Revista Colombiana de Estadística
July 2016, Volume 39, Issue 2, pp. 205 to 227
DOI: http://dx.doi.org/10.15446/rce.v39n2.50151

Bayesian Analysis of the Heterogeneity of Literary Style

Análisis bayesiano de la heterogeneidad del estilo literario

Marti Font[a], Xavier Puig[b], Josep Ginebra[c]

Statistics and O. R. Department, Technical University of Catalonia, Barcelona, Spain

Abstract

We proposed statistical analysis of the heterogeneity of literary style in a set of texts that simultaneously use different stylometric characteristics, like word length and the frequency of function words. The data set consists of several tables with the same number of rows, with the i-th row of all tables corresponding to the i-th text. The analysis proposed clusters the rows of all these tables simultaneously into groups with homogeneous style, based on a finite mixture of sets of multinomial models, one set for each table.

Different from the usual heuristic cluster analysis approaches, our method naturally incorporates the text size, the discrete nature of the data, and the dependence between categories in the analysis. The model is checked and chosen with the help of posterior predictive checks, together with the use of closed form expressions for the posterior probabilities that each of the models considered to be appropriate. This is illustrated through an analysis of the heterogeneity in Shakespeare's plays, and by revisiting the authorship-attribution problem of *Tirant lo Blanc*.

***Key words*:** Authorship, Cluster analysis, Multinomial distribution.

Resumen

Se propone un análisis estadístico para modelar la heterogeneidad del estilo literario en un conjunto de textos, para ello se utilizan simultáneamente diferentes características estilométricas, como longitud de palabra y la frecuencia de palabras *función*. Los datos consisten en varias tablas con el mismo número de filas, donde la fila i-ésima corresponde al texto i-ésimo. El análisis propuesto agrupa las filas de todas estas tablas simultáneamente en

[a]Professor. E-mail: marti.font@upc.edu
[b]Professor. E-mail: xavier.puig@upc.edu
[c]Professor. E-mail: josep.ginebra@upc.edu

grupos de estilo homogéneo, en base a una mezcla finita de modelos multinomiales.

El modelo propuesto tiene la ventaja sobre los análisis de conglomerados heurísticos habituales, de incorporar de forma natural el tamaño del texto, la naturaleza discreta de los datos y la dependencia entre las categorías. El modelo se selecciona y válida con la ayuda de simulaciones de la distribución predictiva a posteriori, junto con el uso de las expresiones en forma cerrada para la probabilidad a posteriori de cada uno de los modelos de mezcla considerados. Todo ello se ilustra a través de un análisis de la heterogeneidad en las obras de Shakespeare, y revisitando el problema de atribución de autoría del texto *Tirant lo Blanc*.

***Palabras clave*:** análisi de conglomerados, atribución, distribución multinomial.

1. Introduction

The statistical analysis of literary style has often been used to characterize the style of texts and authors, and to help settle authorship-attribution problems both in the academic as well as in the legal context. Mendenhall (1887, 1901) has used word length and sentence length to characterize literary style. Other characteristics widely used for this purpose have been the proportion of nouns, articles, adjectives or adverbs; the frequency of use of function words, which are independent of the context, or of characters; and the richness and diversity of the vocabulary used by the author. Good reviews on the statistical analysis of literary style can be found in Holmes (1985, 1994, 1998, 1999) and Stamatatatos (2009).

The range of statistical methods used in this setting is wide, and the most often involve the use of classification tools. In authorship-attribution problems the researcher has a set of candidate authors and a list of texts from each of these authors, and it is necessary to assign texts of an unknown author to one of the authors in the training set by comparing their style to the one of the training texts. In this settings, discriminant analysis should be used. A recent presentation of the Bayesian approach to this problem can be found in Puig, Font & Ginebra (2016).

In the analysis of the heterogeneity of literary style that is tackled in this paper, the setting is a lot less structured because we did not have a reference set of candidate authors and of training texts, and so we used of cluster analysis.

The goal of cluster analysis is to partition observations (texts) into meaningful subgroups, without assuming much about the number of subgroups and about the composition of the groups. Most of the literature on cluster analysis is devoted to continuous data and uses ad hoc heuristic partitioning algorithms that tend to be easy to apply and work well, but do not allow to be assessed cluster uncertainties and do not provide inference based methods to choose the number of clusters and allocate individual observations to clusters. Good introductions to cluster analysis can be found in are Greenacre (1988), Kaufman & Rousseeuw (1990), Gnanadesikan (1997), and Gordon (1999).

Model based clustering assumes that observations come from a population with several subpopulations, and the overall population is modeled through a finite mixture of the subpopulation models. Bayesian model based cluster analysis provides a complete probabilistic framework for the problem by assuming a finite mixture model under which observations belonging to the same cluster have the same distribution, and then the mixed distributions are estimated and observations to these component distributions are assigned. Model based approaches simultaneously group objects and estimate the component parameters, and this avoids the biases appearing if it is done separately. These methods also have the advantage of providing a measure of the uncertainty by allocating individual observations into clusters, and by casting the choice of the number of clusters and hence of component distributions as a statistical model selection problem.

For early examples of the use of Bayesian model based cluster analysis, mostly using mixtures of multivariate normal distributions, see Murtagh & Raftery (1984), Banfield & Raftery (1993), Fernandez & Green (2002), and Fernandez & Green (2002).

To help settle the debate around the authorship of *Tirant lo Blanc*, Giron, Ginebra & Riba (2005) explored the heterogeneity of its style by carrying out a Bayesian model based cluster analysis of word length and of the frequency of the most frequent words in its chapters. The data consisted of two contingency tables of ordered rows, with the i-th row in both tables corresponding to the i-th chapter of the book, and the cluster analysis of the rows of each one of these two tables was carried out separately based on a finite mixture of multinomial models. By using these models to implement a cluster analysis, the texts classified based on the whole vector of word length or of function word counts instead of using only individual counts. This also has the advantage over heuristic and/or normal based clustering approaches in that it naturally incorporates the text size, the discrete nature of the data and the dependence between categories in the analysis.

This analysis, based on finite mixtures of multinomial models, is generalized by:

1. carrying out a single cluster analysis that uses more than one stylometric characteristic and, by treating a set of more than one vector of counts as an observation,

2. by incorporating a model-checking stage that compares the realization of statistics in the data with their realization in predictive simulations from the models, and

3. by providing closed form expressions for the exact calculation of the probabilities of the models considered to be correct, that are to be used to select models. There are used instead of the approximations based on the BIC that are typically used in model based cluster analysis.

The combination of the model-checking and model selection stages will help determine the number of mixture components required by the data, and hence, the number of clusters.

As a by product of the model-checking stage, the analysis allows us to check whether finite mixtures of a small number of purely multinomial models are flexible enough to capture all the variability in the data. If they are not, it would be necessary to resort to more complicated finite mixtures of sets of continuous mixtures of multinomial models instead.

To illustrate the methodology, we used two examples. The first case study explores the heterogeneity of style in the plays in the *first folio edition* of Shakespeare's drama, and in the second, the authorship-attribution problem of *Tirant lo Blanc* is explored. In the examples, the analysis will be mostly exploratory, and we will not assess if the heterogeneities found are linked to differences in authorship could be explained by differences in chronology, genre or topic. Without making a list of candidate authors and of training texts explicit, there is no legitimate statistical way of going beyond proposing tentative explanations for the heterogeneities detected in the corpus.

2. Description of the Data

The methodology presented here combines as many stylometric characteristics is necessary. All the characteristics will involve counting features that have a fixed number of categories. This includes, for example, counting characters, words or sentences of certain lengths, function words, nouns or adjectives. Therefore, data will consist of a set of tables with the same number of rows, with one table for each characteristic. We use *word length* and the count of the most frequent *function words* as illustrating examples.

Early uses of word length can be found in Mendenhall (1901), Mosteller & Wallace (1984), Brinegar (1963), Bruno (1974), Williams (1975), Morton (1978), Smith (1983) and in Hilton & Holmes (1993). Word length is rarely useful in authorship attribution of texts written in English, but Giron et al. (2005) found it to be very useful in the authorship attribution of a text in Catalan. Furthermore, in the first case study, we found that word length is a good discriminator well between Shakespeare's comedies on one side and histories and tragedies on the other. Therefore word length could be useful to detect heterogeneities of style in English texts that are not necessarily linked to differences in authorship.

The frequency of use of function words has proved to be one of the best tools when it comes to discriminating styles. Recent examples of the use of function words can be found in Holmes (1992), Binongo (1994), Oakes (1998), Zhao & Zobel (2005), Miranda-Garcia & Calle-Martin (2007), Luyckx (2010) and Rybicki & Eder (2011).

In those cases in which the analysis of word length and the analysis of word counts separately used lead to very different results, their combination will be problematic. However when they lead to similar results when used separately, as was found to be the case in *Tirant lo Blanc* by Giron et al. (2005), their combination in a single analysis is warranted. This is because it helps to reduce the uncertainty in the classification of texts into clusters.

If a researcher decides to simultaneously analyze word length and function word counts in *Tirant lo Blanc* it leads to the simultaneous analysis of the 487×10 table of word length counts and of the 487×12 table of counts of twelve of the most frequent function words partially presented in Table 1.

TABLE 1: Extract of the table of word length counts in the chapters of *Tirant lo Blanc*, and of the table of counts of twelve of the most frequent function words in them. N_i^1 is the total number of words and $\overline{wl}_i$ the average word length.

Word length counts												
Chapter	1	2	3	4	5	6	7	8	9	10+	N_i^1	$\overline{wl}_i$
1	21	59	44	19	33	20	16	17	9	17	285	4.47
2	53	113	80	49	52	33	28	36	16	16	476	4.14
...	...	...	...	...	...	...	...	...	...	...	...	...
487	48	49	62	53	41	36	21	9	16	13	348	4.20
Most frequent word counts												
Chapter	e	de	la	que	no	l	com	molt	és	jo	si	dix
1	12	15	9	8	1	7	2	1	6	0	3	0
2	26	28	19	9	3	2	3	8	3	1	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...
487	29	13	8	10	2	10	3	9	0	0	0	0

In general, for each chapter i in a book (or act of a play) with $i = 1, \ldots, n$, and each stylometric characteristic, r, with $r = 1, \ldots, R$, one has a vector valued categorical observation, $y_i^r = (y_{i1}^r, \ldots, y_{ik(r)}^r)$, where $k(r)$ denotes the number of categories of the r-th characteristic. This vector, y_i^r, becomes the i-th row of the r-th table considered.

In the *Tirant lo Blanc* example, y_i^1 is the ten dimensional vector of word length counts of its i-th chapter, and y_i^2 is the twelve dimensional vector of function word counts in that chapter. More generally, this leads to a set of R different $n \times k(r)$ tables: one table for each characteristic. The set of all the n rows in the r-th table will be denoted by $y^r = (y_1^r, \ldots, y_n^r)$, and the set of all the R tables will be denoted by $y = (y^1, \ldots, y^R)$. The goal is to cluster the rows of all these tables simultaneously into S different groups with homogeneous style, assuming that the rows in a group are multinomially distributed.

One of the main shortcomings of the heuristic based cluster analysis approach typically used in stylometry, such as the ones based on PCA, k-means or hierarchical methods, is that they implicitly assume data to be continuous or are at least tailored to work best when data is continuous. However, stylometric data is mostly categorical, and the methodology to be used should move in the direction that address the specificities of that kind of data.

Specifically; most of these ad-hoc heuristic methods have problems taking into account that texts of different lengths have different amounts of information regarding the style of their author and, hence, they should be weighted differently in the analysis. These basic methods also have problems taking into consideration the dependence present between the category counts of the same stylometric characteristic.

The cluster analysis proposed next, based on carefully modeling the data probabilistically using mixtures of multinomial models, avoids the continuity assumption, and it naturally weights texts according to text size. Furthermore, by assuming the observations in each cluster to be multinomially distributed, one also naturally takes into account the dependence between counts of categories from the same characteristic.

3. Description of the Multinomial Cluster Model

The i-th row of the r-th table is assumed to be multinomially distributed, $\text{Mult}(N_i^r, \theta_i^r)$, where $\theta_i^r = (\theta_{i1}^r, \ldots, \theta_{ik(r)}^r)$ is such that $\sum_{j=1}^{k(r)} \theta_{ij}^r = 1$, where θ_{ij}^r is the probability of the j-th category for the i-th row and the r-th characteristic, and $N_i^r = \sum_{j=1}^{k(r)} y_{ij}^r$. If all the chapters of the book or acts in the plays shared the same style, one might expect the distribution of all the n rows for any given characteristic to remain the same, in which case they could all be modeled as a random sample of a single $\text{Mult}(N_i^r, \theta^r)$ distribution.

Instead, if the style in the n chapters or acts was not homogeneous, but these chapters grouped themselves in S different styles, maybe because they had been written by S different authors, then the n rows of the r-th table, $y^r = (y_1^r, \ldots, y_n^r)$, could be considered to be conditionally independent and modeled through a finite mixture of S multinomial distributions. They would have the probability density function (pdf):

$$p(y^r \mid \omega, \theta_1^r, \ldots, \theta_S^r) = \prod_{i=1}^{n} \sum_{s=1}^{S} \omega_s \text{Mult}(N_i^r, \theta_s^r), \tag{1}$$

where $\theta_s^r = (\theta_{s1}^r, \ldots, \theta_{sk(r)}^r)$ which determines the distribution of the rows in the s-th cluster of the r-th table, and where $\omega = (\omega_1, \ldots, \omega_S)$ is a set of weights, with $0 \leq \omega_s \leq 1$ and $\sum_{s=1}^{S} \omega_s = 1$, determining the proportion of rows (chapters or acts) belonging to each cluster.

To be able to allocate rows into clusters, which is an essential feature in cluster analysis, it is necessary to introduces a vector of unobserved (latent) categorical variables $\zeta = (\zeta_1, \ldots, \zeta_n)$, where ζ_i takes values in $\{1, \ldots, S\}$, and it is such that $\zeta_i = s$ whenever the i-th row belongs to the s-th cluster. Here, the ζ_i are assumed to be conditionally independent, and hence:

$$p(y^r, \zeta \mid \omega, \theta^r) = \prod_{i=1}^{n} \omega_{\zeta_i} \text{Mult}(N_i^r, \theta_{\zeta_i}^r), \tag{2}$$

where $\theta^r = (\theta_1^r, \ldots, \theta_S^r)$ is the set of multinomial probabilities for the r-th table.

The latent variable ζ assigning chapters or acts into clusters does not depend on r, and hence, it takes a common value for all the stylometric characteristics considered. That is, the i-th rows in all the tables are always allocated into the same cluster. Since the posterior distribution of ζ_i, $\pi(\zeta_i \mid y)$ is the posterior

probability that the i-th row belongs to each cluster, we assign the i-th row to the mode of that distribution. Conditional on $\theta = (\theta^1, \ldots, \theta^R)$ and on ζ, the distribution of the different tables are considered to be independent.

In Bayesian statistics, one needs to choose a distribution for the parameters of the model that captures what one knows about them before observing the data. This is denoted as the prior distribution. Here, that prior distribution will assume that all vectors of probabilities across clusters and tables, θ_s^r for $s = 1, \ldots, S$ and $r = 1, \ldots, R$, are independent, and that the θ_s^r are distributed according to a Dirichlet$(a_{s1}^r, \ldots, a_{sk(r)}^r)$ distribution. The weights ω determine the relative sizes of the clusters that are assumed to be Dirichlet$(b_1, \ldots, b_S)$ distributed, and independent of $(\theta_1, \ldots, \theta_S)$.

Depending on the values chosen for $(a_{s1}^r, \ldots, a_{sk(r)}^r)$, the prior distribution of θ_s^r will be more or less informative. In particular, the expected value of θ_s^r will be $(a_{s1}^r, \ldots, a_{sk(r)}^r)/(\sum_{j=1}^{k(r)} a_{sj}^r)$, and one can choose the a_{sj}^r to reflect the fact that some categories in a table might be known to have larger probabilities than others. Also, the larger $\sum_{j=1}^{k(r)} a_{sj}^r$ the smaller the variances of θ_{sj}^r and the more informative the prior distribution chosen for θ_s^r.

The $R = 1$ and $S = 2$ special case of the model proposed here is the one used in Giron et al. (2005), and the $R = 1$ and any S case is used in Puig & Ginebra (2014). In our examples all the $(a_{s1}^r, \ldots, a_{sk(r)}^r)$ and $(b_1, \ldots, b_S)$ are set to be equal to $(1, \ldots, 1)$, which is equivalent to assuming a uniform distribution on the simplex. Given that the total number of words in the texts will be a lot larger than the number of categories, $k(r)$, the influence of the uniform prior on the conclusions will be a lot weaker than the influence of the data through the likelihood. Choosing a prior with a value for $(a_{s1}^r, \ldots, a_{sk(r)}^r)$ different to $(1, \ldots, 1)$, but with a similar value of $\sum_{j=1}^{k(r)} a_{sj}^r$, will not change the conclusions reached.

Bayesian statistics combines the distribution chosen for the parameters before obtaining the data (the prior distribution) with the data, to compute an updated distribution that incorporates the information contributed by the data. In our case, that updated posterior distribution is too complicated to be computed analytically. Instead, one can update the model and simulate from it with WinBugs (see Lunn, Jackson, Best, Thomas & Spiegelhalter 2013). The convergence of the chains has been assessed through the visual inspection of the sample traces and the monitoring of diagnostic measures. For each model, four chains with different initial values have been run until all of their ergodic means converged to the same values.

4. The choice of the Number of Clusters

A difficulty for the heuristic clustering algorithms is that they often lack a statistically grounded method to determine the number of clusters. Instead, under model based clustering the choice of the number of clusters, S, coincides with the choice of model.

The safest way to build a model is through the iterative use of model checking tools that help discover aspects of reality not adequately captured by the models, and also suggest ways of improving them. Here this will be done through posterior predictive checks.

To help support the model choice, and, hence, the choice of the number of clusters, one can also resort to formal model selection methods that are based on the computation of the posterior probability that each one of the models considered is the one generating the data. Here we will be able to take advantage of the fact that under our finite mixtures of multinomial models setting there is a closed form expression for these model posterior probabilities, and, therefore, one does not have to resort to heuristic model selection criteria instead.

Cluster analysis is useful only when the answer contains a relatively small number of clusters, and, hence, it will typically be better to resolve this issue with an approximate model that has a small number of clusters and explains a large portion of the variability than with a model that is "true" and captures all the variability but requires a large number of clusters.

4.1. Choice of s Through Model-Checking

Bayesian models can be assessed and chosen based on whether it is plausible that they are able to simulate data like the one observed in reality. Following the example of Gelman, Carlin, Stern & Rubin (2004), we will graphically compare the set of R observed tables, with analogous sets of tables simulated from the posterior predictive distribution of the models.

To compare the table with the word length data to the corresponding tables with the data simulated from the predictive distributions of the model, these tables are summarized by the proportion of words of L letters that there are in each chapter or act for $L = 1, \ldots, 9$ and for $L > 9$. We also summarize these tables using average word length, the ratio between the number of words with more than 5 and of less than 6 letters, and by the first correspondence analysis components in each table. To compare the table that has the observed word counts with the corresponding simulated tables, they are summarized through the frequency of the appearance of each one of these words separately, and through the first correspondence analysis components in each table.

A sample of these predictive comparisons will be presented in the first case study. We do not report on the predictive checks for the other example for the sake of brevity. For more examples of posterior predictive checks used to assess Bayesian models in the context of the analysis of literary style, see Font, Puig & Ginebra (2013), and for similar examples in the context of choosing the number of clusters, see Puig & Ginebra (2014a, b).

4.2. Choice of s Through Model Selection

The formal way to select a model is through the posterior probability of each model, $P(M_S \mid y)$, which is the probability that the S-cluster model, M_S, is the

one generating the data, and will be assessed after the data has been observed. It can be computed through:

$$P(M_S \mid y) = \frac{P(M_S)P(y \mid M_S)}{\sum_{s=1}^{S_T} P(M_s)P(y \mid M_s)}, \tag{3}$$

where $P(M_S)$ is the prior probability assigned to M_S, (i.e., the probability that this model is correct, assessed before the data is available), where$P(y \mid M_S)$ is the marginal likelihood of M_S, and where S_T is the largest number of clusters that one is willing to consider. If all models were equally likely a priori, the larger $P(y \mid M_S)$, the more attractive the M_S would be. However there is a big debate on how prior probabilities on model space should be chosen, due to the large difference in complexity between models (see, e.g., Casella, Moreno & Giron 2014).

Most often, computing $P(y \mid M_S)$ exactly is too complicated, and its logarithm is approximated through the BIC, as in Fraley & Raftery (2002). Alternatively, it its possible to estimate $P(y \mid M_S)$ through simulations used to update the model, as in Gelfand & Dey (1994).

However, in our multinomial mixture setting, there is a closed form expression for $P(y \mid M_S)$ that allows us to compute exactly these marginal likelihoods. In this way, it is not necessary to use BIC or MCMC approximations. In this paper, we will compute the marginal likelihoods of the M_S model after conditioning on an estimate of the latent allocation parameter, ζ, which leads to computing posterior probabilities of the models conditional on these $\hat{\zeta}$ through (3.1). The ζ will be estimated by using its marginal posterior mode.

It can be proved that the marginal likelihood under the single cluster model, M_1, is:

$$p(y \mid M_1) = \prod_{r=1}^{R} \frac{\prod_{i=1}^{n} N_i^r!}{\prod_{j=1}^{k(r)} \prod_{i=1}^{n} y_{ij}^r!} \frac{\prod_{j=1}^{k(r)} (\sum_{i=1}^{n} y_{ij}^r)!}{(\sum_{i=1}^{n} N_i^r)!} \text{Dir-Mult}(y_r; \sum_{i=1}^{n} N_i^r, a^r), \tag{4}$$

where y_r is the vector of aggregated counts of the r-th table, $y_r = (\sum_{i=1}^{n} y_{i1}^r, \ldots, \sum_{i=1}^{n} y_{ik}^r)$, and where Dir-Mult$(x; N, a)$ denotes the pdf of a Dirichlet-multinomial distribution with parameters N and $a = (a_1, \ldots, a_k)$ evaluated at $x = (x_1, \ldots, x_k)$,

$$\text{Dir-Mult}(x; N, a) = \frac{N!\Gamma(\sum_{j=1}^{k} a_j)}{\Gamma(N + \sum_{j=1}^{k} a_j)} \prod_{j=1}^{k} \frac{\Gamma(x_j + a_j)}{x_j!\Gamma(a_j)}. \tag{5}$$

In general, the marginal likelihood under the single table S-cluster model, M_S, becomes:

$$\begin{aligned} p(y \mid M_S) = \prod_{r=1}^{R} \frac{\prod_{i=1}^{n} N_i^r!}{\prod_{j=1}^{k(r)} \prod_{i=1}^{n} y_{ij}^r!} \prod_{s=1}^{S} \frac{\prod_{j=1}^{k(r)} \left(\sum_{i=1}^{n} y_{ij}^r I_{[\hat{\zeta}_i = s]}\right)!}{\left(\sum_{i=1}^{n} N_i^r I_{[\hat{\zeta}_i = s]}\right)!} \times \\ \text{Dir-Mult}(y_r^{[\hat{\zeta}_i = s]}; \sum_{i=1}^{n} N_i^r I_{[\hat{\zeta}_i = s]}, a_s^r), \end{aligned} \tag{6}$$

where $I_{[\hat{\zeta}_i=s]}$ denotes the indicator function that is 1 when the i-th observation is estimated to belong to the s-th cluster and it is 0 otherwise, and where $y_r^{[\hat{\zeta}_i=s]}$ denotes the vector of aggregated counts of all the observations estimated to belong to the s-cluster, $y_r^{[\hat{\zeta}_i=s]} = (\sum_{i=1}^n y_{i1}^r I_{[\hat{\zeta}_i=s]}, \ldots, \sum_{i=1}^n y_{ik}^r I_{[\hat{\zeta}_i=s]})$.

In the case studies that follow, these $p(y \mid M_S)$ will be used to assess the S-cluster models when $S = 1, 2, 3, 4$ to help choose the number of clusters. It should be kept in mind that the goal is to explain a large portion of the variability with a small number of clusters. A model that explains all that variability should not be found at the cost of using a lot of clusters.

It is important to note that adopting the formal Bayesian approach to model choice presented here does not help identify where models fail, and when they fail. Hence, computing the posterior probabilities of all the models under consideration in indicated here, does not relieve us of having to check models on the side: as is described in Section 4.1.

5. CaseStudy 1: Shakespeare's Drama

William Shakespeare (1564-1616) is regarded by many to be the greatest writer of English literature. Very little is known about his personal life, which has fueled a debate around the authorship of plays and poems that are attributed to him. Even though only a minority of the experts question his authorship, some claim that the true author of some or all of the works attributed to him could be Francis Bacon, Cristopher Marlowe, Ben Johnson, Sir Walter Raleigh or Edward de Vere. This debate has been on going for more than 150 years, and far too many people have contributed to it for it to be adequately summarized here. For recent overviews of the debate see, for example, Hope (1994, 2010), Edmondson & Wells (2013) or Shahan & Waugh (2013).

Statistical analysis of the literary style in Shakespeare's drama also started a long time ago. Mendenhall (1901) is one of the earliest examples of the use of statistics to compare the style of Shakespeare's plays with the style of some of his contemporaries, such as Marlowe and Bacon. He found that the word length distribution in Shakespeare's plays was extremely close to of Marlowe plays. The list of contributions to the quantitative analysis of the style in texts linked to Shakespeare is too long to be adequately described here.

The type of statistical analysis carried out next is different to the statistical analysis that has carried out so far on Shakespeare's drama in two main ways. The first difference arises from the fact that here we are trying to identify heterogeneities in Shakespeare's drama, irrespective of whether they are linked to authorship differences or not, while the literature on Shakespeare's drama has understandably focused mainly on authorship attribution issues. The second difference with other published statistical analysis of Shakespeare's drama, is that they rely on the use of training texts of undisputed authorship to help determine the authorship of disputed texts, we, however do not.

To explore the heterogeneity of style in Shakespeare's drama, here a cluster analysis is carried out on data we collected on the 35 plays gathered in the first printing of the *first folio edition* of Shakespeare's plays that were published posthumously in 1623. That edition included fourteen comedies, ten histories, and eleven tragedies, and it is the only reliable version for about twenty of these plays. Common wisdom supports the idea that some of the plays, especially the early histories, might have been revised by other writers. *Troilus and Cressida* did not appear in the first printing of that edition and *Pericles and the two noble kinsmen* did not appear in any of its printings. There have not been included in this study even though they are also attributed to Shakespeare.

In the analysis, plays are broken down into five acts each, and, hence, a total of 175 textual units are considered. The goal of the analysis is to check whether acts naturally cluster themselves together into more than one cluster when we takes, into account word length and the frequency of the twenty most frequent function words in them. Hence, data will consist of a 175×10 table with word length counts, and of a 175×20 table with the twenty most frequent word counts.

To choose the number of clusters, we need to assess whether the models involved capture the relevant features in the data. Figure 1, as example of this exercise, compares the observed proportion of words with *one, two, three, nine* and of *more than nine* letters in these 175 acts, the average word length, the ratio of the number of long words, and the short words with the ones corresponding to a sample simulated from the posterior predictive distribution under the one–, the two–, and the three-cluster models. The data plots in the left column of Figure 1 correspond to the actual plays by Shakespeare, while the data plots on the remaining three columns of that figure correspond to data replicates obtained from the three simplest multinomial mixture models.

Figure 2 compares the frequency of *the*, *and*, *I*, *you*, *it*, *your* and *his* that are observed in Shakespeare's plays, in the left column with the corresponding frequencies in a sample simulated from the same multinomial mixture models in the remaining three columns. Figures 1 and 2 also compare the first correspondence analysis component by summarizing the two tables of data considered here with the components summarizing analogous tables obtained by simulating from these models.

Note, for example, that the average word length tends to be smaller and the proportion of one and three lettered words tends to be larger for comedies than for histories or tragedies, while, for example, in there simple of plays the use of the words *I* and *you* tends to be more frequent.

We now have to check whether either one of the one–, two– or three-cluster models considered in Section 3 capture the patterns in Figures 1 and 2 adequately or not. Figures 1 and 2, and many other posterior predictive checks additionally made are, not reported here. However they all indicate that these finite mixtures of multinomial models are able to reproduce most of the variability in the data. To choose among the one–, the two– and the three-cluster models, several of the statistics in Figures 1 and 2 indicate that at least three clusters are needed to capture the variation in the levels of these statistics.

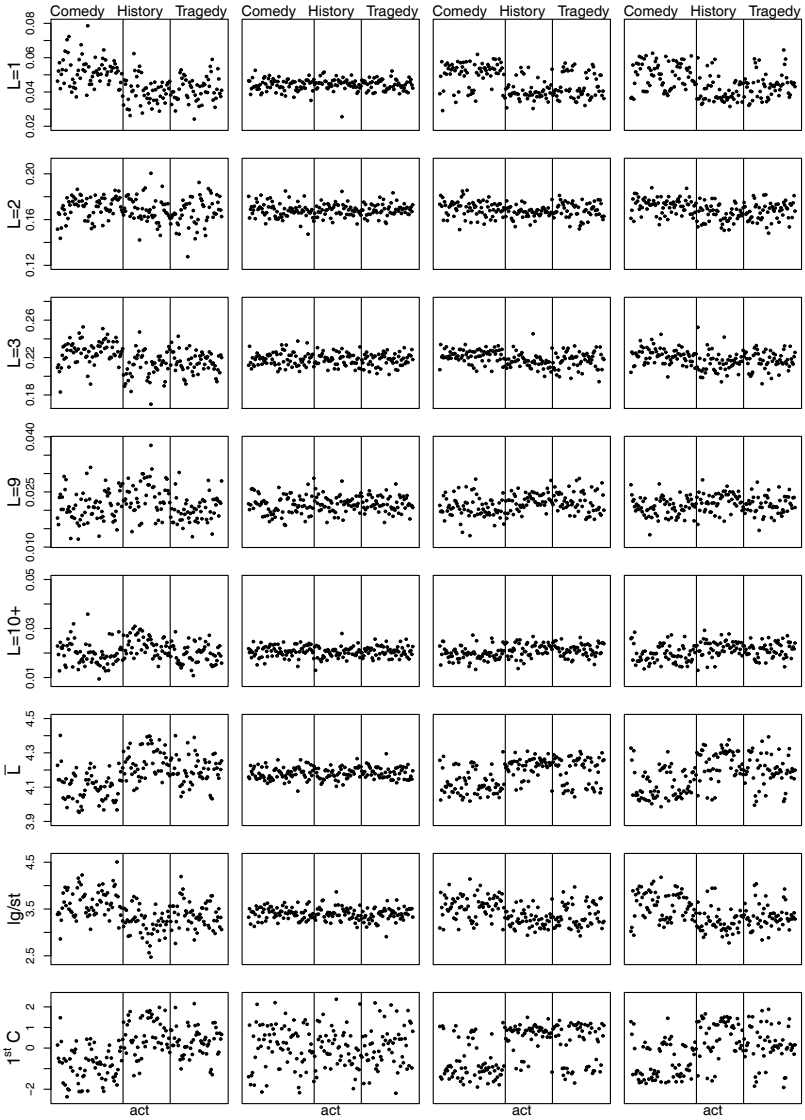

FIGURE 1: The left column shows, proportion whit words of one, two, three, nine and more than nine letters, the average word –lengths, the ratio between the number of long and of short words in the acts of the plays in Shakespeare's drama, and the first correspondence analysis component of the table of word lengths. Next to each of these plots, posterior predictive replicates are shown under the one–, two– and three-cluster models.

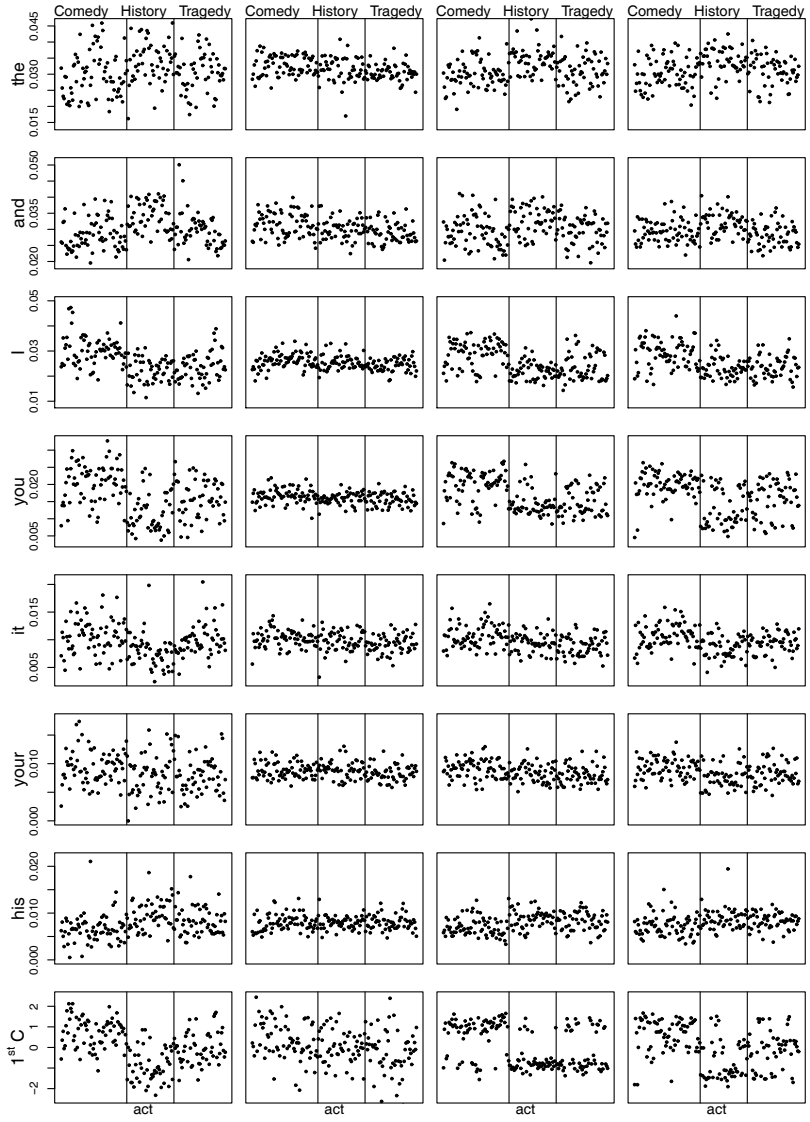

FIGURE 2: The left column shows, the frequency *the*, *and*, *I*, *you*, *it*, *your* and *his* in the as well the plays in the *first folio edition* of Shakespeare, first correspondence analysis component of the table with the twenty most frequent word counts. Next to each of these plots, posterior predictive replicates under the one-, two- and three-cluster models.

Here the natural logarithm of $P(y \mid M_S)$ under the one–, two–, three– and four-cluster models are equal to -25488.4, -23608.0, -22988.9 and -22677.3, respectively. If we compute the posterior probabilities that each one of these four cluster models is the correct one through (3), we have to choose the four-cluster model. However, if we penalizes models with more clusters by assigning them much smaller prior probabilities, as recommended by Casella et al. (2014), we will settle for the two– or three– cluster models. In fact, Figures 1 and 2 indicate that the two- and the three-cluster models already account for most of the variability in the data.

In order to compare the result of the cluster analysis, of the information combining both word length and the use of word counts, with the results of the cluster analysis using only word counts, both analysis are carried out.

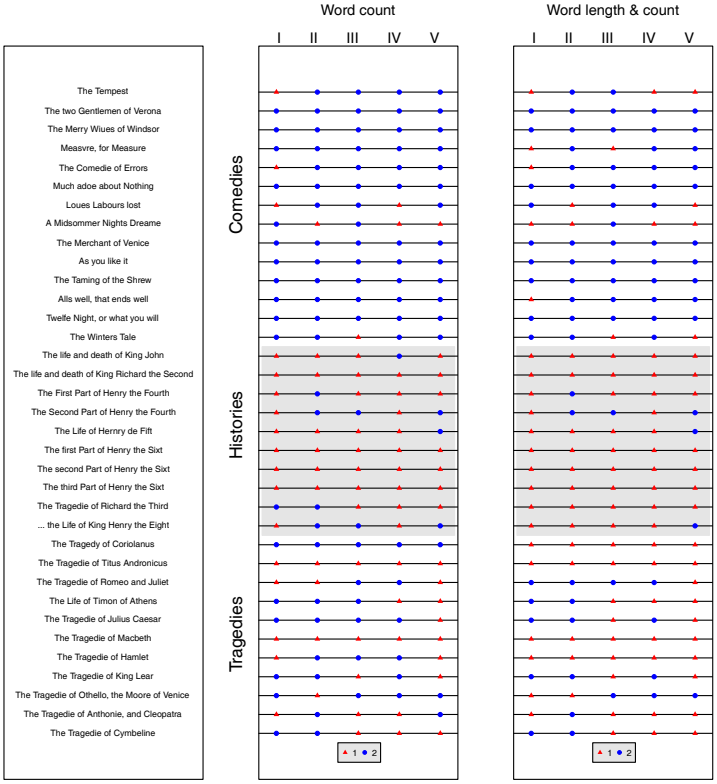

FIGURE 3: Classification of each one of the five acts of each of the plays in Shakespeare's *first folio edition* under the two-cluster model, first using only word counts and second using both word length as well as word counts.

Figure 3 allocates acts into either one of two clusters using the posterior probabilities for ζ_i and uses the two-cluster model. It indicates that the two-cluster analysis classifies acts mostly by genre. In this analysis, most of the acts in comedies fall into Cluster 1, most of the acts in histories fall into Cluster 2, while the acts in tragedies are more or less evenly split across both clusters. An exception

to that rule is that most of the acts in "A Midsommer Nights Dream" are classified under history instead of a comedy. Note that also that all the acts Titus Andronicus and Machbeth tragedies are classified as histories, while the acts of all other tragedies are split between both clusters.

When we compare the result of the analysis that combine word length and word counts with the analysis based only on word counts, we found that only a small number of acts change allocation. The results of both analyses are different and yet, they are similar enough to be able to justify the combination of both characteristics into a single analysis.

Figure 4 allocates acts using clusters under the three-cluster model, again first based only on word counts and second, based on both word counts as well as word lengths. It also appears that the classification of acts into clusters is mostly made by genre, with Cluster 1 being mostly formed by acts in tragedies, Cluster 2 mostly by acts in comedies, and Cluster 3 mostly by acts in histories.

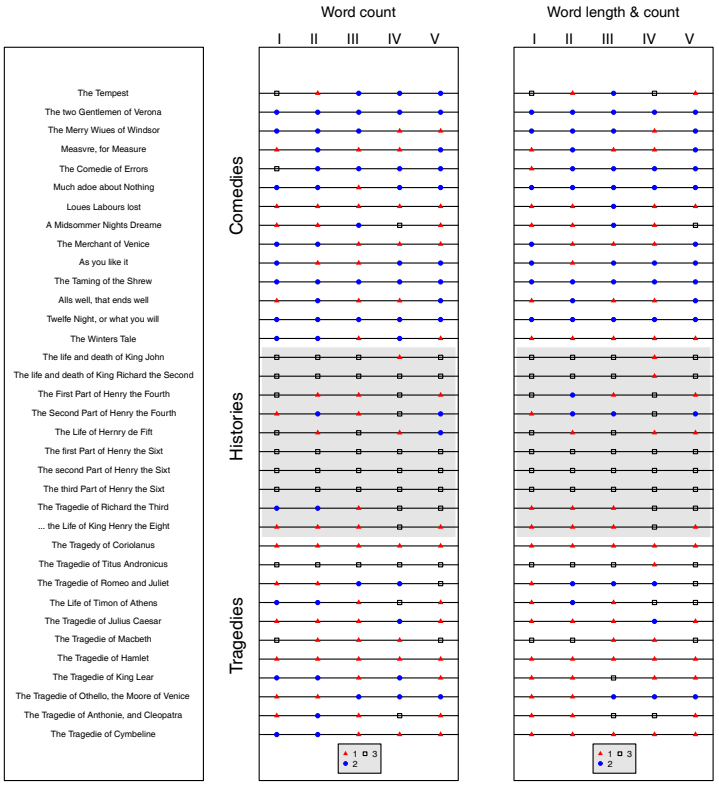

FIGURE 4: Classification of each one of the five acts of each of the plays in Shakespeare's *first folio edition* under the three-cluster model, first using only word counts and second using both word length as well as word counts.

To help interpret the results, Figure 5 presents the first correspondence analysis components, using the method proposed by Greenacre (2007), for the word counts table in the acts of Shakespeare's drama. Acts are stratified first across

genre, which helps emphasize that the heterogeneity of style found in Shakespeare's drama mostly relates to genre. Acts in Figure 5 are also stratified according to their three-cluster classification, which shows how clusters mostly group observations close together in the space of the first correspondence analysis components, and which helps appreciate what changes arise from combining word length and word counts in the analysis instead of just using word counts.

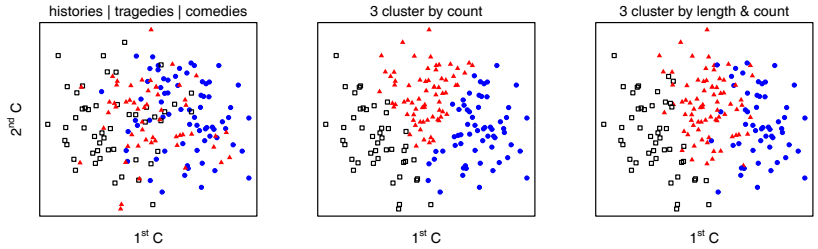

FIGURE 5: First correspondence analysis components of the table of word counts in the Shakespeare's dramas. There are stratified according to genre, and according to the cluster to which the act belongs when using only word counts, and when using both word length as well as word counts.

To help understand what distinguishes clusters different style, Figure 6 presents a sample of the posterior distribution of the multinomial probabilities for word length counts and for the most frequent words under the three-cluster model. Cluster 2, mostly formed by comedies, has the largest proportion of words with *one, two* or *three* letters and the smallest proportion of words with *five, six, seven, eight, nine* or *more than nine* letters. Cluster 2 also has the largest frequencies of *I*, *a*, *you*, *it*, and of *me*, and the smallest frequencies of *and* and of *his*. Clusters 1 and 3 seem to be much more similar in terms of most of the categories considered, with Cluster 3 being especially recognized for having smaller frequencies of *I*, *you*, *it* and *your*, and larger frequencies of *the*, *of* and *with* than the other two clusters.

6. Case study 2: Tirant lo Blanc

Tirant lo Blanc is a chivalry book written in catalan and hailed as be "the best book of its kind in the world" by Miguel de Cervantes. The main body of the book was written between 1460 and 1464, but it was not printed until 1490, and there has been a long lasting debate around its authorship, originating from conflicting information given in its first edition. In 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, take sole responsibility for it,*" at the end of the book it is stated that "*Because of his death, Sir Joanot Martorell 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.*" Over the years, experts have split between the ones favoring the single authorship hypotheses, and the ones backing the hypotheses of a change of author somewhere between chapters 350 and 400.

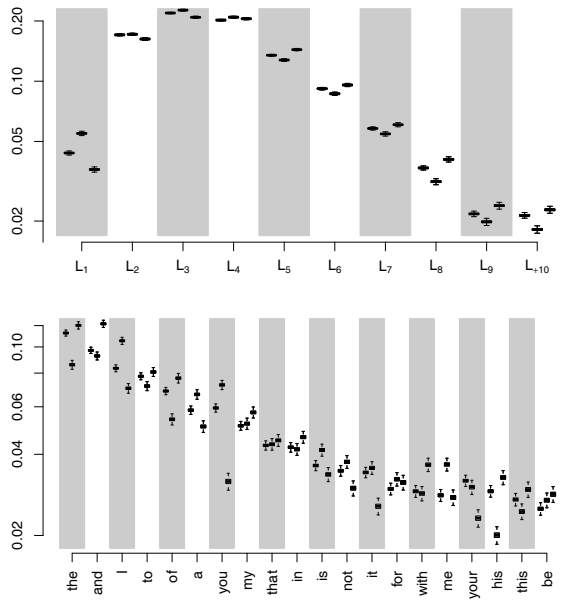

FIGURE 6: Box-plots of a sample of the probabilities for word length, $(\theta_1^{wl}, \theta_2^{wl}, \theta_3^{wl})$, and for word counts, $(\theta_1^{mf}, \theta_2^{mf}, \theta_3^{mf})$, in the three clusters of plays in the *first folio edition* Shakespeare's, all are in a logarithmic scale.

It is generally accepted that the main (and maybe single) author died in 1465, and neither he nor the candidate who was to finish to book left any other texts comparable to this one. An analysis of the diversity of the vocabulary in Riba & Ginebra (2006) found that it is less diverse after chapter 383. Giron et al. (2005) and Riba & Ginebra (2005) carried out a change point and a two-cluster analysis, separately, first for word length and second for the most frequent words. In both cases a stylistic boundary is detected between chapters 371 and 382. This agreement triggered our interest in combining the information in word length with the information in word counts in a single combined analysis.

These papers formally tested for the existence of more than one cluster under each characteristic, by computing the probabilities in (3) under each one of the two tables separately. It was consequently decided that there were two clusters, but it was also conjectured that finite mixtures of Dirichlet-multinomials might be better able to capture the variability in the data than finite mixtures of multinomials.

We carry out a cluster analysis simultaneously based on both the 425×10 table of word length counts as well as on the 425×12 table with the count of the twelve words chosen in Giron et al. (2005) based on the discrimination power between the beginning of the book and its ending. Similarity to that paper, only chapters with more than 200 words are considered. Posterior predictive model checks carried out here similar to the ones in Figures 1 and 2 for Shakespeare's plays indicate that it is also possible to rely on a finite mixture of sets of purely

multinomial models. Hence, the conjecture that one might need mixtures of sets from Dirichlet-multinomial models is not called for.

Figure 7 presents the posterior probability that the i-th row (chapter) belongs to Cluster 1, $\zeta_i = 1$, which is what one needs to classify the chapters of *Tirant lo Blanc* into either one of the two clusters. Cluster 1 mostly includes chapters previous to chapters 375-385, while Cluster 2 mostly includes chapters that come after that boundary; however, there are a fair amount of chapters misclassified by that boundary. This partition of chapters into clusters is similar to the partitions obtained through the analysis carried out in Giron et al. (2005) which considered the two characteristics separately.

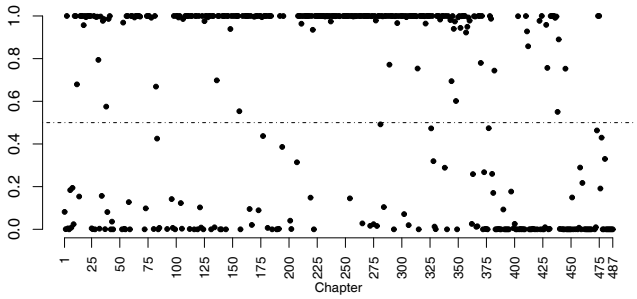

FIGURE 7: Probability that chapters in *Tirant lo Blanc* belong to Cluster 1.

The distribution of the multinomial probabilities under the two-cluster models presented in Figure 8 indicate that *two* and *three* lettered words are more abundant in Cluster 1, while *one, six, seven, eight, nine* and *more than nine* lettered words are more abundant in Cluster 2. That figure also indicates that the words *que*, *no*, *com*, *és*, *jo*, *si* and *dix* are significantly more abundant in Cluster 1, mostly in the first part of the book, while *e*, *de*, *la*, *l'* and *molt* are more abundant in Cluster 2, mostly at the end of the book.

The results presented in this case study are based on the analysis of the counts of twelve words that were selected by Giron et al. (2005). They first under book the analysis with the set of twenty most frequent function words and realized that the main difference in style was between the first four fifths of the book and the last one fifth. They then repeated the analysis with the subset of twelve most discriminating words used here. This sequential approach that starts with about twenty words and then repeats the analysis with the most discriminating words among them helps sharpen the classification power of the method.

Finally, note that different from the previous case study, in this example texts (chapters) are ordered sequentially, and that order is not taken into consideration in the cluster analysis model used here. Puig, Font & Ginebra (2015) propose an alternative analysis that treats the two stylometric variables separately, but incorporates the fact that chapters close together are more likely to belong to the same author than chapters that are far apart. In this case, the results of the analysis are similar.

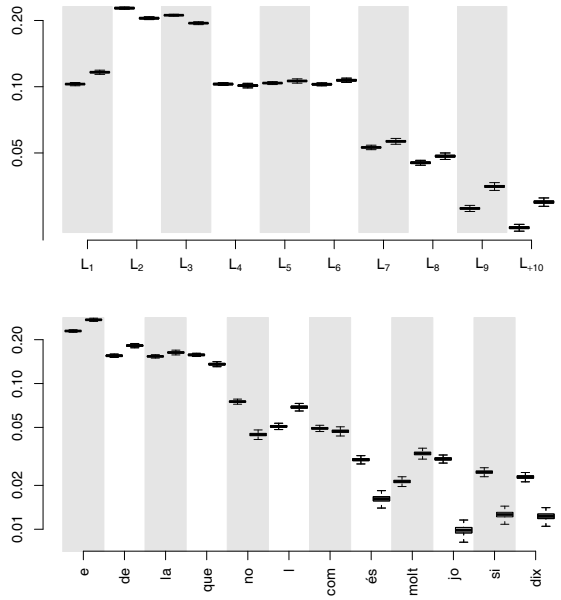

FIGURE 8: Box-plots of a sample of the multinomial probabilities for word length, $(\theta_1^{wl}, \theta_2^{wl})$, and for word counts, $(\theta_1^{mf}, \theta_2^{mf})$, for the two clusters in *Tirant lo Blanc*, all in a logarithmic scale.

7. Final Comments

The paper deals with the analysis of the heterogeneity in literary style, when the researcher does not have a list of candidate authors and of training texts of known authorship to help build the list of best discriminating words needed to determine the authorship of disputed texts. Without them, there is no statistical ground to determine whether the heterogeneities detected are due to authorship, chronology, genre, topic or otherwise.

In the first case study, the results presented are based on twenty of the most frequent words. We also repeated the analysis using only the subset of these words that better according to Figure 6 discriminate between clusters. We consider this sequential approach, starting with about twenty words and then repeating the analysis with the most discriminating words among them, to be very useful. Using far more than twenty words is usually problematic, because included in the analysis; will be many words that do not distinguish between styles and, therefore, hamper the classification power of the algorithm.

When using word length and word counts, our predictive checks indicate that finite mixtures of multinomial models capture most of the variability in the data: this settles the issue raised in Giron et al. (2005). In this setting, one does not need to use hierarchical models, such as the Dirichlet-multinomial models finite mixtures that are used in Puig & Ginebra (2014) to account for any extra variability in the data.

Finally, note that one could still use more sophisticated models that use Dirichlet process mixtures to embed all the S-cluster models into a single model where s becomes a parameter to be estimated. Such an approach would be a lot more demanding computationally than the one taken here, and it would not be easy to implement with WinBugs.

Acknowledgements

This work was funded in part by Grant No. MTM2013-43992-R of the Ministerio de Ciencia e Inovación of Spain.

[Received: April 2015 — Accepted: January 2016]

References

Banfield, J. D. & Raftery, A. E. (1993), 'Model based gaussian and non-gaussian clustering', *Biometrics* **49**, 803–821.

Binongo, J. N. G. (1994), 'Joaquin's Joaquinesquerie, Joaquinesqueri'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.

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

Casella, G., Moreno, E. & Giron, J. (2014), 'Cluster analysis, model selection and prior distributions on models', *Bayesian Analysis* **9**, 613–658.

Edmondson, P. & Wells, S. (2013), *Shakespeare Beyond Doubt: Evidence, Argument, Controversy*, Cambridge University Press, Cambridge.

Fernandez, C. & Green, P. J. (2002), 'Modelling spatially correlated data via mixtures: a bayesian approach', *Journal of the Royal Statistical Society B* **64**, 805–826.

Font, M., Puig, X. & Ginebra, J. (2013), 'A Bayesian analysis of frequency count data', *Journal of Statistical Computation and Simulation* **83**, 229–246.

Fraley, C. & Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis and density estimation', *Journal of the American Statistical Association* **97**, 611–631.

Gelfand, A. E. & Dey, D. K. (1994), 'Bayesian model choice: Asymptotics and exact calculations', *Journal of the Royal Statistical Society, Serie B* **56**, 501–514.

Gelman, A., Carlin, J. C., Stern, H. & Rubin, D. B. (2004), *Bayesian Data Analysis*, 2 edn, Chapman & Hall, New York.

Giron, J., Ginebra, J. & Riba, A. (2005), 'Bayesian analysis of a multinomial sequence and homogeneity of literary style', *The American Statistician* **59**, 19–30.

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

Gordon, A. D. (1999), *Classification*, 2 edn, Chapman and Hall, London.

Greenacre, M. (1988), 'Clustering the rows and columns of a contingency table', *Journal of Classification* **5**, 39–51.

Greenacre, M. (2007), *Correspondence Analysis in Practice*, Chapman and Hall, London.

Hilton, M. L. & Holmes, D. I. (1993), 'An assessment of cumulative control charts for authorship-attribution', *Literary and Linguistic Computing* **8**, 73–80.

Holmes, D. I. (1985), 'The analysis of literary style, a review', *Journal of the Royal Statistical Society, Ser A* **148**, 328–341.

Holmes, D. I. (1992), 'A stylometric analysis of mormon scripture and related texts', *Journal of the Royal Statistical Society* **155**, 91–120.

Holmes, D. I. (1994), 'Authorship attribution', *Computers and the Humanities* **28**, 87–106.

Holmes, D. I. (1998), 'The evolution of stylometry in humanities scholarship', *Literary and Linguistic Computing* **13**, 111–117.

Holmes, D. I. (1999), Stylometry, *in* 'Encyclopedia of Statistical Sciences', Wiley, New York, pp. 721–727.

Hope, J. (1994), *The Authorship of Shakespeare's Plays*, Cambridge: Cambridge University Press, Cambridge.

Hope, J. (2010), *Shakespeare and Language: Reason, Eloquence and Artifice in the Renaissance*, The Arden Shakespeare, London.

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

Lunn, D. J., Jackson, C., Best, N., Thomas, A. & Spiegelhalter, D. (2013), *The BUGS Book. A Practical Introduction to Bayesian Analysis*, Chapman Hall, London.

Luyckx, K. (2010), *Scalability Issues in Authorship Attribution*, University Press Antwerp, Brussels.

Mendenhall, T. C. (1887), 'The characteristic curves of composition', *Science* **9**.

Mendenhall, T. C. (1901), 'A mechanical solution of a literary problem', *The Popular Science Monthly* **60**.

Miranda-Garcia, A. & Calle-Martin, J. (2007), 'Function words in authorship attribution studies', *Literary and Linguistic Computing* **22**, 27–47.

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

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

Murtagh, F. & Raftery, A. E. (1984), 'Fitting straight lines to point patterns', *Pattern Recognition* **17**, 479–483.

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

Puig, X., Font, M. & Ginebra, J. (2015), 'Classification of literary style that takes order into consideration', *Journal of Quantitative Linguistics* **22**, 177–201.

Puig, X., Font, M. & Ginebra, J. (2016), 'A unified approach to authorship attribution and verification', *To appear in The American Statistician* .

Puig, X. & Ginebra, J. (2014), 'A bayesian cluster analysis of election results', *Journal of Applied Statistics* **41**, 73–94.

Riba, A. & Ginebra, J. (2005), 'Change-point estimation in a multinomial sequence and homogeneity of literary style', *Journal of Applied Statistics* **32**, 61–74.

Riba, A. & Ginebra, J. (2006), 'Diversity of vocabulary and homogeneity of literary style', *Journal of Applied Statistics* **33**, 729–741.

Rybicki, J. & Eder, M. (2011), 'Deeper Delta across genres and languages: do we really need the most frequent words?', *Literary and Linguistic Computing* **26**, 315–321.

Shahan, J. M. & Waugh, A. (2013), *Shakespeare Beyond Doubt? Exposing and Industry in Denial*, Llumina Press, London.

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.

Stamatatatos, E. (2009), 'A survey of modern authorship attribution methods', *Journal of the American Society of Information Science and Technology* **60**, 538–556.

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

Zhao, Y. & Zobel, J. (2005), 'Effective and scalable authorship attribution using function words', *Information Retrieval Technology* **3689**, 174–189.