ALTERNATIVE WAYS TO ESTIMATE CHANGE POINTS IN MULTINOMIAL SEQUENCES. APPLICATION TO AN AUTHORSHIP ATTRIBUTION PROBLEM

Alex Riba, Josep Ginebra
Dep. d'Estadística, Universitat Politècnica de Catalunya, Barcelona, Spain.
alex.riba@upc.es, josep.ginebra@upc.es

1 Description of the Problem.

The statistical analysis of literary style is the part of stylometry that compares measurable characteristics in a text that are rarely controlled by the author, with those in other texts. When the goal is to settle authorship questions, these characteristics should relate to the author's style and not to the genre, epoch or editor, and they should be such that their variation between authors is larger than the variation within comparable texts from the same author.

For an overview of the literature on stylometry and some of the techniques involved, see for example Mosteller and Wallace (1964, 82), Herdan (1964), Morton (1978), Holmes (1985), Oakes (1998) or Lebart, Salem and Berry (1998).

Tirant lo Blanc, a chivalry book, is the main work in catalan literature and it was hailed to be "the best book of its kind in the world" by Cervantes in Don Quixote. Considered by writters like Vargas Llosa or Damaso Alonso to be the first modern novel in Europe, it has been translated several times into Spanish, Italian and French, with modern English translations by Rosenthal (1996) and La Fontaine (1993). The main body of this book was written between 1460 and 1465, but it was not printed until 1490.

There is an intense and long lasting debate around its authorship sprouting from its first edition, where its introduction states that the whole book is the work of Martorell (1413?-1468), while at the end it is stated that the last one fourth of the book is by Galba (?-1490), after the death of Martorell. Some of the authors that support the theory of single authorship are Riquer (1990), Chiner (1993) and Badia (1993), while some of those supporting the double authorship are Riquer (1947), Coromines (1956) and Ferrando (1995). For an overview of this debate, see Riquer (1990).

Neither of the two candidate authors left any text comparable to the one under study, and therefore discriminant analysis can not be used to help classify chapters by author. By using sample texts encompassing about ten percent of the book, and looking at word length and at the use of 44 conjunctions, prepositions and articles, Ginebra and Cabos (1998) detect heterogeneities that might indicate the existence of two authors. By analyzing the diversity of the vocabulary, Riba and Ginebra (2000) estimates that stylistic boundary to be near chapter 383.

Following the lead of the extensive literature, this paper looks into *word length*, the use of the most *frequent words* and into the use of *vowels* in each chapter of the book. Given that the features selected are categorical, that leads to three contingency tables of ordered rows and therefore to three sequences of multinomial observations.

Section 2 explores these sequences graphically, observing a clear shift in their distribution. Section 3 describes the problem of the estimation of a suden change-point in those sequences, in the following sections we propose various ways to estimate change-points in multinomial sequences; the method in section 4 involves fitting models for polytomous data, the one in Section 5 fits gamma models onto the sequence of Chi-square distances between each row profiles and the average profile, the one in Section 6 fits models onto the sequence of values taken by the first component of the correspondence analysis as well as onto sequences of other summary measures like the average word length. In Section 7 we fit models onto the marginal binomial sequences to identify the features that distinguish the chapters before and after that boundary. Most methods rely heavily on the use of generalized linear models.

	1	2	3	4	5	6	7	8	9	10+	N_i	$\overline{WL}$	χ^2
Ch.1	21	59	44	19	33	20	16	17	9	17	285	4.47	28.08
2	53	113	80	49	52	33	28	36	16	16	476	4.14	20.13
3	109	274	239	128	112	110	76	51	43	32	1174	4.06	10.30
4	69	150	126	71	60	71	47	32	23	21	670	4.14	7.21
5	119	207	231	123	128	102	61	55	29	34	1089	4.09	11.23
6	69	136	126	69	60	61	37	27	15	15	615	3.96	2.42
7	32	63	51	18	29	28	15	15	19	13	283	4.34	22.69
8	26	52	41	19	27	29	11	16	5	11	237	4.25	10.75
9	23	42	48	16	15	28	12	15	14	10	223	4.48	20.25
10	92	191	190	93	84	72	47	47	27	24	867	4.00	6.61
...	...	...	...	...	...	...	...	...	...	...	...	...	...
480	78	123	150	57	54	65	42	25	34	13	641	4.05	23.29
481	159	282	262	137	124	122	63	71	56	46	1322	4.08	19.34
482	50	47	61	18	32	47	23	32	14	11	335	4.50	49.18
483	158	220	207	80	120	93	65	54	62	50	1109	4.21	72.33
484	59	67	68	37	26	32	15	14	17	6	341	3.82	23.50
485	96	174	106	57	77	86	42	54	24	25	741	4.18	37.46
486	45	88	91	46	40	28	13	30	11	10	402	3.94	16.88
487	48	49	62	53	41	36	21	9	16	13	348	4.20	31.34

Table 1: Part of the 425×10 table of counts of words of each length in each chapter. N_i is the total number of words per chapter, $\overline{WL}$ is the average word length per chapter and the last column, χ^2, gives the contributions of each row to the Chi-squared statistic associated to the table. The Chi-squared statistic to test for independence is $8408,3$

2 Description of the data.

For our study we use the modern edition of *Tirant lo Blanc* by Riquer (1983), excluding from consideration the titles of the chapters and words in italics, that correspond to quotations in latin, and we restrict consideration to the 425 chapters with more than 200 words among the 487 of very unequal lengths.

2.1 Word Length.

We classify words according to their number of letters, with a category for all the words of more than nine letters, and build the corresponding 425×10 contingency table of ordered rows, partially presented in Table 1.

Mendenhall(1887) already used the length of words to discriminate between the writings of Shakespeare, Bacon and Marlowe and Brinegar (1963) used it to argue that Mark Twain did not write the *Quintus Curtius Snodgrass Letters*. Mosteller and Wallace (1964, 84) used it in their study of the *Federalist Papers*. Other authors using the number of letters per word to characterize style are Hilton and Holmes (1993), Williams (1975) and Smith (1983).

If all the book was written by the same author at about the same time, it would be reasonable to expect that all the rows in Table 1 come from a single multinomial distribution. On the other hand, if we could determine that the row profiles suddenly change at a given row, that might explain the existence of a second author that took over in that chapter and completed the book. Thus, the first goal is to determine whether there is a chapter where the distribution of the rows change and what changes in them.

To explore the evolution of these row profiles along the book, Figure 1 presents the sequence of proportions of words of two, three, nine and of more than nine letters, $p_1, \ldots, p_{9+}$. In all these sequences, there is a clear shift in level with words before that shift tending to be shorter than the

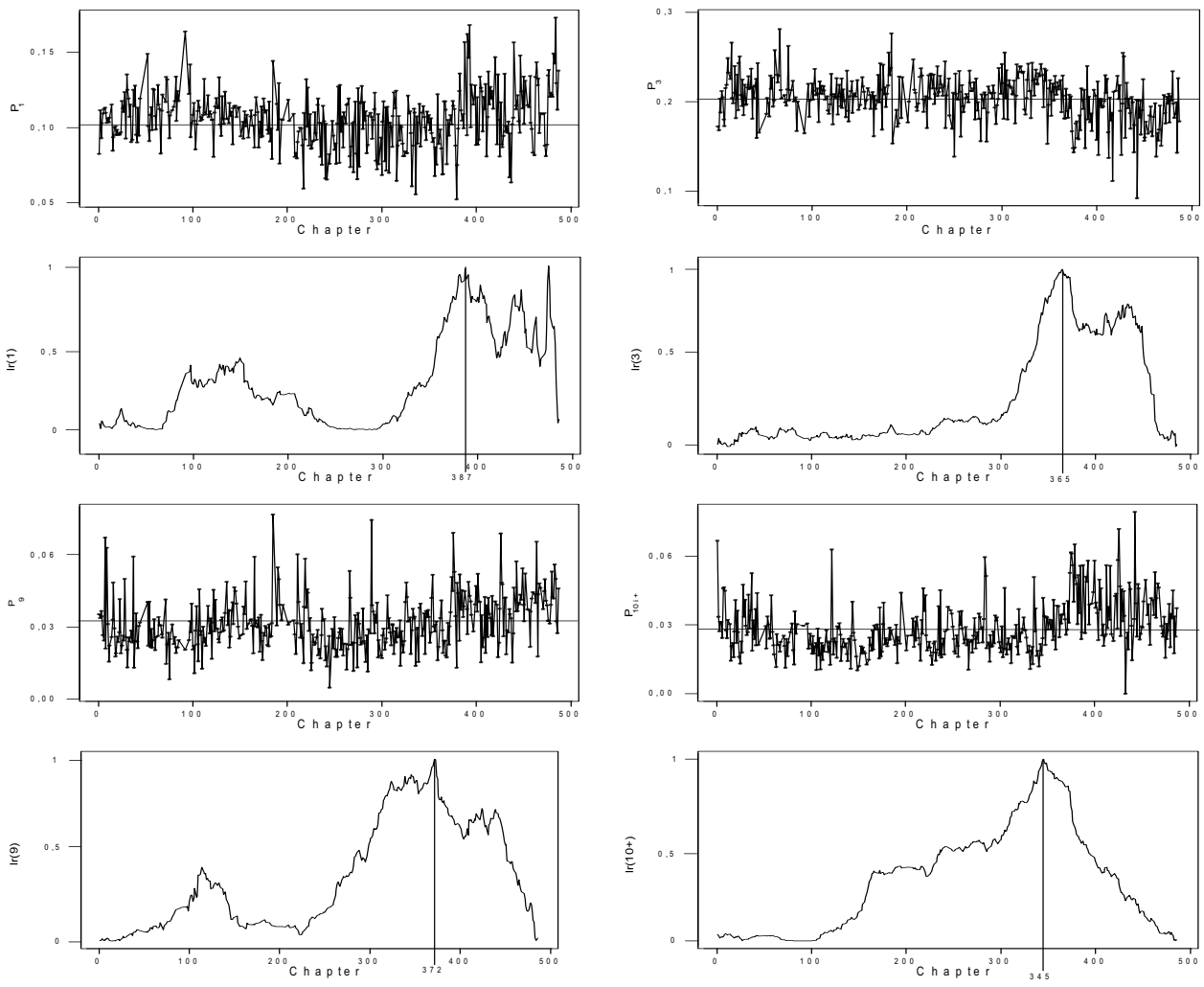

Figure 1: Sequence of the proportion of words of one, two, three, eight, nine and of more than nine letters per chapter, and the values of l_r as a function of r for the three marginal binomial sequences.

words after it.

2.2 Use of the most frequent context free words.

The frequency with which certain context free words are used tends to be rather stable within texts of the same author. In the stylometry literature, these words used to characterize the style of an author and to discriminate it from the style of other authors are often called *function words*. Function words often include articles, pronouns and words chosen among the most frequent ones, as well as conjunctions and prepositions. Some of the many authors that use function words in authorship attribution problems are Ellegard (1962), Mosteller and Wallace, (1964, 82), Morton (1978), Burrows (1987, 92), Burrows and Hassall (1988), Binongo (1994), Lebart, (1994), Ginebra and Cabos (1998) and Peng and Hengartner (2002).

Here, we consider the use of the twenty five most frequent context-free words. The count of the number of appearances of each one of these words in each one or the chapters forms a 425×25

	e	de	la	que	lo	en	a	per	no	...	molt	si	dix
Ch.1	12	15	9	8	10	6	1	4	1	...	1	3	0
2	26	28	19	9	10	12	11	8	3	...	8	3	1
3	66	46	48	53	26	20	22	20	19	...	2	2	4
4	33	29	34	13	9	21	13	11	5	...	8	3	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...
484	31	19	13	12	10	7	15	3	2	...	2	1	2
485	59	66	28	14	12	21	7	8	2	...	6	0	0
486	28	29	14	10	14	13	4	14	1	...	10	0	0
487	29	13	8	10	8	4	4	4	2	...	9	0	0

Table 2: Part of the 425×25 table of counts of each one of the 25 most frequent words in each chapter. The value of the Chi-squared statistic to test for independence is $20972,4$

	a	e	i	o	u
Ch.1	125	191	76	62	41
2	258	269	135	124	74
3	541	707	251	384	200
4	388	363	161	184	112
...	...	...	...	...	...
484	157	208	77	93	40
485	363	542	118	178	103
486	166	274	71	104	48
487	153	212	85	109	60

Table 3: Part of the 425×5 table with the count of vowels in each chapter. The value of the Chi-squared statistic to test for independence is $3786,4$

contingency table of ordered rows partially presented in Table 2. Figure (2) presents the evolution of the proportion of appearances of six of these 25 most frequent words in each chapter, that include the three most frequent words, *e, de, la*, as well as *molt*, *no* and *si*. For graphics presenting the evolution of all 25 words, see Riba (2002).

Finally, we also consider the 425×5 contingency table of counts of each vowel in each chapter, partially presented in Table 3. Although vowels is rarely used in authorship attribution problems, to our surprise we found that this sequence

also shows a slight shift near where the rows of Tables 1 and 2 shift.

3 Change-Point Estimation in a Multinomial Sequence.

Let $y_1, y_2, \ldots, y_n$ be an ordered sequence of mutually independent random variables, with distribution function $F_{\theta_0}(y)$ for all y_i with $i = 1, \ldots, r$, and distribution function $F_{\theta_1}(y)$ for all y_i with $i = r+1, \ldots, n$, where θ_0, θ_1 and r are unknown. Therefore r designates a possible known change point. Testing for the existence of one change-point and the estimation of r has been extensively studied for various univariate distributions. For an approach to the problem based on likelihood, see for example Hinkley (1970, 71), for a non-parametric approach see Bhattacharyia and Johnson (1968) while the Bayesian framework is exposed in Smith (1975, 81), Smith and Cook (1980) and Ferreira (1975). For general reviews see Zacks (1983) and Pettitt (1989). In this literature authors tend to distinguish between the basic problem just described and switching regression models, but we rather treat both problems in the same framework.

Extensions to the estimation of change points in sequences of multivariate distributions exist, but they most often focus on continuous distributions. In our case, for each chapter i with a total of N_i words, one observes a vector valued $y_i = (y_{i1}, y_{i2}, \ldots, y_{il})$, distributed as a $Multinomial(N_i, \pi_i)$,

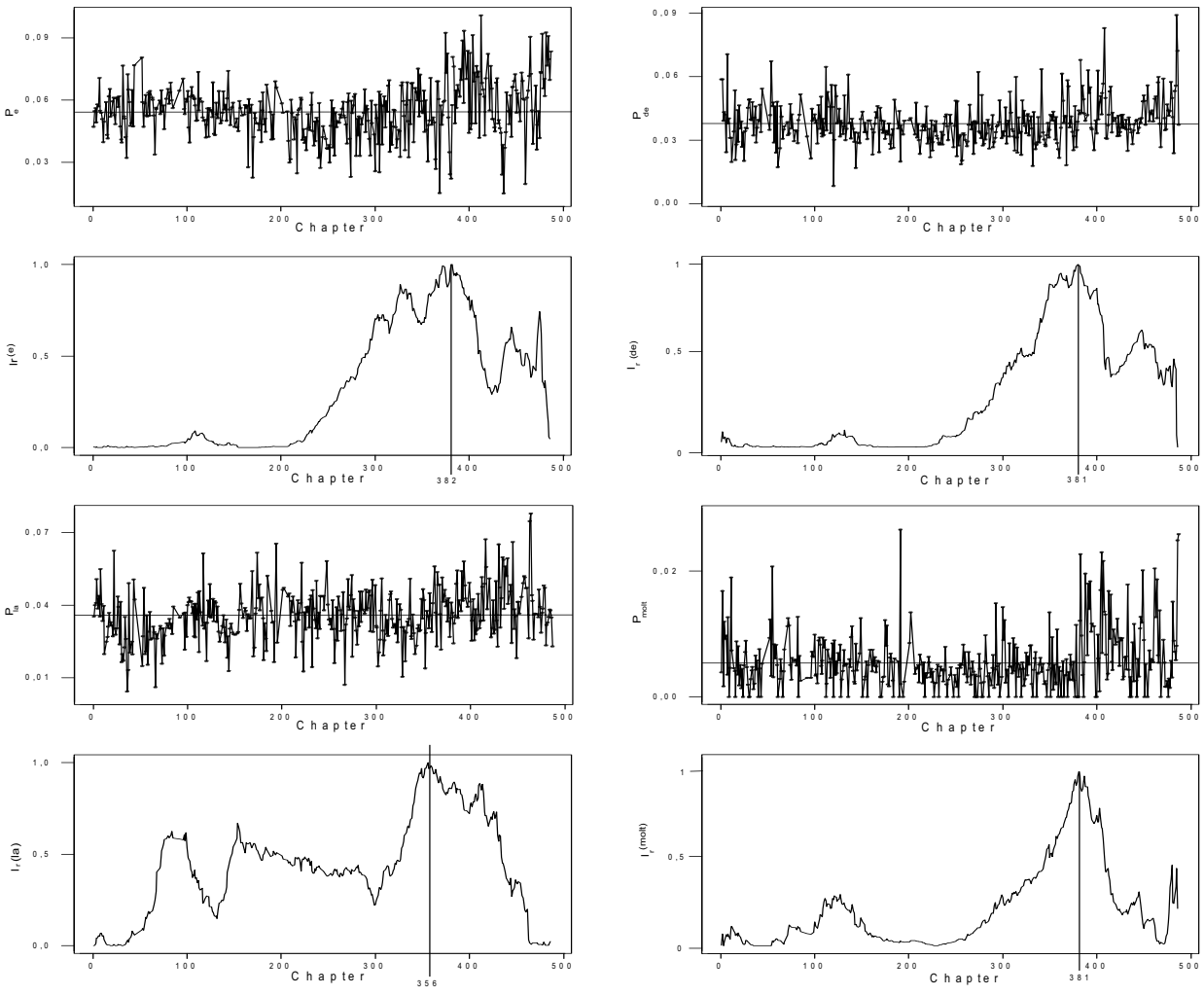

Figure 2: Sequences of the proportion of appearances of the words *e, de, la, molt, no* and *si* in each chapter and the values of l_r as a function of r for the same binomial sequences obtained by marginalizing Table 2.

where $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{il}))$ with l being the number of categories (columns of the contingency table) and with π_{ij} denoting the probability of the j-th category. If all the chapters belong to the same author it is reasonable to expect that π_i stays constant along the whole sequence of 425 chapters. On the the other hand, if there was a change of author in chapter r, one might detect a shift in π_i at $i = r$. Wolfe and Chen (1990) estimate the change point in a multinomial sequence by combining the l solutions to the l change-point problems associated to the l marginal binomial sequences.

Assuming that change to be a sudden one, we propose various ways to estimate change-points in multinomial sequences. The different methods involve the whole multinomial sequence, sequences of Chi-square distances between each row and he average row profile, sequences of summary measures like the average word length or the sequences of values taken by the first component of the correspondence analysis, and the marginal binomial sequences. We find that for most sequences, there is a clear change-point between chapters 371 and 382.

4 Estimation based on models for polytomous data

Our first approach to this basic multinomial change-point estimation problem poses it in terms of switching "regression" models for polytomous data. In particular, we fit generalized linear models for multinomial data that allow π_i to suddenly shift values at $i = r$, and then we estimate r to be the shift point for which the corresponding model fits best the sequence. In order to do that, we fit the extension of the logistic model to polytomous data, described in McCullagh and Nelder (1983) and Hosmer and Lemeshow (1989). By considering category 1 to be the baseline, that model assumes that:

$$log\frac{\pi_{ij}}{\pi_{i1}} = \beta_{r_{j0}} + \beta_{r_{j1}} Ir_i, \tag{1}$$

for $j = 2, \ldots, l$, where Ir_i is an indicator variable such that $Ir_i = 0$ for $i = 1, \ldots, r$ and $Ir_i = 1$ otherwise. Using MINITAB, we estimate $\beta_r = (\beta_{r_{20}}, \ldots, \beta_{r_{l0}}, \beta_{r_{21}}, \ldots, \beta_{r_{l1}})$ for each r in $\{1, \ldots, 424\}$, and record the corresponding maximum likelihood of the model, l_r. The value of r of the model with the largest maximum likelihood,

$$\hat{r} = min\{k : l_k = max_{1 \le r \le 424} l_r\}, \tag{2}$$

will be the maximum likelihood estimate of r.

Figure 3 presents the maximum likelihood values for these 424 models as a function of r for the three multinomial sequences associated to Tables 1, 2 and 3. For the data on word lengths from Table 1, one obtains a global maximum at chapter 371, and a local maximum at chapter 345, with a value of l_{345} that is very close to the global maximum, l_{371}. For Table 2 on the use of function words there is a sharp global maximum at chapter 382, while for the data from Table 3 on the use of vowels, the global maximum is in chapter 371, even though it is a lot less sharp than for Table 1.

5 Estimation based on the Chi-Square sequence

One can also monitor the multinomial sequence $y_1, \ldots, y_n$, with $y_i = (y_{i1}, \ldots, y_{il})$, considered in Section 4 through the sequence of the contribution of each multinomial observation in that sequence to the Chi-square statistic associated to the contingency table,

$$\chi_i^2 = \sum_{j=1}^{l} \frac{(y_{ij} - \hat{y}_{ij})^2}{\hat{y}_{ij}},$$

for $i = 1, \ldots, n$. Figure 4 presents the sequences of such statistic for Tables 1, 2 and 3.

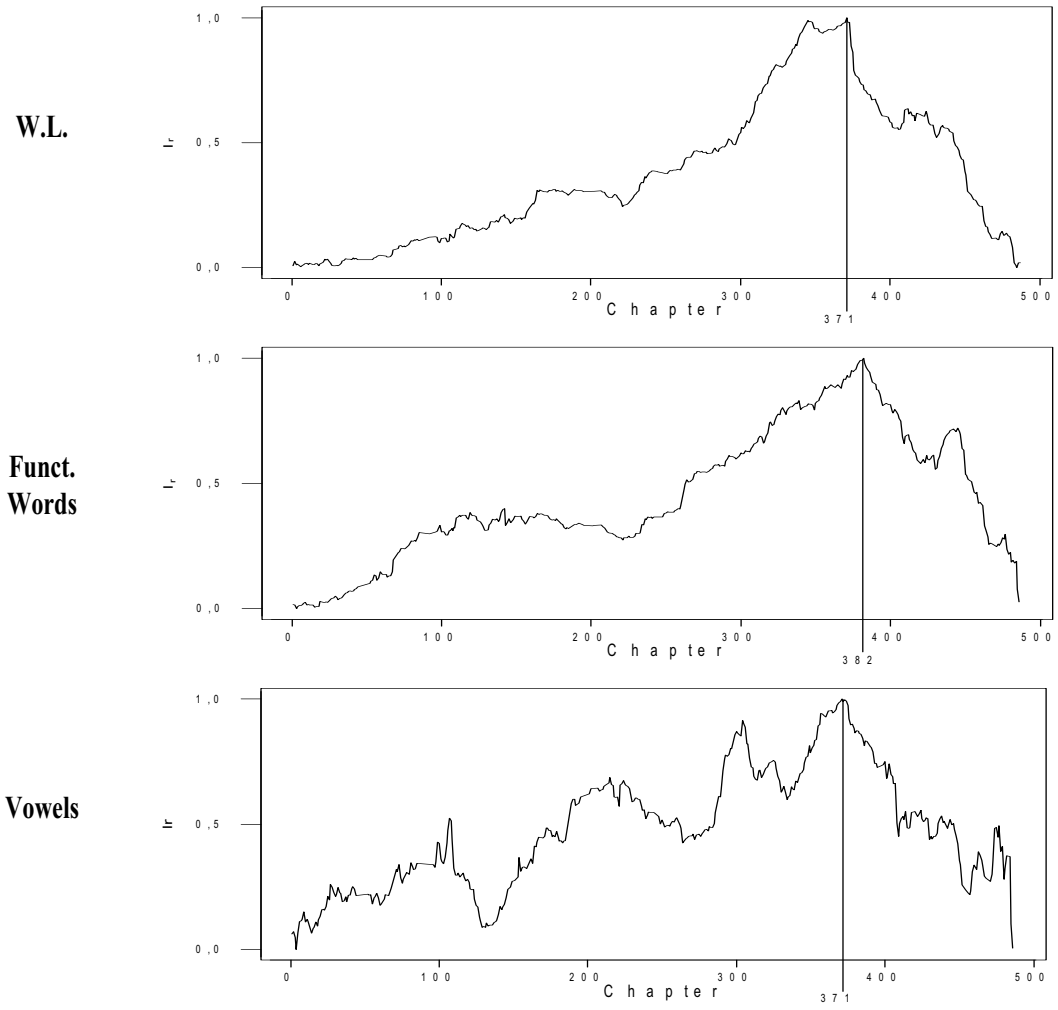

Figure 3: Maximum likelihood values, l_r, for the multinomial model for each r, as a function of r for the data from Table 1 (top), 2 (middle) and 3 (bottom).

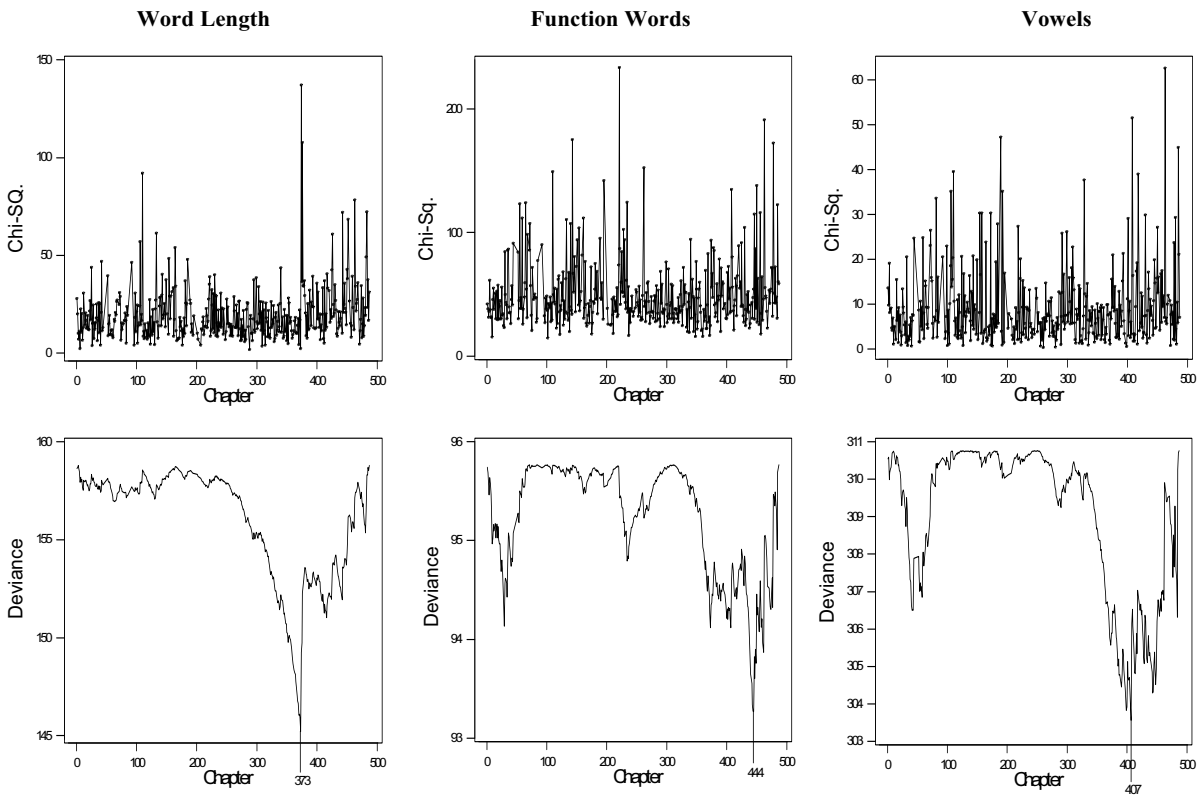

Figure 4: Sequences of the contributions of each observation to the Chi-squared statistic associated to the contingency tables 1, 2 and 3 and the values of l_r as a function of r for the three χ^2 sequences.

By using the same basic idea behind the method in Section 4, now we propose fitting the next gamma model:

$$\chi_i^2 \sim Gamma(\alpha, \beta_i = \frac{1}{\alpha g^{-1}(\beta_{r_0} + \beta_{r_1} Ir_i)}), \tag{3}$$

on this sequence of Chi-square statistics, where again $Ir_i = 0$ for $i = 1, \ldots, r$ and $Ir_i = 1$ otherwise.

This model is fitted for $r = 1, \ldots, 424$, and again r is estimated to be the shift point for which the corresponding model fits best the sequence. This time we use as a goodness of fit statistic the deviance of the fitted model, because it also leads to $\hat{r}$ being the maximum likelihood estimate of the change point. Figure 4 presents the values of the deviance of these models as a function of r for the χ^2 sequences obtained from Tables 1, 2 and 3. We observe that the estimation of the change point although being close, does not completely coincide with the estimation obtained in section 4. note that here one is loosing information by summarizing the whole multinomial observation in a single statistic.

6 Estimation based of summary measures that are approximately normal

It is possible to summarize each row in Tables 1 to 3 in a single summary measure that is approximately normally distributed.

The rows of Table 1 can be summarized through the average word length, $\overline{WL}$. This is possible because the column categories in Table 1 are quantitative. One can not find a similar statistic associated to rows of Tables 2 and 3 because there column categories are not even ordered. Although word length is discrete, the central limit theorem guarantees that their averages are approximately normal. Figure (5) presents the sequence of values of the average word length.

Correspondence analysis, described for example in Greenacre (1993), is an exploratory data analysis tool similar to principal components analysis, that applies to multinomial observations. By simultaneously projecting rows and columns down to a two dimensional space defined by the first two components, one is able to better understand what is it that makes row profiles before the change different from the row profiles after the change.

All the components in correspondence analysis are weighted averages of the l columns, and thus they are also approximately normally distributed. Figure (5) presents the sequence of values of the first principal components for Tables 1, 2 and 3. The shift observed around chapters 371-382 abounds on the relationship between the first components and the stylistic boundary detected in previous Sections.

We use the same model based approach to the estimation of a change-point on these sequences of approximately normally distributed observations by fitting the weighted normal regression model,

$$y_i \sim N(\mu_i = \beta_{r_0} + \beta_{r_1} Ir_i, \sigma_i^2), \tag{4}$$

for $r = 1, \ldots, 424$, where σ_i^2 is proportional to $1/N_i$. r is estimated to be the shift point for which the corresponding model fits best the sequence, using as a goodness of fit statistic the F-statistic from the ANOVA table. Figure (5) presents the values of F_r as a function of r for the three sequences of first components from Tables 1, 2 and 3 and for the average word length.

7 Change point in the marginal Binomial sequences

This approach to the estimation of a change-point can also be used on l marginal binomial sequences, by fitting the the l simple logistic models,

$$y_{ij} \sim Binomial(N_i, \pi_{ij} = \frac{e^{\beta_{r_0} + \beta_{r_1} Ir_i}}{1 + e^{\beta_{r_0} + \beta_{r_1} Ir_i}}), \tag{5}$$

for $r = 1, \ldots, 424$, for each column of the Table, and estimating r to be the one that best fits the corresponding sequence.

In this way, for each table one obtains l estimates for the change point, one for each marginal binomial sequence. These estimates can be combined to obtain a simple estimate for the change point of the whole multinomial sequence. for that, one could use the ideas proposed in Wolfe and Chen (1990) in a slightly different context.

Instead, we use the marginal binomial sequences to identify which characteristics discriminate between styles. Those for which the estimation of the change point agrees with the one obtained for the whole multinomial sequence are the ones that better explain the shift.

Figure (1) presents the values of l_r as a function of r for the binomial sequences corresponding to the use of words of length $1, 2, 9$ and 10 or more in Table 1. We observe how for all those sequences, the change-point estimate comes very close to chapter 371, the estimate obtained for the multinomial sequence. Long words and those words of 1 letter are more abundant at the end of the book, while those of length 3 are more abundant before chapter 371.

Figure (2) presents the values of l_r, the maximum value of the likelihood, as a function of r for the binomial sequences obtained by marginalizing Table 2 on columns for *e, de, la* and *molt*. Note that for all those sequences, the change-point estimate comes very close to chapter 382, the estimate

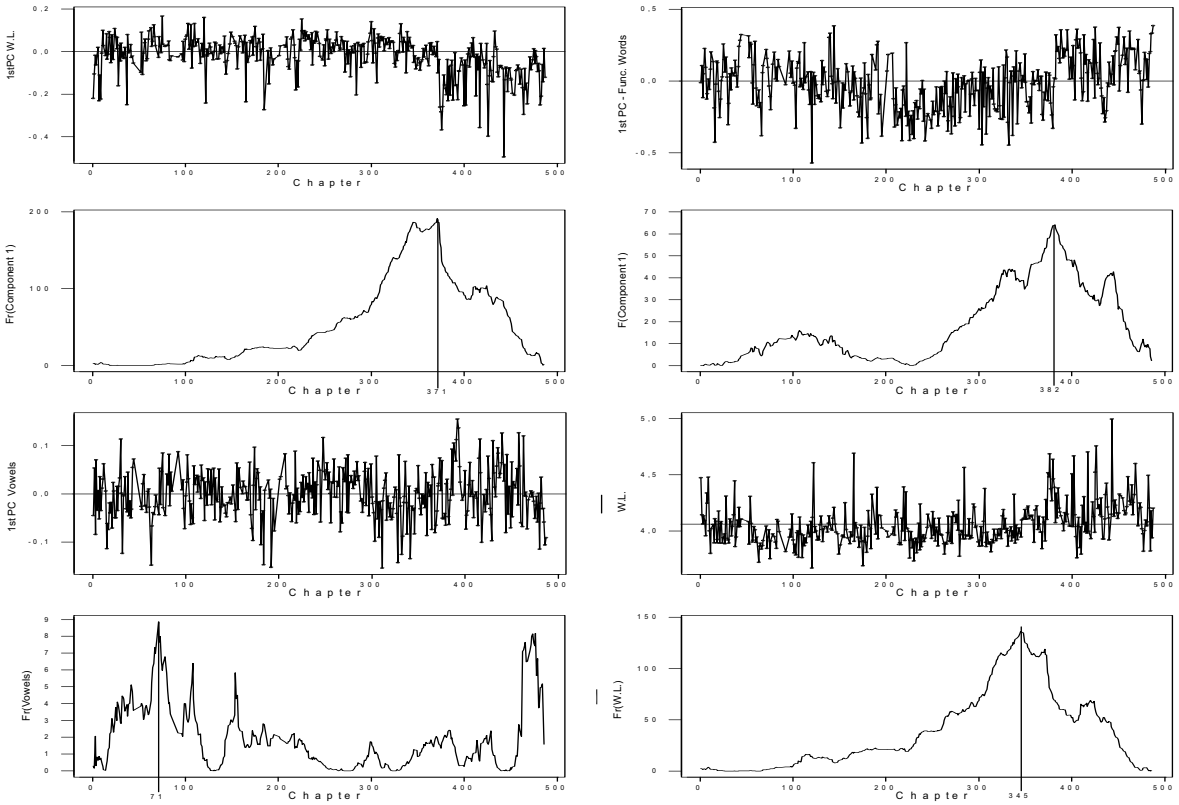

Figure 5: Sequences of values of the first principal components for Tables 1, 2 and 3, the sequence of values of the average word length and the values of F_r as a function of r for the approximately normal sequences. The percentage of the total inertia captured the first components are 27, 19 and 36 respectively.

obtained for the multinomial sequence. It turns that these four words are more abundant at the end of the book. Riba (2002) repeats this analysis for each of the twenty five columns in Table 2 and finds that the use of words like *si, no, com, és, jo*, and *dix* also present a rather sharp boundary around chapter 382, and all of them are significantly less abundant after that boundary than before it.

8 Conclusion and Extensions

Our approach to the estimation of the change point in a multinomial sequence by plotting goodness of fit statistics is very simple to implement, tackles the comparison for chapters of very different sample sizes and it easily extends to the estimation of more than one change point and for sequences of observations from any exponential family distribution.

Note that Figures (1), (2), (4) and (5) seem to indicate that a few of the chapters appearing after the change-point might be miss-classified by that boundary. This could reveal that the second author might have finished the book by filling in chapters at the end of the book. That would be more so if there was an agreement in the chapters that are misclassified by the variables associated to the two contingency tables under consideration.

In order to explore that possibility, Riba (2002) does a cluster analysis of the rows o Tables 1 and 2, using one non-hierarchical algorithm based on the repeated fit of model (1), with the only novelty that now the dummy variable indicates the cluster assigned to each observation and not its position relative to a single boundary point. That analysis indicates that chapters 403, 411, 412, 426 to 429, 431 to 439, 460 and 472 to 475 are very much like the chapters previous to the change-point estimated in Section 3. That is in very close agreement with the results obtained in Riba (2002) through the analysis of the diversity of the vocabulary used in each chapter.

9 Bibliography

Badia, L. (1993) El Tirant en la tardor medieval catalana. In *Actes del Sismposion Tirant lo Blanc*. Barcelona, 35-99

Bhattacharyia, G.K. and Johnson, R.A. (1968) Nonparametric tests for shift at an unknown time point, *Ann. Math. Statist.*, **39**, 1731-1743.

Binongo, J.N.G. (1994) Joaquin's Joaquinesquerie, Joaquinesquerie's Joaquin: a statistical expression of a Filipino writer's style. *Lit. and Linguistic Comput.* **9**, 267-279

Brinegar, C.S. (1963) Mark Twain and the *Quintus Curtius Snodgrass* letters: a statistical test of authorship. *J. Amer. Statist. Ass.*, **58**: 85-96

Burrows, J.F. and Hassall, A.J. (1988) Anne Boleyn and the authenticity of Fielding's feminine narratives. *Eighteenth Century Stud.* **21**: 427-453

Burrows, J.F. (1987) Word patterns and story shapes: the statistical analysis of narrative style. *Lit. and Linguistic Comput.* **2**: 61-70

Burrows, J.F. (1992) Not unless you ask nicely: the interpretative nexus between analysis and information. *Lit. and Linguistic Comput.* **7**: 91-109

Chiner, J.(1993) *El viure novellesc. Biografia de Joanot Martorell.* Ed Marfil, Collecci Universitas.

Coromines, J. (1956) Sobre l'estil i manera de Mart Joan de Galba i els de Joanot Martorell. in *Lleures i converses d'un filleg.* ps . Club Editor. Barcelona, 363-378

Ellegard, A. (1962) *A Statistical Method for Determining Authorship: The Junius Letters, 1769-1772*. University of Gothenburg, Gothenburg

Ferrando, A. (1995) Del Tiran de 1460-1464 al Tirant de 1490. in *Paredes, J.; Nogueras, V.; Snchez, L. Editores. Estudios sobre el Tirant lo Blanc*. Granada. Universidad de Granada. 75-109.

Ferreira, P.E. (1975) A Bayesian analysis of a switching regression model. *J. Amer. Statist. Assoc.*, **70**, 370-74.

Ginebra, J. and Cabos, S. (1998). "Anàlisi estad'istica de l'estil literari; Aproximació a l'autoria del Tirant lo Blanc". *Afers* **29**:185-206.

Greenacre, M.J.(1993) *Correspondence analysis in practice*, Academic Press, New York

Herdan, G. (1964) *Quantitative Linguistics*. Butterworths, London.

Hilton, M.L. and Holmes, D.I. (1993) An Assesment of Cumulative Sum Charts for Authorship Attribution. *Literary and Linguistic Computing*, **8**, 73-80

Hinkley, D.V. (1970) Inference about a change-point in a sequence of random variables, *Biometrika*, **57**, 1-16.

Hinkley, D.V. (1971) Inference in two phase regression, *J. Amer. Statist. Assoc.*, **66**, 736-743.

Holmes, D. (1985). The Analysis of Literary Style. A Review. *J.R.S.S.(A)* **148**: 328-341.

Hosmer, D.W., and Lemeshow, S. (1989).*Applied Logistic Regression* John Wiley and Sons, New york.

La Fontaine, M. (Translator) (1993). *Tirant lo Blanc*, Peter Lang, New York.

Lebart, L. (1994) *Statistique Textuelle*. Dunod, Paris.

Lebart, L., Salem, A. and Berry, L. (1998) *Exploring Textual Data*. Kluwer, Dordrecht.

McCullagh, P., and Nelder, J.A. (1983).*Generalized Linear Models* Chapman Hall, London.

Mendenhall, T.C (1887) The characteristic curves of composition. *Science*, **IX**: 237-249

Morton, A.Q. (1978) *Literary Detection*. Scribners, New York.

Mosteller, F. and Wallace, D.L. (1964) *Inference and Disputed Authorship: the "Federalists"*. Reading, Mass.: Addison-Wesley

Mosteller, F. and Wallace, D.L. (1984) *Applied Bayesian and Classical Inference; the Case of The Federalist Papers*, 2nd Edition. Springer-Verlag, Berlin.

Oakes, M.P. (1998) *Statistics for Corpus Linguistics*. Edimburgh University Press, Edimburgh.

Peng, R. D., and Hengartner, N.W. (2002), Quantitative analysis of literary style, *Am. Stat.*, **56**: 175-185

Pettitt, A. N. (1989) Change-point problem, in *Encyclopedia of Statistical Sciences*, **Vol. S** 26-31

Riba, A. (2002) Homogeneıtat d'estil en el Tirant lo Blanc (in catalan). Unpublished phD thesis, Universitat Politècnica de Catalunya.

Riba, A. and Ginebra, J. (2000). "Riquesa de vocabulari i homogeneitat d'estil en el Tirant lo Blanc". *Revista de Catalunya* .

Riquer, M. de Ed. (1947) J. Martorell i M.J. Galba, *Tirant lo Blanc*, Ed. Selecta, Barcelona.

Riquer, M. (Editor) (1983). *Tirant lo Blanc*, Millors Obres de la Literatura Catalana, volums 99-100, Edicions 62.

Riquer, M. (1990). *Aproximació al Tirant lo Blanc*. Quaderns Crema.

Rosenthal, D. H.(Translator)(1996). *Tirant lo Blanc*, Johns Hopkins University Press.

Smith, A.F.M. (1975) A Bayesian approach to inference about a change point in a sequence of random variables. *Biometrika*, **63**, 407-416

Smith, A.F.M. (1980). Change-point problems: approaches and applications. In *Proceedings*

Smith, M.W.A. (1983) Recent experience and new developments of methods for the determination of authorship. *Ass. For Lit. And Linguist. Bull.*, **11**: 73-82

Smith, A.F.M. and Cook, D.G. (1980) Switching straight lines: A Bayesian analysis of some renal transplant data. *Appl. Statist.*, **29**, 180-189.

Williams, C. B. (1975), Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon, *Biometrika*, **62**: 207-212

Wolfe, D.A. and Chen Y.S. (1990). "The Changepoint Problem in a Multinomial Sequence". *Comm. in Statist., Comput. and Simul.* **19(2)**:603-618.

Zacks, S. (1983) Survey of classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures of testing and estimation. In *Recent Advances in Statistics*, (ed.), pp. 245-269.