Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style

Javier GIRÓN, Josep GINEBRA, and Alex RIBA

To help settle the debate around the authorship of *Tirant lo Blanc*, all words in each chapter are categorized according to their length, and the appearances of certain words are counted, thus forming two contingency tables of ordered rows. A Bayesian multinomial change-point analysis of the sequence of rows, reveals a clear stylistic boundary, estimated to be near chapters 371 and 382. A Bayesian cluster analysis of these rows confirms the existence of that boundary, and reveals a few chapters that are misclassified by the estimated change-point. The statistical evidence supports the hypotheses of one main author writing about four fifths of the book, with a second author finishing the book by filling in material, mainly at the end of it.

KEY WORDS: Gibbs sampler; Multinomial change-point analysis; Multinomial cluster analysis; Stylometry; Word length.

1. DESCRIPTION OF THE PROBLEM

Tirant lo Blanc is a chivalry book written in Catalan, hailed to be "the best book of its kind in the world" by Cervantes in Don Quixote. Considered to be the first modern novel in Europe, (see, e.g., Vargas Llosa 1991), it has two modern English translations. The main body of this book was written between 1460 and 1464, but it was not printed until 1490, and there has been an intense and long-lasting debate around its authorship, originating from conflicting information given in its first edition.

Where in the dedicatory letter at the beginning of the book it is stated that: "*So that no one else can be blamed if any faults are found in this work, I, Joanot Martorell, knight, take sole responsability for it, as I have carried out the task singlehandedly*," in the colophon at the end of the book it is stated that it was written "*by the magnificent and virtuous knight, Sir Joanot Martorell, who because of his death, could only finish writing three parts of it. The fourth part, which is the end of the book, was written by the illustrious knight Sir Martí Joan de Galba. If faults are found in that part, let them be attributed to his ignorance.*" Over the years, experts have split between the ones favoring the single authorship hypotheses, in line with the dedicatory letter, and the ones backing the hypotheses of a change of author somewhere between chapters 350 and 400, in line with the colophon. The majority opinion among experts nowadays seems to lean toward the existence of a single author, even though there is still a rather vocal dissenting minority. For a detailed overview of this authorship debate, see Riquer (1990).

It is well accepted by all medievalists that the main (and maybe single) author, Joanot Martorell, died in 1465 and did not start working on the book before 1460, and that if there were any additions, they would be close to the end of the book and made by the second author much later, when the book was printed in 1490. Through an analysis of the diversity of the vocabulary, Riba and Ginebra (2000) found that it becomes less diverse after chapter 383. Our goal here is to determine whether the style in the book is homogeneous or not, through the statistical analysis of other stylistic features, and if it is not, to identify stylistic boundaries.

Following the lead of the extensive literature on stylometry, (see, e.g., Mosteller and Wallace 1964, 1984; Morton 1978; Holmes 1985; Oakes 1998; Lebart, Salem, and Berry 1998), we analyze the distribution of *word length* and the use of the most frequent *context-free words* in each chapter of the book. Given that the features selected are categorical, that leads to contingency tables of ordered rows, that can be modeled as sequences of multinomial observations. Neither of the two candidate authors left any other texts and therefore we cannot use discriminant analysis to help classify chapters by author.

In Section 2 we explore the sequence of rows of the contingency tables graphically, and find a single shift in their distribution, near the location where the colophon places the change of author. In Sections 3 and 4 we present a Bayesian change-point analysis of these sequences, and a Bayesian cluster analysis for them. No expert conjectures the existence of more than two authors, and the evidence in Section 2 does not suggest otherwise, and therefore we focus our presentation on the partition of the book in only two parts. The fact that the partition obtained through cluster analysis matches closely the partition obtained through the change-point analysis, corroborates the existence of a stylistic boundary, but it also reveals a few isolated chapters that are misclassified by the change-point, something already anticipated in Section 2. Our analysis thus provides evidence in

Javier Girón is Professor, Departamento de Estadística e I. O. Facultad de Ciencias, Universidad de Málaga Campus de Teatinos s/n 29071 Málaga, Spain (E-mail: fj_giron@uma.es). Josep Ginebra is Associate Professor, and Alex Riba is Assistant Professor, Departament d'Estadística, E.T.S.E.I.B., Universitat Politècnica de Catalunya, Avgda. Diagonal 647, 6ª Planta, 08028 Barcelona, Spain (E-mail: josep.ginebra@upc.edu *and* alex.riba@upc.edu). The authors thank the editor, the associate editor, and two referees for their detailed and constructive suggestions, that helped us improve significantly an early version of this article. We are also grateful to Xavi Puig, Anton Espadaler, Lola Badia, and Josep Guia for their comments on our work. This work was partially supported by the Spanish Ministry of Science and Technology grant BFM2001-2327.

 DOI: 10.1198/000313005X21311

Table 1. Part of the 425 × 10 Table of Counts of Words of Each Length in Each Chapter of More than 200 Words. N_i is the total number of words in that chapter, $\overline{wl}_i$ is its average word length, and χ_i^2 is the contribution of that chapter to the overall chi-squared statistic to test for independence.

	1	2	3	4	5	6	7	8	9	10+	N_i	$\overline{wl}_i$	χ_i^2
Ch1	21	59	44	19	33	20	16	17	9	17	285	4.47	28.08
Ch2	53	113	80	49	52	33	28	36	16	16	476	4.14	20.13
Ch3	109	274	239	128	112	110	76	51	43	32	1174	4.06	10.30
Ch4	69	150	126	71	60	71	47	32	23	21	670	4.14	7.21
Ch5	119	207	231	123	128	102	61	55	29	34	1089	4.09	11.23
Ch6	69	136	126	69	60	61	37	27	15	15	615	3.96	2.42
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Ch482	50	47	61	18	32	47	23	32	14	11	335	4.50	49.18
Ch483	158	220	207	80	120	93	65	54	62	50	1109	4.21	72.33
Ch484	59	67	68	37	26	32	15	14	17	6	341	3.82	23.50
Ch485	96	174	106	57	77	86	42	54	24	25	741	4.18	37.46
Ch486	45	88	91	46	40	28	13	30	11	10	402	3.94	16.87
Ch487	48	49	62	53	41	36	21	9	16	13	348	4.20	31.34

favor of a change of author near chapters 371 and 382, in line with the hypothesis backed by the colophon, even though it is not up to us to totally exclude the possibility that the stylistic boundary could be explained by something other than a change of author.

Apart from the pioneering work by Mosteller and Wallace (1964, 1984), we are not aware of the application of Bayesian data analysis techniques to stylometric problems. That is why this case study is also intended as an illustration of the advantages that Bayesian approaches have to offer in these type of applications. For a non-Bayesian analysis of related data, see Ginebra and Cabos (1998) and Riba and Ginebra (2005).

2. DESCRIPTION OF THE DATA

For this study we use the edition of *Tirant lo Blanc* by Riquer. After excluding from consideration the titles of chapters and words in italics, that are quotations in Latin, the book amounts to a total of 398,242 words split into 487 chapters of very unequal lengths. In the analysis, only the 425 chapters with more than 200 words are considered.

Mendenhall (1887) used the length of words to discriminate between the writings of Shakespeare, Bacon, and Marlowe, and Mosteller and Wallace (1964, 1984) used it in their study of the authorship of the *Federalist Papers*. Some of the many other authors who used it to characterize style were Brinegar (1963), Bruno (1974), Williams (1975), Smith (1983), and Hilton and Holmes (1993). Here, all the words in each chapter are classified according to their number of letters, with a single category for all the words of more than nine letters. That leads to a 425×10 contingency table of ordered rows, partially presented in Table 1. For example, the first row in that table indicates that chapter 1 has a total of 285 words, out of which 21 are one-lettered words, 59 are two-lettered words, and so on.

If the book was written by a single author, one could expect that all the rows in Table 1 come from a single distribution. On the other hand, if the row profiles changed suddenly, that could indicate the existence of a second author who took over in that chapter, and completed the book. To explore the evolution of these row profiles, Figure 1 presents the sequence of proportions of words of each length in each chapter, the sequence of average word length per chapter, and the sequence of contributions of each chapter to the usual chi-square statistic for testing the association between the rows and the columns of Table 1. In most of these sequences, there is a clear single shift in level before the last one-fifth of the book, with the average word length and the proportion of single lettered words being larger at the end of the book than elsewhere.

The frequency with which context-free words like articles, pronouns, conjunctions, and prepositions are used tends to be rather stable within texts of the same author, and that explains why they are used to characterize style in authorship attribution problems, (see, e.g., Mosteller and Wallace 1964, 1984; Burrows 1987, 1992; Holmes 1992; Binongo 1994; Peng and Hengartner 2002 among many others). Here, we analyze the number of appearances of each one of the 25 most frequent context-free words in each chapter, partially presented in Table 2 in the form of a 425×25 contingency table of ordered rows. Figure 2 presents the evolution of the frequency of appearance of 15 of these most frequent words, where like in Figure 1, one can appreciate a rather clear single shift in level somewhere between chapters 350 and 400.

The main shift in Figures 1 and 2 is sudden, and therefore it is unlikely that it is a consequence of the evolution in the style of a single author. These plots also indicate that a few chapters might be misclassified by that boundary, a fact that could reveal a second author who might have filled in material at the end of the book, and retouched earlier chapters written by the first author. For a report on a correspondence analysis on these tables that confirms the existence of a stylistic boundary, see Riba and Ginebra (2005).

For each chapter i, one observes the ith vector valued row in Tables 1 and 2, $y_i = (y_{i1}, \ldots, y_{il})$, with a conditional distribution that is assumed to be multinomial,

$$y_i | N_i, \theta_i \sim \text{Mu}_{l-1}(N_i, \theta_i), \tag{1}$$

where $\theta_i = (\theta_{i1}, \ldots, \theta_{il})$, with θ_{ij} being the probability for the jth category in chapter i, and l being the number of categories, (i.e., 10 and 25). Thus, the rows in Tables 1 and 2 are assumed to form sequences of conditionally independent observations, with density function

$$m(y_i | N_i, \theta_i) = \frac{N_i!}{\Pi_{j=1}^{l} y_{ij}!} \Pi_{j=1}^{l} \theta_{ij}^{y_{ij}}. \tag{2}$$

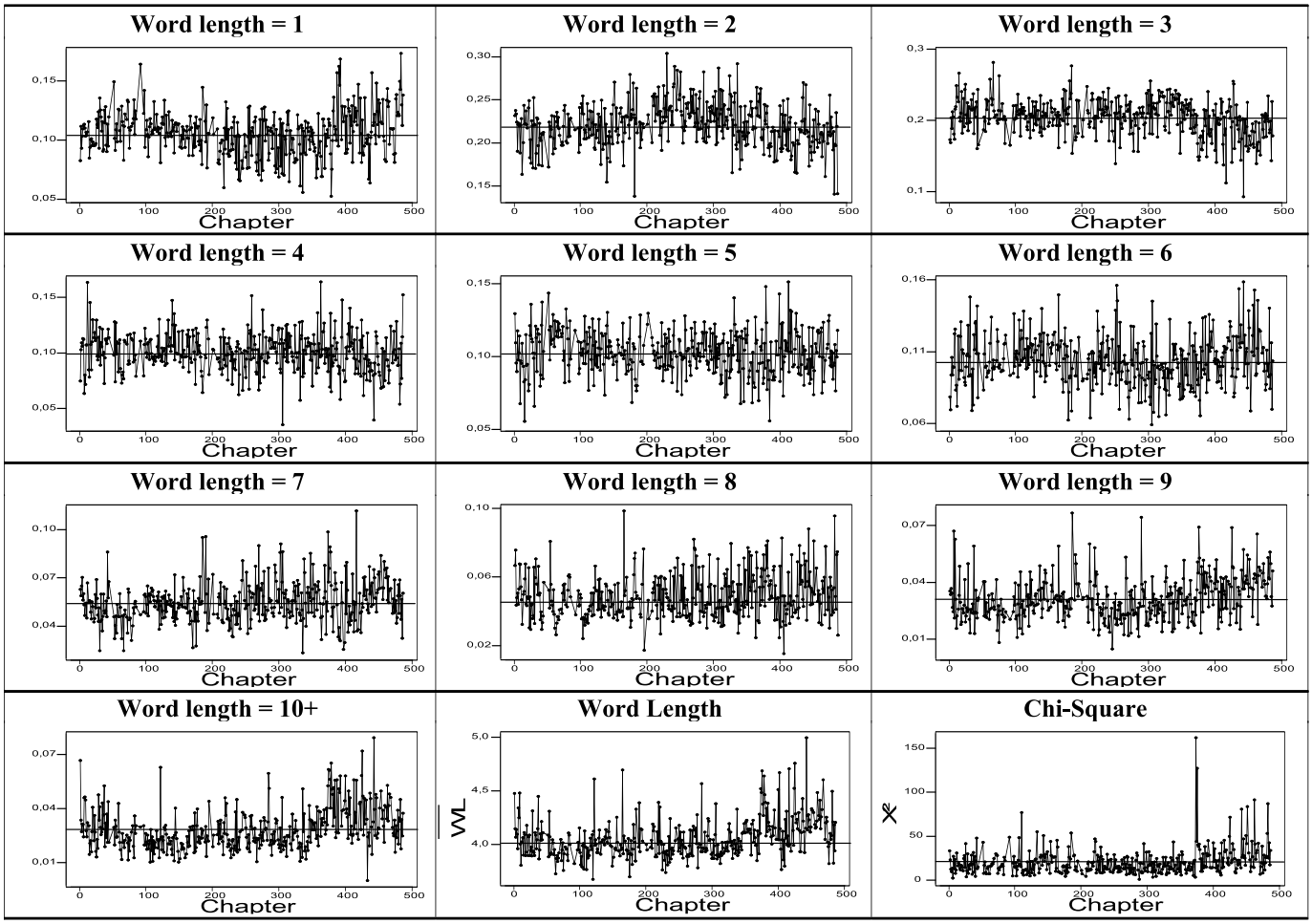

Figure 1. Sequence of proportion of words of one, two, three, four, five, six, seven, eight, nine, and of more than nine letters in each chapter, sequence of average word length, and sequence of the contribution of each chapter to the chi-square statistic, based on Table 1.

In principle, if all the chapters belong to the same author and were written at about the same time, it is reasonable to expect their style to remain the same and thus θ_i to stay approximately constant along the whole sequence of 425 chapters. On the other hand, if one detects a sudden shift in θ_i, that indicates a change in style that might indicate a change of author. In Section 3, we test for the existence and estimate the location of a change-point for the sequences of rows of Tables 1 and 2, and in Section 4 we do a cluster analysis on the rows of these tables.

3. CHANGE-POINT ANALYSIS IN A MULTINOMIAL SEQUENCE

A sequence of conditionally independent ordered random variables, $y = (y_1, \ldots, y_n)$, is said to have a single change-

Table 2. Part of the 425 × 25 Table of Counts of Each of the 25 Most Frequent Words in Each Chapter of More Than 200 Words, with N_i Being the Row Sum and χ_i^2 the Contribution of the Corresponding Chapter to the Overall Chi-Squared Statistic. The words missing are que, lo, en, a, per, l, los, ab, les, d, li, qui, del, se, and gran.

	e	*de*	*la*	···	*no*	*com*	*molt*	*és*	*jo*	*si*	*dix*	···	N_i	χ_i^2
Ch1	12	15	9	···	1	2	1	6	0	3	0	···	100	42.2
Ch2	26	28	19	···	3	3	8	3	1	3	1	···	175	38.2
Ch3	66	46	48	···	19	11	2	9	10	2	4	···	437	31.4
Ch4	33	29	34	···	5	4	8	5	2	3	3	···	245	36.3
Ch5	63	46	42	···	8	16	10	2	1	2	3	···	426	61.4
Ch6	35	15	27	···	7	3	6	1	1	4	1	···	236	37.5
···	···	···	···	···	···	···	···	···	···	···	···	···	···	···
Ch482	31	8	11	···	5	3	3	4	1	0	0	···	119	42.9
Ch483	85	59	39	···	14	9	8	4	3	1	1	···	450	47.9
Ch484	31	19	13	···	2	0	2	0	1	1	2	···	149	30.8
Ch485	59	66	28	···	2	1	6	0	0	0	0	···	298	122.6
Ch486	28	29	14	···	1	3	10	0	0	0	0	···	180	60.1
Ch487	29	13	8	···	2	3	9	0	0	0	0	···	123	58.6

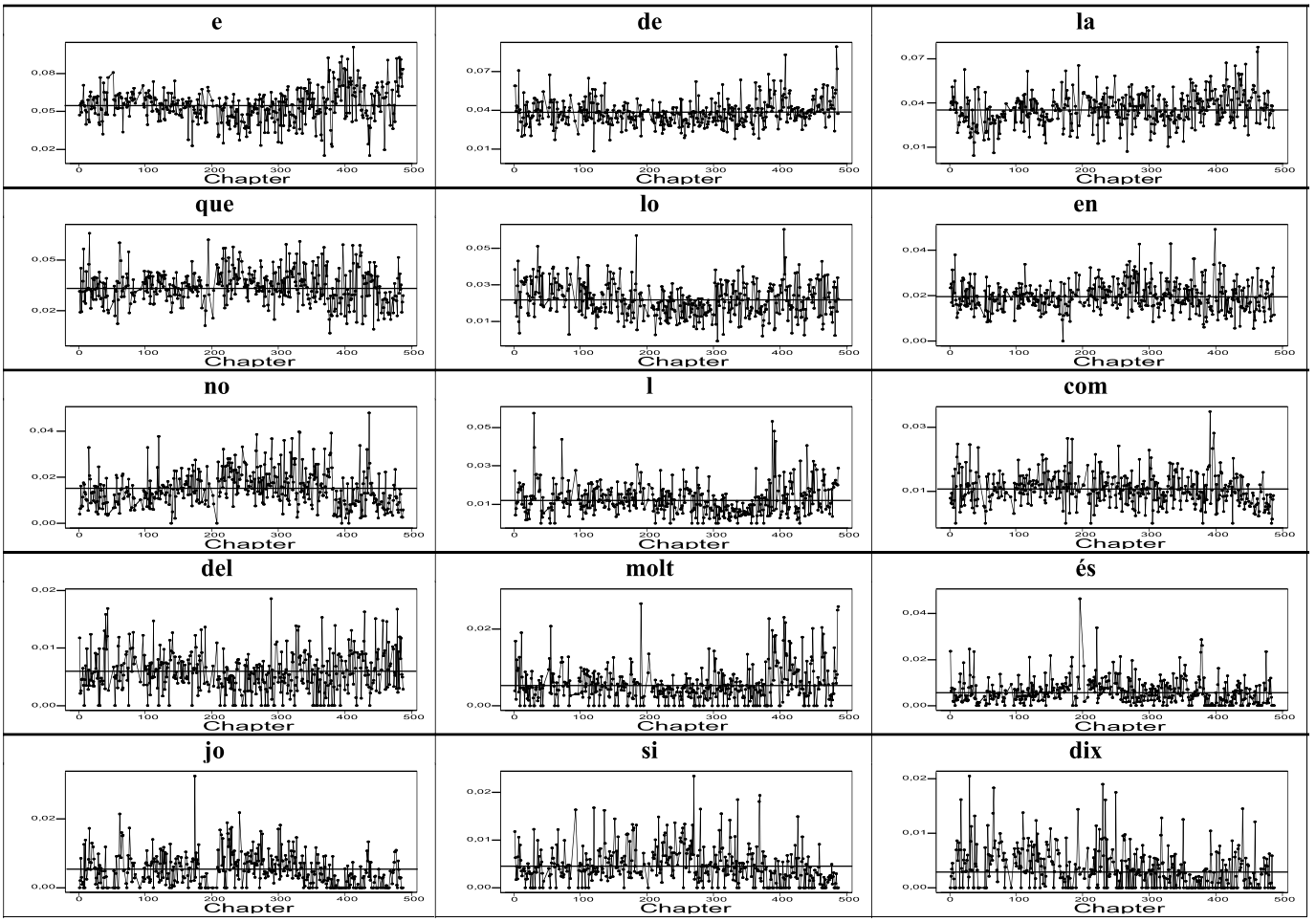

Figure 2. Sequence of the frequency of appearance in each chapter of the words e, de, la, que, lo, en, no, l, com, del, molt, és, jo, si and dix, based on Table 2.

point at r if their distribution function is $F_{\theta_b}(y)$ for $i \leq r$ while it is $F_{\theta_a}(y)$ for $i > r$, where $F_{\theta_b}(y)$ and $F_{\theta_a}(y)$ are different. We consider $F_{\theta_a}(.)$ and $F_{\theta_b}(.)$ to be unknown, but known to belong to the same parametric family. The problem of estimating the location of a change-point has been extensively studied for various distributions and different points of view. For an approach based on likelihood see, for example, Hinkley and Hinkley (1970) and Hinkley (1970, 1971), and for a nonparametric approach see Bhattacharyia and Johnson (1968) or Wolfe and Schechtman (1984). For reviews on this topic see Zacks (1983), Krishnaiah and Miao (1988), and Pettitt (1989).

Early uses of the Bayesian approach to change-point problems, mainly for normal and binomial sequences, can be found, for example, in Sen and Srivastava (1973), Broemeling (1974), Ferreira (1975), Smith (1975, 1980), and Smith and Cook (1980). For Bayesian MCMC based implementations, see Carlin, Gelfand, and Smith (1992), Barry and Hartigan (1993), Stephens (1994), Lee (1998), and Fan and Brooks (2000).

At a conceptual level, the main advantage of Bayesian approaches is that, by transforming the likelihood function of (r, θ_b, θ_a) into a full-fledged posterior probability distribution for (r, θ_b, θ_a), it allows one to test for the existence of a change-point through the posterior probability that a change-point exists, it allows one to estimate the change-point, r, through the posterior marginal distribution for r, and it allows one to investigate what components in θ_b and θ_a do explain that change-point through the posterior marginal distribution for (θ_a, θ_b). All the computations thus involve just the update of a prior distribution on (r, θ_b, θ_a), into a posterior on them, and all the inference questions can be addressed through probability statements on the parameters of interest. That is a lot simpler to implement and interpret than when one addresses them through the classical p values, point estimates, and approximate confidence regions for r, θ_b, and θ_a; note that here the main parameter of interest, r, is discrete-valued, and that makes it even harder to find viable non-Bayesian alternatives to the analysis that follows.

Additional advantages of the Bayesian approach are that it allows one to incorporate prior information about (θ_a, θ_b) and r, and that it makes it easy to deal with short chapters and to combine evidence from different sets of data into a single analysis.

In our setting, we are investigating the existence of a change-point in the sequence of rows of Tables 1 and 2, and we assume that $F_{\theta_a}(.)$ and $F_{\theta_b}(.)$ are multinomial distributions, with unknown θ_b and θ_a, a case rarely considered in the literature. Wolfe and Chen (1990) estimated this change-point through ad-hoc combinations of the maximum likelihood estimates of the change-points for the marginal binomial sequences, while Riba

and Ginebra (2005) estimated it through maximum likelihood for the full multinomial model.

Multiple change-point analysis aims at making inferences about two or more change-points at a time, with a formulation analogous to the one presented in the following for the single change-point problem. Figures 1 and 2 together with subject matter information indicate that, on top of the main change-point, there might be some contamination on both sides of it, fruit of minor interventions of each author on the part mainly written by the other one, thus leading to an undetermined number of secondary change-points of a very different nature than the main change-point estimated here. That is why we decided that it was more appropriate to tackle it through the cluster analysis presented in Section 4.

3.1 Bayesian Formulation of the Change-Point Problem

Let $y = (y_1, \ldots, y_n)$ be a realization of an ordered sequence of conditionally independent multinomial random variables, y_i, with $\theta_i = \theta_b = (\theta_{b1}, \ldots, \theta_{bl})$ for $i \leq r$, and $\theta_i = \theta_a = (\theta_{a1}, \ldots, \theta_{al})$ for $i > r$, and thus, let the likelihood function be proportional to

$$[y|r, \theta_b, \theta_a] = \Pi_{i=1}^{r} m(y_i|N_i, \theta_b)\, \Pi_{i=r+1}^{n} m(y_i|N_i, \theta_a) \propto \Pi_{i=1}^{r} \theta_{b1}^{y_{i1}} \ldots \theta_{bl}^{y_{il}}\, \Pi_{i=r+1}^{n} \theta_{a1}^{y_{i1}} \ldots \theta_{al}^{y_{il}}, \quad (3)$$

with $m(y_i|N_i, \theta_i)$ as in (2). Given y, the main objective in change-point analysis is to infer about r, in order to decide if there is a change-point and if so, where it is located. A secondary objective is to infer about $\theta_b - \theta_a$, or about the coordinate-wise ratio, θ_b/θ_a, in order to understand what it is that changes at the change-point. By assuming a prior distribution on (r, θ_b, θ_a), denoted by $[r, \theta_b, \theta_a]$, the likelihood function can be made into the density of a joint posterior distribution for (r, θ_b, θ_a), denoted by $[r, \theta_b, \theta_a|y]$, through Bayes' theorem that rules that the posterior is proportional to the prior times the likelihood

$$[r, \theta_b, \theta_a|y] \propto [r, \theta_b, \theta_a][y|r, \theta_b, \theta_a], \quad (4)$$

where the proportionality constant is determined so that the posterior is a proper density. A priori, here it is sensible to assume that r, θ_b, and θ_a are independent.

The prior marginal distribution for r, denoted by $[r]$, is supported on $\{1, 2, \ldots, n\}$. In our problem, there is plenty of conflicting subject based information as to where the change-point, r, could be located. Believers in the single authorship hypothesis would place most of the prior probability about r on the value $r = n$, because that indicates a lack of a change-point. Instead, believers in a change of author somewhere between chapters 350 and 400 would spread most of their prior probability about r on the values in $\{300, 301, \ldots, 345, 346\}$, (remember that we have excluded chapters of less than 200 words). Luckily, the evidence in the data in favor of the existence of a boundary is so strong, that we can afford reporting the results under a noninformative prior, "impartial" to both theories. Unless one is certain that the change-point exists, the noninformative prior distribution for r spreads all the prior probability mass uniformly on $\{1, 2, \ldots, n\}$, with $[r = i] = 1/n$ for $i = 1, \ldots, n$; (if one is certain that the change-point exists, then the noninformative prior distribution for its location is uniform on $\{1, 2, \ldots, n-1\}$).

The prior distributions for θ_b and for θ_a, denoted by $[\theta_b]$ and $[\theta_a]$, are supported on the $(l-1)$-dimensional simplex. The conjugate prior distributions for those multinomial parameters are Dirichlet, and thus their density functions are such that

$$[\theta_b] = \mathrm{Di}_{l-1}(\alpha_{b1}, \ldots, \alpha_{bl}) \propto \theta_{b1}^{\alpha_{b1}-1} \ldots \theta_{bl}^{\alpha_{bl}-1}, \quad (5)$$

and

$$[\theta_a] = \mathrm{Di}_{l-1}(\alpha_{a1}, \ldots, \alpha_{al}) \propto \theta_{a1}^{\alpha_{a1}-1} \ldots \theta_{al}^{\alpha_{al}-1}, \quad (6)$$

where $\alpha_b = (\alpha_{b1}, \ldots, \alpha_{bl})$ and $\alpha_a = (\alpha_{a1}, \ldots, \alpha_{al})$ are assumed to be known. These distributions allow one to incorporate many kinds of prior information about θ_b and θ_a. For example, if one knows that the values for θ_{bj} (and for θ_{aj}) are likely to be similar for all j, as in Table 1, one would chose a prior for θ_b such that $\alpha_{b1} = \alpha_{b2} = \cdots = \alpha_{bl}$, and similarly for θ_a. If instead, one knows that the values for θ_{bj} (and for θ_{aj}) are likely to be ordered from larger to smaller, as in Table 2, then it is natural to assume a prior for θ_b with $\alpha_{b1} \geq \alpha_{b2} \geq \cdots \geq \alpha_{bl}$, and similarly for θ_a. The noninformative prior distribution for θ_b (and for θ_a) is the uniform on the $(l-1)$-dimensional simplex, that is a $\mathrm{Di}_{l-1}(\alpha)$ with $\alpha = (1, \ldots, 1)$. We report the results based on uniform priors, but when we tried informative priors of the type described earlier, we obtained very similar results, because samples are large enough to make posterior distributions insensitive to the prior choice.

Under the conjugate prior setting, the density of the joint posterior distribution becomes

$$[r, \theta_b, \theta_a|y] \propto [r]\theta_{b1}^{\alpha_{b1}(r)-1} \ldots \theta_{bl}^{\alpha_{bl}(r)-1} \theta_{a1}^{\alpha_{a1}(r)-1} \ldots \theta_{al}^{\alpha_{al}(r)-1}, \quad (7)$$

where $\alpha_{bj}(r) = \alpha_{bj} + \sum_{i=1}^{r} y_{ij}$ and $\alpha_{aj}(r) = \alpha_{aj} + \sum_{i=r+1}^{n} y_{ij}$, and thus it has the same form as the prior distribution. The densities for the marginal posterior distributions of r and of (θ_b, θ_a), denoted by $[r|y]$ and $[\theta_b, \theta_a|y]$, can then be obtained through integration of $[r, \theta_b, \theta_a|y]$, after normalizing the left hand side of (7), which leads to

$$[r|y] \propto [r] \frac{\prod_{j=1}^{l} \Gamma(\alpha_{bj}(r))}{\Gamma(\sum_{j=1}^{l} \alpha_{bj}(r))} \frac{\prod_{j=1}^{l} \Gamma(\alpha_{aj}(r))}{\Gamma(\sum_{j=1}^{l} \alpha_{aj}(r))}, \quad (8)$$

where $\Gamma(.)$ is the gamma function, and

$$[\theta_b, \theta_a|y] = \sum_{r=1}^{n} [r|y]\, \mathrm{Di}_{l-1}(\alpha_{b1}(r), \ldots, \alpha_{bl}(r))\, \mathrm{Di}_{l-1}(\alpha_{a1}(r), \ldots, \alpha_{al}(r)). \quad (9)$$

By simulating large enough samples from the joint posterior distribution, $[r, \theta_b, \theta_a|y]$, any characteristic of the marginal posterior distributions, $[r|y]$ and $[\theta_b, \theta_a|y]$, can be calculated to any degree of accuracy. That can be accomplished either by direct simulation from (8) and (9), or by the use of the Gibbs sampler algorithm that does not require closed form expressions for the posterior density functions. The basic idea behind the Gibbs sampler is to construct a Markov chain that has $[r, \theta_a, \theta_b|y]$ as its equilibrium distribution, by iteratively sampling from the conditional distributions

$$[r|y, \theta_b, \theta_a] = \frac{[y|r, \theta_b, \theta_a][r]}{\sum_{r=1}^{n} [y|r, \theta_b, \theta_a][r]}, \quad (10)$$

$$[\theta_b|y, r, \theta_a] = [\theta_b|y, r] \sim \text{Di}_{l-1}(\alpha_{b1}(r), \ldots, \alpha_{bl}(r)), \quad (11)$$

and

$$[\theta_a|y, r, \theta_b] = [\theta_a|y, r] \sim \text{Di}_{l-1}(\alpha_{a1}(r), \ldots, \alpha_{al}(r)). \quad (12)$$

That is, one first picks arbitrary initial values (θ_b^0, θ_a^0), and then one successively makes random drawings r^i from $[r|y, \theta_b^{i-1}, \theta_a^{i-1}]$, θ_b^i from $[\theta_b|y, r^i, \theta_a^{i-1}]$ and θ_a^i from $[\theta_a|y, r^i, \theta_b^i]$, starting from $i = 1$. Repeated iteration of this cycle of random variate generation produces a Markov chain, $(r^i, \theta_b^i, \theta_a^i)$, with the desired equilibrium distribution. For a general description and many illustrations of the use of this algorithm and of related MCMC methods in Bayesian analysis, see, for example, Gelfand and Smith (1990), Casella and George (1992), Chib and Greenberg (1995), Besag, Green, Higdon, and Mengersen (1995), Robert and Casella (1999) and references therein. For a detailed description of its implementation on sequential multinomial data, see Liu (1994), and for implementations on change-point problems, see Carlin et al. (1992) and Stephens (1994).

3.2 Inferences about the Location of the Change-Point, *r*

Assuming that the change-point, r, occurs at the last chapter of the book, $r = n$, is the same as assuming that there is no change-point, and therefore testing for the nonexistence of a change-point is equivalent to testing for $H_0 : r = n$ against $H_1 : r \neq n$. One of the attractive features of the Bayesian approach is that it treats both hypotheses symmetrically, and that it allows one to make a decision based on the values of the posterior probabilities of H_0 and H_1, $\Pr(H_0|y) = [r = n|y]$ and $\Pr(H_1|y) = 1 - \Pr(H_0|y)$, that can be easily evaluated from the samples of $[r|y]$ obtained through the Gibbs sampler or through direct computation from (8). The noninformative prior for testing assumes $\Pr(H_0) = \Pr(H_1) = 1/2$, and because H_1 is composite and one needs to be impartial to all $n - 1$ possible change-point locations, that part of prior probability is spread uniformly on $\{1, \ldots, n-1\}$, with $[r = i] = 1/2(n-1)$ for $i = 1, \ldots, (n-1)$. For the sequence of rows of Tables 1 and 2, $\Pr(H_0|y)$ is extremely close to 0, and therefore the evidence in favor of the existence of a change-point is very strong.

Table 3. Marginal Posterior Probability for r, [r|y], Under Independent Uniform Priors for r, θ_b, and θ_a, for the Sequence of Rows in Table 1 and for the Sequence of Counts of the Words e, de, la, que, no, l, com, és, molt, jo, si and dix in Table 2.

Chapter	Word length (Table 1)	Use of words (Subset of Table 2)
1 to 343	.0000	.0000
344	.0003	.0000
345	.0158	.0000
346	.0032	.0000
347	.0042	.0000
348	.0017	.0000
349–367	.0000	.0000
368	.0001	.0000
369	.0007	.0000
370	.0012	.0000
371	**.9655**	.0000
372	.0038	.0000
373	.0033	.0000
374 to 377	.0000	.0000
378	.0000	.0003
379	.0000	.0278
380	.0000	.0919
381	.0000	.0729
382	.0000	**.8070**
383	.0000	.0001
384–487	.0000	.0000

In order to estimate the location of the change-point, r, in Bayesian analysis one typically reports the posterior distribution of r, $[r|y]$. By providing a whole posterior probability distribution for r, the Bayesian approach allows one to locate the change-point at the same time that one assesses the uncertainty associated with that location. That is a big advantage relative to providing just point estimates, especially given that r is discrete and therefore it is not clear how one would compute the approximate confidence intervals called for in non-Bayesian approaches.

Given the conflicting nature of the theories about r, for estimation purposes we present the posterior of r under the uniform prior on $\{1, \ldots, n\}$. That avoids the suspicion that our results could be biased by our prior guess on the location of that change-point. Note that, at this point, we are assuming that $[r = n] = 1/n$ instead of $[r = n] = 1/2$, as we assumed for testing $H_0 : r = n$. That is what is typically called for in noninformative Bayesian analysis, when one shifts focus from testing point null hypothesis to the estimation of the parameter. In Bayesian practice, often the only thing that changes when one shifts from testing to estimation is the prior used.

Table 3 presents the marginal posterior, $[r|y]$, computed exactly from (8), for Table 1 and for a subset of Table 2 that considers only the 12 columns that best explain the change-point in the sense described in Section 3.3. The marginal posterior under the complete Table 2 is such that $[r = 382|y] = .9441$, which is even more concentrated than the one for the subset of Table 2. The mode of these change-point posteriors—chapters 371 and 382—play the role of estimates for r, and stand at about three quarters of the total number of chapters, thus placing the change in style near where the colophon and some specialists places the change of author. The fact that the posterior probability for $r = n$ is negligible abounds on the evidence in favor of the existence of at least one change-point.

The strong peakedness of these marginal posterior distributions indicates that the evidence in favor of a change in style near their modes is strong; if instead of a uniform prior for r, we had used a prior giving more credibility to the colophon, the posterior distribution would be even more peaked around these modes. The fact that the two posterior distributions for r are so concentrated 11 chapters away from each other might be a sign that the data call for a slightly refined analysis that adds some uncertainty through the use of a hierarchical model, as described in Section 5.

3.3 Inference about the Multinomial Parameters, (θ_a, θ_b)

Once the existence and the location of the change-point is settled, the question arises as to what components in $\theta_i = (\theta_{i1}, \ldots, \theta_{il})$ do change and change the most in that change-point. To answer that, one possibility would be to analyze each marginal binomial sequence separately. Instead, we use the posterior for $\log(\theta_{bj}/\theta_{aj})$; our rationale for choosing $\log(\theta_{bj}/\theta_{aj})$ instead of $\theta_{bj} - \theta_{aj}$ or $\text{logit}(\theta_{bj}) - \text{logit}(\theta_{aj})$, is the fact that the

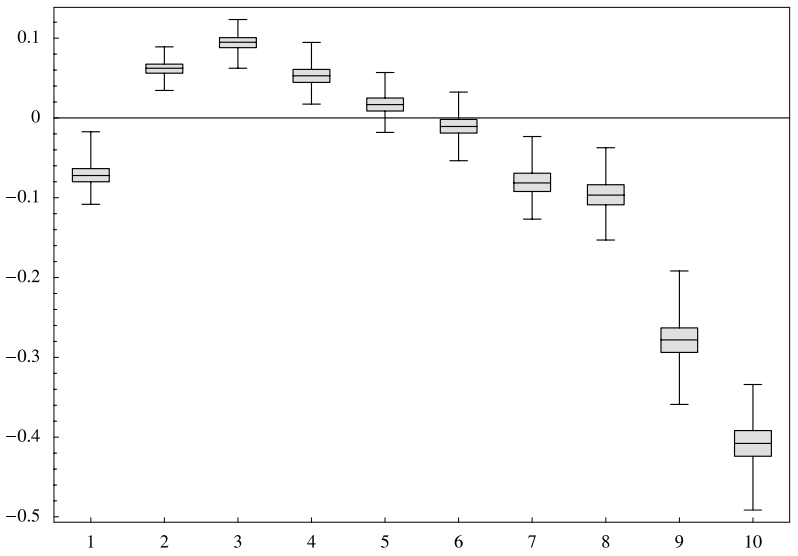

Figure 3. Boxplot of a posterior sample of [log(θ_{bj}=θ_{aj})|y] for Table 1.

Bayesian discriminant function for comparing two multinomial models with *known* vector parameters θ_b and θ_a, is the corresponding log-likelihood ratio and therefore it is a linear function with coefficients $\log(\theta_{bj}/\theta_{aj})$, (see Gilbert 1968; Moore 1973).

Figures 3 and 4 present boxplots of samples of $[\log(\theta_{bj}/\theta_{aj})|y]$ for $j = 1, \ldots, l$, obtained through simulation of the marginal posterior for θ_b and θ_a, (9), for the data in Tables 1 and 2, respectively. The further the value 0 is from the box, the more relevant the corresponding category is for the purpose of discriminating between chapters before and after the change-point. Figure 3 indicates that two-, three-, four-, and five-lettered words are more abundant before the change-point, while one-, six-, seven-, eight-, nine-, and ten-or-more lettered words are more abundant after that change-point.

Figure 4 indicates that words *que*, *no*, *com*, *és*, *jo*, *si* and *dix* are more abundant before the change-point, while *e*, *de*, *la*, *l'* and *molt*, are more abundant after that change-point. The posterior marginal distribution for r in the third column of Table 3, is obtained after restricting consideration to this subset of 12 columns of Table 2, that "best explain" the change-point; they are the ones with $E[\log(\theta_{bj}/\theta_{aj})|y]$ lying farther away from 0.

We have also explored the existence of two or three change-points through a multiple change-point analysis, (see Stephens 1994), but everything points rather towards the existence of minor contamination on both sides of chapters 371–382, as a consequence of sparse interventions by the author responsible of

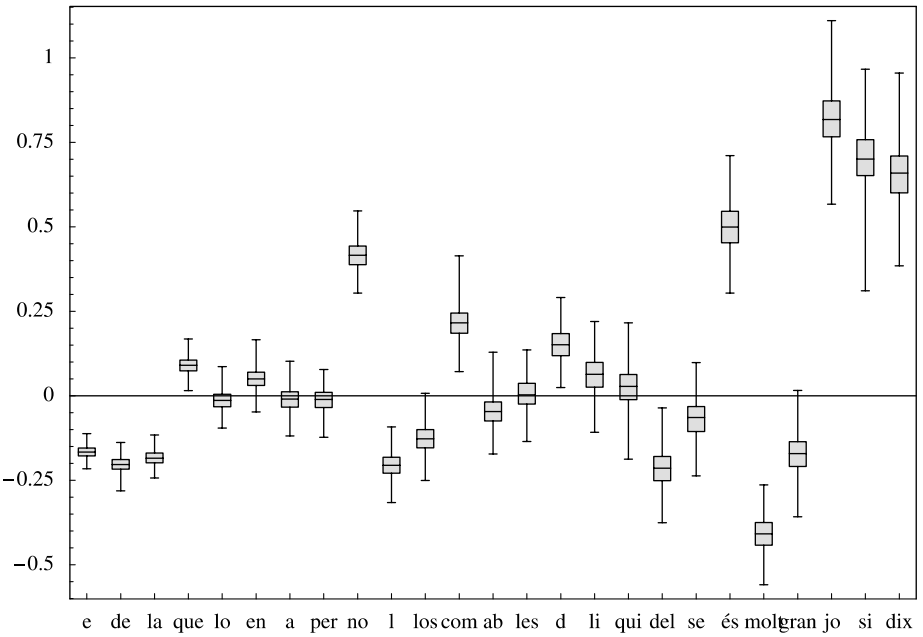

Figure 4. Boxplot of a posterior sample of [log(θ_{bj}=θ_{aj})|y] for the complete Table 2.

the other part of the book. The cluster analysis of Section 4 aims at uncovering that contamination, while at the same time corroborating the existence of a main single change-point. Doing the cluster analysis on the tables including only the columns that best explain the boundary, should help reclassify chapters misclassified by the change-point.

4. CLUSTER ANALYSIS OF MULTINOMIAL OBSERVATIONS

In change-point analysis, one partitions a sequence into subsequences of observations that are more homogeneous than the whole sequence, forcing consecutive observations to belong to the same group. Instead, cluster analysis partitions a set of observations into groups that are more homogeneous than the whole, but without imposing any type of order restriction when forming the groups.

Most of the literature on cluster analysis is geared toward continuous data, and it uses ad hoc heuristic partitioning algorithms, (see, e.g., Kaufman and Rousseeuw 1990; Gnanadesikan 1997; or Gordon 1999). Cluster analysis can be given a probabilistic formulation by assuming a mixture model in which any two observations that belong to the same cluster have the same distribution, and then estimating the mixture density and assigning observations to component densities, (see, e.g., McLachlan and Basford 1988 or Banfield and Raftery 1993).

Binder (1978) and Bernardo and Girón (1988) are early examples of Bayesian cluster analysis built on the finite mixture model formulation. For MCMC-based Bayesian cluster analysis for continuous data, see Robert (1996) and Bensmail, Celeux, Raftery, and Robert (1997), and for an example involving both continuous and categorical variables, see Dellaportas (1998). Like in our case, all these implementations assume the number of clusters known. More flexible and sophisticated approaches to estimating the number of clusters, at the same time that they partition the data in clusters, were presented by Green (1995), Phillips and Smith (1996), and Quintana and Iglesias (2003).

Besides all the conceptual and practical advantages of the Bayesian approach listed in Section 3, one specific advantage of the Bayesian cluster analysis is that it allows for the allocation of each observation into one cluster through the posterior probability that the observation belongs to each cluster. That provides an extremely useful assessment of the uncertainty associated to that allocation, that is something lacking in non-Bayesian approaches that provide only yes/no classifications of observations to clusters.

We are interested in clustering the rows of Tables 1 and 2, to partition chapters in two internally homogeneous subsets. Greenacre (1988) tackled that problem through distance based heuristic partitioning algorithms, while Riba and Ginebra (2005) formulated it in terms of mixtures of multinomial distributions, and solves it through the maximum likelihood approach.

4.1 Bayesian Formulation of the Cluster Analysis Problem

When $y = (y_1, \ldots, y_n)$ is such that y_i has a $\text{Mu}_{l-1}(N_i, \theta_1)$ distribution with probability p and a $\text{Mu}_{l-1}(N_i, \theta_2)$ distribution with probability $1 - p$, and they are conditionally independent, then their likelihood function is proportional to

$$[y|p, \theta_1, \theta_2] = \Pi_{i=1}^{n}(p * m(y_i|N_i, \theta_1) + (1-p) * m(y_i|N_i, \theta_2)), \quad (13)$$

where $m(y_i|N_i, \theta)$ is defined in (2), $\theta_1 = (\theta_{11}, \ldots, \theta_{1l})$ and $\theta_2 = (\theta_{21}, \ldots, \theta_{2l})$. This likelihood is a sum of 2^n terms, each corresponding to a different partition of observations into clusters, and the classical likelihood and Bayesian analysis become unfeasible for anything other than for very small n. Besides, the main goal in cluster analysis is not to estimate (p, θ_1, θ_2), but to allocate observations into clusters, a feature missing in (13).

To avoid this combinatorial explosion and to help assign observations to clusters, Dempster, Laird, and Rubin (1977) introduced unobserved classification indicators, $z = (z_1, \ldots, z_n)$, such that if y_i belongs to the first component of the mixture, $z_i = 1$, and if y_i belongs to the second one, $z_i = 0$. This leads to $[y_i|z_i = 1]$ being $\text{Mu}_{l-1}(N_i, \theta_1)$, $[y_i|z_i = 0]$ being $\text{Mu}_{l-1}(N_i, \theta_2)$, $[z_i|p]$ being Bernoulli(p), and

$$[y, z|p, \theta_1, \theta_2] = [y|p, \theta_1, \theta_2, z][z|p] = \Pi_{i=1}^{n}(p * m(y_i|N_i, \theta_1))^{z_i}((1-p) * m(y_i|N_i, \theta_2))^{1-z_i}. \quad (14)$$

By considering (y, z) instead of just y, the mixture form of the likelihood disappears, and that simplifies the computations. Besides, in cluster analysis, z becomes the main parameter of interest, because the expected value of the posterior distribution of each of its components, $[z_i|y]$, is the posterior probability that the ith observation belongs to cluster 1, and thus it is what one needs to assign observations to clusters.

The mixture model specified by (13), presents a permutation-type of nonidentifiability, because $[y|p, \theta_1, \theta_2] = [y|1 - p, \theta_2, \theta_1]$, and therefore the value of the likelihood function at (p, θ_1, θ_2) is always equal to its value at $(1 - p, \theta_2, \theta_1)$. That means that the observed sequence, y, does not allow one to distinguish between these two possibilities. The simplest way to solve this label-switching problem is by imposing the restriction that $p \geq .5$. For a general treatment of this problem, see Swartz, Haitovsky, Vexler, and Yang (2004).

Once the constraint on p has been established, imposing constraints on θ_1 and θ_2 to assure identifiability is unnecessary, and it is sensible to assume that, a priori, p, θ_1, and θ_2 are independent. For reasons analogous to the ones given in Section 3.1, it is convenient to assign a distribution $\text{Di}_{l-1}(\alpha_{11}, \ldots, \alpha_{1l})$ for the prior marginal of θ_1, and a distribution $\text{Di}_{l-1}(\alpha_{21}, \ldots, \alpha_{2l})$ for the one of θ_2. The conjugate prior distributions for the mixing proportion, p, which take into account the constraint $p \geq .5$, are Beta distributions lower truncated at .5, denoted by trunc-Beta(a_1, a_2) and with density

$$[p] \propto p^{a_1 - 1}(1-p)^{a_2 - 1} I_{[.5,1]}(p), \quad (15)$$

where $I_{[.5,1]}(p)$ is equal to 1 when $p \in [.5, 1]$ and is 0 otherwise. An expert who believes in the single authorship hypotheses would choose a truncated Beta distribution with most of its probability mass near 1, while an expert who believes in the colophon of the book would choose one with most of its probability mass around $p = .75$ or .8, coherent with their belief that the second author wrote about one-fourth or one-fifth of the book. We report the results under the assumption that the prior on p is uniform on $[.5, 1]$, that is a trunc-Beta$(1, 1)$ with

$E(p) = .75$, and that the priors for θ_1 and θ_2 are also uniform. We have repeated the analysis assuming various other mildly informative priors for p, θ_1, and θ_2, obtaining basically the same results, because the sample sizes involved are large enough to render inferences insensitive to the prior choice.

The posterior distribution for (p, θ_1, θ_2) conditional on y and z, can be obtained from

$$[p, \theta_1, \theta_2|y, z] \propto [y, z|p, \theta_1, \theta_2][p, \theta_1, \theta_2] = [y|p, \theta_1, \theta_2, z][\theta_1][\theta_2][z|p][p]. \quad (16)$$

By augmenting the data through the introduction of z, conditionally on y and z, the conjugate structure is preserved, because the posterior distribution for (p, θ_1, θ_2) given y and z is again the product of two updated Dirichlet distributions for θ_1 and θ_2 and an updated trunc-Beta distribution for p, and p, θ_1 and θ_2 are still conditionally independent,

$$[p, \theta_1, \theta_2|y, z] \propto \theta_{11}^{\alpha_{11}(z)-1} \dots \theta_{1l}^{\alpha_{1l}(z)-1} \theta_{21}^{\alpha_{21}(z)-1} \dots \theta_{2l}^{\alpha_{2l}(z)-1} p^{a_1(z)-1}(1-p)^{a_2(z)-1} I_{[.5,1]}(p), \quad (17)$$

where $a_1(z) = a_1 + \sum_{i=1}^n z_i$, $a_2(z) = a_2 + \sum_{i=1}^n (1 - z_i)$, $\alpha_{1j}(z) = \alpha_{1j} + \sum_{i=1}^n z_i y_{ij}$ and $\alpha_{2j}(z) = \alpha_{2j} + \sum_{i=1}^n (1 - z_i) y_{ij}$. However, the posterior distribution of (p, θ_1, θ_2) given y, is an intractable mixture of 2^n terms of the form (17), whose weights are $[z|y]$.

Instead of computing the posterior marginals, $[z|y]$, $[p|y]$, and $[\theta_1, \theta_2|y]$, one can resort to the Gibbs sampler or to alternative MCMC sampling methods for mixture models, like the ones considered by Lavine and West (1992), Diebolt and Robert (1994), and Robert and Casella (1999, p. 422). To obtain samples from $[p, \theta_1, \theta_2, z|y]$ through the Gibbs sampler, the first step is to simulate the classification variables, z_i, from Bernoulli distributions with probability of success

$$[z_i = 1|y, p, \theta_1, \theta_2] = \frac{p\, m(y_i|N_i, \theta_1)}{p\, m(y_i|N_i, \theta_1) + (1-p)\, m(y_i|N_i, \theta_2)}, \quad (18)$$

as the z_i's are conditionally independent given y, p, θ_1, θ_2, and the second step is to generate θ_1, θ_2, and p independently from

$$[\theta_1|y, z, p, \theta_2] = [\theta_1|y, z] \sim \text{Di}_{l-1}(\alpha_{11}(z), \dots, \alpha_{1l}(z)), \quad (19)$$

$$[\theta_2|y, z, p, \theta_1] = [\theta_2|y, z] \sim \text{Di}_{l-1}(\alpha_{21}(z), \dots, \alpha_{2l}(z)), \quad (20)$$

and

$$[p|y, z, \theta_1, \theta_2] = [p|z] \sim \text{trunc-Beta}(a_1(z), a_2(z)). \quad (21)$$

By picking initial values for θ_1, θ_2, and p, and by iterating this random variate generation cycle, one produces a Markov chain with equilibrium distribution $[p, \theta_1, \theta_2, z|y]$.

4.2 Inference about the Mixing Proportion, p

The problem of testing whether the chapters group themselves into one or two clusters can be viewed as one of model selection. We need to decide between model M_0, with

$$M_0 : \{[y|\theta] = \Pi_{i=1}^n m(N_i, \theta), [\theta]\}, \quad (22)$$

under which all observations belong to the same multinomial distribution and therefore there is a single cluster, and model M_1, with

$$M_1 : \{[y|p, \theta_1, \theta_2] = \Pi_{i=1}^n (p * m(y_i|N_i, \theta_1) + (1-p) * m(y_i|N_i, \theta_2)), [p, \theta_1, \theta_2]\}, \quad (23)$$

under which the conditional distribution for y_i is a mixture of two multinomial distributions, and therefore under which observations belong to either one of two clusters. The Bayesian approach decides between the single cluster and the two-cluster models through the posterior probabilities for M_0 and M_1, $\Pr(M_0|y)$, and $\Pr(M_1|y)$, or what is the same, through the posterior odds ratio

$$\frac{\Pr(M_0|y)}{\Pr(M_1|y)} = \frac{\Pr(M_0)}{\Pr(M_1)} \times \frac{\int [y|\theta][\theta]\, d\theta}{\int\int\int [y|p, \theta_1, \theta_2][p][\theta_1][\theta_2]\, dp\, d\theta_1\, d\theta_2}, \quad (24)$$

which is the prior odds times the Bayes factor, that is the ratio of the marginal densities of the data y given M_0 and M_1, and thus it can be regarded as an extension of the likelihood ratio statistic. The computation of the numerator is straightforward, but the computation of the denominator is analytically unfeasible; Chib (1995) offered an efficient estimation procedure when data augmentation is coupled with Gibbs sampling, as it is our case. Assuming that $\Pr(M_0) = \Pr(M_1) = 1/2$, that the prior for θ is uniform and that a priori p, θ_1, and θ_2 are independent, that p is uniform on $[.5, 1]$ and θ_1 and θ_2 are uniform on the simplex, the values of $\Pr(M_0|y)$, and of $\Pr(M_0|y)/\Pr(M_1|y)$ for Tables 1 and 2 are extremely close to 0, and therefore the evidence in favor of the existence of two clusters is very strong.

Figure 5 presents the histograms of samples of $[p|y]$ obtained through the Gibbs sampler. The posterior density at $p = 1$, when the mixture collapses to one term, is very small, in accordance with $\Pr(M_0|y)$ being very small. For Table 1, $E[p|y]$ is estimated to be .7399, thus attributing about 74% of the chapters to the first author, in close agreement with the change-point being estimated as chapter 371, that is located at 75.3% of the total. For the subset of Table 2, $E[p|y]$ is estimated to be .7037, relatively close to the result for Table 1 and to the change-point being located in chapter 382, at 77.9% of the total.

4.3 Allocation of the Chapters into Clusters

The posterior probability that y_i belongs to the first cluster is $E[z_i|y]$, and the one that it belongs to the second cluster is $1-E[z_i|y]$, quantities that are easily estimable through the Gibbs samples. Once they are available, it is most natural to allocate each observation to the cluster with largest posterior probability; if $E[z_i|y] > .5$ one assigns y_i to cluster 1, and otherwise one assigns it to cluster 2. Using this rule for Table 1, 319 chapters are attributed to the first author, which represents 75.06% of the 425 chapters considered, and only 75 chapters are misclassified according to the ordering implied by the change-point, of which 38 are attributed to the second author but are located before chapter 371, and 37 are attributed to the first author but are located after it.

For Table 2, only 256 chapters are attributed to the first author, which represents 60.24% of the 425 chapters. Nevertheless, after removing the columns with less "discriminatory power," and

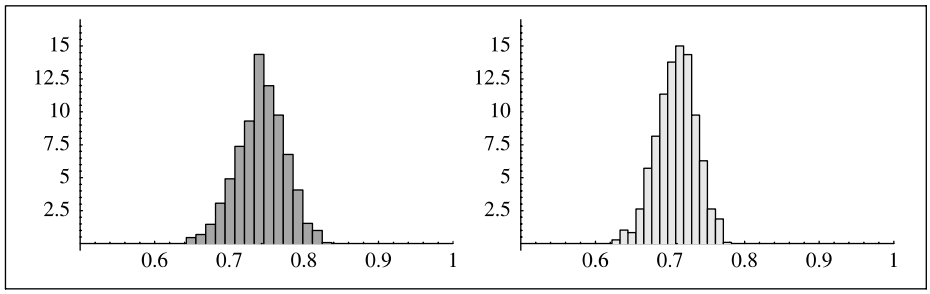

Figure 5. Histogram of samples of [p|y], under independent uniform priors for p (on [.5,1]), θ_1, and θ_2, for the rows of Table 1 and for the rows of the subset of Table 2 that restricts consideration to the columns e, de, la, que, no, l, com, és, molt, jo, si, and dix.

repeating the cluster analysis for the remaining ones, (namely, *e*, *de*, *la*, *que*, *no*, *l*, *com*, *és*, *molt*, *jo*, *si*, and *dix*), one finds 304 chapters attributed to the first author, that is 71.53% of the total. In this case, 91 chapters are misclassified by the change-point, of which 59 are attributed to the second author but located before chapter 382, while 32 are attributed to the first author but located after it.

Interestingly, the number of chapters misclassified by a hypothetical change-point is *minimized* exactly at the change-point estimates in Table 3. Not surprisingly, the marginal posteriors for (θ_1, θ_2) are similar to the ones for (θ_a, θ_b), and boxplots of samples of $[\log(\theta_{1j}/\theta_{2j})|y]$ are almost identical to Figures 3 and 4.

Given that the cluster analysis does not take into account the order of the observations, the close match between cluster and change-point partitions is very strong evidence in favor of the existence of a sudden change in style near chapters 371 and 382. Instances where the cluster and the change-point classification of a chapter disagree, might reveal chapters by an author in the part mainly written by the other one. In particular, chapters at the end of the book, like 403, 410 to 412, 424, 427 to 429, 431 to 439, 472 to 478, and 480 might actually have been written by Martorell, while earlier chapters like 2, 4, 28, 52 to 54, 84, 107, 144, 190, 191, 272, 306, and 349 might involve Galba. Observe that the density of misclassified chapters is larger after chapters 371 and 382 than before them.

The absence of long streaks of consecutive chapters that are misclassified by the main change-point is more in line with authors sparsely filling in material in the part mainly written by the other one, than with the hypothesis of the existence of a given fixed (small) number of change-points, and it thus vindicates the choice of cluster analysis instead of multiple change-point analysis.

5. FINAL COMMENTS

The statistical analysis presented here consistently detects a change in style near chapters 371–382, with a few chapters being misclassified by that change-point. That agrees very closely with the boundary detected in chapter 383 through the statistical analysis of the evolution of the diversity of the vocabulary reported by Riba and Ginebra (2000), and it is very much in line with the hypothesis supported by all the experts that attribute more credibility to the colophon of the book than to its dedicatory letter. Nevertheless, we want to make it clear that, even though our analysis supports the hypothesis of the existence of two authors, others might try to explain that stylistic boundary otherwise, and therefore we do not pretend to be closing this authorship debate whatsoever.

In practice, the assumption that the observations in the same change-point or cluster group are conditionally independent and identically distributed as a $\text{Mu}_{l-1}(N_i, \theta)$, with $\theta_i = \theta$ for all i in the group, might not be appropriate. Often observations in the same group have similar, but not identical, values of θ_i, and that leads to the data being more dispersed than anticipated by the model that assumes $\theta_i = \theta$ for all i in the group. Bayesian hierarchical models cover for that; for example, in the change-point hierarchical setting, one would assume that the θ_i before the change-point are exchangeable and come from a $\text{Di}_{l-1}(\alpha_b)$, and those after the change point are also exchangeable and come from a $\text{Di}_{l-1}(\alpha_a)$, where α_b and α_a are independent random variables with a known distribution. That would most likely add some uncertainty on the posterior distributions in Table 3. On this extension see, for example, Gelman, Carlin, Stern, and Rubin (2004, p. 128) or Quintana (1998).

Between change-point problems, that force consecutive observations to belong to the same group, and cluster analysis problems, that assign observations to groups without taking order into consideration, there is a range of problems where one incentives but does not force consecutive observations to belong to the same group. When authors intervene exchangeably all over the book, this can be modeled through hidden Markov models, where a Markov chain rules which multinomial distribution is active for each observation. For a Bayesian treatment of these models, see Robert, Celeux, and Diebolt (1993), Chib (1996), Robert and Casella (1999), and Robert, Ryden, and Titterington (2000). In our setting though, it is not realistic to assume that authors are exchangeable, and one would need more sophisticated models.

[Received February 2004. Revised September 2004.]

REFERENCES

Banfield, J. D., and Raftery, A. E. (1993), "Model Based Gaussian and non-Gaussian Clustering," *Biometrics*, 49, 803–821.

Barry, D., and Hartigan, J. A. (1993), "A Bayesian Analysis of Change-Point Problems," *Journal of the American Statistical Association*, 88, 309–319.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997), "Inference in Model-Based Cluster Analysis," *Statistics and Computing*, 7, 1–10.

Bernardo, J. M., and Girón, J. (1988), "A Bayesian Approach to Cluster Analysis," *Qüestiió*, 12, 97–112.

Besag, J., Green, P. J., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3–66.

Bhattacharyia, G. K., and Johnson, R. A. (1968), "Nonparametric Tests for Shift at an Unknown Time Point," *Annals of Mathematical Statistics*, 39, 1731–1743.

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65, 31–38.

Binongo, J. N. G. (1994), "Joaquin's Joaquinesquerie, Joaquinesquerie's Joaquin: A Statistical Expression of a Filipino Writer's Style," *Literary and Linguistic Computing*, 9, 267–279.

Brinegar, C. S. (1963), "Mark Twain and the *Quintus Curtius Snodgrass* Letters: A Statistical Test of Authorship," *Journal of the American Statistical Association*, 58, 85–96.

Broemeling, L. D. (1974), "Bayesian Inferences About a Changing Sequence of Random Variables," *Communications Statistics Theory and Methods*, 3, 243–255.

Bruno, A. M. (1974), *Toward a Quantitative Methodology for Stylistic Analysis*, Berkeley, CA: University of California Press.

Burrows, J. F. (1987), "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style," *Literary and Linguistic Computing*, 2, 61–70.

——— (1992), "Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information," *Literary and Linguistic Computing*, 7, 91–109.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992), "Hierarchical Bayesian Analysis of Change Point Problems," *Applied Statistics*, 41, 389–405.

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

——— (1996), "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models," *Journal of Econometrics*, 75, 79–97.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 48, 327–335.

Dellaportas, P. (1998), "Bayesian Classification of Neolithic Tools," *Applied Statistics*, 47, 279–297.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Diebolt, J., and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society*, Ser. B, 56, 363–375.

Fan, Y., and Brooks, S. P. (2000), "Bayesian Modelling of Prehistoric Corbelled Domes," *The Statistician*, 49, 339–354.

Ferreira, P. E. (1975), "A Bayesian Analysis of a Switching Regression Model," *Journal of the American Statistical Association*, 70, 370–374.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), London: Chapman Hall.

Gilbert, E. S. (1968), "On Discrimination Based on Qualitative Variables," *Journal of the American Statistical Association*, 63, 1399–1418.

Ginebra, J., and Cabos, S. (1998), "Anàlisi Estadística de l'estil Literari; Aproximació a l'autoria del Tirant lo Blanc" (in Catalan), *Afers*, 29, 185–206.

Gnanadesikan, R. (1997), *Methods of Statistical Data Analysis of Multivariate Observations* (2nd ed.), New York: Wiley.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Greenacre, M. (1988), "Clustering the Rows and Columns of a Contingency Table," *Journal of Classification*, 5, 39–51.

Gordon, A. D. (1999), *Classification* (2nd ed.), London: Chapman and Hall.

Hilton, M. L., and Holmes, D. I. (1993), "An Assessment of Cumulative Control 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.

——— (1971), "Inference in Two Phase Regression," *Journal of the American Statistical Association*, 66, 736–743.

Hinkley, D. V., and Hinkley, E. A. (1970), "Inference About the Change-Point in a Sequence of Binomial Variables," *Biometrika*, 57, 477–488.

Holmes, D. I. (1985), "The Analysis of Literary Style. A Review," *Journal of the Royal Statistical Society*, Ser. A, 148, 328–341.

——— (1992), "A Stylometric Analysis of Mormon Scripture and Related Texts," *Journal of the Royal Statistical Society*, Ser. A, 155, 91–120.

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding Groups in Data*, New York: Wiley.

Krishnaiah, P. R., and Miao, B. Q. (1988), "Review About Estimation of Change-Points," in *Handbook of Statistics, Vol. 7*, eds. P. R. Krishnaiah and C. R. Rao, Amsterdam: Elsevier Science, pp. 375–402.

Lavine, M., and West, M. (1992), "A Bayesian Method for Classification and Discrimination," *Canadian Journal of Statistics*, 20, 451–461.

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

Lee, C. B. (1998), "Bayesian Analysis of a Change-Point in Exponential Families with Applications," *Computational Statistics and Data Analysis*, 27, 195–208.

Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 95, 958–966.

McLachlan, G. J., and Basford, K. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

Mendelhall, T. C. (1887), "The Characteristic Curves of Composition," *Science*, IX, 237–249.

Moore, D. H. (1973), "Evaluation of Five Discrimination Procedures for Binary Variables," *Journal of the American Statistical Association*, 68, 399–404.

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

Mosteller, F., and Wallace, D. L. (1964, 1984), *Applied Bayesian and Classical Inference; the Case of The Federalist Papers* (1st and 2nd ed.), Berlin: Springer-Verlag.

Oakes, M. P. (1998), *Statistics for Corpus Linguistics*, Edinburg: Edinburgh University Press.

Peng, R. D., and Hengartner, N. W. (2002), "Quantitative Analysis of Literary Style," *The American Statistician*, 56, 175–185.

Pettitt, A. N. (1989), "Change-Point Problem, in *Encyclopedia of Statistical Sciences*, Volume S, New York: Wiley, pp. 26–31.

Phillips, D. B., and Smith, A. F. M. (1996), "Bayesian Model Comparison via Jump Diffusions," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman Hall, pp. 215–239.

Quintana, F. A. (1998), "Nonparametric Bayesian Analysis for Assessing Homogeneity in $k \times 1$ Contingency Tables with Fixed Right Margin Totals," *Journal of the American Statistical Association*, 93, 1140–1149.

Quintana, F. A., and Iglesias, P. L. (2003), "Bayesian Clustering and Product Partition Models," *Journal of the Royal Statistical Society*, Ser. B, 65, 557–574.

Riba, A., and Ginebra, J. (2000), "Riquesa de Vocabulari i Homogeneitat d'estil en el Tirant lo Blanc" (in catalan), *Revista de Catalunya*, 13, 99–118.

——— (2005), "Change-Point Estimation in a Multinomial Sequence and Homogeneity of Literary Style," *Journal of Applied Statistics*, 32, 61–74.

Riquer, M. (1990), *Aproximació al Tirant lo Blanc* (in Catalan), Barcelona: Quaderns Crema.

Robert, C. P. (1996), "Mixtures of Distributions: Inference and Estimation," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 441–464.

Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New

York: Springer-Verlag.

Robert, C. P., Celeux, G., and Diebolt, J. (1993), "Bayesian Estimation of Hidden Markov Models: A Stochastic Implementation," *Statistics and Probability Letters*, 16, 77–83.

Robert, C. P., Ryden, T., and Titterington, D. M. (2000), "Bayesian Inference in Hidden Markov Models Through the Reversible Jump Markov Chain Monte Carlo Method," *Journal of the Royal Statistical Society*, Ser. B, 62, 57–75.

Sen, A. K., and Srivastava, M. S. (1973), "On Multivariate Tests for Detecting Change in Mean," *Sankhya A*, 35, 173–186.

Smith, A. F. M. (1975), "A Bayesian Approach to Inference About a Change-Point in a Sequence of Random Variables," *Biometrika*, 63, 407–416

——— (1980), "Change-Point Problems: Approaches and Applications," in *Proceedings of the 2nd Bayesian Meeting in Valencia*, eds. J. Bernardo and M. De Groot, pp. 83–98.

Smith, A. F. M., and Cook, D. G. (1980), "Switching Straight Lines: A Bayesian Analysis of Some Renal Transplant Data," *Applied Statistics*, 29, 180–189.

Smith, M. W. A. (1983), "Recent Experience and New Developments of Methods for the Determination of Authorship," *Association for Literary and Linguistic Computing Bulletin*, 11, 73–82.

Stephens, D. A. (1994), "Bayesian Retrospective Multiple-Changepoint Identification," *Applied Statistics*, 43, 159–178.

Swartz, T., Haitovsky, Y., Vexler, A., and Yang, T. (2004), "Bayesian Identifiability and Misclassification in Multinomial Data," *The Canadian Journal of Statistics*, 32, 285–302.

Vargas Llosa, A. (1991), *Carta de Batalla por Tirant lo Blanc*, (in Spanish), Barcelona: Seix Barral.

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," *Communications in Statistics, Computations and Simulation*, 19, 603–618.

Wolfe, D. A., and Schechtman, E. (1984), "Nonparametric Statistical Procedures for the Change Point Problem," *Journal of Statistical Planning and Inference*, 9, 389–396.

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: Herman Chernoff Festschrift*, New York: Academic Press, pp. 245–269.