Extended nearest shrunken centroid classification: A new method for open-set authorship attribution of texts of varying sizes

G. Bruce Schaalje and Paul J. Fields
Department of Statistics, Brigham Young University, Provo, UT, USA

Matthew Roper
Neal A. Maxwell Institute for Religious Scholarship, Brigham Young University, Provo, UT, USA

Gregory L. Snow
Statistical Data Center, LDS Hospital Intermountain Health Care, Salt Lake City, UT, USA

Abstract

The nearest shrunken centroid (NSC) methodology, originally developed for high-dimensional genomics problems, was recently applied in a stylometric study. Although NSC has many advantages, stylometric problems usually differ from genomics problems in several important ways: texts are of a wide range of sizes, a large series of texts are often the subjects for classification, and most importantly the set of candidate authors cannot usually be assumed to be closed. Consequently, naïve application of NSC methodology can produce misleading results. We extend the NSC methodology for more general application to stylometry. Reanalysis of the Book of Mormon using the open-set NSC method produced dramatically different results from a closed-set NSC analysis.

Correspondence:
G. Bruce Schaalje, Department of Statistics, Brigham Young University, Provo, UT 84602, USA.
E-mail:
schaalje@byu.edu

1 Introduction

Many, possibly most, authorship attribution problems are 'open games' (Burrows 2002) in which it is not realistic to unquestionably assume that the author of the text in question is contained in the candidate set (Koppel *et al.*, 2009). One purpose of this article is to show that failure to recognize or deal with this issue produces highly misleading results.

In a recent example, Jockers *et al.* (2008) applied a novel technology, nearest shrunken centroid (NSC) classification (Tibshirani *et al.*, 2002, 2003), as well as the well-established delta method (Burrows, 2002, 2003; Hoover, 2004; Argamon, 2008) to the investigation of a specific theory of authorship of the Book of Mormon. This attribution problem is almost certainly an open game given the provenance of the Book of Mormon (for a brief discussion of the Book of Mormon authorship controversy see

Literary and Linguistic Computing, Vol. 26, No. 1, 2011.
doi:10.1093/llc/fqq029 Advance Access published on 18 January 2011

Appendix A), but Jockers *et al.* (2008, p. 470) treated it as a closed-set classification problem with a very small set of candidate authors.

NSC classification is one of several statistical methods lately developed for ultra high-dimensional genomics problems (hundreds or thousands of variables measured on relatively few individuals). The application of these genomic methods in stylometry is an attractive trend because stylomery itself is a high-dimensional problem: hundreds of literary features measured for relatively few blocks of text. The fields of stylometry and genomics, though seeming to have very little in common, actually have a history of cross-fertilization relative to statistical methodology (Giron *et al.*, 2005; Sims *et al.*, 2009).

While NSC and related methodologies have much potential in stylometric analysis (Jockers and Witten, 2010), the use of the closed-set NSC method in the Book of Mormon study (Jockers *et al.*, 2008) points to the need for extensions to the method. The primary difficulty is that basic NSC requires the training author set to be closed. In addition, a goodness-of-fit test for the model has not been developed, the model requires test texts to be of the same size, and problems of multiplicity have not been discussed in connection with the method. These issues would not usually be problems in genomics investigations using NSC, but they are important issues in most authorship attribution studies. Relatively simple modifications to the closed-set NSC method can extend it to overcome these problems.

Burrows' delta method is a well-used tool in authorship attribution, but it can be strengthened. For example, it produces probability rankings rather than actual authorship probabilities so that the selected author must be referred to carefully as the 'least unlikely' candidate (Burrows, 2003). Text size plays no role in the calculation of delta. The threshold for false-positive attribution (Hoover, 2004) is based on empirical evidence rather than theory. The delta method is more difficult to generalize than the related 'quadratic delta' method (Argamon, 2008). The extended version of NSC proposed in this article, which is closely related to quadratic delta, deals with these issues.

In this article, we review details of the closed-set NSC classification model, discuss problems with its naïve application to stylometric problems, and propose extensions to it that will make it a much more useful tool in stylometrics. We then apply it to an open-set problem, the Book of Mormon authorship data, to demonstrate its usefulness and highlight its differences from closed-set NSC.

2 The NSC Classification Model

To understand the closed-set NSC method, consider a vector x of r literary features (for example, the relative frequencies of r non-contextual words) that has a different distribution for each of m candidate authors. Let $f_i(x)$ denote the joint density function of x for author i. Let the true author of some new text be unknown, even though it is known that one of the m candidates is the true author. Burrows (2002) calls this situation the 'closed game'. This assumption is critical. If met, the posterior probability that author k is the author of the new text is

$$p(k|x^*) = \frac{f_k(x^*)\pi_k}{\sum_{i=1}^{m} f_i(x^*)\pi_i} \qquad (1)$$

where x^* denotes the vector of features for the new text and π_i denotes the prior probability that author i is the author of the new text. The denominator in Equation (1), called the normalizing constant, transforms the ratios into meaningful authorship probabilities. In practice, the densities have to be estimated from training data of $n_i, i = 1, \ldots, m$, samples from the m authors, respectively. A common method of classification is to assign the text to the author with the highest posterior probability; that is, choose author $\hat{h}$ where

$$\hat{h} = \text{argmax}_i[\hat{p}(i|x^*)] \qquad (2)$$

and $\hat{p}(i|x^*)$ is the posterior probability of author i based on estimated densities. Using estimated densities, the procedure should approximately minimize classification errors.

The NSC procedure of Tibshirani *et al.* (2003) is a special case of this procedure in which the densities are assumed to be multivariate normal and the

features are assumed to be mutually independent. Thus, after simplification,

$$p(k|x^*) = \frac{\pi_k e^{-\frac{1}{2}\sum_{j=1}^{r}\left((x_j^* - \mu_{kj})/\sigma_j\right)^2}}{\sum_{i=1}^{m}\pi_i e^{-\frac{1}{2}\sum_{j=1}^{r}\left((x_j^* - \mu_{ij})/\sigma_j\right)^2}} \quad (3)$$

where μ_{ij} is the mean for author i and feature j, and σ_j^2 is the common variance for feature j across authors.

This model is closely related to the 'quadratic delta' stylistic distance of Argamon (2008) and thus is also closely related to Burrows' delta. In fact, if the prior probabilities were specified to be equal, the author with the highest posterior probability would also be the author with the lowest quadratic delta. The normalizing constant and the prior probabilities (see Equation 3) transform the probability ranking based on quadratic delta into a normalized (actual) probability.

The NSC procedure estimates σ_j^2 with s_j^2, the pooled within-author variance for feature j. One of the innovative features of NSC is that it estimates μ_{ij} by thresholding and shrinking $\bar{x}_{ij}$, the sample mean of the training data for feature j and author i, toward $\bar{x}_j$, the mean for feature j across authors. Shrinkage allows information about the whole set of candidate authors to be used in estimating the true vector of means (centroid) for each author. The shrinkage equation is

$$\tilde{x}_{ij} = \bar{x}_j + q_{ij} \times \operatorname{sign}\left(\bar{x}_{ij} - \bar{x}_j\right) \times \max\left[\left(\frac{\left|\bar{x}_{ij} - \bar{x}_j\right|}{q_{ij}} - \Delta\right), 0\right], \quad (4)$$

$$q_{ij} = \left(s_j + \delta\right)\sqrt{1/n_i + 1/n}, \quad (5)$$

where Δ is a shrinkage parameter, $\delta = \text{median}(s_j, j = 1, \ldots, r)$, and $n = \sum_{i=1}^{m} n_i$. Hence,

$$\hat{p}(k|x^*) = \frac{\pi_k e^{-\frac{1}{2}\sum_{j=1}^{r}\left((x_j^* - \tilde{x}_{kj})/s_j\right)^2}}{\sum_{i=1}^{m}\pi_i e^{-\frac{1}{2}\sum_{j=1}^{r}\left((x_j^* - \tilde{x}_{ij})/s_j\right)^2}} \quad (6)$$

Genomics researchers do not universally agree that the NSC method minimizes classification errors (Dabney, 2005), but the method has been widely and successfully used.

3 An Extension to the NSC Model to Deal with an Open Candidate Set

The reliability of any authorship attribution method depends on its ability to handle open-set problems (Koppel *et al.*, 2009). Juola (2006, p. 289) referred to this as the 'none-of-the-above' authors scenario. Burrows (2002, p. 268) felt that one of the greatest advantages of the delta method was its capability 'of distinguishing the most likely candidates from a large group and also, where no candidate lays sufficient claim, of indicating that it might be wise to look further afield'. He also recommended that the delta method be used initially as a 'prelude to tests in the ''closed form'' ' (Burrows, 2002, p. 277).

On the surface, the delta method would seem to only work for closed games because it gives probability rankings rather than actual probabilities. As Burrows (2003) and Hoover (2004) pointed out, the candidate author whose style is 'least unlike' the test text based on the delta value need not be very 'like' the test text. Empirically, however, it often turns out that if the smallest delta value is a great deal smaller than delta values for all other candidate authors, the chosen author is likely the true author. Based on this observation, the open game can become a closed game. To aid in making this determination, Burrows' (2003) method standardizes the delta values and then imposes a threshold (e.g. −1.9) on the resulting delta-z values. If the smallest delta-z is less than the threshold, the set is considered to be closed and the attribution is taken seriously. Otherwise, the candidate set is considered to be open and the attribution is considered to be a false positive.

The problem with this approach to open-set classification is that it is empirical, and the threshold depends on the nature of the corpus (Hoover, 2004). It could be fooled by a corpus in which

one false author is much more similar to a test text than all other false authors, thus falsely producing small delta-z values.

We now propose a theoretical rather than empirical extension of the closed-set NSC procedure that allows it to be useful in open-set situations. We first emphasize that naïve use of closed-set NSC classification in open-set situations will always produce inflated and misleading values for the posterior probabilities. To see this, assume without loss of generality that author m is the author of the new text. The posterior probability of author k can be written as

$$p(k|x^*) = \frac{f_k(x^*)\pi_k}{f_m(x^*)\pi_m + \sum_{i=1}^{m-1} f_i(x^*)\pi_i}. \qquad (7)$$

If author m is left out of the candidate set, the denominator in Equation (7) decreases and naïve use of the posterior probability formula yields

$$p'(k|x^*) = \frac{f_k(x^*)\pi_k}{\sum_{i=1}^{m-1} f_i(x^*)\pi_i} > p(k|x^*) \qquad (8)$$

Thus, when the true author is not included, the normalizing constant is too small and all the calculated posterior probabilities are inflated. The posterior probabilities lose their meaning, as does the associated classification. Even conceding that the candidate author with the highest posterior probability has the smallest quadratic delta stylistic distance to the new text, there is no reason to attach any meaning to the classification results because the smallest distance might still be very large. The inflation of the posterior probabilities is insidious because the inflated naïvely calculated probabilities tend to instill false confidence about the classification results.

A simulation illustrates this problem. We simulated one literary feature for seven training authors and one test author. The means of the candidate authors were assumed to follow a normal distribution with grand mean μ and between-author standard deviation (σ_B) one to four times larger than the within-author standard deviation (σ_W). For each author, the literary feature was assumed to follow the normal distribution with the selected mean and constant standard deviation σ_W. The discrepancy between the mean of the test author and μ varied from $0\sigma_W$ to $10\sigma_W$. Twenty-five training samples were taken from each of the training author distributions, and one sample was taken from the test author distribution. Based on the restricted maximum likelihood estimates of σ_B and σ_W, empirically shrunken means (Rencher and Schaalje, 2008, p. 500), and equal prior probabilities for all candidate authors, Equation (6) was used to calculate posterior probabilities that the test text was written by each of the training authors. The maximum of these posterior probabilities was averaged over 1,000 simulation runs. Consistent with Equations (7) and (8) but contrary to what one might naively expect, as the discrepancy between the mean of the test author and the grand mean of the training authors increased, the incorrectly calculated maximal posterior probabilities approached one (Fig. 1). As Equations (7) and (8) predict, the most highly inflated posterior probabilities for training authors occurred when the style of the test author was most different from the training authors. Thus, the unmodified NSC procedure can only be used for author attribution when it is known that the actual author is among the candidate authors, the classical example being the Federalist study (Mosteller and Wallace, 1964; Burrows, 2002).

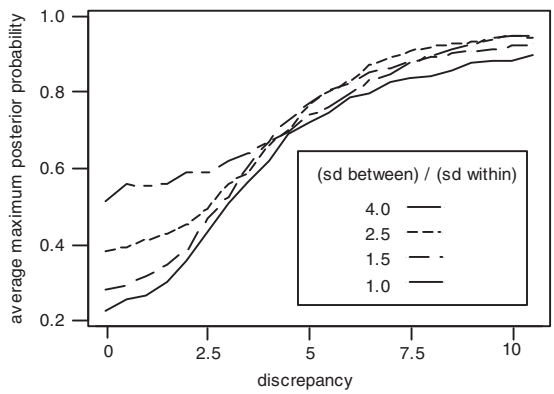

Fig. 1 Naively calculated posterior probabilities of authorship when the test author is not among the set of training authors. Maximum posterior probabilities were averaged over 1,000 simulation runs

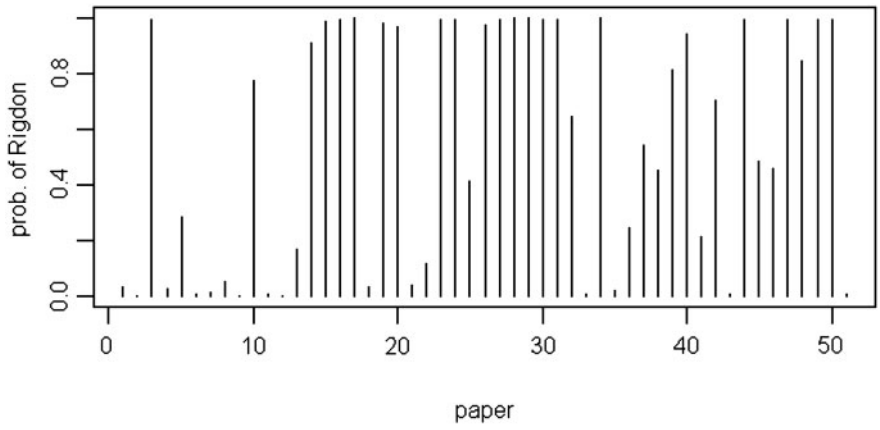

Fig. 2 Posterior probabilities of authorship of 51 Alexander Hamilton Federalist papers by Sidney Rigdon. Probabilities were calculated using closed-set NSC classification in which Hamilton was not among the candidate authors in the training data

To further illustrate the problem of probability inflation, we created an artificial authorship attribution problem in which the style of the test texts was deliberately chosen to be far different from those of all training authors. As training data, we computed 130 literary features for word blocks from six nineteenth-century authors connected with early Mormonism: Joseph Smith, early Sidney Rigdon (1831–46), late Sidney Rigdon (1863–73), Solomon Spalding, Oliver Cowdery, and Parley P. Pratt (see Appendices A and B). As test data, we calculated the same features for the 51 Federalist papers authored by Alexander Hamilton. We then naively used the closed-set NSC procedure to calculate posterior probabilities and classifications for the Hamilton texts (as if they were anonymous). The 130 literary features included relative frequencies of 93 non-contextual words, 35 word-pattern ratios (Hilton 1990), and 2 vocabulary richness measures (Holmes 1992).

Early or late Rigdon was falsely chosen as the author of 28 of the 51 Hamilton texts with inflated posterior probabilities ranging as high as 0.9999 (Fig. 2). Pratt was falsely chosen as the author of 12 of the papers, and Cowdery was falsely chosen as the author of the remaining 11 papers. These results dramatically demonstrate the danger of misapplying closed-set NSC.

One message of this example is that before applying closed-set NSC to any authorship attribution problem, an initial examination of the data must be carried out to see if the x^* vectors for the test texts are reasonably near the distribution of x vectors in the training set for at least one of the candidate authors. A dimension reduction procedure such as principal components analysis (PCA) or a high-dimensional dynamic visualization program such as GGobi (Buja *et al.*, 2003) must be used for this even though some information is lost when visualizing a high-dimensional data set in two or three dimensions. A principal components plot (Fig. 3) shows, as expected, that the test (Hamilton) texts are highly distinct from those of every author in the training data; thus, one expects that inflation of posterior authorship probabilities will be a serious problem for naïve authorship attribution of these test texts using closed-set NSC. Although the test authors do not appear very distinct in Fig. 3, a principal components plot (not shown) of only the training author data shows that they are reasonably distinct.

Note that principal components plotting is not a competitor with NSC classification; it is not 'a test of authorship but only of comparative resemblance' (Burrows, 2003, p. 8). Principal components is an unsupervised learning tool, useful in discovering

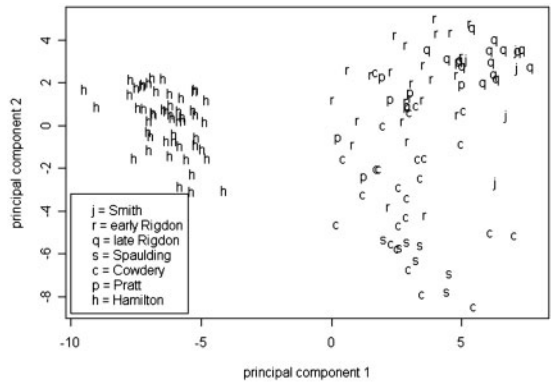

Fig. 3 First two principal component scores of 130 authorship features for texts written by Smith, Rigdon, Spalding, Cowdery, Pratt, and Hamilton. Although these components only account for 20.3% of the total variability, the large distinction between the Hamilton texts and those of the training authors indicates that closed-set NSC should not be used to classify the test (Hamilton) texts

general clustering patterns of the data. NSC is a supervised learning technique in which information about the known clustering structure of the data is used to classify new individuals. NSC is designed to optimally use authorship information in classification of new texts. In the present artificial classification case, the PCA plot clearly shows that closed-set attribution of the test (Hamilton) papers to the19th-century authors in question would be highly misleading.

After applying the closed-set NSC procedure, a goodness-of-fit check (Gelman *et al.*, 2004) should be carried out to validate the posterior probabilities and associated classifications. One way this can be done is to compute several posterior predicted vectors for each test text using the NSC posterior probabilities, and then compare the posterior predicted vectors with the observed vectors for the test texts. If the observed and predicted vectors do not generally agree, the classifications are not valid; closed-set NSC was used when the open-set extension should have been used.

Given a set of NSC posterior probabilities $\hat{p}(i|x^*), i = 1, \ldots, m$ for a new text with x^* as the vector of features, posterior predicted vectors are random draws from the finite mixture distribution

$$\hat{f}_*(y^*) = \sum_{i=1}^{m} \hat{f}_i(y^*)\hat{p}(i|x^*). \tag{9}$$

To demonstrate this procedure, we trained the closed-set NSC model on the Smith, early Rigdon, late Rigdon, Spalding, Cowdery, and Pratt data, classified the Hamilton texts, and then generated 10 posterior predicted vectors for each of the Hamilton texts using the finite mixture distribution. We then used principal components analysis to visually compare the observed and predicted vectors (see Fig. 4, left panel). Even though the first two components account for only 6.4% of the total variability, it is obvious that the posterior predictions are completely distinct from the observed test texts (Hamilton), and thus the closed-set NSC classification results are invalid. More detail on this goodness-of-fit procedure is given elsewhere (Schaalje and Fields, 2011).

An important extension to NSC classification is to allow an open set, i.e. the possibility that the test texts might not be authored by any of the candidate authors. We propose that this can be done by positing an unobserved author for each test text in addition to the observed candidates in the training data. We propose an unobserved author with a distribution of literary features just barely consistent with the test text. Thus, as a straightforward extension of the NSC classification model, we suggest that posterior probabilities for the candidate authors be calculated as

$$\hat{p}(k|x^*) = \frac{\pi_k e^{-\frac{1}{2}\sum_{j=1}^{r}\left((x_j^* - \tilde{x}_{kj})/s_j\right)^2}}{\sum_{i=1}^{m} \pi_i e^{-\frac{1}{2}\sum_{j=1}^{r}\left((x_j^* - \tilde{x}_{ij})/s_j\right)^2} + \pi_{m+1} e^{-\frac{1}{2}\sum_{j=1}^{r} a_j^2}} \tag{10}$$

where

$$a_j = \min\left(\max_i \left|\frac{x_j^* - \tilde{x}_{ij}}{s_j}\right|, \lambda\right),$$

and λ is a tuning constant representing the maximum allowed distance for components of x^* from

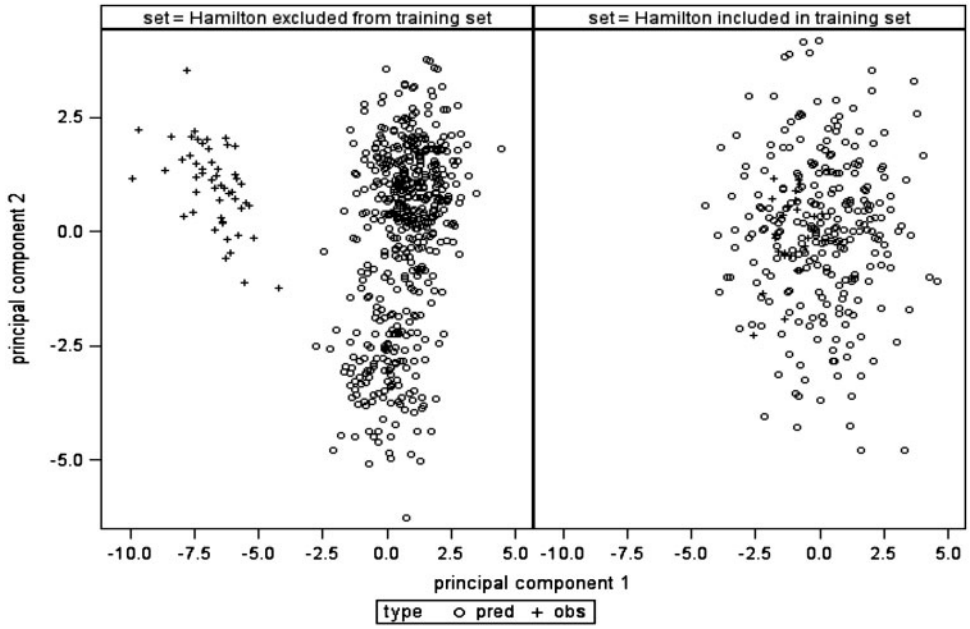

Fig. 4 On the left, principal components plot of observed test vectors and posterior predictions using the closed-set NSC model for the Hamilton texts with Hamilton excluded from the training set. The distinction between observed and predicted vectors indicates that the closed-set model is invalid. On the right, principal components plot of observed test vectors and posterior predictions using the closed-set NSC model for the Hamilton texts with Hamilton included in the training set. Agreement of observed and predicted vectors indicates that the closed-set model is valid

corresponding components of the centroid of an unobserved or latent author. The approximate posterior authorship probability (actually a lower bound) of an unobserved author is

$$p(m+1|x^*) = \frac{\pi_{m+1} e^{-\frac{1}{2}\sum_{j=1}^{r} a_j^2}}{\sum_{i=1}^{m} \pi_i e^{-\frac{1}{2}\sum_{j=1}^{r} \left((x_j^* - \tilde{x}_{ij})/s_j\right)^2} + \pi_{m+1} e^{-\frac{1}{2}\sum_{j=1}^{r} a_j^2}}. \quad (11)$$

A lower limit for λ is 3 since, for a standard normal vector $\boldsymbol{z}$ of dimension ν,

$$\text{prob}(|z_i| < \lambda, i = 1, \ldots, \nu) = \theta \text{ if } \lambda = \left|\Phi^{-1}\left(\frac{1-\theta^{1/\nu}}{2}\right)\right|$$

where $\Phi^{-1}(.)$ is the inverse cumulative Student's t distribution function. As ν increases to 100, λ increases to about 3 when $\theta = 0.8$. More detail on the selection of λ is given by Schaalje and Fields (2011).

Applying this extended model to the Hamilton texts with Smith, early Rigdon, late Rigdon, Spalding, Cowdery, and Pratt as training authors, only 2 of the test texts were assigned to early or late Rigdon, while the remaining 49 were assigned to an unobserved author (obviously Hamilton) (Fig. 5).

As a further test of the open-set NSC procedure, in addition to Rigdon, etc., we included Hamilton as a training author represented by the first 25 Hamilton papers. We classified the remaining 26 Hamilton papers as test texts. We first used the closed-set model. All 26 Hamilton test texts were correctly assigned to Hamilton; none was assigned to an unobserved author. The goodness-of-fit procedure (Fig. 4, right panel) indicated that the closed-set model was valid. We then used the

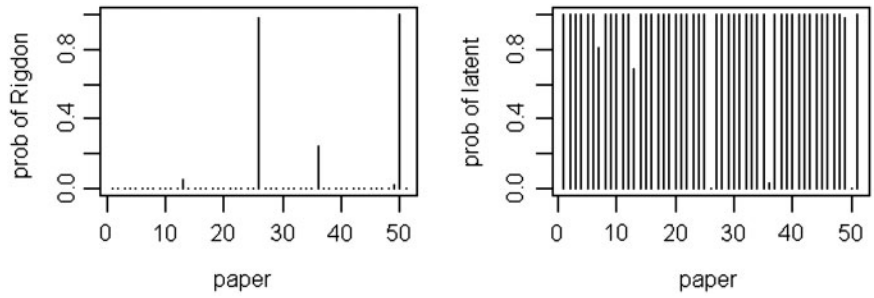

Fig. 5 Posterior probabilities of authorship of 51 Hamilton Federalist papers by Rigdon and latent authors. Probabilities were calculated using an open-set NSC classification method in which Hamilton was not among the candidate authors, but a latent author was allowed for each text

open-set model. All 26 Hamilton test texts were still correctly assigned to Hamilton. Hence, when the actual author was included in the training set, the allowance for an unobserved author as in Equation (10) did not appear to compromise the ability of open-set NSC to correctly attribute authorship.

It is important to note that the open-set NSC procedure does not indicate how many unobserved authors there are. All we know is that if an unobserved author is selected for a test text, one unobserved author is most probable as the author of that text. There could be as many unobserved authors as the number of test texts, or as few as one. A clustering procedure would provide some information as to the total number of unobserved authors. A related issue is that the goodness-of-fit procedure (Equation 9 and Fig. 4) could be applied to results from open-set NSC, but predicted vectors could only be produced for texts assigned to observed authors because very little is known about the unobserved author(s).

4 An Extension to the NSC Model to Deal with Heterogeneous Test Text Sizes

Choosing text sizes in stylometric studies is a balancing act between conflicting aims (Peng and Hengartner, 2002). Jockers *et al.* (2008) made the case that in classification studies it is more important to break writings into meaningful samples to avoid imposing bias on authorship features than to worry about text sizes. While this is a good point, it is not a justification for ignoring text size. Jockers *et al.* (2008) took it to the extreme in using training texts ranging from 114 to 17,979 words, and test texts ranging from 95 to 3,752 words. The measurement of 100 or more word frequencies on texts of less than 100 words is almost sure to produce unreliable measurements (Holmes and Kardos, 2003). For the delta procedure, Burrows (2003, p. 21) found that 'with texts of fewer than two thousand words in length... the test gradually becomes less effective'. Others have worked with texts of 1,000, 5,000, and 10,000 words (Larsen *et al.*, 1980; Hilton, 1990; Holmes, 1992). Because the NSC method produces probabilities of authorship, it does not make sense to ignore the reliability of measurements from the texts when calculating probabilities. One expects that small texts must have greater uncertainty about authorship than large texts.

To investigate this, we divided the early Rigdon texts into blocks of size 100, 200, 500, 1,000, 2,000, 3,000, 4,000, and 5,000 words, and then calculated between-text variances for the text sizes. There were clear relationships between text size and variance for non-contextual word frequencies (Fig. 6) as well as word-pattern ratios and vocabulary richness measures (not shown).

A relationship between text size and an authorship feature is common in linguistic studies

(Heaps, 1978). The observed relationships of variance to text size for non-contextual words (Fig. 6) are as expected. If a non-contextual word occurred at random with a constant intensity equal to the mean per word, μ_x, the Poisson model would apply; the variance–text size relationship would be $\sigma_x^2 = \mu_x/w$, where x is the relative frequency of the non-contextual word and w is the text size. If the intensity was not constant, one might expect the Smith (1938) law to apply: $\sigma_x^2 = \mu_x/w^b$, where b is the 'heterogeneity index'. We used a slightly more general model: $\sigma_x^2 = c\mu_x^a/w^b$ where a, b, and c are constants. Empirical values for the constants were obtained by fitting the variance model jointly to squared deviations from the mean for all word frequencies for the Hamilton and late Rigdon data. The fitted model is superimposed in Fig. 6. Similar models applied for vocabulary richness measures and word-pattern ratios.

Using these relationships, the posterior probability formula can be adjusted to take into account heterogeneous test text sizes. For example, if training text sizes were all greater than 1,000 so that their variances would be nearly equal (Fig. 6), but test texts were highly heterogeneous in size, posterior probabilities could be calculated as

$$p(k|x^*) = \frac{\pi_k e^{-\frac{1}{2}\sum_{j=1}^{r}\left((x_j^* - \tilde{x}_{kj})/s(\tilde{x}_{kj}, w^*)\right)^2}}{\sum_{i=1}^{m}\pi_i e^{-\frac{1}{2}\sum_{j=1}^{r}\left((x_j^* - \tilde{x}_{ij})/s(\tilde{x}_{ij}, w^*)\right)^2} + \pi_{m+1}e^{-\frac{1}{2}\sum_{j=1}^{r}a_j^2}} \quad (12)$$

where

$$a_j = \min\left(\max_i\left|\frac{x_j^* - \tilde{x}_{ij}}{s(\tilde{x}_{ij}, w^*)}\right|, \lambda\right)$$

and

$$s^2(\tilde{x}_{ij}, w^*) = \begin{cases} \frac{0.259\tilde{x}_{ij}^{1.1147}}{w^{*0.661}} \text{ for non-contextual word frequencies} \\ \frac{200\tilde{x}_{ij}}{w^*} \text{ for word-pattern ratios} \\ 65000\tilde{x}_{ij}\left(\frac{1}{w^*} + 40\right) \text{for vocabulary richness (R)} \end{cases}$$

If the training text sizes were highly heterogeneous, the procedure could be generalized but would be

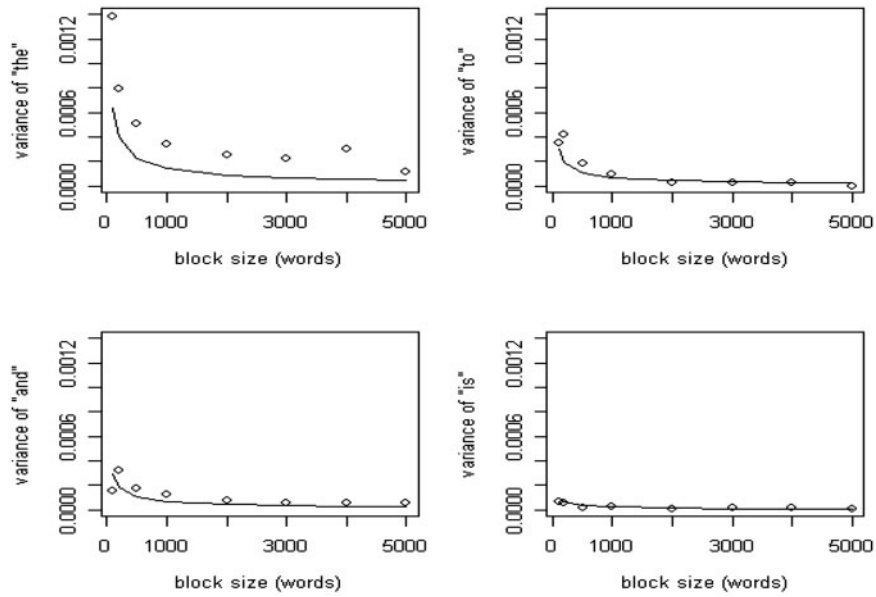

Fig. 6 Between-text variances for relative frequencies of 'the', 'to', 'and', and 'is' for early Rigdon texts of various sizes. The lines are based on a model for variance as a function of the mean and text size

more complicated. Shrinkage and thresholding would be much more complex. Hence, we recommend in general that the training data involve only texts of at least 1,000 words because feature-specific variances do not change greatly with text size beyond 1,000. Within limits, the problem of training text size variation can be dealt with simply by compositing shorter texts of known authorship to create training texts of at least 1,000–2,000 words. Hoover (2004), in fact, found that combining several texts 'helps to improve accuracy' of authorship attribution.

To illustrate the effects of both extensions (Equations 10 and 12) to the NSC method, we applied the closed-set NSC method and the two extensions to 95 'revelations' attributed to Sidney and Phebe Rigdon between 1863 and 1873, decades after Rigdon had left the Mormon movement. The test texts ranged in size from 60 to 4,128 words. We used the Smith, Cowdery, Spalding, Pratt, and early Rigdon texts as the training data, and specified informative priors based on the fact that Smith, Cowdery, Spalding, and Pratt had all died long before 1863. The closed-set NSC model attributed the texts mainly to Rigdon and Smith (Fig. 7), the open-set NSC model attributed most of the texts to latent authors, and the fully expanded NSC model attributed the texts to Rigdon, Smith, and latent authors. The point here is that open-set NSC without adjustments for test texts sizes is inadequate if some of the test texts are very small.

5 Extensions to the NSC Model to Deal with Multiplicity

A disturbing feature of classification analysis when the set of test texts is large is that test texts on the

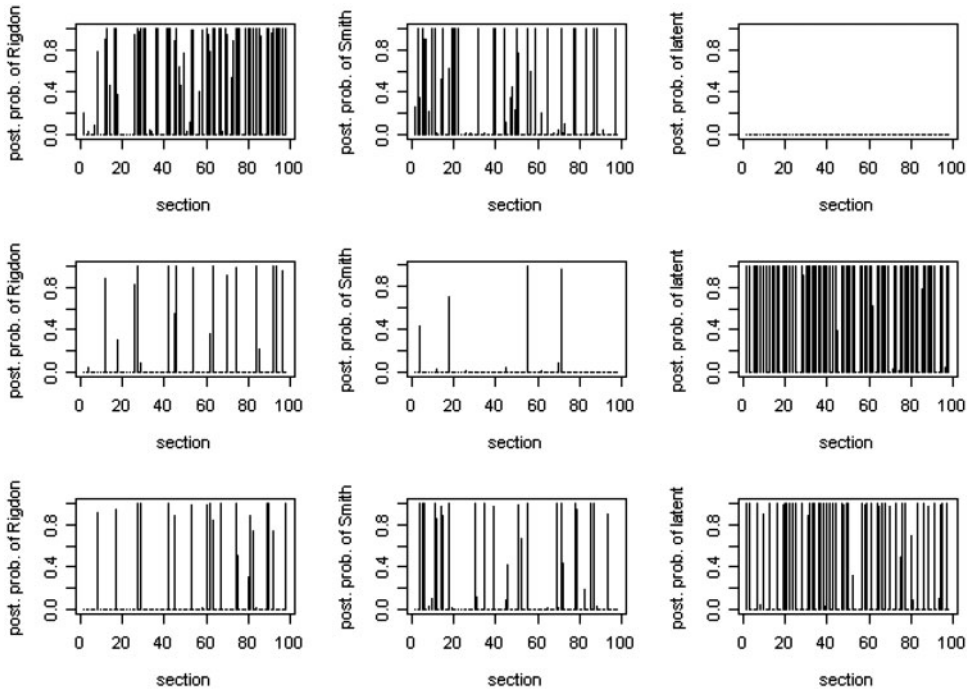

Fig. 7 Posterior probabilities of authorship of 95 Sidney and Phebe Rigdon texts. The upper panel used the closed-set NSC classifier, the middle panel used the open-set NSC classifier, and the lower panel used the open-set NSC classifier adjusting for text size. Prior probabilities of authorship were set to 0.9 for Rigdon, 0.02 for Smith, 0.02 for Cowdery, 0.02 for Spalding, 0.02 for Pratt, and 0.02 for an unobserved author

stylistic fringe of the distribution for the true author can occur by chance, and may therefore 'stray' into the distribution of a nearby author. This explains why 2 of the 51 Hamilton texts were assigned to Rigdon (Fig. 5), and might partially explain why 21 of the 95 late Rigdon texts classified strongly as writings of Smith (Fig. 7) even though Smith had died 20 years earlier. Historians who study this period would be hard-pressed to imagine any way that Rigdon could have retained otherwise unknown Smith texts.

The same problem was observed by Hoover (2004) with regard to the delta method. He noted (Hoover, 2004, p. 460) that for particular sets of authors and texts, 'false attributions are a serious possibility'. Burrows (2002, p. 281) similarly cautioned that the 'the system for distinguishing between insiders and outsiders is not foolproof' because of its dependence on probabilities rather than absolutes.

This problem, which is exacerbated by heterogeneity in text sizes, is an example of the multiplicity or multiple comparisons problem in statistics (Benjamini and Hochberg 1995). One not completely satisfactory solution would be to composite all of the test texts into one or a few large texts, and then classify those texts. We combined the Sidney texts into two large texts, combined the joint Phebe–Sidney texts into one large text, and combined the Phebe texts into one text. Assigning realistic prior probabilities, the first Sidney text was classified to an unobserved author, the second Sidney text was assigned to Cowdery, the joint text was assigned to Cowdery, and the Phebe text was assigned to an unobserved author. These results indicate, at a minimum, that the authorship style of the late Rigdon texts was different from that of Rigdon's earlier writings. This may be due to genre differences, the passage of time, or the interposition of editors. In any case, the cause of the difference is not germane to this study.

In the case of authorship attribution of a collection of texts suspected to be written by different authors, the texts cannot be composited. The posterior probabilities and classification results alone must then be interpreted as a whole. Individual classifications must not be overinterpreted.

As an illustration of this issue, we continued the simulation used in Fig. 1. For the simple case of five authors, one literary feature, 25 textual samples of the same size from each author, and $\sigma_B/\sigma_W = 5$, we simulated the NSC classification of 100 texts by a single author. A wide range of posterior probabilities occurred (Fig. 8).

Another solution to the multiplicity problem in the case that the texts are sequential (as they are for the late Rigdon texts) would be to 'smooth' the posterior probabilities (Cleveland, 1981). The posterior probabilities are not likely to change abruptly or unpredictably from text t to text $t+1$. Hence, sequential information can be used by the smoothing technique to remove some of the effects of multiplicity. The smoothed posterior probabilities for the simulated data give a more consistent idea of authorship evidence for the texts as a group (Fig. 8). We are continuing our research on the most effective way to implement this idea, especially with sequential texts that are unequally sized.

6 Revised Attribution of Book of Mormon Chapters Using Open-Set NSC

We now apply the open-set NSC method to Book of Mormon data in order to demonstrate its usefulness

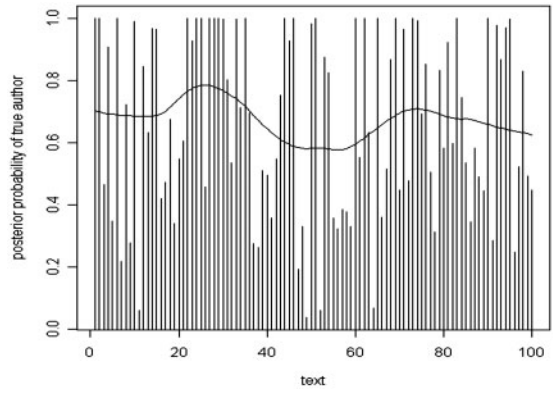

Fig. 8 NSC posterior probabilities of authorship of 100 simulated texts of the same size by the same author. A lowess smoother (with smoothing parameter 1/3) is superimposed

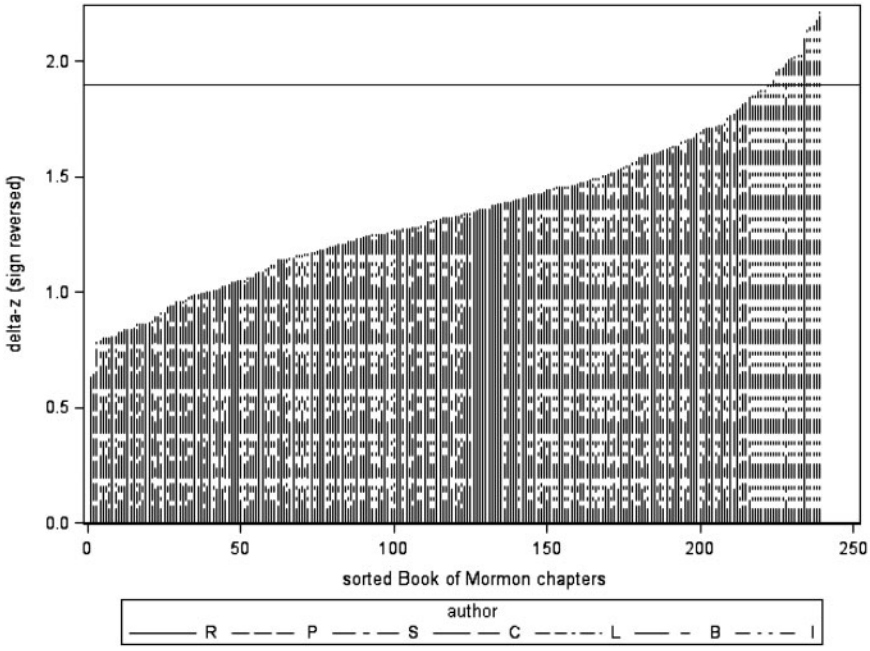

Fig. 9 Burrow's sign-reversed delta-*z* values for the 239 chapters of the Book of Mormon using data from Jockers *et al.* (2008). The chapters are sorted by delta-*z* values, and the line style indicates the author with the smallest delta-*z* value. The horizontal line is at 1.9, the threshold used by Burrows (2003)

and highlight its differences from closed-set NSC. Jockers *et al.* (2008) applied both closed-set NSC and the Burrows' delta method to the problem of authorship attribution of the 239 chapters of the Book of Mormon. Their set of candidate authors was closed, and some training texts were as small as 114 words. Some of the test texts (Book of Mormon chapters) were as small as 95 words.

Using closed-set NSC, Jockers *et al.* (2008) attributed 37% of the chapters to Rigdon, 28% to Isaiah/Malachi, 20% to Spalding, 9% to Cowdery, 5% to Pratt, and 1% to Longfellow. In contrast, using open-set NSC, we conclude that 73% of the chapters cannot be reliably attributed to any of the candidate authors. We first note that Jockers *et al.* (2008) bolstered their NSC attributions by claiming close agreement between attribution results due to Burrows' delta and those due to closed-set NSC. That these stylistic measures would nominally agree well numerically is not surprising because Burrows' delta stylistic distance is closely related to the quadratic delta stylistic distance (Argamon, 2008) upon which NSC is based.

However, there actually is strong disagreement between the closed-set NSC results and the delta results. This is because delta-*z* scores should not be taken seriously unless they are very small (i.e. very negative). Burrows (2003) found that a threshold of −1.9 separated most false positives from true attributions for a set of 17th-century poets. Jockers *et al.* (2008) failed to do this. In the Jockers *et al.* (2008) study, only 16 of the 239 chapters had delta-*z* values as small as −1.9 (Fig. 9). Ten of these 16 chapters were essentially verbatim quotations of Isaiah/Malachi, and all 10 were correctly attributed to Isaiah/Malachi. Four additional chapters were attributed to Isaiah/Malachi and the others to Rigdon and Spalding. The remaining 223 chapters had large delta-*z* values and were thus apparently false positive. Hence, the delta results of Jockers *et al.* (2008) actually say little more than what is already uncontroversial about Book of Mormon

authorship: that some of the chapters are quotations of Isaiah and Malachi. The delta-*z* results do not, in fact, attribute sizeable percentages of the chapters to Rigdon, Spalding, or Cowdery.

We subjected the 1830 Book of Mormon (see Appendix B) to a new chapter-by-chapter classification analysis using the open-set NSC procedure with adjustments for sizes of test texts (Equations 10 and 12). We dealt with the issue of variable training text sizes by compositing training texts to obtain new training texts of at least 2,000 words. We dealt with the multiplicity issue by smoothing the posterior authorship probabilities across chapters. Using the data of Jockers *et al.* (2008), we tested our computer code for closed-set size-invariant NSC classification. Jockers *et al.* (2008) did not present their shrinkage parameter, but using a shrinkage parameter of $\Delta = 0.44$ we obtained, with the exception of one chapter, identical classification results.

In most respects, our analysis was similar to the seven-author study of Jockers *et al.* (2008). As authorship features, we used relative frequencies of the 110 words employed by Jockers *et al.* (2008) in their seven-author study. As training texts, we used texts by early Rigdon, late Rigdon, Spalding, Cowdery, Pratt, and Isaiah (see Appendix B). We also added Joseph Smith as a candidate author, represented solely by texts in his own hand, largely the same texts used by Jockers (2011) as the standard of Smith's style. Admittedly most of these writings were personal letters and so may exhibit a genre-related difference from religious writings. All training texts contained 2,000–6,000 words. This sometimes necessitated the compositing of smaller texts, but efforts were made to ensure that logical divisions such as chapters were not split. We did not include Barlow or Longfellow texts, but we did the analysis with and without the Hamilton texts as a negative control.

To choose tuning constants, we applied cross-validation to the training author texts. The correct classification percentage reached 88.1% with the tuning constants Δ (Equation 4) and λ (Equation 10) equal to 2 and 4, respectively. We extracted authorship features for each text using a PERL script (available from the authors), and used the SAS system (SAS Institute Inc., 2004) to carry out all NSC analyses. Graphs were generated using both SAS and R.

We first applied principal components analysis to the training and test texts (Fig. 10). The Book of Mormon texts clustered separately from the training texts, which were all in the lower left quadrant. The high variability among Book of Mormon texts is partially due to the fact that several of the chapters were small and partially due to differences among texts associated with different purported authors of the books of the Book of Mormon (Larsen *et al.*, 1980; Hilton, 1990). As expected from the classification results for the Rigdon texts (Fig. 7), the late Rigdon texts clustered distinctly from the early Rigdon texts. As a result, we did not combine them into a single Rigdon corpus.

For comparison to the results of Jockers *et al.* (2008), we initially carried out naive classification with the closed-set NSC procedure. The results were comparable to those of Jockers *et al.* (2008). Of the 239 chapters, 68 (28.5%) were classified to Isaiah, 96 (40.2%) were classified to early or late Rigdon, 37 (15.5%) were classified to Spalding, 29 (12.1%) were classified to Smith, 4 (1.7%) were classified to Cowdery, and 5 (2.1%) were classified to Pratt. However, the goodness-of-fit check (see Equation 9 and Fig. 4) showed that the posterior predictions did not match the observed test data; hence the naïve closed-set NSC classification results are not valid.

We then carried out the extended open-set NSC procedure. Of the 239 Book of Mormon chapters, 175 (73.2%) were classified to one or more unobserved authors, 35 (14.6%) were classified to Isaiah, 17 (7.1%) were classified to early or late Rigdon, 8 (3.4%) were classified to Smith, 2 (0.8%) were classified to Spalding, and 2 (0.8%) were classified to Cowdery (Fig. 11). Seventeen of the 20 Book of Mormon chapters that were essentially verbatim quotations of Isaiah or Malachi were correctly classified. Chapters classified to Rigdon, Smith, or Cowdery appeared to occur essentially at random in the sequence of chapters. The texts classified to Rigdon, Smith, or Cowdery were on the fringe of the Book of Mormon cluster, and the classifications were likely due to multiplicity. Hence, we conclude that based on relative frequencies of the 110

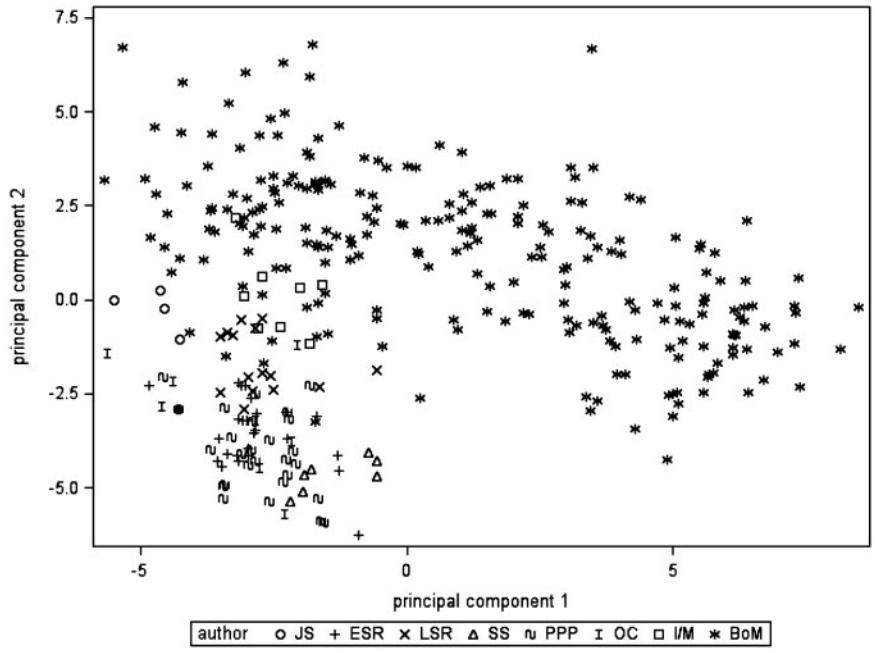

Fig. 10 First two principal component scores of 110 authorship features for texts written by Smith, early Rigdon, late Rigdon, Spalding, Pratt, Cowdery, Isaiah, and Book of Mormon texts

common (mostly non-contextual) words of Jockers *et al.* (2008), 80% of the non-Isaiah chapters of the Book of Mormon are dissimilar in style from the authors in the candidate set.

We carried out several variations of this analysis. We treated the entire Rigdon corpus as of single authorship, treated 'and it came to pass that' as a single word, classified sequential 5,000-word blocks from the Book of Mormon rather than chapters, excluded 15 nouns from the list of 110 words (Jockers *et al.*, 2008), included the Hamilton texts as a negative control, and included vocabulary richness measures in addition to the word frequencies. In all cases similar results were obtained. The similarity of the results confirms that the precise choice of function words and other authorship features is not crucial (Koppel *et al.*, 2009) to the conclusion that the non-Isaiah chapters of the Book of Mormon are dissimilar in style from the authors in the candidate set.

Consistent with previous analyses of the Book of Mormon, this analysis shows that based on several sets of stylometric measures, there is little stylometric support for the Spalding–Rigdon theory of Book of Mormon authorship. Less than 9% of the non-Isaiah chapters were attributed Rigdon or Spalding, and those were randomly distributed throughout the text consistent with multiplicity. The writing styles throughout the book do not credibly match Rigdon, Spalding, or any of the other candidates, as claimed by Jockers *et al.* (2008).

In future studies with the Book of Mormon, we intend to adjust for the deliberate archaic language used throughout. We also intend to supply measures of uncertainty for the estimated posterior authorship probabilities.

7 Conclusions

While the NSC procedure is a useful tool for classification analysis of high dimensional data, it must be used carefully in stylometric problems. Application of NSC to most stylometric problems will benefit from relatively simple extensions to allow for an unobserved author and adjustments

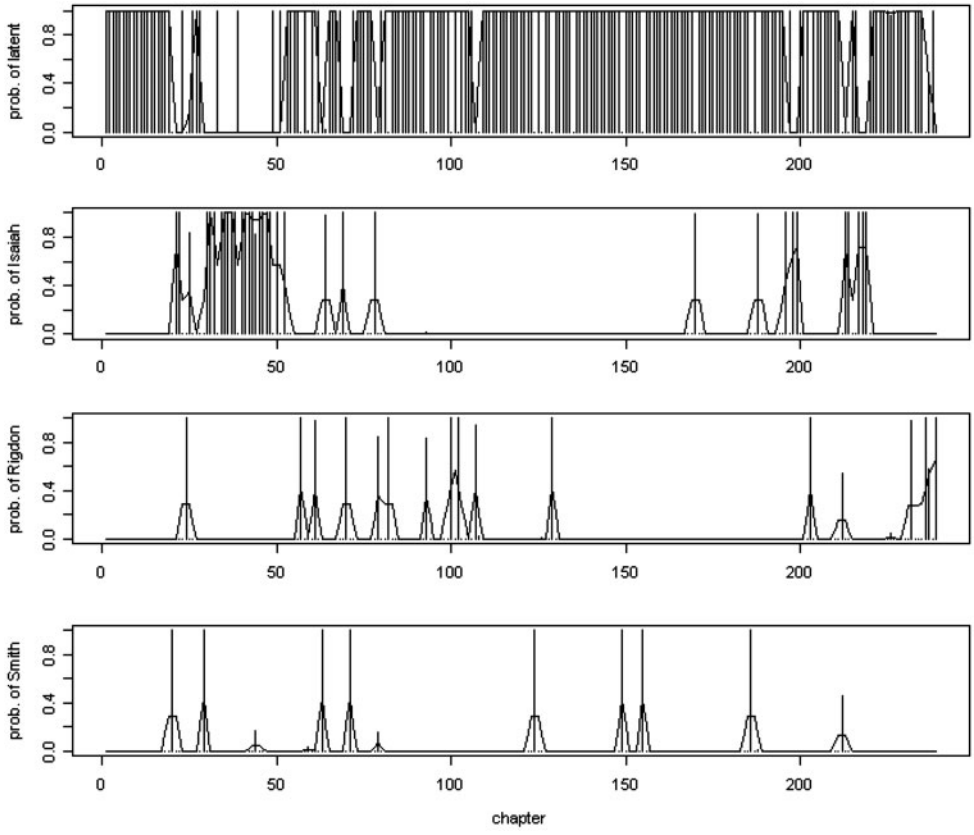

Fig. 11 Posterior probabilities of authorship for 239 Book of Mormon chapters using the open-set NSC model. Lowess smoothers (with smoothing parameter 1/10) are superimposed to reduce the effects of multiplicity. Calculations were based on relative frequencies of 110 words, most of which were non-contextual

to deal properly with test texts of varying size. NSC classifications should be augmented with goodness-of-fit tests to validate the classifications and smoothers to alleviate multiplicity problems. This is especially true in order that the posterior probabilities will be sensible quantities, but also in order that the classifications based on the posterior probabilities will be credible.

There is remarkably good agreement for attribution of Book of Mormon chapters between open-set NSC and Burrows' delta method with its threshold for eliminating false positives. However, open-set NSC has advantages in that it produces posterior authorship probabilities rather than probability rankings, it can be modified to deal with heterogeneous sizes of test texts, and its open-set capabilities are based on theory rather than simply empirical evidence. When using open-set NSC, or any authorship attribution procedure, results need to be interpreted in light of the stylistic features used, the genres of the training texts, and other subtle properties of the problem. Even so, the method has great potential for future careful use in authorship attribution.

Acknowledgements

We thank Mark Jones and Patrick Staples for technical assistance and Brittany Morgan for excellent transcription of texts. We thank several colleagues for comments, advice and reviews of drafts of this

article. The Neal A. Maxwell Institute for Religious Scholarship and the Department of Statistics, Brigham Young University provided funding for this project.

Appendix A: Authorship of the *Book of Mormon*

The Book of Mormon has been controversial since its publication in 1830. The book, held to be scripture by members of the Church of Jesus Christ of Latter-day Saints, tells of the migrations of three groups of people to an American land of promise and describes the dealings of God with them and their descendants. Joseph Smith claimed that a divine messenger led him to the location of the record, which had been written on metal plates by ancient prophets. Although primarily a religious record, Smith claimed it related genuine historical events and that he translated it into English by the 'gift and power of God'.

Given the nature of these claims, it is unsurprising that the origins of the Book of Mormon have been controversial. Unwilling to believe Smith's explanation, critics have proposed various alternative theories (Kirkham, 1959; Midgley, 1997; Givens, 2002). Early critics argued that Smith alone was responsible for the work (Campbell, 1831). Subsequent writers were less sure, arguing that the book must have been the product of someone with more ability such as one of Smith's early associates Oliver Cowdery or Sidney Rigdon. These theories had little basis other than speculation. In 1834, using testimony gathered by a disgruntled former member, Howe (1834) speculated that the Book of Mormon was based on an unpublished novel by former clergyman Solomon Spalding. He conjectured that Rigdon had somehow acquired the unpublished manuscript, added religious material, and concocted the Book of Mormon.

Parley P. Pratt, an early convert to Mormonism, encountered the Book of Mormon 5 months after its publication (Pratt, 1985). Impressed by the book, he was baptized shortly thereafter into the church established by Smith. Later that year, Pratt, Cowdery, and other missionaries traveled several hundred miles from New York to Mentor, Ohio, where they introduced the Book of Mormon to Pratt's friend Rigdon, a talented Campbellite preacher. After initial resistance, Rigdon was baptized. Pratt, Rigdon, Rigdon's wife Phebe, and their children all affirmed that this was Rigdon's first encounter with the Book of Mormon. In later years, after his estrangement from Mormon society, Rigdon likewise maintained that he had nothing to do with the origins of the book (Roper, 2005).

The recovery of an original Spalding manuscript in 1884, the only one of any significance proven to exist, led many critics to abandon the Spalding–Rigdon theory (Roper 2005, 2009). A minority continue to argue that the Book of Mormon was based upon a hypothesized second now-lost Spalding manuscript. The study by Jockers *et al.* (2008) is an effort to provide stylometric evidence for the Spalding–Rigdon theory.

The Jockers *et al.* (2008) study is the latest in a series of stylometric investigations into Book of Mormon authorship. Previous studies used multivariate statistical tools together with relative frequencies of non-contextual words (Larsen *et al.*, 1980), word-pattern ratios (Morton, 1978; Hilton, 1990), and vocabulary richness measures (Holmes, 1992). The studies used different text sizes and different methodologies, but together they suggest that the Book of Mormon is the product of multiple authorship in which the styles of the various authors are characterized by non-contextual word usage and word-pattern ratios, but not by vocabulary richness. The styles do not match those of any modern candidates that have been suggested.

The Jockers *et al.* (2008) study applied closed-set NSC classification (Tibshirani *et al.*, 2002, 2003) to texts characterized by relative frequencies of 110 mostly non-contextual words. Disagreeing with previous studies, they concluded that the majority of the chapters of the Book of Mormon were written either by Rigdon or Spalding.

Appendix B: Texts and Source Materials

Book of Mormon: we copied the text of the 1830 edition from http://purl.stanford.edu/ir:rs276tc2764, and used chapters defined by the

1981 edition of the *Book of Mormon*. This chapter structure was invented in 1879 and was not part of the original Book of Mormon.

Sidney Rigdon (1831–46): we created a corpus by transcribing 54 newspaper articles and 1 oration signed by Rigdon between 1831 and 1846. The articles appeared in the *Ohio Star, Evening and Morning Star, Latter-day Saints Messenger and Advocate, Elders' Journal, Quincy Whig, Times and Seasons,* and *The Peoples Organ*. These were composited into 25 texts ranging in size from 2,214 to 8,747 words. The texts were composited in chronological order, and no article was split between two texts.

Sidney Rigdon (1863–73): we created a corpus by transcribing all the sections from the *Book of the Revelations of Jesus Christ to the Children of Zion through Sidney Rigdon, Prophet, Seer and Revelator*, located in Folders 11 and 12, Stephen Post Papers, Special Collections, Harold B. Lee Library, Brigham Young University, Provo, UT. For use in extending the NSC method to account for different sample sizes, these sections were used without compositing, including the sections jointly or singly authored by Phebe Rigdon (as an unobserved author). For use as training texts, the sections authored by Sidney Rigdon or jointly by Sidney Rigdon and Phebe Rigdon were composited into 15 texts ranging in size from 3,678 to 6,784 words. The texts were composited in chronological order, and no section was split between two texts.

Solomon Spaulding: we copied the Spalding manuscript from http://www.mormonstudies.com/spaldg.htm as had Jockers *et al.* (2008). We split the manuscript into seven texts ranging is size from 4,030 to 6,767 words, sequenced as in the Spalding manuscript. No chapter was split between two texts.

Joseph Smith: we created a corpus by compositing holograhic portions of Smith's personal letters and history. These were obtained from Jessee (1989). There were five texts ranging in size 2,081 to 3,362 words, and they were sequenced in chronological order. No letter or article was split between two texts.

Parley P. Pratt: texts for Pratt were taken from Crawley (1990) and Pratt (1985). Twenty-six texts were composited from Crawley (1990) and 2 texts were composited from Pratt (1985). The texts ranged in size from 3,443 to 5,994 words.

Oliver Cowdery: we created a corpus for Cowdery by compositing lengthy personal letters that Cowdery wrote to W. W. Phelps between 1834 and 1835. We composited these letters into eight texts ranging in size from 3,342 to 6,167 words. The texts were composited in chronological order, and no letter was split between two texts.

Isaiah: we used the King James version of the Book of Isaiah from http://etext.virginia.edu/toc/modeng/public/KjvIsai.html. The chapters were composited chronologically into eight texts ranging in size from 3,903 to 5,429 words. No chapter was split between two texts.

Alexander Hamilton: the texts for Hamilton came from http://etext.virginia.edu/hamilton/. The papers were not composited. They ranged in size from 960 to 5,716 words.

References

Argamon, S. (2008). Interpreting Burrows's delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, **57**(1): 289–300.

Buja, A., Lang, D. T., and Swayne, D. F. (2003). GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Journal of Computational and Graphical Statistics*, **43**(4): 423–44.

Burrows, J. F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Burrows, J. F. (2003). Questions of authorship and beyond. *Computers and the Humanities*, **37**: 5–32.

Campbell, A. (1831). Delusions. *Millennial Harbinger*, **2**(2): 85–96.

Cleveland, W. S. (1981). LOWESS: a program for smoothing scatterplots by Robust locally weighted regression. *The American Statistician*, **35**: 54.

Crawley, P. L. (1990). *The Essential Parley P. Pratt*. Salt Lake City, UT: Signature Books.

Dabney, A. R. (2005). Classification of microarrays to nearest centroids. *Bioinformatics*, **21**(22): 4148–54.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edn. New York: Chapman & Hall/CRC.

Giron, J., Ginebra, J., and Riba, A. (2005). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *American Statistician*, **59**(1): 19–30.

Givens, T. L. (2002). *By the Hand of Mormon: The American Scripture that Launched a New World Religion*. Oxford: Oxford University Press, pp. 155–84.

Hilton, J. L. (1990). On verifying worprint studies: Book of Mormon authorship. *BYU Studies*, **30**: 89–108.

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

Holmes, D. I. and Kardos, J. (2003). Who was the author? An introduction to stylometry. *Chance*, **16**(2): 5–8.

Hoover, D. L. (2004). Testing Burrows's delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Howe, E. D. (1834). *Mormonism Unvailed (sic)*. Painesville, OH: Telegraph Press.

Jessee, D. C. (1989). *The Papers of Joseph Smith, Volume 1 Autobiographical and Historical Writings, ''History (1832)''*. Salt Lake City, UT: Deseret Book Company.

Jockers, M. L. (2011). Testing authorship in the personal writings of Joseph Smith using NSC classification. *Literary and Linguistic Computing*, **26**(in press).

Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**(2): 215–23.

Jockers, M. L., Witten, D. M., and Criddle, C. S. (2008). Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, **23**(4): 465–91.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.

Kirkham, F. W. (1959). *A New Witness for Christ in America*, 2 vols. Salt Lake City, UT: Utah Printing.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1): 9–26.

Larsen, W. A., Rencher, A. C., and Layton, T. (1980). Who wrote the Book of Mormon? An analysis of wordprints. *BYU Studies*, **20**: 225–51.

Midgley, L. C. (1997). Who really wrote the Book of Mormon? The critics and their theories. In Reynolds, Noel B. (ed.), *Book of Mormon Authorship Revisited*. Provo, UT: Foundation for Ancient Research and Mormon Studies, pp. 101–39.

Morton, A. Q. (1978). *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. New York: Charles Scribner's Sons.

Mosteller, F. and Wallace, D. L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer.

Peng, R. D. and Hengartner, N. W. (2002). Quantitative analysis of literary styles. *American Statistician*, **56**(3): 175–85.

Pratt, P. P. (1985). *Autobiography of Parley P. Pratt*. Salt Lake City, UT: Deseret Book, pp. 18–27.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear Models in Statistics*, 2nd edn. New York: Wiley Interscience, pp. 499–500.

Roper, M. (2005). The mythical 'Manuscript Found'. *FARMS Review*, **17**(2): 7–140.

Roper, M. (2009). Myth, memory, and 'Manuscript Found'. *FARMS Review*, **21**(2): 179–223.

SAS Institute Inc. (2004). *SAS/IML 9.1 User's Guide*. Cary, NC: SAS Institute Inc.

Schaalje, G. B. and Fields, P. J. (2011). Open-set nearest shrunken centroid classification. *Communications in Statistics – Theory and Methods*, **40**(in press).

Sims, G. E., Jun, S., Wu, G. A., and Kim, S. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Science*, **106**(8): 2677–82.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields in agricultural crops. *Journal of Agricultural Science*, **28**: 1–23.

Tibshirani, R., Hastie, T., Narasimham, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Science*, **99**(10): 6567–72.

Tibshirani, R., Hastie, T., Narasimham, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with application to DNA microarrays. *Statistical Science*, **18**(1): 104–17.