CAPÍTOL 15

LÍNIES FUTURES DE RECERCA

Índex Capítol

CAPÍTOL 15

LÍNIES FUTURES DE RECERCA

En aquest capítol llistem tres vies a través de les que ens proposem continuar l'estudi de l'homogeneïtat d'estil del *Tirant*. En l'apartat 15.1 revisem algunes de les distribucions de vocabulari que s'han fet servir en la literatura estadística. En la tesi hem deixat la base de dades preparada per poder ajustar-les. El càlcul de distàncies entre distribucions ajustades ens permetrien agrupar els capítols en clusters. En l'apartat 15.2 descrivim algunes de les dificultats que se'ns presenten a l'hora de calcular distàncies entre aquestes distribucions. Finalment, en l'apartat 15.3 donem un exemple d'aplicació dels models de Markov amagats per a modelar el llenguatge com a procés estocàstic i per detectar punts de canvi.

15.1 Ajust de distribucions de vocabulari

L'ús de models matemàtics per descriure la distribució de freqüències de V_d, el nombre de paraules que apareixen exactament d vegades en el text, ha despertat l'interès d'estadístics des de fa molts anys. La variable aleatòria d, que indica el nombre de vegades que una paraula és usada, és discreta i té l'origen a $d=1$. En la taula 15.1 hi ha un tros de les distribucions de vocabulari per als capítols del 362 al 400. La primera columna senyala el número de capítol, la segona, N_i, la llargada del capítol, i les següents V_d, per $d=1,2,...,17$. Per limitacions d'espai no hem posat totes les altres columnes.

Zipf (1932) va ser el primer en trobar que existeix una relació entre el nombre d'ocurrències d i les seves freqüències V_d. Proposa la relació rank-grandària:

$$V_d \cdot d = k,$$

essent k una constant. Sigui p_d la probabilitat de que una paraula aparegui d vegades en un text i N la llargada del text, llavors la distribució de probabilitats per la d basada en aquesta relació és:

$$p_d = \frac{k}{N}\frac{1}{V_d^2},$$

que és un cas particular de la distribució zeta, o distribució de Pareto discreta. Aquesta "llei", coneguda com *llei de Zipf*, ha sigut criticada a la literatura. La crítica més coneguda és la de Herdan (1966), que argumenta que l'ajust és molt dolent en la cua de la distribució, que correspon a la freqüència de les paraules més abundant.

Yule (1944) va conjecturar que la distribució correcta per a les freqüències de paraula hauria de ser una mescla de distribucions de Poisson, però mai no va ser capaç de provar-ho. Un ajust bo és pot obtenir emprant el que en literatura estadística és coneguda com la distribució de Waring. En reconeixement al treball novedós de Herdan (1964) en aplicar aquesta distribució a la freqüència d'ús de paraules, la distribució també s'ha conegut en la literatura lingüística com el model de Waring-Herdan.

La distribució de probabilitats de Waring –Herdan ve donada per:

$$p_d = \frac{(x-a)a(a+1)...(a+d-2)}{x(x+1)(x+2)....(x+d-1)},$$

on els paràmetres x i a han de caracteritzar la llargada del text, l'abast del vocabulari emprat i la dispersió o riquesa del vocabulari. Si V és el nombre de paraules diferents, el nombre esperat de paraules que apareixen exactament d vegades és $V \cdot p_d$. Herdan va proposar com a estimadors:

$$\hat{a} = \left[\frac{V}{V-V_1} - \frac{V}{N} - 1\right]^{-1},$$

$$\hat{x} = \frac{\hat{a}V}{V-V},$$

on N és la llargada del text i V_1 és el nombre de paraules que apareixen una sola vegada.

Muller (1969) mostra com aquest model és raonablement bo per texts amb $1000<N<100.000$, tot i que Herdan va cometre un error en els seus càlculs que va ser demostrat per Dolphin (veure Muller, 1975), qui va corregir l'estimador per:

$$\hat{a} = \left[1 - \frac{V_1}{V}\right] \cdot \left[\frac{N}{V} - 1\right] \cdot \left[\frac{NV_1}{V^2} - 1\right]^{-1}.$$

L'error de Herdan va donar a Muller i Dolphin la idea d'avaluar l'abast total del lèxic d'una autor, calculant V_0, és a dir, el nombre de paraules que l'autor coneix i que no apareixen cap vegada en el text en estudi. Les lleis de Waring-Herdan i Dolphin-Muller han estat considerades com els millors models existents per a ajustar corbes de vocabulari.

Efron i Thisted (1976) empren un model empirical-bayes paramètric i un de no paramètric per determinar quantes paraules coneixia Shakespeare, és a dir, quin era V_0 de l'escriptor, a partir de la distribució de freqüències d'ús de cadascuna de les *31.534* paraules diferents que apareixen en alguna de les seves obres. Mostren que és versemblant una fita inferior per a V_0 de 35.000 noves paraules. Thisted i Efron (1987)

examinen la consistència en l'ús de paraules d'un poema recentment descobert i atribuït a Shakespeare amb el corpus bibliogràfic del mateix Shakespeare, usant un model empirical bayes no paramètric. També analitzen poemes de poetes elisabetians (Jonson, Marlowe i Donne) i quatre poemes atribuïts sense dubtes al mateix Shakespeare. Conclouen que el poema s'ajusta raonablement bé a l'ús de paraules de Shakespeare.

Sichel (1975) reprèn la idea de Yule d'una mescla de distribucions de Poisson i proposa una nova família de distribucions com a model per a les freqüències d'ús de paraules. La distribució de probabilitats ve donada per:

$$p_d = \frac{(2\alpha/\pi)^{1/2} e^{\alpha}}{\exp\{\alpha[1-(1-\theta)^{1/2}]\}-1} * \frac{(\frac{1}{2}\alpha\theta)^{r}}{d!} K_{d-1/2}(\alpha) \qquad d = 1,2,....,$$

on p_d és la probabilitat que una paraula qualsevol aparegui exactament *d* vegades en un text, $K_{d-1/2}(\alpha)$ és la funció modificada de Bessel de segon ordre, i α i θ es defineixen de manera que $\alpha > 0$ i $0 < \theta < 1$.

Amb la seva distribució va obtenir un ajust molt bo per a nombroses dades de comptatges de paraules preses de diferents texts, però no va donar una interpretació als paràmetres. Pollatschek i Radday (1981) apliquen el model de Sichel a alguns texts bíblics hebreus, i mostren com els dos paràmetres de la distribució poden ser interpretats com a riquesa de vocabulari i concentració de vocabulari, respectivament, on α representa el pendent de el cap de la distribució i θ el pendent de la cua.

Sichel va afrontar el problema, ja proposat per Yule, de si considerar només paraules d'un determinat tipus, per exemple noms o pronoms, a l'hora d'obtenir la distribució del vocabulari. Conclou que quan s'utilitzen tot tipus de paraules no ens hauríem de sorprendre de trobar anomalies, producte de la superposició de poblacions completament diferents. Pollatschek i Radday inclouen tot tipus de paraules en considerar que per l'hebreu era adequat, però noten que per l'anglès els articles determinats i indeterminats són tan predominants que la seva inclusió pot provocar distorsions en les estimacions, en ser la cua de la distribució molt llarga. Holmes (1992) usa estimacions dels paràmetres de la distribució de Sichel com a variables per a una anàlisi en components principals i cluster, en la serva recerca sobre l'autoria del llibre dels Mormons i els *Federalist Papers*.

Delcourt (1981) va proposar models log-lineals amb 2 i 3 paràmetres obtenint ajustos molt bons per a texts en francès, en els que les categories gramaticals van ser tingudes en compte, quedant encara per provar el model a altres llenguatges.

En un futur proper ens proposem abordar l'anàlisi de l'homogeneïtat d'estil del *Tirant* mitjançant l'ajust de distribucions de vocabulari: distribució de Sichel, de Waring-Herdan i de Zipf, estimant el punt de canvi en les seqüències formades per els estimadors dels paràmetres de la distribució en els capítols i en els blocs.

Cap	**N_i**	**d**																	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**...**
362	446	187	31	6	2	4	3	4	1	1	0	0	2	0	0	0	1	0	...
363	244	117	14	4	9	0	0	1	1	1	0	0	1	0	0	1	0	0	...
364	349	132	27	10	5	4	2	1	3	0	1	1	1	0	0	0	0	1	...
365	261	113	22	5	7	0	1	1	0	0	2	0	0	0	2	0	0	0	...
366	579	192	46	12	10	1	2	5	1	2	0	1	0	0	0	0	0	1	...
367	579	192	46	12	10	1	2	5	1	2	0	1	0	0	0	0	0	1	...
368	331	134	28	7	1	2	4	2	1	1	0	0	1	0	0	0	0	0	...
369	463	181	34	14	5	2	1	4	2	2	1	2	0	0	0	1	0	0	...
370	505	192	32	10	2	4	3	1	4	1	0	1	0	1	0	0	0	0	...
371	750	191	42	21	17	3	4	5	5	0	2	1	2	0	1	1	0	1	...
372	892	287	65	13	9	6	3	3	5	3	3	0	0	2	0	0	1	1	...
373	388	165	30	5	4	2	4	2	1	0	0	0	0	0	0	0	1	0	...
374	1441	442	69	37	16	8	10	3	2	2	2	4	1	0	0	1	0	1	...
375	357	164	26	3	6	1	2	1	2	1	0	1	0	0	0	1	0	0	...
376	752	260	56	14	9	4	7	1	1	0	1	1	3	2	0	1	0	0	...
377	303	144	18	8	2	2	1	2	0	1	0	0	0	1	1	0	0	0	...
378	523	198	43	11	5	1	1	1	2	1	2	2	0	0	2	2	1	0	...
379	230	114	18	4	3	0	1	1	3	1	1	0	0	0	0	0	0	0	...
380	414	189	24	9	9	2	0	2	2	1	2	0	1	0	0	0	1	1	...
381	272	134	16	6	4	3	3	0	1	1	1	0	1	0	0	0	0	0	...
382	330	125	21	11	3	3	2	1	3	3	0	0	0	0	1	0	0	0	...
383	309	115	25	8	3	3	1	2	1	1	1	0	0	0	0	0	0	0	...
384	512	152	38	13	7	4	3	0	2	2	1	0	1	1	0	1	1	0	...
385	287	116	22	7	3	2	4	1	0	1	0	0	0	0	2	0	1	0	...
386	588	184	35	12	3	5	5	3	1	5	0	1	0	1	0	0	0	2	...
387	1639	358	77	38	23	11	12	5	2	6	1	1	2	3	3	0	1	0	...
388	357	114	21	9	2	5	3	2	1	1	2	0	0	0	1	1	0	0	...
389	173	68	11	8	4	1	2	0	0	0	1	0	0	0	0	0	1	0	...
390	454	148	25	16	6	1	1	2	2	3	1	1	0	0	1	0	0	2	...
391	333	122	13	10	1	2	4	2	1	0	0	0	0	0	0	1	1	0	...
392	287	96	19	9	4	3	0	3	1	1	1	0	1	0	0	1	0	0	...
393	327	111	22	6	3	2	1	3	2	1	0	0	0	0	1	0	0	0	...
394	1059	298	55	24	11	7	2	6	8	2	0	3	0	2	1	1	1	0	...
395	257	98	15	6	5	3	3	2	0	1	0	1	0	0	0	0	0	0	...
397	356	128	21	8	6	1	3	1	1	2	1	0	1	0	1	0	0	0	...
399	203	81	16	7	2	2	1	0	1	0	2	0	0	0	0	0	0	1	...
400	362	180	17	6	7	3	3	1	1	0	0	1	1	0	0	1	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figura 15.1: tros de la distribució de vocabulari per als capítols 362 al 400. La primera columna senyala el número de capítol, la segona, N_i, la llargada del capítol, i les següents V_d, per $d=1,2,...,17$. Per limitacions d'espai no hem posat totes les altres columnes.

15.2 Agrupació de les distribucions de vocabulari

En el capítol 7 s'ha descrit com l'objectiu de l'anàlisi cluster és el d'agrupar objectes de manera que els que els que han estat col·locats un mateix grup siguin "similars" entre sí i "diferents" als dels altres grups. Els mètodes d'agrupació han de ser completament numèrics, i es basen en mesures de similaritat o de dissimilaritat calculables a partir de les dades. En els capítols anteriors hem agrupat els blocs de 1000 paraules i els capítols del *Tirant* mitjançant algorismes no jeràrquics basats en la distància euclídea, en el capítol 13 on les dades són contínues, i en la distància χ^2, en els capítols 8, 11 i 12 on les dades es troben en forma de taula de contingència.

Quan els objectes a agrupar són les distribucions de vocabulari com les de la taula 15.1 ens trobem amb el problema de trobar una mesura de dissimilaritat adequada. La distància χ^2 no serveix perquè, com observem en la taula 15.1, la majoria de les cel·les contenen un zero. La majoria de les mesures de divergència entre poblacions pateixen el mateix problema. La distància de Kublack-Leibler entre les distribució de probabilitats f_l i f_m es defineix com:

$$I_{lm} = \sum_{i=1} \log\left(\frac{f_{li}}{f_{mi}}\right) f_{li},$$

on el sumatori és sobre totes les categories i f_{ki} és la distribució de probabilitats per a la població k-èssima, per $k=1,2,...,n$. Com que no disposem de les distribucions de probabilitats teòriques, si les aproximem mitjançant les distribucions estimades:

$$\hat{f}_{ki} = \frac{y_{ki}}{\sum_{i=1} y_{ki}},$$

clarament, molts dels valors de $\hat{f}_{ki}$ seran iguals a 0 i la distància no es pot calcular.

Una possible solució consisteix en fer servir les distribucions de vocabulari llistades a l'apartat 15.1, prenent en lloc dels valors teòrics desconeguts, f_{ki}, els valors ajustats. Per exemple, si la ajustem a la població k-èssima la distribució de Waring-Herdan, els valors ajustats són:

$$\hat{f}_{kd} = \hat{p}_{kd} = \frac{(\hat{x}_k - \hat{a}_k)\hat{a}_k(\hat{a}_k + 1)...(\hat{a}_k + d - 2)}{\hat{x}_k(\hat{x}_k + 1)(\hat{x}_k + 2)....(\hat{x}_k + d - 1)},$$

on $\hat{x}_k$ i $\hat{a}_k$ s'han definit a 15.1.

Quan volem agrupar les distribucions de vocabulari pels capítols ens trobem amb el problema afegit que la distribució depèn, a més a més del capítol *i*, de la llargada N_i. Així, en el capítol 13 hem comprovat com, per exemple, el logaritme del nombre de paraules que surten una vagada en el text (els *hapax legomena*, V_1) creix linealment amb el logaritme de la llargada del text. Aquest fet ens obliga a trobar alguna manera d'estandaritzar les distribucions de vocabulari, de forma que la distància només depengui del grau de divergència entre distribucions i no les diferències entre llargades del text.

15.3 Dependència del llenguatge

Es pot modelar l'ordre d'aparició del vocabulari i d'altres unitats lingüístiques estudiant el llenguatge com a procés estocàstic. En el capítol 2 s'ha descrit breument com l'ajust de Models de Markov Amagats (HMM) permet modelar la dependència en forma de sèrie temporal discreta. En els HMM s'assumeix que cada element de la sèrie prové d'una entre *m* distribucions discretes, i on la regla per decidir quina distribució és activa en un punt donat és una cadena de Markov no observada. Normalment es trien models amb un nombre d'estats, *m*, petit, i com a distribució discreta la Poisson, la binomial o la multinomial. El mètode, però, pot ser generalitzat de forma molt senzilla per a altres distribucions. Ajustar aquests models implica estimar els paràmetres de les *m* distribucions i la matriu de transició de la cadena de Markov no observada.

Com a exemple, es pot considerar la sèrie de llargades de paraula $\{S_{ti}: t \in n_i\}$ en el capítol i-èssim, on s'assumeix que S_{ti} està generada per una entre *m* diferents distribucions de Poisson truncades, de manera que no es pugui donar el valor *0*. S'assumeix també que, en un punt *t* donat, es tria la distribució mitjançant una cadena de Markov no observada de *m* estats,

$$C^{(T)} = \{C_{ti}: t=1,...., n_i\},$$

de manera que, condicionat a $C^{(T)}$, les variables aleatòries $\{S_{ti}: t=1,...., n_i\}$ són mútuament independents i on:

$$P(S_{ti} = s \mid C_{ti} = k) = \frac{\lambda_{ki}^{s} e^{-\lambda_{ki}} / s!}{1 - e^{-\lambda_{ki}}}$$

és la probabilitat que la Poisson truncada per a l'estat k-èssim doni *s*, i on λ_{ki} és el paràmetre de la distribució per al capítol i-èssim.

Per a estudiar l'homogeneïtat d'estil, caldrà analitzar l'evolució al llarg del *Tirant* tant dels estimadors dels paràmetres λ_k, per $k=1,2,...,m$, com de les probabilitats de transició de la cadena de Markov.