Stylometric Analyses using
Dirichlet Process Mixture Models

Paramjit S. Gill
I. K. Barber School of Arts and Sciences
University of British Columbia Okanagan
Kelowna, BC, Canada, V1V 1V7
E-mail: paramjit.gill@ubc.ca

Tim B. Swartz
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC, Canada, V5A 1S6
E-mail: tim@stat.sfu.ca

Summary. Stylometry refers to the statistical analysis of literary style of authors based on the characteristics of expression in their writings. We propose an approach to stylometry based on a Bayesian Dirichlet process mixture model using multinomial word frequency data. The parameters of the multinomial distribution of word frequency data are the "word prints" of the author. Our approach is based on model-based clustering of the vectors of probability values of the multinomial distribution. The resultant clusters identify different writing styles that assist in author attribution for disputed works in a corpus. As a test case, the methodology is applied to the problem of authorship attribution involving the Federalist papers. Our results are consistent with previous stylometric analyses of these papers.

Keywords: Bayesian methods, Clustering, Computational linguistics, Dirichlet process priors, Disputed authorship, Federalist papers, Multinomial distribution

1 Introduction

Stylometry deals with the statistical and computer analysis of literary style using the characteristics of expression. For more than 150 years, a wide variety of statistical techniques have been proposed to compare literary styles and address the problem of authorship attribution. The methods range from goodness-of-fit tests, principal component analysis, discriminant analysis and classical clustering methodologies to sophisticated artificial intelligence and neural networks models. Specific stylometric applications have involved religious texts (*Old and New Testaments*, *Book of Mormon*), Old English manuscripts, political essays (*Federalist papers*), popular literature (*The Royal Book of Oz*), Shakespearean plays, American Civil War letters, plagiarism in students' essays and the authenticity of legal documents. Holmes (1999) provides a good review and history of research in this field.

Pioneering work on stylometry was based on word-length and sentence-length distributions. However, it was often the case that these variables were unable to discriminate between authors (Mosteller and Wallace 1984). Consequently, many stylometric analyses include other variables such as vocabulary richness, the proportion of nouns, the proportion of adjectives, the number of one-letter and two-letter words, etc. (Holmes 1999). For the purpose of author attribution, successful methods rely on identifying and exploiting an author's "word prints". The methodologies involve words, called *function words*, which are non-contextual (i.e. topic-independent) but serve as useful indicators of an author's unconscious stylistic preferences while writing on any topic. Commonly used function words are conjunctions, prepositions and articles that have little meaning by themselves but are used to define relationships between content words in a sentence. Data extracted from texts consist of the frequencies of occurrence of various function words.

The question of authorship falls into the general statistical problem of classification where objects that are "similar" are grouped together in clusters. In stylometry, the objects are texts and the clusters consists of texts that are deemed to have been written by the same author. Sometimes prior knowledge is available of the form where the authorship is known for some of the texts under consideration.

In our approach, we view the frequencies of function words as samples from an underlying multinomial distribution. The use of the multinomial distribution is an important aspect of our methodology and differs from analyses based on the multivariate normal distribution (e.g. Holmes and Forsyth 1995). When dealing with frequency data, percentages sum to 100% and induce negative correlations between the frequencies of function words. In the normal setting, the negative correlations are not quite handled correctly even when a non-diagonal variance matrix is permitted. Furthermore, the discrete nature of the data is not taken into consideration when analyses are based on the normal distribution.

Our approach also differs from most stylometric analyses in that it is Bayesian. Unlike deterministic clustering algorithms, this implies that there is a quantifiable uncertainty in the resultant clusters of texts. We account for the uncertainty by expressing probabilities associated with clusters. Although methods of fuzzy analysis (Dunn 1977) provide membership coefficients for individual objects, classical (as opposed to model-based) agglomerative approaches do not provide probability assessments for clustering. Clustering probabilities are sometimes available using Bayesian mixture models (Liu, Zhang, Palumbo and Lawrence 2003). However, these methods typically rely on Markov chains which often require fine tuning to promote mixing in the Markov chain. A mixture model based on generalized Dirichlet multinomial distributions to cluster count data with applications to digit recognition is discussed in Bouguila (2008).

In Section 2, we describe the underlying Bayesian model used in our approach. The data on function word frequencies are assumed to arise from multinomial distributions. The parameters of the multinomial distributions are therefore characteristics of an author's writing style, and clustering is carried out on these parameters. As there exists uncertainty in the parameters, it follows that there is uncertainty in the clustering which leads to the calculation of posterior clustering probabilities. In Gill, Swartz and Treschow (2007), stylometric clustering is done in a two-stage process whereby parameter output is generated from a Markov chain, and the output is then fed into a standard clustering algorithm. In the proposed approach, the clustering is more natural as the clustering is an inherent part of the statistical model. This is accomplished via the Dirichlet process whose support is restricted to discrete distributions (Ferguson 1974) and is therefore well-suited for clustering. Using Dirichlet process priors also allows weaker prior assumptions by going from a parametric to a semiparametric framework. Our approach makes various prior assumptions that are appropriate to the stylometric context. We also suggest a pragmatic approach to the selection of function words. Our methodology and analysis is computational and we discuss aspects of the computation in Section 3. Some theory and discussion is provided regarding the effect of increasing the number of function words. In Section 4, a detailed simulation study provides insights regarding the interplay of the various input parameters of the algorithm. The simulation study suggests conditions when the approach may be effective. In Section 5, the methods are applied to the well-studied problem concerning the authorship of the Federalist papers. In the analyses, we indicate how prior knowledge can be utilized by grouping texts in appropriate ways and by varying the concentration parameter used in the Dirichlet process. A concluding discussion is provided in Section 6.

2 Bayesian model

A principle of stylometric analysis is that authors use high-frequency function words unreflectively in their writings. These function words occur regardless of context, and hence, differential rates of usage form a basis for distinguishing authorship. Function words are typically prepositions, conjunctions, articles and common verbs. The choice of function words is typically determined by a subject matter expert. Given the selection of K function words, we assume that frequency counts are obtained for each of N texts. Let X_{ik} denote the frequency of function word k in the ith text, $k = 1, \ldots, K$, $i = 1, \ldots, N$. This gives rise to the model

$$X_i = (X_{i1}, \ldots, X_{iK})^T \sim \text{Multinomial}(n_i\,;\, p_{i1}, \ldots, p_{iK}) \tag{1}$$

where $n_i = \sum_k X_{ik}$ is the total number of function words in the ith text, the multinomial vectors are assumed independent over $i = 1, \ldots, N$ and $\sum_k p_{ik} = 1$. The multinomial distribution is a natural distribution for use in this problem. However, in a clustering context, we are not aware of any model-based mixture approaches that use the multinomial distribution. Perhaps this is due to the fact that there is no clear choice for a dissimilarity measure between the multinomial vectors $X_1, \ldots, X_N$ in classical partitioning and hierarchical clustering algorithms. Traditional stylometric methods are often based on large sample multivariate normal approximations to the multinomial distribution for word frequency data (Holmes and Forsyth 1995).

The unknown parameters in model (1) are the p_{ik}'s which represent the word prints of the author of the ith text. In a standard Bayesian analysis based on (1), it is typical to assign independent flat priors to $p_i = (p_{i1}, \ldots, p_{iK})^T$ for $i = 1, \ldots, N$. In the given application to stylometry, we suggest that it is more appropriate to assume that the p_i's are grouped in latent clusters representing the underlying authors. We proceed by using a Dirichlet process mixture model whereby

$$\begin{aligned} p_i \,|\, G &\overset{\text{iid}}{\sim} G \\ G \,|\, \alpha, G_0 &\sim \mathcal{DP}(\alpha, G_0) \end{aligned} \tag{2}$$

for $i = 1, \ldots, N$. In non-technical language, (2) states that the parameters p_i are independent and are distributed according to G. However, the distribution G is unknown and this is the semiparametric aspect of the Dirichlet process. Furthermore, (2) states that the distribution G arises from a distribution of distributions where $E(G) = G_0$ and the concentration parameter $\alpha > 0$ is such that larger (smaller) values of α imply that G is closer (further) in distributional distance to G_0. Again, an important but not so obvious feature of (2) is that the support of G is limited to discrete distributions, and this is fundamental to the clustering of the p_i's. We also note that the Dirichlet process does not require the specification of the number of clusters. Once a theoretical curiosity, the Dirichlet process and its extensions are finding increasing application areas in nonparametric statistics. For a clear exposition of a Dirichlet process mixture model in an applied random-effects setting together with references to the technical literature concerning the Dirichlet process, see Ohlssen, Sharples and Spiegelhalter (2007).

To complete the Bayesian model, we require the specification of the baseline distribution G_0 and the remaining prior distributions. In our problem, we suggest the baseline distribution

$$G_0 \sim \text{Dirichlet}(a_1, ..., a_K) \tag{3}$$

where the Dirichlet is appropriately defined on the $(K-1)$-dimensional simplex. Although setting $a_1 = \cdots = a_K = 1$ in (3) may be interpreted as noninformative, we take the view that the individual writing styles p_i arise from a superpopulation of writing styles. We therefore take an empirical Bayes approach and set $a_k = mq_k$ where q_k is the proportion of times that function word k appears across all manuscripts (i.e. $q_k = \sum_i X_{ik} / \sum_i n_i$) for $k = 1, \ldots, K$. The specification implies that the kth component of the Dirichlet distribution has expectation q_k, $k = 1, \ldots, K$. The determination of m is based on variability considerations where we equate the "empirical variance" $s^2 = \sum (q_i - \overline{q})^2 / K$ with the theoretical variance of the kth component of the Dirichlet distribution. This leads to $(m+1)s^2 = q_k(1 - q_k)$, and summing over $k = 1, \ldots, K$ gives

$$m = \frac{\sum q_k(1 - q_k)}{\sum (q_k - \overline{q})^2} - 1.$$

Finally, in various applications, the treatment of the concentration parameter α has sometimes proved problematic as inferences may be sensitive to α (Dorazio 2009). Accordingly, we treat the concentration parameter as a tuning parameter and investigate its effect under various fixed settings. In our examples, we set values of α ranging from 1 to 100. This is demonstrated in the simulation study of Section 4.

2.1 Choice of function words

Earlier, we glossed over the problem of determining function words. In fact, this is a serious issue which has a direct impact on the success of any stylometric approach.

A temptation and standard strategy is to look at the texts in question, and choose non-contextual function words that have good discriminating power in differentiating texts between authors who are known to be distinct. In classical analyses, such an approach typically violates the statistical procedures used in the analyses. The reason is that statistical procedures such as t-tests are unconditional tests, and unconditional tests do not allow the user to first look at the data, and then use the information to determine aspects of the test (i.e. determine the function words). This standard stylometric strategy therefore argues for a Bayesian approach. In Bayesian statistics, all inferences proceed from the posterior distribution, and the posterior describes parameter uncertainty conditional on the observed data. Bayesian procedures are therefore conditional procedures which allow a user to make decisions on how to proceed based on an initial inspection of the data.

In our Bayesian approach to stylometry, we determine function words based on data inspection. Suppose that we have K non-contextual function words initially chosen by a subject matter expert. Often the subject matter expert may do something as simple as list the most frequent words and eliminate contextual words from the list. As discussed at the end of Section 3, we do not want K large, and therefore, we attempt to reduce the list of non-contextual function words. Suppose further that we have N texts in the stylometric analysis. Given the kth function word under consideration and texts i and j, we calculate the statistic

$$z = \frac{\hat{p}_{ik} - \hat{p}_{jk}}{\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\tilde{p}(1-\tilde{p})}} \tag{4}$$

where $\hat{p}_{lk} = X_{lk}/n_l$ is the proportion of occurrence of function word k in text l, $\tilde{p} = (X_{ik}+X_{jk})/(n_i+n_j)$ is the pooled proportion of the given function word in text i and text j, and n_l is the total number of function words in text l, $l = i, j$. The statistic z in (4) is calculated for all $K\binom{N}{2}$ combinations of function words and pairs of texts. We sort $|z|$ and use the largest values to determine the reduced set of function words. This is a simple and effective method of determining function words, and it is a legitimate procedure when implemented as part of a Bayesian analysis.

3 Computational algorithm

In Bayesian analyses, the posterior distribution provides the full description of uncertainty for all that is unknown. Accordingly, it is often convenient to describe the posterior using summaries such as posterior means and posterior standard deviations. However, in most Bayesian applications, the complexity and dimensionality of the posterior prevents the analytic calculation of summary measures which take the form of integrals. In these problems, it is common to take a sampling approach to approximate the integrals, and Markov chain methods are often the sampling method of choice.

The Dirchlet process mixture model presented in Section 2 is a model whose complexity seems to demand an analysis via Markov chain methods. Our first attempt at computation involved an implementation via the software package WinBUGS (Spiegelhalter, Thomas and Best 2003). WinBUGS analyses are appealing as the programmer need only specify the statistical model, the prior and the data. In WinBUGS, the Markov chain calculations are done in the background whereby the user is supplied with Markov chain output. As the Dirichlet process prior is not one of the available WinBUGS distributions, an approximation of the prior was considered as carried out in Ghosh, Gill, Muthukumarana and Swartz (2010). The approximation is based on a truncation of the "stick-breaking" sum as given in the Sethuraman (1994) construction of the Dirichlet process. We use this approach as a confirmatory procedure to ensure that we have qualitatively correct results. Our preferred procedure which we describe next does not require the tuning parameter involved in the stick-breaking truncation.

An alternative approach involves the direct programming of a Markov chain for the Dirichlet process mixture model of Section 2. Neal (2000) describes various Markov chain algorithms for the analysis of Dirichlet process mixture models. Our model has a conjugate structure (i.e. Dirichlet baseline distribution with multinomial data) and this facilitates the development of algorithms. We implement Algorithm 2 given by Neal (2000) which involves the introduction of latent variables describing the cluster membership of the writing styles $p_1, \ldots, p_N$. Compared to various alternative MCMC algorithms, Algorithm 2 is thought to mix well and lead to practical convergence in realistic computing times for a variety of problems (Neal 2000). Specifically, let c_i denote the latent class of the ith text and let n_{ic} be the size of latent class $c = 1, \ldots, G$ when text i is excluded from consideration. Then Algorithm 2 from Neal (2000) translates into iterations of the following two steps:

- For $i = 1, \ldots, N$: generate c_i according to $\Pr[c_i = c] = r_c$ where

$$r_c \propto \begin{cases} 0 & \text{if } n_{ic} = 0 \\ \left(\frac{\alpha}{N-1+\alpha}\right) \frac{\Gamma(a_1+\cdots+a_K)}{\Gamma(a_1)\cdots\Gamma(a_K)} \frac{\Gamma(a_1+X_{i1})\cdots\Gamma(a_K+X_{iK})}{\Gamma(a_1+\cdots+a_K+X_{i1}+\cdots+X_{iK})} & \text{if } c = G+1 \\ \left(\frac{n_{ic}}{N-1+\alpha}\right) p_{c1}^{X_{i1}} \cdots p_{cK}^{X_{iK}} & \text{otherwise .} \end{cases} \tag{5}$$

  If c_i belongs to an existing class c_j, then set $p_i = p_j$. If c_i does not belong to an existing class, then generate $p_i \sim \text{Dirichlet}(a_1 + X_{i1}, \ldots, a_K + X_{iK})$.

- For each latent class $c = 1, \ldots, G$: sample the corresponding $p_c \sim \text{Dirichlet}(a_{c1}, \ldots, a_{cK})$ where $a_{ck} = a_k + \sum_{i:c_i=c} X_{ik}$.

The primary inferential question for stylometry concerns the attribution of authorship. To address this question, consider a single iteration of the Markov chain which yields parameter output $p_1, \ldots, p_N$. For each pair of texts (i, j), we check whether $p_i = p_j$ and increment its counter accordingly. After many iterations, we obtain the proportion of time that p_i and p_j are equal, and this provides an estimate of the probability that authors of text i and j are the same.

In (5), the concentration parameter α is fixed and we view it as a tuning parameter for the algorithm. However, it may be more satisfying to consider a fully Bayes procedure where a prior distribution is assigned to α. Neal (2000) states that the random α case can be handled. Gill and Casella (2009) consider the case where α is discrete. In the continuous case, let $\pi(\alpha)$ denote the prior density of α. Then it is not difficult to show that the full conditional distribution for α is

$$[\alpha \mid \cdot] \propto \frac{\Gamma(\alpha)}{\Gamma(\alpha+N)} \alpha^G \pi(\alpha) \tag{6}$$

where G is the number of latent classes. We recognize that (6) is a nonstandard density which does not readily admit variate generation. We therefore imbed a Metropolis step in the Gibbs sampling algorithm (5) where $q(\alpha)$ is chosen as the proposal density for α. The Metropolis step then proceeds by generating $u \sim \text{Uniform}(0, 1)$ and generating α according to the proposal distribution. We denote the previous value of α in the Markov chain as α_*. We use the generated α from the proposal as the next iterate in the Markov chain unless

$$u > \frac{\Gamma(\alpha)}{\Gamma(\alpha_*)} \frac{\Gamma(\alpha_* + N)}{\Gamma(\alpha + N)} \left(\frac{\alpha}{\alpha_*}\right)^G \frac{\pi(\alpha)}{\pi(\alpha_*)} \frac{q(\alpha_*)}{q(\alpha)}, \tag{7}$$

in which case, we set $\alpha = \alpha_*$. As a particular case, we choose the prior density $\pi(\alpha)$ as $\text{Uniform}(0, \theta_0)$ for some prescribed θ_0. Then noting that $\Gamma(\alpha)\alpha^G/\Gamma(\alpha+N)$ is increasing in α, we choose the proposal density $q(\alpha) = 2\alpha/\theta_0^2$ on $(0, \theta_0)$. Generating from this proposal is straightforward via $\alpha = \theta_0\sqrt{v}$ where $v \sim \text{Uniform}(0,1)$ and inequality (7) reduces to

$$u > \frac{\Gamma(\alpha)}{\Gamma(\alpha_*)} \frac{\Gamma(\alpha_* + N)}{\Gamma(\alpha + N)} \left(\frac{\alpha}{\alpha_*}\right)^{G-1}.$$

Initially, our intuition suggested that by increasing the number K of function words, the discriminating power of the methodology would improve, and better stylometric inferences would be obtained. This is what is generally observed in multivariate analyses where increasing the number of variables tends to increase the opportunity to discriminate. However, our investigations suggest that this is not the case, and in fact, very large values of K lead to unreasonably small numbers of clusters. Therefore, the user needs to strike a balance between adding function words that are truly discriminatory versus adding function words that create noise in the analysis. To get a sense why this happens, consider for simplicity the noninformative prior $a_1 = \cdots = a_K = 1$. Then using the probabilities in (5), we investigate the probability that a new cluster is formed in the Markov chain. Letting Q denote a constant and using Stirling's approximation to the log-gamma function, we have

$$\begin{aligned}
\log\left(\frac{r_{G+1}}{r_c}\right) &= \log\alpha + \log\Gamma(K) + \sum_k \log\Gamma(X_{ik}+1) \\
&- \log\Gamma(K + \sum_k X_{ik}) - \log n_{ic} - \sum_k X_{ik}\log p_{ck} \\
&\approx \log(\alpha/n_{ic}) + (K-1)\log(K-1) - (K-1) + \sum_k X_{ik}\log X_{ik} - \sum_k X_{ik} \\
&- (K - 1 + \sum_k X_{ik})\log(K - 1 + \sum_k X_{ik}) + (K - 1 + \sum_k X_{ik}) - \sum_k X_{ik}\log p_{ck} \\
&\approx Q + (K-1)\log(K-1) - (K - 1 + \sum_k X_{ik})\log(K - 1 + \sum_k X_{ik}) \\
&\leq Q + (K-1)\log(K-1) - (K - 1 + \sum_k X_{ik})\log(K-1) \\
&= Q - \sum_k X_{ik}\log(K-1) \\
&\to -\infty
\end{aligned}$$

as $K \to \infty$. This implies that new clusters do not form in the Markov chain when the number of function words becomes excessively large.

4 Simulation Study

We now provide a comprehensive simulation study on the effect of various input parameters to the stylometric clustering algorithm.

4.1 The baseline data, model and analysis

We consider a stylometric problem involving $N = 10$ texts and three clusters. We generated data such that texts $i = 1, \ldots, 5$ belong to the first cluster, texts $i = 6, 7, 8$ belong to the second cluster and texts $i = 9, 10$ belong to the third cluster. Specifically, for texts $i = 1, \ldots, 10$, we generated data

$$(X_{i1}, \ldots, X_{iJ})^T \sim \text{Multinomial}(\tilde{n}_i\,;\, p_{i1}, \ldots, p_{iJ})$$

where $J = 21$ and we initially considered a total text size of $\tilde{n}_i = 5000$ words. Note that $p_{iJ} = 1 - \sum_{j=1}^{J-1} p_{ij}$ corresponds to the frequency of non-function words such that the number of function words that can be used in model (1) is $K = 2, \ldots, J-1$ leading to $n_i = \sum_{k=1}^{K} X_{ik}$.

The multinomial parameters for texts $i = 1, \ldots, 5$ in the first cluster were set as follows:

p_{i1}	p_{i2}	$p_{i3} \cdots p_{i5}$	$p_{i6} \cdots p_{i8}$	p_{i9}	p_{i10}	$p_{i11} \cdots p_{i20}$
0.060	0.030	0.025	0.020	0.010	0.005	0.002

.

For texts $i = 6, 7, 8$ in the second cluster, we set

$$p_{ij} = \begin{cases} a_2 p_{1j} & j \text{ odd} \\ b_2 p_{1j} & j \text{ even} \end{cases} \tag{8}$$

where $a_2 = 0.9$ and $b_2 = 1.1$ were chosen to differentiate the second cluster from the first cluster at the 10% level. For texts $i = 9, 10$ in the third cluster, we set

$$p_{ij} = \begin{cases} a_3 p_{1j} & j \text{ odd} \\ b_3 p_{1j} & j \text{ even} \end{cases} \tag{9}$$

where $a_3 = 1.1$ and $b_3 = 0.9$ were chosen to differentiate the third cluster from the first cluster at the 10% level. With these settings, the first cluster “lies” somewhere between the second and third clusters.

The generated data are intended to portray a realistic stylometric problem. We have considered texts which are of standard essay length (5000 words) and the function word frequencies p_{ik}

correspond to the usage of common non-contextual words such as “the” and “as”. Note that the frequencies decrease as we increase the number of function words; this is realistic as most English function words have frequencies less than 2%. Table 1 provides the data counts $X_{i1}, \ldots, X_{i10}$ for the first 10 function words for each of the texts $i = 1, \ldots, 10$.

Text i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	X_{i6}	X_{i7}	X_{i8}	X_{i9}	X_{i10}
1	316	142	116	131	124	95	111	82	41	20
2	271	145	118	154	132	99	105	101	56	17
3	307	161	114	123	124	96	115	110	43	22
4	322	137	124	127	132	106	94	86	41	35
5	298	150	127	125	134	95	91	95	57	33
6	268	161	127	129	117	111	96	102	56	31
7	245	188	112	131	116	116	90	106	41	24
8	268	158	112	151	129	109	85	116	39	26
9	347	139	123	121	130	73	105	85	52	27
10	356	142	114	107	126	66	108	60	58	23

Table 1: The frequency counts $X_{i1}, \ldots, X_{i10}$ corresponding to the first 10 function words of the baseline data from Section 4.1.

Having generated the baseline data, we completed the specification of the baseline stylometric model by using the continuous prior $\alpha \sim \text{Uniform}(0, 3)$ and using the first $K = 2$ function words.

Table 2 provides the resultant posterior probabilities of pairwise clustering between texts. Using a probability threshold of 0.5 for clustering, Table 2 suggests three clusters with text memberships $\{1, 4, 9, 10\}$, $\{2, 3, 5, 6, 8\}$ and $\{7\}$. At first glance, the clustering results may appear poor. However, upon close inspection of the first two columns of Table 1 (recall $K = 2$), the clustering appears sensible and reflects the variation of the data generation mechanism. The posterior probabilities in Table 2 are reasonable in the sense that they are not too close to either 0 or 1. For example, texts 2, 3 and 5 are nearly clustered with texts 1 and 4. We repeated the exercise by generating several datasets under the same input conditions, and in each of the cases, the clustering was “incorrect” with a number of misclassified texts.

Text i	Text j									
	1	2	3	4	5	6	7	8	9	10
1	0.00	0.43	0.46	0.67	0.53	0.23	0.01	0.26	0.64	0.64
2	0.43	0.00	0.64	0.35	0.60	0.55	0.11	0.58	0.29	0.29
3	0.46	0.64	0.00	0.39	0.63	0.53	0.09	0.55	0.32	0.31
4	0.67	0.35	0.39	0.00	0.47	0.16	0.00	0.19	0.71	0.71
5	0.53	0.60	0.63	0.47	0.00	0.44	0.06	0.47	0.41	0.40
6	0.23	0.55	0.53	0.16	0.44	0.00	0.26	0.66	0.11	0.11
7	0.01	0.11	0.09	0.00	0.06	0.26	0.00	0.23	0.00	0.00
8	0.26	0.58	0.55	0.19	0.47	0.66	0.23	0.00	0.14	0.14
9	0.64	0.29	0.32	0.71	0.41	0.11	0.00	0.14	0.00	0.75
10	0.64	0.29	0.31	0.71	0.40	0.11	0.00	0.14	0.75	0.00

Table 2: The posterior probability of clustering between texts i and j for the baseline data of Section 4.1 where $K = 2$ and $\alpha \sim \text{Uniform}(0, 3)$.

4.2 The effect of the function word probabilities p_{ik}

Hopefully, the methodology is better able to distinguish differences between texts when there is in fact greater differentiation between texts. To explore this, we changed the differentiation parameters in (8) and (9) to $a_2 = 0.8$, $b_2 = 1.2$, $a_3 = 1.2$ and $b_3 = 0.8$ corresponding to an increased 20% level of differentiation. We generated four datasets and the clustering results were as follows:

- $\{1, 2, 3, 4, 5\}$, $\{6, 7, 8\}$, $\{9, 10\}$
- $\{1, 2, 3, 4, 5, 8\}$, $\{6\}$, $\{7\}$, $\{9, 10\}$
- $\{1, 2, 3, 4, 5, 6\}$, $\{7, 8\}$, $\{9, 10\}$
- $\{1, 3, 4, 5\}$, $\{2\}$, $\{6, 7, 8\}$, $\{9, 10\}$

Although the clustering is “perfect” only in the first case, the overall clustering performance here is better than in the case of 10% differentiation. This provides evidence that our stylometric methodology makes better decisions when authors are more heterogeneous in their writing styles.

4.3 The effect of the DP concentration parameter α

One of the contributions in the paper is the specification and implementation of a continuous prior for α. In the analysis of section 4.1, the posterior mean of α was 1.6. If we instead use a discrete Uniform prior for α (Gill and Casella 2009), we observe that the posterior means can differ substantially. For example, using a discrete Uniform$(1, \ldots, 20)$ prior, the posterior mean of α is 5.1. The clustering results also differ with texts apportioned to six clusters according to $\{1, 4\}$, $\{2, 3\}$, $\{5\}$, $\{6, 8\}$, $\{7\}$ and $\{9, 10\}$. This is expected as expanding the limit of support for the concentration parameter α is known to increase the numbers of clusters. However, the good news is that changing the prior for the nuisance parameter α did not greatly change the pairwise clustering probabilities. In Table 3, we provide the posterior pairwise clustering probabilities using the baseline data given in Section 4.1 with $K = 2$ and $\alpha \sim$ Discrete Uniform$(1, \ldots, 20)$. The entries in Table 3 and Table 2 are qualitatively similar in the sense that for every text, we have the same rank order for the row probabilities.

Text i	Text j									
	1	2	3	4	5	6	7	8	9	10
1	0.00	0.32	0.35	0.52	0.41	0.16	0.00	0.19	0.48	0.48
2	0.32	0.00	0.51	0.26	0.47	0.43	0.06	0.45	0.20	0.19
3	0.35	0.51	0.00	0.28	0.49	0.41	0.05	0.43	0.21	0.21
4	0.52	0.26	0.28	0.00	0.35	0.11	0.00	0.13	0.54	0.54
5	0.41	0.47	0.49	0.35	0.00	0.34	0.03	0.36	0.28	0.27
6	0.16	0.43	0.41	0.11	0.34	0.00	0.17	0.52	0.07	0.07
7	0.00	0.06	0.05	0.00	0.03	0.17	0.00	0.15	0.00	0.00
8	0.19	0.45	0.43	0.13	0.36	0.52	0.15	0.00	0.08	0.08
9	0.48	0.20	0.21	0.54	0.28	0.07	0.00	0.08	0.00	0.57
10	0.48	0.19	0.21	0.54	0.27	0.07	0.00	0.08	0.57	0.00

Table 3: The posterior probability of clustering between texts i and j for the baseline data of Section 4.1 where $K = 2$ and $\alpha \sim$ Discrete Uniform$(1, \ldots, 20)$.

We also consider the effect of Gamma(a, b) priors which appear well-suited to the stylometric application. Unlike Uniform distributions with truncated support that needs to be specified, Gamma distributions are defined on $(0, \infty)$. Another appealing feature of the Gamma family is that the Metropolis step (7) can be avoided. Instead, Escobar and West (1995) provide a direct expression for the posterior distribution of α as a mixture of two Gamma's involving an auxilliary parameter η. Dorazio (2009) provides a method of selecting the Gamma parameters a and b such that the number of possible clusters $1, \ldots N$ is apriori equiprobable. In our simulated baseline data with $N = 10$, Dorazio's (2009) method provides $a = 0.525$ and $b = 0.046$. In this case, the clustering methodology yields the partition $\{1\}$, $\{2\}$, $\{3\}$, $\{4, 9, 10\}$, $\{5\}$, $\{6\}$, $\{7\}$ and $\{8\}$, which as anticipated, contains more clusters than when using the Uniform$(0, 3)$ prior. However, if we were apriori more inclined towards fewer clusters, we might consider changing the clustering threshold from 0.5 to 0.4. With the lower clustering threshold, the posterior pairwise clustering probabilities yield the partition $\{1, 4, 9, 10\}$, $\{2, 3, 5, 6, 8\}$ and $\{7\}$ which is the same partition when using the Uniform$(0, 3)$ prior.

Further, we note that using various fixed values of α (e.g. $\alpha = 5.0$) does not result in markedly different clustering.

4.4 The effect of the text sizes $\tilde{n}_i$

Our intuition is that increasing text sizes leads to more information which in turn leads to improved clustering. To explore this conjecture, we generated a new dataset according to the specifications of Section 4.1 except that each of the text sizes is increased tenfold from $\tilde{n}_i = 5,000$ words to $\tilde{n}_i = 50,000$ words.

The pairwise clustering results corresponding to the new dataset are presented in Table 4. From Table 4, we observe "perfect" clustering (i.e. $\{1, 2, 3, 4, 5\}$, $\{6, 7, 8\}$, $\{9, 10\}$) and we note that the probability assessments are more definitive than in Table 2. These observations suggest that our stylometric methodology may provide better results when we study larger texts.

4.5 The effect of the number of function words K

In this section, we consider the effect of the number of function words K on clustering. Although our intuition suggests that larger K leads to more information and therefore improved clustering, we will see that the effect of K is less straightforward.

To investigate the effect of K, we ran the stylometric clustering algorithm on the baseline data of Section 4.1 using $K = 6$. The posterior pairwise probabilities are reported in Table 5 and lead to the two clusters $\{1, 2, 3, 4, 5, 6, 7, 8\}$ and $\{9, 10\}$. This is an improvement over the analysis of Section 4.1

Text i	Text j									
	1	2	3	4	5	6	7	8	9	10
1	0.00	0.90	0.88	0.88	0.92	0.01	0.00	0.00	0.00	0.00
2	0.90	0.00	0.86	0.85	0.90	0.01	0.00	0.00	0.00	0.00
3	0.88	0.86	0.00	0.90	0.89	0.02	0.00	0.00	0.00	0.00
4	0.88	0.85	0.90	0.00	0.88	0.02	0.00	0.00	0.00	0.00
5	0.92	0.90	0.89	0.88	0.00	0.01	0.00	0.00	0.00	0.00
6	0.01	0.01	0.02	0.02	0.01	0.00	0.84	0.77	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.86	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.77	0.86	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00

Table 4: The posterior probability of clustering between texts i and j based on larger text sizes of $\tilde{n}_i = 50,000$ words where $K = 2$ and $\alpha \sim \text{Uniform}(0, 3)$.

where $K = 2$ was used and clusters $\{1, 4, 9, 10\}$, $\{2, 3, 5, 6, 8\}$ and $\{7\}$ were obtained. We note that the pairwise probabilities in Table 5 are more affirmative than those in Table 2. This reflects one of the idiosyncrasies regarding the effect of increasing K. We remark that when $K = 10$, we obtain the same clustering results ($\{1, 2, 3, 4, 5, 6, 7, 8\}$ and $\{9, 10\}$) except that the pairwise probabilities become even more affirmative. For example, the posterior probability of clustering between text 1 and text 2 increases from 0.76 with $K = 6$ to 0.81 with $K = 10$.

When we set $K = 20$ and use the baseline dataset from Section 4.1, all of the 10 texts clustered together with posterior pairwise probabilities of 1.0. This is clearly problematic, and confirms the theoretical derivation from Section 3 that suggests that increasing K leads to the formation of fewer clusters. Although we have reported results only in the case of the dataset from Section 4.1, we have seen these patterns in all of the datasets that we have generated.

The complex role of K in our clustering algorithm does not appear to be generally well-known. The reason for this may be that most applications of the DP in hierarchical models (2) involve a parameter (p_i in our case) that is univariate. For example, often an experimenter has a univariate parameter that represents means of sub-populations. In our case, the parameter p_i has dimension $K - 1$, and it is the increased dimensionality which impedes the formation of additional clusters.

Text i	Text j									
	1	2	3	4	5	6	7	8	9	10
1	0.00	0.76	0.95	0.97	0.95	0.75	0.68	0.75	0.22	0.21
2	0.76	0.00	0.80	0.77	0.81	0.97	0.91	0.98	0.01	0.01
3	0.95	0.80	0.00	0.96	0.97	0.79	0.72	0.79	0.18	0.18
4	0.97	0.77	0.96	0.00	0.96	0.75	0.68	0.75	0.21	0.21
5	0.95	0.81	0.97	0.96	0.00	0.79	0.72	0.79	0.18	0.17
6	0.75	0.97	0.79	0.75	0.79	0.00	0.93	0.98	0.00	0.00
7	0.68	0.91	0.72	0.68	0.72	0.93	0.00	0.93	0.00	0.00
8	0.75	0.98	0.79	0.75	0.79	0.98	0.93	0.00	0.00	0.00
9	0.22	0.01	0.18	0.21	0.18	0.00	0.00	0.00	0.00	0.99
10	0.21	0.01	0.18	0.21	0.17	0.00	0.00	0.00	0.99	0.00

Table 5: The posterior probability of clustering between texts i and j for the baseline data of Section 4.1 where $K = 6$ and $\alpha \sim \text{Uniform}(0, 3)$.

When a new function word is added, will it assist in discovering new clusters? If the function word is truly discriminatory, it may. However, if the new function word is not discriminatory (i.e. $p_{i,K+1} \approx p_{j,K+1}$), then because the p_{ik}'s lie on a simplex (i.e. $p_{i1} + \cdots + p_{iK} = 1$), a new function word reduces the magnitude of $\mid p_{ik} - p_{jk} \mid$ for $k = 1, \ldots, K$.

It therefore appears that we need to strike a balance between the additional discriminatory information provided by increasing the number K of function words versus the tendency of forming fewer clusters. This highlights the importance of choosing a “good” but small set of function words in stylometric analyses.

5 The Federalist papers

In 1788, 85 articles were compiled and published as *The Federalist.* The main purpose of these essays (most of which had appeared earlier in newspapers) was to persuade citizens of the State of New York to ratify the new Constitution of the United States. It is widely assumed that the authors of these papers were limited to Alexander Hamilton, John Jay and James Madison. The overall consensus amongst scholars is that Hamilton was the author of 51 of the papers, Madison wrote 14, Jay wrote 5, and 3 were jointly written by Hamilton and Madison. The authorship of the remaining 12 papers (referred to as the *disputed papers*) is unknown, but each of these disputed papers is widely thought to be the work of either Hamilton or Madison. Mosteller and Wallace (1963, 1984) analyzed the distribution of function words extracted from the Federalist papers and concluded that Madison was the author of all 12 disputed papers. The Federalist papers are now considered a test case for new methods of authorship attribution and the papers are freely available from both the Project Gutenberg and the Library of Congress websites. An interesting account of earlier, though unsuccessful, efforts for determining authorship of the Federalist papers is given by Mosteller and Wallace (1984).

To investigate our methodology with respect to the Federalist papers, we amalgamate all of the 51 Hamilton papers into a single text. We also amalgamate the 14 Madison papers into a single text. We eliminate the 5 Jay texts from the analysis as the disputed texts are believed to be written by either Hamilton or Madison. We also eliminate the 3 joint papers as it is unclear that the writing styles of these texts ought to exhibit the style of Hamilton, the style of Madison or some intermediate style. Thus we have $N = 1 + 1 + 12 = 14$ texts under consideration where the text sizes of the disputed papers are much smaller than the text sizes of the two amalgamated texts.

The word frequency data for each of the papers were extracted using WordSmith tools (Scott 1998). We started with a list of 125 function words gathered from Tables 2.5, 2.6 and 2.7 of Mosteller and Wallace (1963). We emphasize that the starting list of function words was not based on any screening. As observed in the simulation study in Section 4, selection of a small number of function words is very crucial for stylometric analysis. In the case of Federalist papers we are interested in identifying the authorship of disputed papers between Hamilton and Madison. Therefore, it seems sensible that the function words used be such that they differ in the frequency of occurrence in the known works of Hamilton and Madison. Using the Z-test methodology of Section 3 on the amalgamated Hamilton and Madison federal papers, we ended up with 10 words with the largest $|Z|$ values. These words, ordered according to decreasing $|Z|$ value, are listed in Table 6.

ANOTHER	ALSO	ANY	AND	AS	ON	ARE	VOICE	AN	ALL

Table 6: The 10 function words used in the analysis of the Federalist papers.

We ran the Dirichlet process mixture algorithm using the selected 10 function and the continuous prior $\alpha \sim$ Uniform$(0, 3)$ and obtained the same conclusions as Mosteller and Wallace (1963, 1984). That is, two clusters were obtained with all 12 of the disputed papers clustering with Madison and the second cluster consisting of the amalgamated Hamilton text. The strength of membership of the disputed papers with Madison was strong. The smallest posterior probability of membership amongst the 12 disputed texts into the "Madison cluster" was 0.70.

6 Discussion

In this paper, we have developed an algorithm for the stylometric clustering of texts. Some of the features of the approach include suggestions on the selection of function words, the incorporation of the multinomial distribution and a semiparametric Bayesian framework based on the Dirichlet process. The Dirichlet process is well suited to the stylometric problem as clustering is a by-product of model development and posterior probability assessments can be obtained with respect to clusters. The approach also enables a user to incorporate prior knowledge concerning authorship by amalgamating texts which are believed to have the same author.

The prior for the function word probabilities p_{ik} and prior for the DP concentration parameter α have been proposed for the stylometric problem in hand. However, we note that apart from these priors, the proposed algorithm has much greater applicability as a general clustering tool for categorical (i.e. multinomial) data.

In an honest appraisal concerning the suitability of the algorithm for stylometric problems, one keeps in mind the adage that "the more that is assumed, the more that can be inferred". In the case of the DP mixture model, very little is assumed. It is therefore comforting that our approach replicates long-standing views concerning the authorship of the Federalist papers. In stylometric problems where there is controversy concerning authorship, it remains to be seen whether the proposed approach can differentiate texts. As investigated in Section 4, this is partly a function of text size and the "differentiability" between texts.

With the need to keep the number of function words K at a reasonable level, the approach may be best suited to stylometric problems where the potential number of authors (i.e. clusters) is relatively small. Ultimately, as discussed in some detail in the paper, the existence of a small set of good discriminating function words provides the best chance for success.

7 References

Bouguila, N. (2008). Clustering of count data using generalized Dirichlet multinomial distributions, *IEEE Transactions on Knowledge and Data Engineering*, 20, 462-474.

Dorazio, R.M. (2009). On selecting a prior for the precision parameter of Dirichlet process mixture models, *Journal of Statistical Planning and Inference*, 139, 3384-3390.

Dunn, J.C. (1977). Indices of partition fuzziness and the detection of clusters in large data sets, In M. Gupta (Ed.), *Fuzzy Automata and Decision Processes*, 271-284 (New York: Elsevier).

Escobar, M.D. & West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, 90, 577-588.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures, *Annals of Statistics*, 2, 615-629.

Gill, J. & Casella, G. (2009). Nonparametric priors for ordinal Bayesian social science models: specification and estimation, *Journal of the American Statistical Association*, 104, 453-464.

Gill, P.S., Swartz, T.B. & Treschow, M. (2007). A stylometric analysis of King Alfred's literary works, *Journal of Applied Statistics*, 34, 1251-1258.

Ghosh, P., Gill, P.S., Muthukumarana, S. & Swartz, T.B. (2010). A semiparametric Bayesian approach to network models using Dirichlet process priors, *Australian and New Zealand Journal of Statistics*, 52, 289-302.

Holmes, D.I. & Forsyth, R.S. (1995). The Federalist revised: New directions in author attribution, *Literary and Linguistic Computing*, 10, 111-127.

Holmes, D.I. (1999). Stylometry, *Encyclopedia of Statistical Sciences: Update Volume 3*, 721-727 (New York: Wiley).

Liu, J.S., Zhang, J.L., Palumbo, M.J. & Lawrence, C.E. (2003). Bayesian clustering with variable and transformation selections, In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith & M. West (Ed.), *Bayesian Statistics 7*, 249-276 (Oxford: Oxford University Press).

Mosteller, F. & Wallace, D. L. (1963). Inference in an authorship problem. A comparative study of discrimination methods applied to the authorship of the disputed Federalist papers, *Journal of the American Statistical Association*, 58, 275-309.

Mosteller, F. & Wallace, D.L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers* (New York: Springer-Verlag).

Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics*, 9, 249-265.

Ohlssen, D.I., Sharples, L.D. & Spiegelhalter, D.J. (2007). Flexible random-effects models using semi-parametric models: Applications to institutional comparisons, *Statistics in Medicine*, 26, 2088-2112.

Scott, M. (1998) *WordSmith Tools Manual, version 3.0* (Oxford: Oxford University Press).

Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica*, 4, 639-650.

Spiegelhalter, D. Thomas, A. & Best, N. (2003). *WinBUGS (Version 1.4) User Manual*, Cambridge: MRC Biostatistics Unit.

8 ACKNOWLEDGMENTS

The authors were partially supported by research grants from the Natural Sciences and Engineering Research Council of Canada. The authors are appreciative of the comments made by an anonymous referee. The comments have lead to an improvement in the manuscript.