Measuring style with the authorship ratio
An invariant metric of lexical similarity

Edward J. L. Bell, Damon Berridge and Paul Rayson
Department of Maths & Stats / Computing Department
Lancaster University, UK
e.bell@comp.lancs.ac.uk, {d.berridge, p.rayson}@lancaster.ac.uk

Abstract

Stylometry is the study of the computational and mathematical properties of style. The aim of a stylometrist is to derive stylometrics and models based upon those metrics to quantitatively gauge stylistic propensities. This paper presents a method of formulating a stylistic distance function via a weighted ratio of lexical stylometrics, the higher the ratio the more the styles diverge. The coefficients of the distance function are estimated using Powell's conjugate gradient method (Powell, 1964) on a 4 million word corpus of 19th Century literature. The distance metric proves accurate over 30,000 binary comparisons and rivals the discernment aptitude of established techniques (Labbé and Labbé, 2001). Previous metrics have suffered from sample-size dependencies, the metric proposed here is resilient to such bias.

1 Introduction

A distance metric δ is a function of 2 vector parameters with the following properties, all of which are intuitive when thought of in terms of Euclidean geometry.

$$\begin{aligned} \textit{non-negativity} \quad & \delta : x\ y \rightarrow \{ z \mid z \in \mathbb{R}, z \geq 0 \} \\ \textit{identity} \quad & \delta(x,y) = 0 \text{ iff} x = y \\ \textit{symmetry} \quad & \delta(x,y) = \delta(y,x) \\ \textit{subadditivity} \quad & \delta(x,y) \leq \delta(x,z) + \delta(z,y) \end{aligned}$$

A textual distance function operates on linguistic data and although simple by definition, dealing with language in a statistically valid manner is hard. Language is generally high dimensional, non-independent, non-identically distributed and dependent on the sample size. Here we deal with high dimensionality by using frequency spectra, rely on existing statistics to deal with (or ignore) the non-independent and non-identical distribution problems, and explicitly control for residual sample-size variation. The extra effort required in forming a distance metric is worthwhile because they not only classify but provide the degree of similarity between two linguistic samples. This interpretability leads to improved comprehension especially when combined with other methodologies such as clustering. Typical black-box machine learning techniques do not provide the transparency needed in a broad range of linguistic applications. Textual distance functions can be used in many areas including; stylometry, stylistics, stylochronometry, language acquisition and corpus linguistics; essentially any study where the sampling unit of interest is a block of text rather than individual words.

1.1 Data

The corpus under study comprises 248 fictional works by 103 authors and totals around 40 million words. The works range from the late 18th Century to the early 20th Century with most texts

centred in the 19th Century. The mean number of words per book is approximately 20,000 with the largest being Frances Burney's 5 volume 1796 work *'Camilla: Or, A Picture of Youth'*. All books were taken from the Chadwyck-Healey Literature Collection[1].

Distance metrics are calculated using two samples of text, therefore the total number of possible comparisons can be expressed as a binomial coefficient where n is the number of books in the corpus and k is 2 due to the comparisons being binary. The ratio of comparisons between same and different authors depends upon the distribution of authors in the dataset; in this case the number of different author comparisons will always exceed the number of same author comparisons. The data are split into three cumulative sets to assess how well the proposed method performs on smaller training samples. The criteria for observation inclusion are governed by the number of books that observation's author has in the collection.

ID	*min*	N_b	N_a	N_c
A	1	248	103	30628
B	2	184	39	16836
C	4	140	19	9730

Table 1: *dataset and subset characteristics*

Table 1 shows attributes for each of the datasets; *min* is the minimum number of books an author requires to be included in the sample; N_b is the total number of books; N_a is the total number of unique authors and N_c is the number of binary comparisons where

$$N_c = \frac{N_b!}{2!(N_b - 2)!}$$

In the remainder of this paper, section 2 reviews related work; section 3 deals our methodology; section 4 describes our initial exploratory analysis on a selection of the most prolific authors in the corpus (Austen, Dickens, Hardy and Trollope) which leads to the development and evaluation of the authorship ratio in section 5. Finally we conclude with section 6 and describe possible areas of future work.

2 Background

The origins of quantitative stylistics are commonly thought to reside with Mendenhall (1887) when he attempted to find differences in sentence length distributions for works by Bacon, Marlowe and Shakespeare. Although Mendenhall did not work alone, his peers include Mascol (1888) and Sherman (1888). Yule (1939) was also an early pioneer and used sentence lengths during his investigation of *De Imutatione Christi*.

No discussion of stylometry can proceed without mentioning Zipf's empirical law (Zipf, 1932). Zipf's law is the most cited example of a quantitative lexical relation and the structure most parametric stylistic models are based upon. In standard word frequency distribution notation; N is equal to the number of vocabulary tokens; V is equal to the number of vocabulary types; V_m is the number of vocabulary types with frequency m; C is a normalising constant and a is a free parameter which can be estimated but is usually fixed. Zipf's Law in standard notation can be expressed as,

$$V_m = \frac{C}{m^a}$$

In English the most frequent word is 'the', therefore it commonly has a V_m equal to 1 and a very high m (m $\approx .07 \cdot V$). Conversely, the words which only occur once, known as 'hapax legomena', have an m equal to 1 and a very high V_m ($V_1 \approx V/2$) hence the distribution is characterised by a 'large number of rare events'. The most frequent vocabulary tokens are usually function words acting as syntactic glue (articles, pronouns, conjunctions etc.). Zipf's law implies, on untransformed axes, distributions of lexical occurrences are characterised by an inverse *'J'* shape, this structure is known as a LNRE distribution (Baayen, 2001).

The classic authorship attribution studies are in historically disputed cases such as the origins of Shakespearean work or determining the most probable author of the unknown Federalist papers (Mosteller and Wallace, 1964; Fung et al., 2003). Non-traditional applications include the detection of plagiarism in educational environments (Clough, 2000) and the identification of anonymous authors; either on the Internet (de Vel et al., 2001); authors who write under a pseudonym or even in forensic stylistic cases such as the Derek Bentley trial (McMenamin, 2002). For a historical perspective on the development of stylometry see Holmes (1998).

2.1 Stylometrics

Style is a latent property of language meaning it cannot be measured directly. The rarity and regularity of linguistic constructs can be quantified, and they provide an indirect indication of the underlying stylistic profile. Measures of style are commonly referred to as stylometrics. An ideal stylometric would capture innate subconscious stylistic attributes unique to a given author and would transcend all forms of textual variation. In practice no single indicator can measure style in its entirety. Stylometrics must be compounded to capture the diversity of stylistic expression and variation is still a pressing problem (Rudman, 1998).

Using the frequency of linguistic constructs is an often applied stylometric method but frequency, by definition, is a function of the sample size. The same is true of the sample probabilities, sometimes referred to as relative frequency, for any feasible text size. Stylometrists have so much uncontrollable variation to contend with, eliminating as many forms as possible is a prudent course. Furthermore raw frequency-based metrics lack transparency, interpreting pure frequency is unwieldy at best, it is wiser to make use of specialised statistics and models.

We selected the stylometrics based on a number of criteria, the most important of which is sample size independence or having the property of 'invariance'. Samples of text are naturally heterogeneously sized so any stylometric which incorporates sample size variation is going to be a poor choice (Tweedie and Baayen, 1998). Sample sizes can be homogenised if a loss of information is acceptable but there is no guarantee the bias will be equal just because the sample size is equal due to the non-random nature of language.

The second criterion, transparency or 'ease of interpretation', is important because the comprehension of style is surely a stylometrist's most laudable goal. If stylometrics do not have strong subject-matter relevance practitioners are blind. The value of a stylometric must correspondence to some interpretable aspect of an author's stylistic profile. Understanding what each stylometric value implies is key to interpreting stylistic idiosyncrasies. Burrows (2003) expressed a similar sentiment towards opaque classification methods such as neural networks and discriminatory analysis; "...they lack the transparency so useful for exploring the evidence".

Finally, stylometric models should seek parsimony. All redundant degrees of freedom need to be eliminated from consideration otherwise the true source of interest is obscured. An attribution system ought to strive for absolute lucidity whilst maintaining ontological minimality to reduce the number of assumptions it makes. So called 'feature proliferation' has seen a resurgence of late, a trend noticed by other researchers (Argamon et al., 2007). Classifying style by considering

every possible permutation of grammatical minutiae is undoubtedly going to produce good results but little is learnt in the process. In the words of nominalist logician William of Ockham in accordance with his eponymous law; "plurality should not be posited without necessity".

3 Methodology

Two techniques will be used to assess the suitability of our proposed algorithm, intertextual distance (Labbé and Labbé, 2001; Labbé, 2007) and support vector machines. The intertextual distance uses the intersection of lexical frequencies as a similarity measure and has proved useful for exploratory studies but it's classification ability has not been quantified. Support vector machines (SVMs) are excellent classifiers but are not interpretable in the statistical sense. For successful applications of SVMs to authorship attribution see Hirst and Feiguina (2007) or Diederich et al. (2003). It is hoped the composite distance measure detailed here will hold the middle ground and attain high classification accuracy whilst providing exploratory capabilities.

All documents were split into words using standard word boundary regular expressions, sentences were parsed using the *NLTK* (Loper and Bird, 2002) *Punkt tokenizer* (Kiss and Strunk, 2006). The parametric LNRE models were fitted using Baayen's *lexstat toolkit* [2]. SVM classification was performed using *libSVM* (Chang and Lin, 2001). The intertextual distance and the non-parametric stylometrics were implemented based upon canonical definitions in the literature.

3.1 Non-parametric statistics

Non-parametric statistics do not make assumptions about the source of data. They are distribution-free and as a result no unknown values have to be estimated in order to proceed with their calculation. The first non-parametric stylometric, mean sentence length, has a rich history of use (Mendenhall, 1887; Yule, 1939; Sichel, 1974) and is known to be invariant. Average sentence length (denoted by M) is also easy to interpret. How well it segregates authors is a contentious issue but discrimination aptitude is not of prime importance as classification proficiency can be synthesised by compounding weak stylometrics into a strong model.

K (Yule, 1944) measures word repeat rates in a sample size independent manner, the higher the value the more the author repeats words (Equation 1). The i/n term is the sample probability estimate for token i hence the square of i/n is the probability of sampling the same token twice (assuming independence). Multiplying the sample probability by V_i, the number of types with token frequency i, results in the expected value of V_i's repeat rate. All expected values are summed and standardised with the sample size.

$$K = -\frac{1}{N} + \sum_{i=0}^{N} V_i \left(\frac{i}{n}\right)^2 \quad (1)$$

Other constants are very similar (Simpson, 1949; Herdan, 1955) but only K will be considered here because it was the first to be published and the measure which has gained most traction in the literature. Besides, including multiple word repeat rate measures would invalidate the notion of parsimony. Some research has shown K to be an inadequate discriminator on its own (Hoover, 2003) but as part of a system it excels (Smith and Kelly, 2002; Somers and Tweedie, 2003).

3.2 Parametric models

Parametric models make assumptions about the distribution the data are drawn from. As a result the distribution's parameters have to be estimated using methods such as maximum likelihood or

optimisation of a fitness statistic. Usually only a single parameter is of interest; the remaining are known as nuisance parameters. In this study all parameters are estimated using downhill simplex minimisation (Nelder and Mead, 1965) with a C_1 cost function (Sichel, 1986; Baayen, 2001). The cost function attempts to match the expected value of selected vocabulary tokens with the model's estimates of those tokens. The generalised Zipf model (Orlov, 1983) is a parametric manifestation of Zip's law with two free parameters

$$E[V] \;=\; \frac{Z}{log(pZ)}\frac{N}{N-Z}\,log(N/Z)$$

The maximum sample probability p is equal to the frequency of the most common word divided by the text length. The parameter of interest Z can be interpreted as the minimal text size where Zipf's law holds. Z is calculated by extrapolating the sample frequency distribution in accordance with Zipf's law, therefore Z remains invariant across varying text lengths.

The final and most complex stylometric is a parameter from the generalised Gauss-Poisson model (Sichel, 1975). Sichel's model is a three parameter continuous distribution, where $g(\pi)$ is equal to the proportion of words in the population with probability $\pi \pm \epsilon$ for small ϵ.

$$g(\pi) \;=\; C\pi^{\gamma-1}exp\left\{-\frac{\pi}{c}-\frac{b^2c}{4\pi}\right\}$$

The normalising constant, C, is a function of the parameters and the modified Bessel function of the second kind order γ; for more details see Sichel (1975) or Baayen (2001). The γ term is a shape parameter and traditionally fixed at -0.5 to aid computational tractability although there are methods to estimate it. The b and c terms are measures of lexical richness linked to the density decay rates. Smaller values of b and c equate to a larger number of types. Neither b or c is perfectly invariant but c is preferable as it shows less of a sample size dependency because it is responsible for the decay rate of the dominant tokens.

4 Exploratory analysis

We began our investigation with a exploratory analysis in order to better understand the measures described in section 3. Fitting a regression model with the stylometrics as explanatory variables and a dummy binary variable as the response will help determine the utility of the selected measures. The response is binomially distributed and equal to 1 for the author of interest and 0 for all other authors. The model coefficients can be interpreted as the log ratio of each criterion's 'affinity' with a given author. A positive coefficient reflects an increased probability of that variable being pronounced in the author's work; a negative coefficient implies the opposite.

Whilst fitting binomial models it became apparent the data were problematic. Some of the models did not converge due to 'perfect separation' in the response. Perfect separation implies a non-concave likelihood function tending to infinity; therefore the estimation procedure only finds a maximum when the curve reaches a sufficiently flat area, resulting in extremely inflated coefficients and standard errors. To remedy this a method of bias-reduced binomial regression was employed (Firth, 1993), for more details see the *brglm* R package (Kosmidis, 2008).

Table 2 shows the significant coefficients from the stylometric binomial models (β) and their corresponding p-values (p). All explanatory variable were transformed via natural logarithms to ensure effect linearity and homoscedasticity (variance homogeneity). Cells are empty where a stylometric was far from significant at the 5% level and therefore not critical to the identification of a given author. The grouping of significant stylometrics is interesting but coincidental; each author is identified by two stylometrics and each stylometric identifies two authors.

		Austen	Dickens	Hardy	Trollope
M	β	.	7.583	-21.369	.
	p	.	0.02213	0.0088	.
K	β	-32.703	.	.	23.231
	p	0.0221	.	.	0.00402
C	β	.	-7.703	.	7.782
	p	.	0.00243	.	0.00198
Z	β	-5.857	.	7.219	.
	p	0.0228	.	0.0272	.

Table 2: *Significant stylometrics and their corresponding log-odds*

For Dickens the M coefficient is positive which implies his work is characterised by high sentence densities. The log of M resides in the range $[2.7, 3.6]$ and for each point increase the probability a work was penned by Dickens increases by a factor of $e^{7.583}$. C is negative which implies a decrease in probability by a factor of $e^{-7.703}$, given that lower values of C are associated with high vocabulary densities, his work appears lexically rich. Trollope has a positive coefficient for C and therefore a smaller vocabulary ($e^{7.782}$). Trollope also has a large K coefficient so his work typically has a low type-token ratio. Austen's extremely large negative K coefficient shows her work is characterised by very low word repeat rates ($e^{-32.703}$). The magnitude of the value is due to Austen dominating the lower ranges of K for the authors included in the model; with more authors the coefficients should stabilise. Austen's Z coefficient is also negative therefore she has small expected texts sizes ($e^{-5.857}$). Hardy's work is associated with large values of Z ($e^{7.219}$) but very small average sentence lengths, for each point increase in M the probability a work was penned by Hardy decreases by a factor of $e^{-21.369}$.

Overall, Dickens exhibits the most complex style, Austen and Trollope the simplest and Hardy resides somewhere in-between. The most interesting finding is the characterisation of each author by a unique set of stylometric values, with more authors this may not continue but the segregation in Table 2 bodes well for the formation of a classification system.

5 Classification

In this section we develop the authorship ratio (AR), a method of judging stylistic distance. The purpose of AR is to provide an interpretable stylistic metric rather than operate as a pure classification algorithm; but discernment aptitude is key to AR's success. If AR possesses a high level of classification accuracy then practitioners can have confidence in the estimated distances.

All stylometrics are standardised to have mean 0 and unit variance to mitigate scaling bias and normalised using log transformations. Three statistics are used to judge the effectiveness of each model: accuracy, sensitivity and specificity. Accuracy is the overall success rate, sensitivity is the accuracy of comparisons between the same author and specificity is the accuracy of different author comparisons. All classification results presented in this section are the product of stratified K-fold cross-validation where $K = 10$. The observations are first split into negative (n) and positive (p) classes, each class is then partitioned into K subsets. Same author comparisons are treated as positive cases and different author comparisons as negative cases. The classification algorithm is validated on a single partition from each of the classes ($V_i = p_i \cup n_i$) and trained on

the remaining data ($T_i = V_i^c$). This process repeats for all K validation sets and therefore makes use of all the data to provide a better estimate of the generalisation error.

5.1 Authorship ratio

For each author a weighted sum of stylometrics is estimated (a linear predictor) where λ represents the weights or coefficients and X the stylometrics (Equation 2). The λ_0 constant acts as the intercept or scaling term. The ratio of two linear predictors will be equal to 1 for identical works and diverge from 1 for non-identical works. To ensure the metric adheres to the definition of a distance measure, 1 is taken away from the absolute value of the ratio; therefore identical texts will have an authorship ratio (AR) of 0 and non-identical texts will have a larger positive AR proportional to the difference in the stylometric values. In addition to the stylometrics, a sample size covariate is included (the λN term) in an effort to control any residual sample size variation. To ensure symmetry in the metric space it must hold that either $N_i \geq N_j$ or $N_j \geq N_i$ for all comparisons.

$$AR_{ij} = \delta_{ij} = \left| 1 - \frac{\lambda_0 + \lambda_1 N_i + \sum_{z=0}^{n} \lambda_{z+2} X_{iz}}{\lambda_0 + \lambda_1 N_j + \sum_{z=0}^{n} \lambda_{z+2} X_{jz}} \right| \tag{2}$$

The AR estimation algorithm optimises the coefficients so that the authorship ratio is maximised for comparisons between texts by different authors and minimised for comparisons between books by the same author. To form a cost function and measure the classification success rate, a discrimination boundary must be defined (denoted by α). If the authorship ratio is less than α the algorithm considers the texts to be penned by the same author, if AR is above or equal to α the books are considered to be penned by different authors.

$$\begin{aligned} max(\delta_{ij}) &\geq \alpha \quad i \neq j \\ min(\delta_{ij}) &< \alpha \quad i = j \end{aligned}$$

The coefficients of the predictors are constant across all works. They are estimated using Powell's conjugate gradient method (Powell, 1964). The Nelder-Mead method (Nelder and Mead, 1965) was considered but Powell's method performed better, perhaps due to the non-linearity of the search space or the large number of parameters. Faster algorithms are available but they require derivatives of the function being optimised. In our case the classification function is rather computationally complex so calculating derivatives is non-trivial.

Powell's method attempts to minimise the classification error via measured modifications of the predictor coefficients. The error is normally $1 - accuracy$ but AR will use Matthew's Correlation Coefficient (Matthews, 1975). Matthew's Correlation Coefficient (MCC) should perform better than accuracy alone because it accounts for sensitivity (Equation 3). The class distribution in the training data is unbalanced so optimising the coefficients based on accuracy alone would result in an unbalanced classifier.

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{3}$$

In Equation 3, tp represents the number of true positives, fp the number of false positives, tn the number of true negatives and fn the number of false negatives. An MCC score of 1 demonstrates perfect classification, 0 is average classification and -1 is perfect misclassification. Powell's method minimises a function, to maximise the MCC score we simply minimise -MCC so that optimal classification occurs when -MCC = -1.

Table 3 shows the results of AR classification across the three nested datasets defined in Table 1. The best results were achieved by setting the discrimination boundary α to 0.5. The complete set of all authors (A) was predicted with 84% accuracy and with 50% sensitivity, given that over 97% of the cases are negative comparisons the high sensitivity demonstrates the applicability of MCC to unbalanced classification. The full dataset consists of 30628 comparisons therefore around 25862 observations were correctly predicted by the algorithm. Most misclassifications were caused by three factors; short works whereby the author's stylistic profile did not have time to converge; temporal variation whereby works penned by the same author spanning large time periods exhibited dramatically different styles and similarity of stylometric values between distinct authors. All of these errors could potentially be fixed by introducing addition stylometrics or specialised models.

	A	B	C
accuracy	84.44	83.48	81.52
sensitivity	50.05	48.33	55.83
specificity	85.16	84.85	83.23

Table 3: *Authorship-ratio results*

As the number of observations increases the accuracy and specificity improve. The sensitivity is highest for the smallest sample because the proportion of negative to positive comparisons is smaller. The sensitivity seems to drop uncharacteristically for dataset B but this may just be due to random variation. For the complete dataset the 50% sensitivity is acceptable considering only 3% of the sample is composed of positive cases.

5.2 Intertextual distance

The intertextual metric (Labbé and Labbé, 2001) is an established method of judging stylistic distance, it is essentially a Bray-Curtis distance with the addition of a standardising constant. For two samples of text i and j of size N_i and N_j where $0 < N_i \leq N_j$ and given that F_{ix} and F_{jx} represent the frequency of token x in sample i and j respectively, Labbé's intertextual distance can be expressed as,

$$\delta(i,j) \quad = \quad \frac{\sum \left| F_{ix} - F_{jx}\frac{N_i}{N_j} \right|}{\sum F_{ix} + \sum F_{jx}\frac{N_i}{N_j}} \tag{4}$$

The N_i/N_j term is present to standardise the frequency of the larger sample to that of the smaller sample. Labbé (2007) found empirically that this measure was a function of the sample size, this can be corroborated by considering how sample size increases will affect the value of δ.

The frequency count vectors are a function of the sample size. For any given document of size N_i it is self-evident that $\sum_{x=1}^{n} F_{ix} = N_i$. Therefore the denominator of Equation (4) is equal to $2N$ because the second frequency term is standardised to the sample size of the first. For δ to be sample size independent the difference between frequencies, the numerator, would have to remain proportional to $2N$. This is extremely unlikely owing to the LNRE property of lexical distributions and the non-independent nature of word occurrences. To prevent this dependency Labbé proposed the 'sliding window' method (Labbé, 2007).

The sliding window method attempts to remove sample size variation by calculating intertextual distances for fixed sized contiguous blocks and then taking the mean of all blocks to be the

overall intertextual distance. The idea being that the block mean should be invariant because each block is equally sized. The block size B is calculated via the greatest common divisor (gcd).

$$B \;=\; gcd(N_i - \epsilon_i, N_j - \epsilon_j)$$

Commonly there is no sufficiently sized *gcd* for both text lengths and a few characters have to be removed so that an adequate common divisor can be found (denoted here by ϵ). Once the block size is known samples i and j of size N_i and N_j are split into S_i and S_j subsamples of the size B.

$$\left.\begin{aligned} S_i &= \frac{N_i - \epsilon_i}{B} \\ S_j &= \frac{N_j - \epsilon_j}{B} \end{aligned}\right\} where\ S_i < S_j$$

The sliding window intertextual distance is then calculated by summing the non-sliding window intertextual distances for all possible combinations of blocks and standardising with the total number of blocks.

$$\Delta_{ij} = \frac{1}{S_i S_j} \sum_{x=1}^{S_i} \sum_{z=1}^{S_j} \delta(B_{ix}, B_{jz})$$

Where B_{ix} represents block x from text i and B_{jz} block z from text j. The sliding window method implicitly assumes that any sample size bias present in two samples will be equal because the size of the samples is equal.

5.3 Sample size dependencies

We will now explore how the sliding window method and the proposed authorship ratio mitigate sample size dependencies. The data used for testing was arbitrarily selected from the literary corpus with the only criterion being that the texts are of a reasonable length, in this case the compared texts are the complete works of Austen and Dickens. To test for invariance the distance metrics will be calculated on cumulative subsamples ranging from 10% of the data to 100%. If the metrics are free of sample size bias they should show minimal variation and if any variation is exhibited, it should not be systematically dependent on the sample size.

The intertextual distance sliding window method will be calculated in the same manner as the non-sliding-window method whilst accommodating the mean subsample procedure, i.e. each cumulative subsample will be split into partitions and the mean of all partitions will be taken as the intertextual distance for that cumulative subsample.

Figure 1 shows the resulting cumulative subsample plots for both distance metrics. All distance measures apart from the sample size covariate authorship ratio look to be dependent on the sample size. Each distance measure's dependency on the sample size can be quantified using Spearman's non-parametric rank correlation coefficient. The null hypothesis of the Spearman rank test states the population correlation coefficient ρ is equal to 0. A low p-value indicates the data is unlikely to have occurred under that hypothesis, therefore the data must have come from a population with a non-zero correlation coefficient.

Table 4 shows that both intertextual distances have significant negative correlation with N. The authorship ratio shows a significant positive correlation with N but less than both the intertextual distances. Clearly the sample size covariate authorship ratio performs best. It shows a minor non-significant positive correlation with N and the standard deviation is tiny; 3 degrees of magnitude less than the other authorship ratio, 2 less than the intertextual window and 1 less than the raw intertextual distance. The standard deviation for the basic authorship ratio is extremely large so it appears that including a sample size covariate greatly reduces variability.

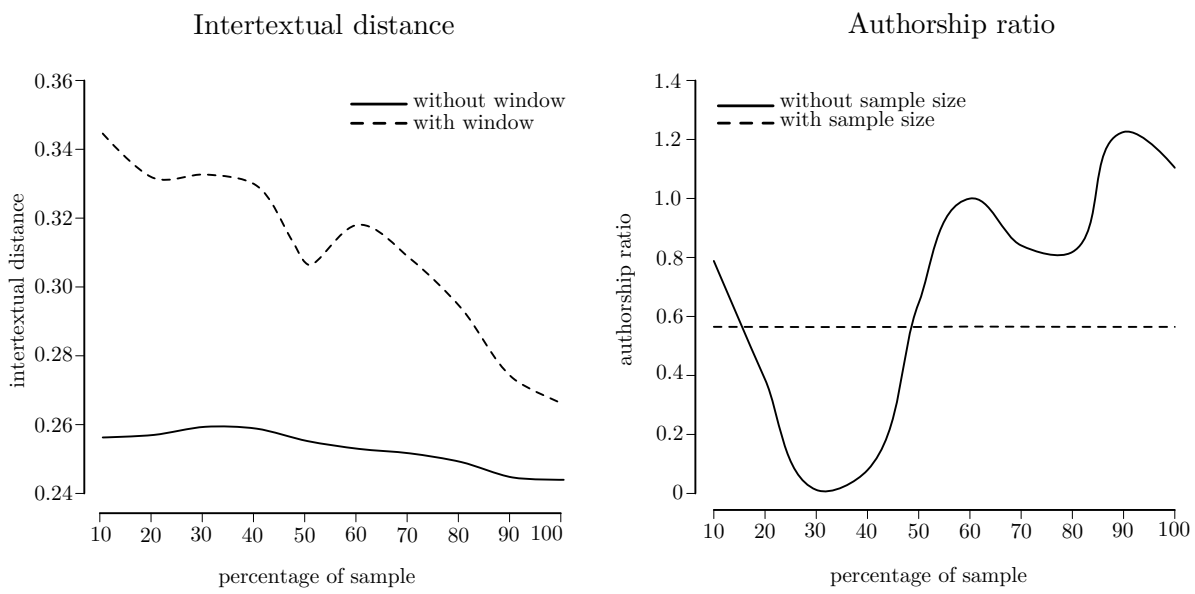

Figure 1: *Empirical sample size bias of distance metrics for texts by Dickens and Austen*

δ	ρ	p-value	std.dev
intertextual	-0.89090	0.00138	0.00552
intertextual window	-0.95151	0.00001	0.02616
AR	0.78182	0.01072	0.41220
AR sample size	0.28484	0.41710	0.00046

Table 4: *Correlation between distance metrics and sample size for texts by Dickens and Austen*

5.4 Support vector machines

To compare the discrimination ability of the distance metrics both will be classified with a radial basis function (RBF) support vector machine (Vapnik, 1999). Support vector machines (SVM) separate a response variable into two partitions by constructing an n-dimensional feature-space hyperplane, in this case the partitions will be 'same author' and 'different authors'. SVMs attempt to maximise the margin between two partitions so that the generalisation error is minimal and over-fitting is prevented. SVMs excel at classification but besides accuracy and error statistics they are rather opaque hence the need for a more interpretable stylistic distance function.

In SVM terminology the radial basis function is known as a kernel, the kernel transforms a non-linear relationship into a linear one (Cortes and Vapnik, 1995) by replacing every support vector dot product with the kernel function. The support vectors are the observations on the hyperplane margins. The RBF kernel has a hyperparameter traditionally denoted by γ, it needs to be estimated to ensure the most optimal accuracy is achieved. Methods of statistical estimation are too computationally intensive to use with a SVM, instead γ is estimated using a simple grid search over the likely parameter space using the *e1071* R package *libSVM* bindings.

To aid the SVM with the unbalanced data we specify a prior weight. SVMs make use of weighting through error penalisation so that misclassifications of the dominant class are considered less important than misclassifications of the recessive class. The magnitude of the class weighting is proportional to the magnitude of the class bias within each dataset.

The SVM classifying the authorship ratio will operate on the individual stylometric components rather than the distance function itself; primarily because classifying the data twice will produce bad results but also it will allow us to independently assess the classification power of the stylometrics. The RBF SVM results are shown in Tables 6 and 5, both hyperparameters were estimated to be 1.

	A	B	C
accuracy	91.52	91.30	91.06
sensitivity	83.73	84.88	91.55
specificity	91.68	81.67	91.61

Table 5: *Weighted intertextual distance classified with RBF SVM*

	A	B	C
accuracy	93.82	91.82	90.34
sensitivity	78.42	77.83	84.95
specificity	94.11	92.37	90.72

Table 6: *Weighted authorship ratio classified with RBF SVM*

The two distance metrics are similarly accurate but the intertextual distance has higher sensitivity; although as shown in Section 5.3, the intertextual distance incorporates systematic sample size bias so the accuracies are inflated. The support vector machine achieves better accuracy than the native authorship ratio estimation algorithm (Table 3). The SVM's higher accuracy could be caused by a range of factors; the AR algorithm may sacrifice classification power for interpretability; some of the stylometric classification power may not be exploited by the AR algorithm; or the SVM could be incorporating sample size bias so the results are inflated.

6 Conclusion and future work

In this paper we have investigated whether is is possible to define a metric that provides an interpretable measure of distance between authorship styles. Out test data consisted of 40 million words of 19th Century literature. The authorship ratio proposed in this paper proved a good discriminator and was resilient to sample size bias. But it was shown the SVM performed better when classifying the same data with the same features. Ideally we would like to exploit the geometric elegance and error bounds guarantees of an SVM whilst providing some form of distance statistic. The raw intertextual distance is much simpler to calculate than the authorship ratio but heavily dependent on the sample size. Using the sliding window method does not seem to lessen this dependency, with the Austen and Dickens data the sliding window method actually showed a stronger correlation with the sample size.

The authorship ratio misclassifications were primarily caused by three types of error; short samples whereby the author's stylistic profile did not have time to converge; similarity of stylometrics for distinct authors and stylochronometric development whereby works written by a single author appear different due to chronological changes in the author's style.

Some of these errors could be fixed by introducing additional stylometrics, perhaps statistics that capture non-lexical data such as syntactic metrics (Köhler and Altmann, 2000). Syntactic stylometrics seem a promising approach because syntax is further removed from content and closer to subconscious author-specific cognitive behaviour. The low sensitivity of the authorship ratio is a cause for concern but this is more a function of the class distribution than a specific weakness in the algorithm. Incorporating some form of prior weighting instead of using the MCC cost function may improve the low score for unbalanced datasets. Finally, it may be beneficial to train the distance metric on cumulative samples of data. Instead of treating each text as a single unit, each text could be partitioned into n cumulative subsamples so that the sample size covariate is better adjusted to variation in text length.

Notes

[1] http://collections.chadwyck.co.uk/home/home_c19f.jsp
[2] http://www.ualberta.ca/~baayen/software.html

References

Argamon, S., C. Whitelaw, P. Chase, S. Dhawle, S. Hota, N. Garg, and S. Levitan. (2007). "Stylistic text classification using functional lexical features". *Journal of the American Society of Information Science*, 7, 91–109.

Baayen, H. R. (2001). *Word Frequency Distributions*. Text, Speech and Language Technology. Springer.

Burrows, J. F. (2003). "Questions of authorship: Attribution and beyond". *Computers and the Humanities*, 37.

Chang, C. C. and C. J. Lin. "LIBSVM: a library for support vector machines", 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Clough, P. "Plagiarism in natural and programming languages: an overview of current tools and technologies. Research Memoranda: Department of Computer Science, University of Sheffield, UK", 2000.

Cortes, C. and V. N. Vapnik. (1995). "Support vector networks". In *Machine Learning*, 273–297.

de Vel, O., A. Anderson, M. Corney, and G. Mohay. (2001). "Multi-topic e-mail authorship attribution forensics". In *Proc. Workshop on Data Mining for Security Applications, 8th ACM Conference on Computer Security (CCS 2001)*.

Diederich, J., J. Kindermann, E. Leopold, and G. Paass. (2003). "Authorship attribution with support vector machines". *Applied Intelligence*, 19.

Firth, D. (1993). "Bias reduction of maximum likelihood estimates". *Biometrika*, 80, 27–38.

Fung, G., O. Mangasarian, and J. Jay. (2003). "The disputed federalist papers: SVM feature selection via concave minimization". In *In TAPIA 03: Proceedings of the 2003 conference on Diversity in computing*, 42–46. ACM Press.

Herdan, G. (1955). "A new derivation and interpretation of yule's characteristic k". *Zeitschrift for Angewandte Mathematik und Physik (ZAMP)*, 6(4), 332–339.

Hirst, G. and O. Feiguina. (2007). "Bigrams of syntactic labels for authorship discrimination of short texts". *Lit Linguist Computing*, 22(4), 405–417.

Holmes, D. I. (1998). "The Evolution of Stylometry in Humanities Scholarship". *Lit Linguist Computing*, 13(3), 111–117.

Hoover, D. L. (2003). "Another perspective on vocabulary richness". *Computers and the Humanities*, 37(2), 151–178.

Kiss, T. and J. Strunk. (2006). "Unsupervised multilingual sentence boundary detection". *Computational Linguistics*, 32(4), 485–525.

Köhler, R. and G. Altmann. (2000). "Probability distributions of syntactic units and properties". *Journal of Quantitative Linguistics*, 7, 189–200.

Kosmidis, I. "brglm: Bias reduction in binomial-response GLMs", 2008. `http://cran.r-project.org/web/packages/brglm/`.

Labbé, C. and D. Labbé. (2001). "Intertextual distance and authorship attribution – Corneille and Moliere". *Journal of Quantitative Linguistics*, 8(19), 213–231.

Labbé, D. (2007). "Experiments on authorship attribution by intertextual distance in English". *Journal of Quantitative Linguistics*, 14(48), 33–80.

Loper, E. and S. Bird. "NLTK: the natural language toolkit", 2002. `http://www.nltk.org`.

Mascol, C. (1888). "Curves of Pauline and pseudo-Pauline style I". *Unitarian Review*, 539–546.

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". *Biochimica et Biophysica Acta - Protein Structure*, 405(2), 442–451.

McMenamin, G. R. (2002). *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton, FL: CRC Press.

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

Mosteller, F. and D. Wallace. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Massachusetts, series in behavioral science (quantitative methods).

Nelder, J. and R. Mead. (1965). "A simplex method for function minimization". *Computer Journal*, 308–313.

Orlov, J. K. (1983). "Ein modell der haufigkeitsstruktur des vokabulars". In *Studies on Zipf's Law, eds. H Guiter, MV Arapov*, 154–233. Bochum: Brockmeyer.

Powell, M. J. D. (1964). "An efficient method for finding the minimum of a function of several variables without calculating derivatives". *The Computer Journal*, 7(2), 155–162.

Rudman, J. (1998). "The State of Authorship Attribution Studies: Some Problems and Solutions". *Computers and the Humanities*, 31, 351–365.

Sherman. (1888). *Principle of sentence length as an indicator of style and attribution*.

Sichel, H. S. (1974). "On a distribution representing sentence-length in written prose". *Journal of the Royal Statistical Society. Series A (General)*, 137(1).

Sichel, H. S. (1975). "On a distribution law for word frequencies". *Journal Of The American Statistical Association*, 70, 542–547.

Sichel, H. S. (1986). "Word frequency distributions and type-token characteristics". *Mathematical Scientists*, 11, 45–72.

Simpson, E. H. (1949). "Measurement of diversity". *Nature*, 163, 688.

Smith, J. A. and C. Kelly. (2002). "Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works". *Computers and the Humanities*, 36, 411–430.

Somers, H. and F. J. Tweedie. (2003). “Authorship attribution and pastiche”. *Computers and the Humanities*, 37(23), 407–429.

Tweedie, F. J. and H. R. Baayen. (1998). “How variable may a constant be? measures of lexical richness in perspective”. *Computers and the Humanities*, 32, 323–352.

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer.

Yule, G. U. (1939). “On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship”. *Biometrika*, 30, 363–390.

Yule, G. U. (1944). *The Statistical Study Of Literary Vocabulary*. Cambridge University Press, Cambridge.

Zipf, G. K. (1932). *Selected studies on the principle of relative frequency in language*. Harvard University Press, Cambridge, MA.