítol, l'ús de les paraules més abundants i els índexs de diversitat.

Els gràfics de sumes acumulades són més eficaços que els gràfics de mitjanes per detectar petits canvis en la mitjana, i són particularment eficaços quan la grandària de la mostra és *N=1*, quan $\bar{y}_i = y_i$. Per al gràfic CUSUM, l'ARL és de, aproximadament, 500, superior a l'ARL dels gràfics vistos anteriorment. Aquest resultat implica que els gràfics de sumes acumulades donen falsos senyals d'alarma menys freqüentment.

Quan el procés està en estat de control, amb el nivell al voltant del valor objectiu μ_0, la suma acumulada varia aleatòriament al voltant del *0*. Un desplaçament sistemàtic de la suma acumulada, S_m, cap a valors positius o negatius indica una evidència de canvi en la mitjana del procés. Existeixen procediments de decisió formals per determinar si un augment (o disminució) en la suma acumulada senyala un canvi en el procés.

Per a més detalls sobre els gràfics de sumes acumulades veure Page (1961), Johnson (1961), Johnson i Leone (1962a, 1962b, 1962c), Ewan (1963) i Lucas (1976). Pignatiello i Runger (1990) proposen dos mètodes per construir gràfics CUSUM multivariants, dels quals un és millor en termes de ARL, i que tenen bon comportament per detectar una gran varietat de canvis en la mitjana d'un procés Normal multivariant.

5.3 Gràfics de Control per dades Multivariants

En algunes situacions pot interessar controlar simultàniament dues o més característiques de la mateixa unitat. Si aquestes variables no estan relacionades entre sí, és raonable analitzar-les per separat, realitzant un gràfic de control univariant per cadascuna d'elles.

Tot i així, controlar les variables de forma independent pot donar resultats enganyosos: en monitoritzar simultàniament l variables estant el procés en estat de control, la probabilitat de que algun gràfic mostri un punt fora de límits és

$$\alpha' = 1 - (1-\alpha)^l ,$$

on α és aquesta probabilitat per a un sol gràfic. Observem com α' creix ràpidament en augmentar l, per exemple, per α=0,003 i l=5, α'=0,015.

Sovint ens trobem les l variables estan correlacionades. Si es dóna aquesta situació el càlcul d'aquesta probabilitat no és vàlid, i no existeix una manera senzilla d'obtenir-lo.

5.3.1 Gràfics de control T^2 per a dades Normals

En l'anàlisi de l'homogeneïtat d'estil en texts, fem servir el gràfic de control multivariant per variables normalment distribuïdes en l'estudi de la riquesa i diversitat del llenguatge, per a la que es disposa, per cada tros de text, de deu índexs de diversitat que, tot i que quantifiquen aspectes diferents, són complementaris i estan fortament correlacionats entre ells.

Hotteling (1947) va ser el primer en proposar un gràfic de control que monitoritzés l variables Normals correlacionades, $\boldsymbol{y_i}=(y_{1i},y_{2i},...,y_{li})$. És a dir, en cada un dels n períodes de monitorització es pren una mostra de grandària N, i per a cada una de les unitats s'observen els valors que prenen les variables $\boldsymbol{y_{ik}}=(y_{1ik}, y_{2ik}, ..., y_{lik})'$, amb $i=1,...,n$ punts mostrals en el gràfic; i $k=1,...,N$ unitats en la mostra. La hipòtesi del mètode és que les observacions y_{ik} es distribueixen segons una Normal multivariada L'estadístic a representar en el gràfic de control és:

$$T_i^2 = N(\bar{y}_i - \bar{\bar{y}})' S^{-1} (\bar{y}_i - \bar{\bar{y}}),$$

on $\bar{\boldsymbol{y}}_{\boldsymbol{i}} = (\bar{y}_{1i},...,\bar{y}_{li})$, amb:

$$\bar{y}_{ji} = \frac{1}{N}\sum_{k=1}^{N} y_{jik} \ ,$$

per $i=1,...,n,$ i per $j=1,...l$, i $\bar{\bar{\boldsymbol{y}}} = (\bar{\bar{y}}_1, \bar{\bar{y}}_2, ..., \bar{\bar{y}}_l)'$ és el vector de mitjanes històriques per al període en estat de control, i és un estimador dels valors nominals per a cada variable. $S=\{S_{jh}\}$ és la matriu de covariances mostrals de les l variables a controlar, on S_{jh} és la covariança entre la variable j i la variable h, es a dir,

$$S_{jh} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{N-1} \sum_{k=1}^{N} (y_{jik} - \bar{y}_{jk})(y_{hik} - \bar{y}_{hk}).$$

Si $j=h$, $S_{jj} = S_j^2$ és la variança mostral de la variable j-èssima.

L'estadístic T_i^2 es distribueix segons una distribució T^2 de Hotteling, amb l i n-l graus de llibertat i el límit superior de control ve determinat pels percentils d'aquesta distribució. La distribució T^2 de Hotteling està relacionada amb la F a través de la relació:

$$T^2_{\alpha,l,n-l} = \frac{l(n-1)(N-1)}{nN-n-l+1} F_{\alpha,l,nN-n-l+1}.$$

Això permet que els límits de control es puguin calcular a partir dels percentils de la distribució F. Normalment cal estimar $\bar{\bar{y}}$ i S a partir de mostres preliminars, preses quan el procés se suposa que està en estat de control.

Quan les mostres són de grandària $N=1$, no es pot estimar la variabilitat intrínseca del procés, que s'estima a partir dels rangs mòbils, tal i com es fa pel gràfic I.

5.3.2 Gràfics de Control per dades Multinomials

Quan les unitats poden classificar-se en una entre l categories, sigui $\boldsymbol{\pi}=(\pi_1,\pi_2,...,\pi_l)$ el vector que indica les probabilitats de que una unitat sigui classificada en la categoria $1,2,...,l$, amb:

$$\sum_{j=1}^{l} \pi_j = 1,$$

denotem com $\mathbf{y_i}=(y_{1i},y_{2i},....,y_{li})$ el vector que conté el nombre de observacions per cadascuna de les l categories per a la mostra i-èssima, on:

$$N_i = \sum_{j=1}^{l} y_{ji}$$

és la grandària de la mostra per aquest període, i considerem $\mathbf{y_i}$ distribuït $Mult(N_i,\boldsymbol{\pi})$.

En el cas de l'estudi d'homogeneïtat d'un text literari podem classificar la seqüència de paraules en una entre vàries categories en funció, per exemple, de la seva llargada o de la funció gramatical. Analitzant si les proporcions en cada categoria són estables o no podem descobrir l'existència de variacions en l'estil literari. En el capítol 8, per exemple, on s'han classificat les paraules en deu categories en funció de la llargada, tenim que y_{1i} denota el nombre de paraules d'una lletra en el capítol i-èssim, y_{2i} és el nombre de paraules de dues lletres, ..., y_{10+i} indica el nombre de paraules de deu lletres i més.

L'aproximació que habitualment s'ha usat per monitoritzar una seqüència multinomial ha consistit en utilitzar el gràfic de control Chi-Quadrat. El gràfic consisteix en monitoritzar l'estadístic:

$$\chi_i^2 = \sum_{j=1}^{l} \frac{(y_{ij} - N_i\pi_j)^2}{N_i\pi_j}. \qquad (5.3)$$

Quan el procés es troba en estat de control, la distribució assimptòtica de χ^2 és una Chi-Quadrat amb l-1 graus de llibertat. Per tant, el límit superior de control es pot obtenir a partir dels percentils d'aquesta la distribució. En el cas de $l=2$, χ_i^2 és el quadrat de l'estadístic monitoritzat per a la distribució binomial.

En realitat, les probabilitats π_j no seran mai conegudes. En alguns casos es pot disposar, de bones estimacions obtingudes amb dades del passat. Nosaltres, en no disposar de dades del passat, estimarem les probabilitats π_j a partir de les dades a monitoritzar.

El gràfic de control χ^2 coincideix amb una adaptació del gràfic T^2 al cas multinomial. Això és fàcil de comprovar, ja que si s'aproxima una multinomial per una Normal es té que:

$$T_i^2 = (y_i - N_i\hat{\pi})' S^{-1} (y_i - N_i\hat{\pi}),$$

on $S = \{S_{rs}\}$ és la matriu $l \times l$ de covariances entre les categories,

$$S_{rs} = \mathrm{cov}(Y_r, Y_s) = \begin{cases} N\hat{\pi}_r(1-\hat{\pi}_r) & r = s \\ -N\hat{\pi}_r\hat{\pi}_s & r \neq s. \end{cases}$$

Sota la única restricció que les probabilitats π_j siguin positives, la matriu de covariances de la distribució multinomial té rang *l-1* i, per tant, no és invertible. La inversa generalitzada més simple és:

$$S_Y^- = diag\left\{\frac{1}{N\pi_j}\right\},$$

que té rang *l*. A partir de la inversa generalitzada pot es veure que l'estadístic T_i^2 a monitoritzar queda en:

$$T_i^2 = \left((y_{i1} - N_i\hat{\pi}_1),...,(y_{il} - N_i\hat{\pi}_l)\right) \begin{pmatrix} \frac{1}{N_i\hat{\pi}_1} & & 0 \\ & \cdots & \\ 0 & & \frac{1}{N_i\hat{\pi}_l} \end{pmatrix} \begin{pmatrix} y_{i1} - N_i\hat{\pi}_1 \\ \cdots \\ y_{il} - N_i\hat{\pi}_l \end{pmatrix} = \sum_{j=1}^{l} \frac{\left(y_{ij} - N_i\hat{\pi}_j\right)^2}{N_i\hat{\pi}_j},$$

que és exactament igual a l'estadístic Chi-quadrat en el cas que el valor esperat sigui $N_i\hat{\pi}_j$. Per tant, el gràfic de control Chi-quadrat i el T^2 pel cas de distribució multinomial amb probabilitats π_j conegudes coincideixen.

5.4 Anàlisi de Correspondències

L'anàlisi de correspondències és un mètode gràfic per a l'anàlisi de dades multinomials. En la tesi el farem servir en l'anàlisi de l'ús de paraules i de la llargada de paraula i de frase.

El punt de partida són les dades organitzades en forma de taula de contingència, on les *n* files corresponen, en l'estudi estilomètric del *Tirant*, a trossos de text i les *l* columnes corresponen a les categories en que s'ha classificat la unitat estilística que analitzem. En cada cel·la de la taula hi comptem el nombre d'ocurrències, y_{ji}, en la categoria j-èssima pel tros de text, bloc o capítol, i-èssim. Observem que els vectors d'observacions $\boldsymbol{y_i}=(y_{1i},y_{2i},....,y_{li})$ estan distribuïts *Multinomial* $(N_i,\boldsymbol{\pi_i})$, on N_i és la suma dels elements de

la fila i-èssima i π_i el vector amb les probabilitats que una unitat de la fila i-essima sigui assignada a cadascuna de les *l* categories.

En la taula 5.1 hi ha un exemple de taula de contingència que s'analitzarà, posteriorment, en el capítol 8: les files són els capítols de més de 200 paraules del *Tirant*, mentre que les columnes corresponen a les *l=10* categories en que s'ha subdividit la llargada de paraula. La suma dels elements d'una fila són, en l'exemple de la taula 5.1, els totals de paraules per capítol (o llargada del capítol) mentre que la suma dels elements d'una columna és el total de paraules en el *Tirant* que tenen una llargada donada.

Llargada / Capítol	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10 +**	N_i
1	21	59	44	19	33	20	16	17	9	17	255
2	53	113	80	49	52	33	28	36	16	16	476
3	109	274	239	128	112	110	76	51	43	32	1174
4	69	150	126	71	60	71	47	32	23	21	670
5	119	207	231	123	128	102	61	55	29	34	1089
6	69	136	126	69	60	61	37	27	15	15	615
7	32	63	51	18	29	28	15	15	19	13	283
8	26	52	41	19	27	29	11	16	5	11	237
9	23	42	48	16	15	28	12	15	14	10	223
10	92	191	190	93	84	72	47	47	27	24	867
...	...	...	...	...	...	...	...	...	...	...	...
480	78	123	150	57	54	65	42	25	34	13	641
481	159	282	262	137	124	122	63	71	56	46	1322
482	50	47	61	18	32	47	23	32	14	11	335
483	158	220	207	80	120	93	65	54	62	50	1109
484	59	67	68	37	26	32	15	14	17	6	341
485	96	174	106	57	77	86	42	54	24	25	741
486	45	88	91	46	40	28	13	30	11	10	402
487	48	49	62	53	41	36	21	9	16	13	348
Total	42336	86561	80953	40271	41245	41007	21351	18219	12012	9972	393927
Perfil Promig	0,108	0,220	0,206	0,102	0,105	0,104	0,054	0,046	0,031	0,025	

Taula 5.1: Part de la distribució de llargades de paraula, mesurada en nombre de lletres, pels capítols del *Tirant*. Les files representen els capítols de més de 200 paraules, mentre que les columnes el nombre de paraules de *k* lletres, per *k=1,2,...,10+*. La última columna *(N_i)* són els totals per fila, és a dir, la llargada dels capítols, mentre que la última fila són els totals per categoria, és a dir, el nombre total de paraules de *k* lletres en el *Tirant*. La última fila s'obté dividint els totals per columna pel gran total: 393927. Aquestes dades s'analitzaran en el capítol 8.

El concepte més bàsic en l'Anàlisi de Correspondències és el de *perfil*. Un perfil fila s'obté dividint totes les entrades y_{ji} pel seu total, N_i. La taula 5.2 mostra el perfil fila pel primer capítol. Les entrades de la taula 5.2 són, per tant, la proporció d'ocurrències de paraules en la categoria j-èssima en el primer capítol, i que obtenim dividint el contingut de les cel·les de la primera fila de la taula 5.1 pel valor de la primera fila per la columna N_i. En general, les cel·les en el perfil fila són la proporció d'unitats en la categoria respecte al total de la fila. Això fa que els perfils fila siguin comparables, mentre que les files en general no ho són, en tant en quant els totals, N_i, són diferents. El perfil columna s'obté de forma anàloga, dividint el contingut de les cel·les d'una columna pel seu total.

1	2	3	4	5	6	7	8	9	10 +
0,0824	**0,2314**	**0,1725**	**0,0745**	**0,1294**	**0,0784**	**0,0627**	**0,0667**	**0,0353**	**0,0667**

Taula 5.2: Perfil fila pel primer capítol. El valor s'obté dividint el valor observat per cada categoria pel total dels valors observats per la seva fila, és a dir, per la llargada total del capítol (255 paraules pel capítol 1).

Si no hi diferències estilístiques entre els trossos de text, s'espera trobar que els perfils fila són molt semblants entre sí, i molt semblants al perfil promig, $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, ..., \hat{\pi}_l)$, obtingut dividint els totals per columna pel total, i que és la proporció de paraules en cadascuna de les categories pel conjunt de tot el llibre, i les úniques diferències seran degudes a causes aleatòries. La hipòtesi nul·la, H_0, de "no diferència" entre perfils fila també es coneguda com la hipòtesi de independència entre files i columnes, o hipòtesi de homogeneïtat. Si la H_0 fos certa, la proporció de paraules de j lletres, π_{ji}, seria la mateixa en tots els capítols del *Tirant*. Per tant, si la H_0 fos certa, el valor esperat per la cel·la *(i,j)* de la taula 5.2, $\hat{y}_{ij}$, seria el nombre total de paraules de la fila i multiplicat per la proporció promig de la columna j, $\hat{\pi}_j$:

$$\hat{y}_{ji} = N_i \frac{\sum_{i=1}^{n} y_{ji}}{\sum_{i=1}^{n}\sum_{j=1}^{l} y_{ji}} = N_i \frac{\sum_{i=1}^{n} y_{ji}}{\sum_{i=1}^{n} N_i} = N_i \hat{\pi}_j .$$

Per veure si la hipòtesi és creible, es comparen els valors observats en cadascuna de les cel·les, y_{ji}, amb els seus valors esperats suposant que la H_0 fos certa. Aquesta comparació es fa a través de la distància χ^2, que es defineix com:

$$\chi^2 = \sum_i \sum_j \frac{(Observat_{ji} - Esperat_{ji})^2}{Esperat_{ji}} = \sum_i \sum_j \frac{\left(y_{ji} - \hat{y}_{ji}\right)}{\hat{y}_{ji}}, \qquad (5.4)$$

on la suma és sobre totes les cel·les. Aquesta definició és la suma, per totes les files, de l'estadístic χ_i^2 definit a (5.3), i que havíem fet servir per als gràfics Chi-quadrat.

Com més gran sigui la contribució de la cel·la *(i,j)*, més allunyada es troba la proporció de paraules en la categoria j-èssima en el capítol i de la proporció promig de la columna j. Com més gran sigui la distància χ^2, més gran és la discrepància entre els valors observats i els esperats, més lluny estan els perfils fila observats del perfil fila promig i, per tant, menys evidència hi ha de que la hipòtesi nul·la sigui certa. La contribució relativa de cada cel·la a la distància total indica quines són les principals responsables d'aquesta discrepància.

L'anàlisi de correspondències és una representació gràfica que identifica en què són diferents les files (o columnes) d'una taula de contingència. Volem descobrir quins perfils fila (o columna) s'assemblen a quins perfils fila (o columna). El problema es que els perfils fila estan situats en un espai de dimensió l, i els perfils columna en un espai de dimensió n, on és difícil veure què queda a la vora de què. En el cas de l'exemple de la taula 5.2, $n=425$ capítols de més de 200 paraules, i $l=10$ categories per a la llargada de paraula.

L'anàlisi de correspondències identifica un subespai de dimensió dos que queda a la vora de tots els n perfils fila, i els hi projecta per veure com s'hi agrupen. Una mesura de la qualitat de la projecció ve donada pel *percentatge de inèrcia* representat per

cadascuna de les dues direccions que composen el subespai. Com més gran és el percentatge de inèrcia, millor és la projecció. Passant d'un espai de dimensió *l* a un de dimensió 2 es perd la informació de en quina direcció queden els perfils originals fora d'aquest subespai i a quina distància, de manera anàloga a com una fotografia del perfil d'un cotxe no ens diu res de la seva amplada. El resultat es representa en un gràfic asimètric per files (o columnes) o bé via el gràfic simètric.

El *gràfic asimètric per columnes* projecta totes les columnes sobre el subespai de dimensió dos i cada fila es representa el punt on es projectaria el perfil columna que tingués un 1 en aquella fila i un 0 per totes les altres files. El centre de coordenades *(0,0)* del diagrama correspon al perfil columna promig, de manera que la situació de cada perfil columna respecte al centre indica el seu grau de desviació respecte a aquest perfil promig. Les desviacions principals són les de dreta a esquerra, corresponents al que s'anomena primera direcció principal o primera component. Les distàncies entre els punts columna donen una idea aproximada de les distàncies χ^2 entre els perfils columna. Com més a la vora un perfil columna estigui d'un vèrtex fila, més gran és la proporció en aquella fila per aquella columna. Com més dispersió entre els perfils columna, més gran és la distància χ^2 entre els perfils columna i, per tant, menys homogeneïtat hi ha entre columnes. En aquests *gràfics asimètrics per columnes*, (o *per files*), les distàncies entre els vèrtex fila (vèrtex columna) no són interpretables, i tenen el problema que els punts columna (punts fila) acostumen a quedar apilotats en el centre del diagrama, i és molt difícil de veure-hi res. De forma anàloga es pot construir un *gràfic asimètric per files*.

En els *gràfics simètrics*, les distàncies entre fila i fila i les distàncies entre columna i columna aproximen les distàncies χ^2 entre files i entre columnes, però les distàncies entre files i columnes no indiquen el grau d'associació entre fila i columna i per tant no són interpretables. En canvi, punts fila i columna apareixen igualment repartits per tota l'àrea del gràfic. El que sempre val, tant pel *gràfic simètric* com pels *asimètrics*, és interpretar una dimensió cada vegada, fent servir primer les posicions relatives dels punts fila (columna) per donar un nom descriptiu a aquell eix i a continuació observant les posicions dels punts columna (fila) relatives al mateix eix. Ginebra i Cabos (1998) donen més detalls sobre aquests gràfics, Greenacre (1993) és una molt bona referència sobre l'anàlisi de correspondències.

Les figures 5.1 i 5.2 mostren les projeccions de files i columnes en el millor subespai de dimensió 2 i els gràfics asimètrics per files i per columnes, per l'Anàlisi de Correspondències fet sobre la taula de contingència de capítols per llargada de paraula.

Com que els perfils fila es troben en un espai de dimensió *l*, tindrem *l* eixos principals, dels quals n'hem considerat només els dos primers per definir aquest espai. Les coordenades dels perfils fila en termes dels eixos principals són les coordenades principals de les files. El millor subespai de dimensió *k* està definit pels *k* primes eixos principals. Si projectem els perfils fila sobre millor subespai de dimensió *k*, llavors les coordenades principals de les files també ho són per aquest subespai.

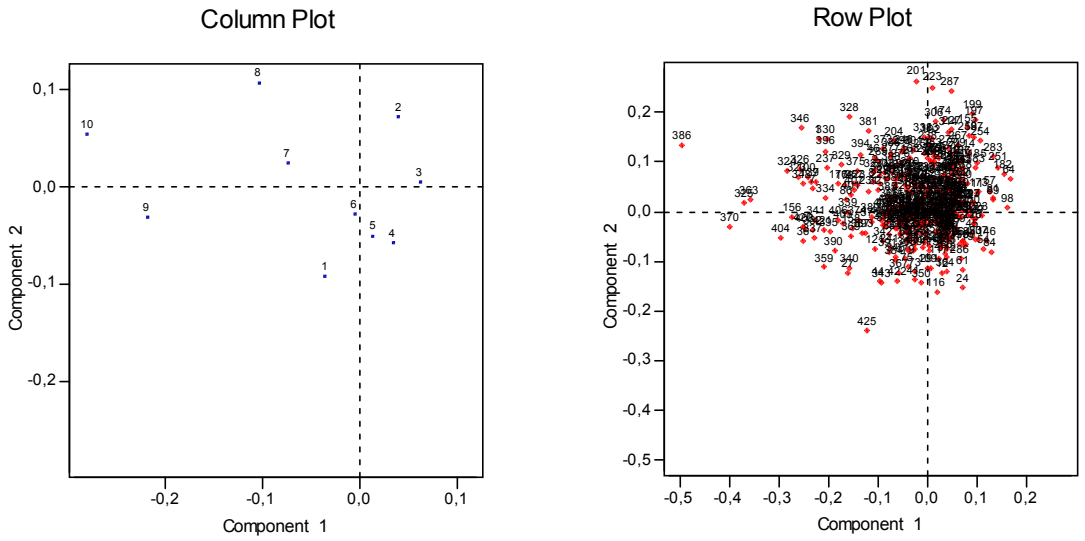

Figura 5.1: Projecció dels perfils fila i columna en el subespai de dimensió 2, per l'anàlisi de correspondències fet sobre la taula 5.1.

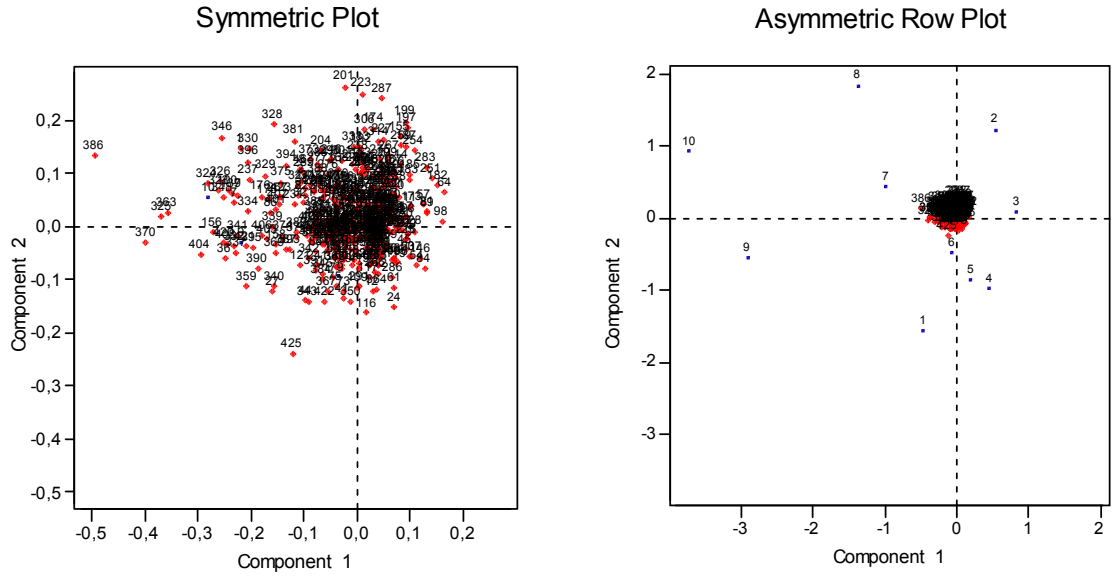

Figura 5.2: Gràfic asimètric per columnes i per files, per l'anàlisi de correspondències fet sobre la taula 5.1