

Mahalanobis distance with an adapted within-author covariance matrix: An authorship verification experiment

Shunichi Ishihara 

Australian National University, Speech and Language Laboratory, Canberra, Australia and Linguistics Program, School of Culture, History and Language, College of Asia and the Pacific, Canberra, ACT, 2600, Australia

Abstract

The rotated delta, which is argued to be a theoretically better-grounded distance measure, has failed to receive any empirical support for its superiority. This study revisits the rotated delta—which is more commonly known as the Mahalanobis distance in other areas—with two different covariance matrices that are estimated from training data. The first covariance matrix represents the between-author variability, and the second the within-author variability. A series of likelihood ratio-based authorship verification experiments was carried out with some different distance measures. The experiments made use of the documents arranged from a large database of text messages that allowed for a total of 2,160 same-author and 4,663,440 different-author comparisons. The Mahalanobis distance with the between-author covariance matrix performed far worse compared to the other distance measures, whereas the Mahalanobis distance with the within-author covariance matrix performed better than the other measures. However, superior performance relative to the cosine distance is subject to word lengths and/or the order of the feature vector. The result of follow-up experiments further illustrated that the covariance matrix representing the within-author variability needs to be trained using a good amount of data to perform better than the cosine distance: the higher the order of the vector, the more data are required for training. The quantitative results also infer that the two sources of variabilities—notably within- and between-author variabilities—are independent of each other to the extent that the latter cannot accurately approximate the former.

1 Introduction

In any task involving identification, classification or verification of objects and/or items, the essential part is an assessment of how similar or different the objects or items are. For this purpose, various distance

measures have been devised and tested in stylometric studies, particularly those concerning the authorship of text source (Burrows, 2002; Hoover, 2004a; Argamon, 2008; Smith and Aldridge, 2011; Ishihara, 2021).

The most well-known measure in stylometric studies is undoubtedly Burrows's (2002) Delta; its

effectiveness and robustness have been demonstrated for a variety of texts from different genres and languages, and Burrows's Delta has been successfully and widely endorsed in relevant studies (Hoover, 2004b; Rybicki and Eder, 2011; AbdulRazzaq and Mustafa, 2014; Þorgeirsson, 2018). At the same time, since Burrows (2002), several variations have been proposed to better deal with the unique characteristics of linguistic texts (e.g. the dependency of features and the high dimensionality of feature vectors), expecting to result in superior identification and discrimination performance (Hoover, 2004b; Argamon, 2008; Smith and Aldridge, 2011; Eder, 2015). While pointing out some inappropriate assumptions inexplicitly posited by Burrows's Delta,¹ Argamon (2008) proposed the rotated delta (a non-axis parallel quadratic delta) as a theoretically suitable derivate. The rotated delta considers the different degrees of dependency between words in their frequencies for measuring a distance, and it is expected to have a superior performance (see Section 2 for a mathematical exposition of the suitability of the rotated delta). However, Argamon's argument for the rotated delta was not empirically supported (Jannidis *et al.*, 2015; Evert *et al.*, 2017), unexpectedly unveiling a considerably poor performance for authorship attribution experiments in comparison to other distance measures. That is, the theoretical correctness of the rotated delta failed to be substantiated by experiments; other distance measures outperformed the rotated delta. This study revisits the rotated delta, which is based on Mahalanobis distance—or, as it is also called in statistical pattern recognition, a 'hyper-quadratic distance' (McLachlan, 2004).

For example, in authorship verification, a pair of documents (e.g. *A* and *B*, each of which may be represented as a vector consisting of the measurements of stylometric features) are assessed to answer the question of whether *A* and *B* were written by the same author or by different authors. Naturally speaking, the more similar they are, the more likely they are, *ceteris paribus*, to have been written by the same author. If each individual has their own unique writing style, the more distinctive it is between individual authors, the more accurately the authorship analysis (including identification and verification) can be performed. As such, the variability between authors (or inter-author variability) plays a key role in authorship

analysis. However, there is another variable, which is equally important to the between-author variability but nonetheless often overlooked, particularly within-author variability (or intra-author variability). Variations in the writing style of (mainly literary) authors—for example, within the same text or across different texts over time—have been studied in computational stylistics as an approach to gain a better understanding of authorship and textual style (McKenna and Antonia, 1996; Craig, 1999a, 1999b; Stewart, 2003; Hoover and Corns, 2004). It is widely assumed that change in an author's writing style is inevitable over time (Stamou, 2007; Juola, 2008; Hoover, 2017), and this within-author variability makes authorship analysis more challenging (Hoover, 2003, 2017). That said, many studies also show that such within-author variations are compatible with that author's overall distinctiveness from other authors (Burrows, 1987a; McKenna and Antonia, 1996; Hoover, 2003). The crucial point is to accurately model the within-author variability because the individual author model that is trained by multiple texts enhances the accuracy of authorship analysis (Burrows, 2002; Hoover, 2004b, 2017).

If within-author variability is constant, the bigger the between-author variability, the better the results of the authorship analysis. Likewise, if the between-author variability is constant, the smaller the within-author variability, the higher the accuracy with which authorship analysis can be carried out. As such, the between- and within-author variabilities interact with each other for authorship analysis, but it is not clear for writing styles to what extent relationships that exist among individuals are the same or different to those that exist within individuals. In other words, it is not certain as to what extent and how within- and between-author variabilities are interrelated so that, for example, one could be estimated from the other.

In the Mahalanobis distance (or rotated delta), an accurate estimate of the covariance matrix—which is the representation of the within-author variability as to, for example, how words are used and how they are interrelated—determines the accuracy of the system. In particular, when only one text is available for a given author, the covariance matrix of that author needs to be estimated from training data, with the unavoidable assumption of within-author variability being homogeneous across authors. In this case, there

are at least two possible approaches to obtaining the covariance matrix depending on the underlying assumptions with respect to the relationship between between- and within-author variability.

In this study, we compare the performance of the likelihood ratio-based authorship verification systems built on different distance measures, including Burrows's Delta, the cosine distance, and the Mahalanobis distance. For general introductions to the concept of the likelihood ratio, refer to Pawitan (2001) and Robertson *et al.* (2016). Refer also to Ishihara (2017, 2021) for the application of the likelihood ratio to authorship verification. For the Mahalanobis distance, two different types of covariance matrices were trialled based on two different assumptions. One is that the within- and between-author variabilities share a large extent of similarities; thus, the former can be well approximated on the basis of the latter. The other is that the within-author variability is substantially different from the between-author variability; thus, the former cannot be approximated on the basis of the latter.

It may seem logical to think that the within- and between-variabilities would not be related; thus, the second assumption is not valid. However, this point will be specifically touched on in Section 2.1, where it is argued that the second assumption could be valid depending on features. Nonetheless, the current study will provide insights into these two assumptions for linguistic text data. The two assumptions will be described in Section 2.2, and the advantage of the first assumption in statistical modelling over the second will be discussed.

The current study is built on Ishihara (2021), using the same database for experiments and the same approach for calculating likelihood ratios. Like other previous studies, Ishihara (2021) demonstrated the efficacy of cosine distance over other distance measures, including Burrows's Delta and Euclidian distance, to estimate forensic likelihood ratios for linguistic text evidence. For the sake of comparison, the results of cosine distance and Burrows's Delta were used as baseline cases in the current study for assessing the performance of the Mahalanobis distance. However, the database prepared by Ishihara (2021) was used differently to fit the purpose of the current study (see Section 3.1 for the different use of the database). Hence, the likelihood ratios were recalculated for the cosine distance and Burrows's Delta in this study.

Using a large existing database of electronically generated text messages, in Ishihara (2021), three pairs of documents (differing in word length: 700, 1,400, and 2,100 words) were arranged from texts composed by 2,160 authors, resulting in 2,160 same-author and 4,663,440 different-author comparisons. For estimating likelihood ratios, the measured properties (relative frequencies of words) of the documents in comparison (i.e. each document is modelled with a bag-of-words approach; see Section 3.3 for details) are first compared in terms of how similar or different they are, and the degree of similarity or difference is quantified by the target distance measures as a score. In a subsequent process, the score is converted to a likelihood ratio based on models built by training data. The entire database (2,160 same-author and 4,663,440 different-author comparisons) is used as training data for the conversion models, as well as to test the accuracy of the system. The accuracy of the system with the different distance measures is then assessed by means of the equal error rate (EER).

This article is structured as follows. Following Argamon (2008), Section 2 recapitulates the characteristics of Burrows's Delta, thus, clarifying why the rotated delta is considered to be theoretically appropriate for textual data. The superior performance of the cosine distance is also accounted for by referring to Evert *et al.* (2017). The Mahalanobis distance, on which the rotated delta is based, is introduced after emphasizing the importance of the within-author variance as well as the between-author variance. Section 3 next describes the experimental procedure, including a detailed explanation of the database, the text pre-processing, the bag-of-words model with word frequencies, the approach for estimating likelihood ratios, and the assessment metric. The results of the experiments are described and discussed in Section 4. A summary of the results then concludes the article (see Section 5).

2 Details of Different Distance Measures

Argamon (2008) provides mathematical insight into Burrows's Delta (D_B), explaining the probabilistic interpretation of the geometric distance while pointing

out the discrepancies between the statistical assumptions and the nature of textual data—particularly in terms of the distributions of word frequencies and the dependency between them. Evident from Equation (1), Burrows’s Delta is the sum of the absolute distance of each dimension (i) of the z -scored vectors ($z(A)$ and $z(B)$) that represent two documents (A and B) in comparison, and the sum is averaged by the number of dimensions (n):

$$D_B(A, B) = \frac{1}{n} \sum_{i=1}^n |z(A_i) - z(B_i)| \quad (1)$$

The z -score is obtained by subtracting out the mean frequency (μ_i) of word i and dividing it by the standard deviation (σ_i). Thus, Equation (1) can be reformulated as Equation (2):

$$\begin{aligned} D_B(A, B) &= \frac{1}{n} \sum_{i=1}^n |z(A_i) - z(B_i)| \\ &= \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i} \right| \\ &= \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - B_i}{\sigma_i} \right| \end{aligned} \quad (2)$$

If the author of document B has multiple documents that are represented as multivariate vectors, and vector B_i is the mean of the documents, the distribution of the word frequencies of the documents is

assumed to have a multivariate Laplace distribution. That is, measuring the geometric distance between two documents represented by the z -scores of the multivariate features (e.g. word frequencies) is the same as calculating the likelihood of one document against the model of the other document, which is based on a Laplace distribution. Use of the z -score normalization, which adjusts the weight of each value by the standard deviation (σ) of the samples, assumes a Gaussian distribution of the data. This contradicts the assumption of Burrows’s Delta regarding the data distribution, which is a Laplace distribution. As for word-frequency distributions, Jannidis *et al.* (2015) reported that patterns of incidence, in this context, are better modelled by a Gaussian distribution than a Laplace distribution, at least for frequently occurring words. This observation can be confirmed with the data in this study. The distributions of the word frequencies given in Fig. 1 (the second, fourth, and sixth most frequent words in the entire database)—which were collected from 4,320 documents, each of them consisting of approximately 2,100 words—are better approximated by a Gaussian model than a Laplace model. However, it is important to point out that the distributions of less frequent words start showing more skewed distributions. This is due to the fact that word frequencies are based on discrete data (i.e. occurrence counts), which are considered to follow a binomial or Poisson type distribution (Sichel, 1975).

Multiple discrepancies arise when Burrows’s Delta is applied for textual data comparison. The

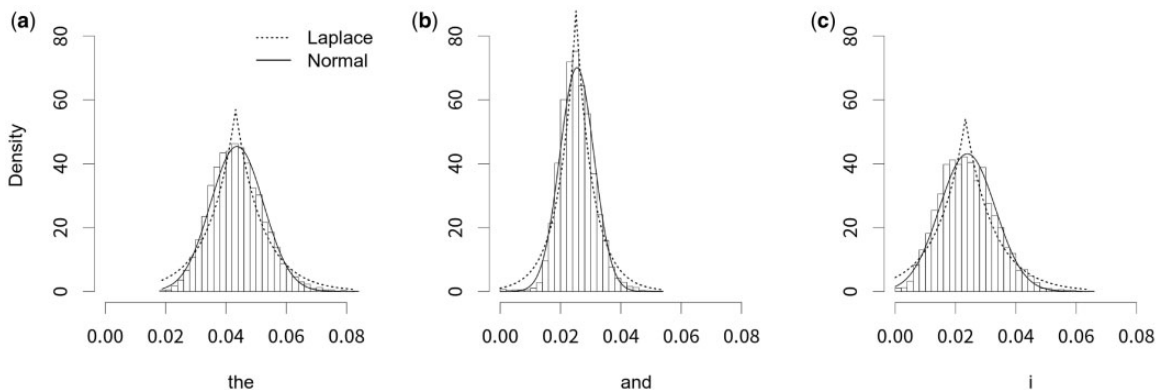


Fig. 1 Histograms of the frequencies of (a) ‘the’, (b) ‘and’ and (c) ‘i’, calculated from 4,320 documents of 2,100 words; all approximations have Gaussian and Laplace distributions

assumptions made by Burrows's Delta and the z-score normalization for the distribution of data are different; the former assumes a Laplace distribution and the latter a normal distribution. However, when it comes to linguistic text data, many of the typical features used are discrete (e.g. word frequencies), being likely to exhibit a negatively skewed distribution—neither a Laplace nor a normal distribution.

To rectify this inappropriate assumption of a Laplace distribution in Burrows's Delta, Argamon proposed the quadratic delta (D_Q) based on Euclidean distance, which assumes a Gaussian distribution of data (see Equation (3)):

$$D_Q(A, B) = \sum_{i=1}^n \frac{1}{\sigma_i^2} (z(A_i) - z(B_i))^2. \quad (3)$$

In addition to Argamon's proposals, some other distance measures have been put forward as possible improvements of Burrows's Delta (Hoover, 2004a; Smith and Aldridge, 2011; Eder, 2015). Among them, the cosine distance (or cosine similarity measure) is an important one, as it has been demonstrated to work better than Burrows's Delta (Jannidis *et al.*, 2015; Evert *et al.*, 2017). Smith and Aldridge (2011) also reported a substantial improvement in correctly identifying the authors of poems when the cosine measure is adapted to their authorship attribution tests. The cosine distance is represented using a dot product and magnitude as in Equation (4):

$$D_C(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2} \sqrt{\sum_i B_i^2}}. \quad (4)$$

The cosine distance is commonly used in information retrieval and text mining (Chowdhury, 2010) in which the angular similarity measure has been proven to reliably work with text data that are modelled with large word vectors (Frigui and Nasraoui, 2004). Mathematically speaking, the cosine distance between two vectors is proportional to the squared Euclidean

distance between the vectors that are normalized by the L_2 norm, expressed as $\|A\|_2 = \|B\|_2 = 1$. Thus, the squared Euclidean distance between the vectors is expressed as in Equation (5):

$$\begin{aligned} D_E^2(A, B) &= \|A - B\|^2 \\ &= (A - B)^T (A - B) \\ &= \|A\|^2 + \|B\|^2 - 2A^T B \\ &= 2 - 2A^T B \\ &= 2(1 - \cos(A, B)) \end{aligned} \quad (5)$$

As such, the squared Euclidean distance between two vectors ($\|A - B\|^2$) is a monotonic function of the angle of the vectors ($2(1 - \cos(A, B))$). Evert *et al.* (2017) reported in their study that the experimental result based on the quadratic delta (D_Q) measured from the normalized vectors by the L_2 norm is virtually identical to the experimental result of the cosine distance (D_C). Based on the substantial improvement resulted from the normalized feature vector by the L_2 norm (i.e. the transformation of the feature vectors to a uniform length of 1), Evert *et al.* (2017, p. ii14) concluded that 'the difference in direction rather than in length of the vectors is decisive for authorship attribution'.

Argamon (2008) pointed out another discrepancy between the statistical assumptions of Burrows's Delta and the nature of textual data. Burrows's Delta assumes the independence of words regarding their frequencies, but that assumption does not stand up to scrutiny (Argamon, 2008; Evert *et al.*, 2017). Different degrees of dependency between words can be well observed in Fig. 2, where two-dimensional plots of some paired word frequencies of the present study are given. All of the words used in Fig. 2 belong to the top ten most frequent words in the database. The correlation coefficients were calculated from the variance-covariance matrices included in Fig. 2. The correlation between the words given in Panel A is the weakest ($r = -0.02502$), and the correlation given in Panel C is the strongest ($r = 0.44547$) out of the three; Panel B is in-between ($r = -0.21605$).

Argamon (2008) proposed the rotated delta (D_R), which is a non-axis parallel quadratic delta, to

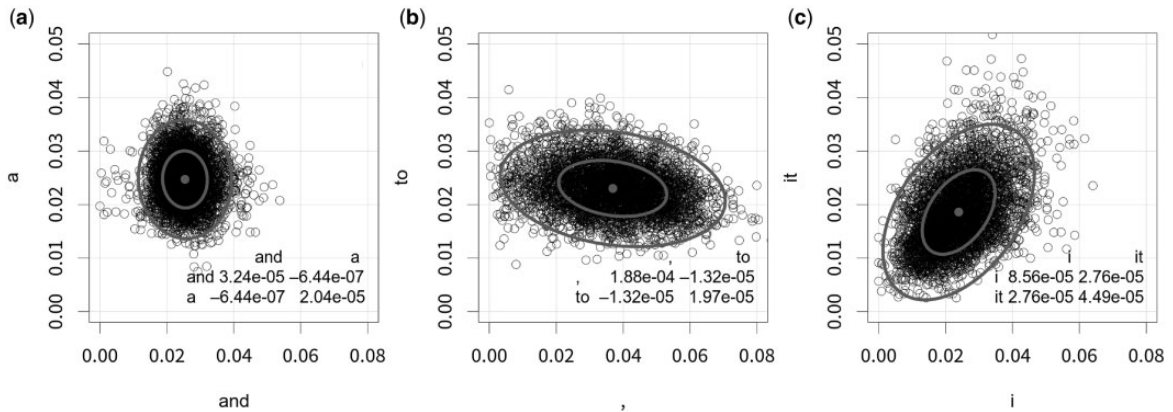


Fig. 2 Scatter plots of the word frequencies of the paired words (a=[‘and’, ‘a’]; b=[‘,’, ‘to’]; c=[‘i’, ‘it’]) with 50% and 95% confidence levels. The ellipse centre is marked with a grey dot. A variance–covariance matrix is also included in the panels

appropriately deal with this dependence on word frequency. D_R is expressed in Equation (6):

$$D_R(A, B) = (A - B)^T S^{-1} (A - B) = \sum_{i=1}^n \sum_{j=1}^n (A_j - B_j) (S^{-1})_{ij} (A_i - B_i) \quad (6)$$

Here, S denotes the covariance matrix, and $(S^{-1})_{ij}$ denotes the i , the j th element of the (S^{-1}) covariance matrix. The rotated delta is based on Mahalanobis distance. The basic idea is to rotate the distribution of word frequencies by a covariance matrix into a space where they are maximally independent before measuring the Euclidean distance. The rotated delta also assumes a Gaussian distribution.

The theoretical superiority of the rotated delta arises from the very form of the Mahalanobis distance. That is, the covariance term—which is absent from the formulations of the other distance measures considered in this research—allows one to account for cross-dependencies either on a within- or on a between-author basis. A fortiori, one would expect the Mahalanobis distance to provide more flexibility in applying different weightings and to perform relatively better. However, the outcome of the experiments carried out by Jannidis *et al.* (2015) and Evert *et al.* (2017) failed to substantiate the theoretical correctness of the rotated delta in that it constantly performed worst in comparison to other distance

measures, with the cosine distance being the best performer. This study will revisit the rotated delta or the Mahalanobis distance, with particular focus on the selection of the covariance matrix (S) and its underlying assumptions.

2.1 Variability

In any task relating to classification, discrimination, and identification, two kinds of variabilities need to be taken into consideration to ensure accurate performance: between-source and within-source variabilities. In the context of authorship analysis, these are between-author (or inter-author) variability and within-author (or intra-author) variability (McMenamin, 2002). The former variable (between-author variability) is straightforward to understand, and is linguistically related to the concept of *idiolect*. As a result of this, McMenamin (2002, p. 53) stated that ‘no two individuals use and perceive language in exactly the same way, so there will always be at least small differences in the grammar each person has internalized to speak, write and respond to other speakers and writers’.

Even within the same individual, small changes in emotion, social context (e.g. writing formally or informally) and physiologic state (e.g. health), or the use of different communication platforms (e.g. email and Twitter), can cause significant variability in writing styles. Due to this, the same person would not write

in exactly the same way, even when they are asked to write on the same topic on more than one occasion.

Despite the fact that the two sources of variability are equally important to any authorship analysis task, to the best of our knowledge, the extent and nature of within-author variability in writing styles has been studied far less than the extent and nature of between-author variability. It is neither clear in what way(s) between-author and within-author variabilities are related. Fundamentally, the within-author variability must be lower in magnitude than the between-author variability to make authorship analysis feasible. Judging by the success of authorship analysis, this assumption must be empirically correct (Rocha *et al.*, 2017; Kestemont *et al.*, 2018; Stamatatos *et al.*, 2018; Altamimi *et al.*, 2019).

In the area of speaker recognition including both humans and machines (which is probably the field that comes closest to authorship analysis), some studies demonstrate that voice spaces for individual speakers are structured similarly to population voice spaces (Kreiman *et al.*, 2017; Lee *et al.*, 2019). From the viewpoint of statistical modelling, this implies that within-speaker variability can be well approximated based on information from between-speaker variability. In a state-of-the-art automatic speaker recognition system, in fact, each speaker model is built by adapting the background model (i.e. speaker-independent model) (Reynolds *et al.*, 2000, p. 25), which is based on large between-speaker data (Hansen and Hasan, 2015).

2.2 Mahalanobis distance and adapted covariance matrix

When an unsourced document is compared with a set of documents written by a specific author (e.g. Author Q) to observe the extent to which the former is similar to or different from the styles of Author Q, a Mahalanobis distance can be calculated as a measure of similarity/difference between the unsourced document and the mean vector of the documents written by that author with the covariance matrix obtained from the set of documents. That is, if multiple documents are available for an author, it is possible to directly obtain a covariance matrix, which is unique to that author; the more documents, the better it can capture within-author variance. However, when one document (A) is compared against another one (B)

(i.e. in the case of a one-to-one comparison), it is impossible to directly calculate the covariance of the author (e.g. the author of document B). In this case, it is unavoidable to adapt a covariance matrix from training data, with the inevitable assumption that within-author variability is constant across different authors.

Any databases that can be used for authorship studies (e.g. authorship attribution, verification, and classification) need to have at least two documents from the same authors to successfully implement and test same-author comparisons; only then will it be possible to assess whether the system can correctly identify that the documents are, indeed, written by the same author. This study deals with the case in which only two documents are available from each author. Thus, suppose the training data are word frequencies calculated from two documents ($n=2$) that were contributed from each of m authors; on this assumption, there are altogether $N (=n \times m)$ documents in the training data. Each document is represented by a th order of the feature vector. Denote the training data as matrix x_{ij} , $i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$.

In the case of a one-to-one comparison, there are two possible ways of obtaining a covariance matrix from training data, depending on different underlying assumptions. The first covariance matrix (S_B) can be obtained from the training data (x_{ij}) in the manner expressed in Equation (7):

$$S_B = \frac{\sum_{i=1}^m (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T}{m-1}, \quad (7)$$

where

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij}$$

That is, the S_B is the covariance matrix of the pooled mean feature vectors of the documents belonging to the same authors in the database. This essentially quantifies the between-author variability based on the assumption that the within-author variability can be well approximated by adapting the between-author variability.

The second covariance matrix S_W is estimated from the training data (x_{ij}) (Equation (8)):

$$S_w = \frac{\sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T}{N - m} \quad (8)$$

Here, the S_W is the covariance matrix, estimated as the overall mean of the pooled covariance matrices from each author. This essentially quantifies the within-author variability based on the assumption that it is substantially different from the between-author variability; thus, the former cannot be estimated using the latter.

On theoretical grounds, the second assumption should be easier to make explicit than the first one. However, statistically speaking, the first assumption is more advantageous than the second one because the covariance matrix (S_B) can be estimated from a good amount of data (represented by m), whereas the covariance matrix (S_W) is only the overall mean of the individual author's variance; this is based only on a minimal amount of data ($n=2$). Another point of relevance is the relationship between the between-author and within-author variabilities. To what extent are they similar to each other? If they share a good degree of similarity, the first assumption may bring about a better result than the second one.

This study carries out a series of likelihood ratio-based authorship verification tests using Burrows's Delta, the cosine distance and the Mahalanobis distance to empirically investigate the theoretical appropriateness of the Mahalanobis distance (or the rotated delta). Two different covariance matrices, expressed by Equations (7) and (8), were tested with the Mahalanobis distance. That is, the following four types of distance measures were compared for their performance:

- Burrows's Delta
- Cosine distance
- Mahalanobis distance with the adapted between-author covariance matrix
- Mahalanobis distance with the adapted within-author covariance matrix.

The word length of a document is a factor that has some bearing on the performance of the authorship

verification system; longer document lengths are naturally better. To investigate this point with respect to the different distance measures, documents with different word lengths (approximately 700, 1,400, and 2,100 words) were prepared from 2,160 authors of the database in this study. A pair of documents ($n = 2$) is available for each of the 2,160 authors ($m = 2,160$) for each of the document lengths. The efficacy of the distance measures was assessed as a function of word length. See Section 3.1 for details of the database.

A bag-of-words model was employed to represent each document with word frequencies based on the most frequent words selected from the entire database. The number of most frequent words was varied in the bag-of-words model to examine how it affects the performance of the distance measures. See Section 3.3 for details of the feature vectors.

3 Experiments

3.1 Database and comparison

A series of authorship verification experiments was carried out using the likelihood ratio as a discriminant function. This was done using the same database prepared by Ishihara (2021) in which same-author and different-author documents were drawn from a part of the Amazon Product Data Authorship Verification Corpus² (Halvani *et al.*, 2017). This is itself based on the existing Amazon Product Data Corpus³ (He and McAuley, 2016). The Amazon Product Data Corpus is a compilation of product reviews (e.g. ratings, text, and helpfulness votes) and metadata (e.g. descriptions, category information, price, brand, and image features) from Amazon, including 142.8 million review texts, collected from May 1996 to July 2014. For the purpose of authorship verification studies, only reviewers' identification and their associated review texts (altogether 21,534 review texts) were extracted from a total of 3,228 reviewers and subsequently stored as the Amazon Product Data Authorship Verification Corpus. Many of the reviewers contributed six or more reviews on different topics. Sizes of review texts were equalized to around 4kB, which corresponds to lengths of approximately 700 words each.

From the Amazon Product Data Authorship Verification Corpus, those reviewers ($n=2,160$)

who composed six or more reviews were singled out. The first six reviews of each, all written on different topics, were used in [Ishihara \(2021\)](#). Authorship verification considers two competing hypotheses based on the comparison of pairs of documents; they are known as the same-author and different-author hypotheses. To simulate pairs of comparisons under each of the hypotheses, the six reviews were first separated into two groups: the first three reviews and the last three reviews. Three documents that differed in word length (around 700, 1,400, and 2,100 words) were generated based on the three review texts of each group by concatenation (shown in [Fig. 3](#)).

Evident in [Fig. 3](#), the first review text of each group was used as is (i.e. as a document of 700 words). The first and second review texts were concatenated into a document of 1,400 words, and then all three review texts were combined into a document totalling 2,100 words. Documents of different word length were prepared to test the correlation between the amount of data (the word length of a document) and the performance of the system.

Hence, each of the 2,160 reviewers (=authors) included in the experiment had three sets of documents differing in word length (either 700, 1,400, or 2,100 words). The database contained 4,320 review documents in total for each of the three word lengths. Two documents for each word length belonging to

each of the 2,160 authors allowed us to generate 2,160 same-author comparisons and 4,663,440 ($=_{2160}C_2 \times 2$) different-author comparisons for each word length.

In [Ishihara \(2021\)](#), the database of 2,160 authors was partitioned into the test, background, and development databases, each of which consists of 720 authors; this partitioning of the database is a standard approach to avoid the overestimation of forensic likelihood ratios. However, the same partitioning was not performed in the current study. That is, the same-author and different-author comparisons generated from the entire database were used both as data for training a score-to-likelihood ratio conversion model and for testing the performance of the authorship verification system (see [Section 3.4](#) for details of the conversion model and testing scheme). The use of the entire database without partitioning was necessary to investigate how much data is required to reliably estimate the within-author covariance matrix (see [Section 4.2](#) for details).

3.2 Tokenization

The documents stored in the database were tokenized into word tokens with the `tokens()` function in the `quanteda` library ([Benoit et al., 2018](#)) of R Statistical Package. The `tokens()` function was used in the default setting. All characters were changed to lower case, and

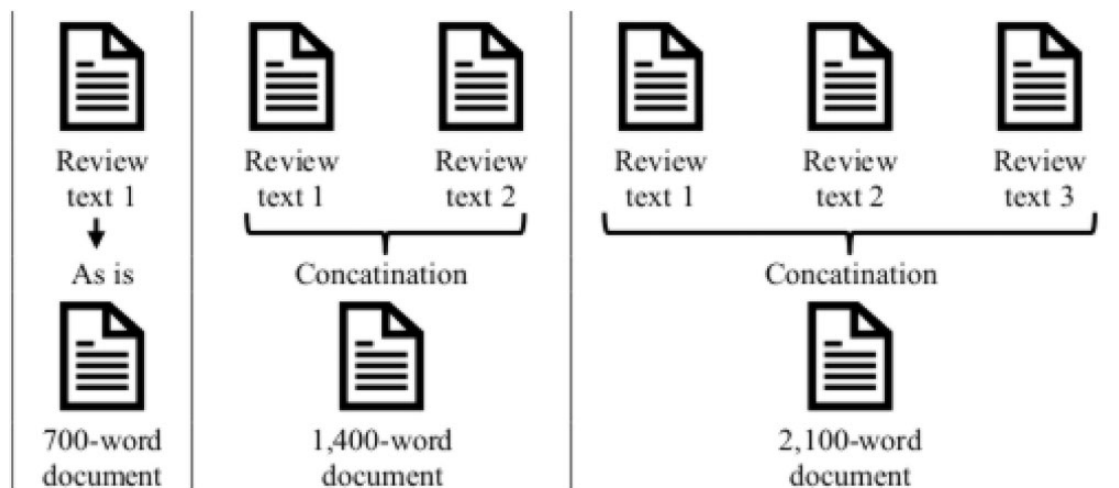


Fig. 3 Concatenation of different review texts for generating documents of different word lengths. Adapted from [Ishihara \(2021\)](#)

no stemming algorithm was applied. Further, punctuation marks were not removed and, as such, were counted as single word tokens.

3.3 Bag-of-words model with most frequent words

A bag-of-words model is a simple representation of text data that is widely used in natural language processing and related tasks. In this model, a text is represented as a set of words with the frequencies of their occurrences. It has been well reported that the bag-of-words model is an effective approach for authorship analysis tasks (Diederich *et al.*, 2003; Rocha *et al.*, 2017).

The 500 most frequent words appearing in the entire dataset were selected as components of the bag-of-words model for documents of 1,400 and 2,100 words, but 420 of the most frequent words for documents of 700 words. Then, the word frequencies of the model were calculated for each document. The total number of word tokens appearing in the entire dataset was 1,0749,013.

Deciding whether to number (N_{MFW}) the most frequent words in the bag-of-words model was based on arbitrary criteria (Burrows, 1987b; Stamatatos, 2006; Koppel *et al.*, 2007). However, an optimal selection may depend on the amount of data, the type of language, and the genre of the text data (e.g. Twitter, email, web content)—all of which must be empirically decided. In the experiments, the size of the bag-of-words vector was incremented from $N_{\text{MFW}} = 20$ to $N_{\text{MFW}} = 420$, or 500 by 20, to observe how the performance of the system fluctuates according to the order of the feature vector.

The word frequencies of the bag-of-words vector were normalized over the entire database in such a way that the mean of the frequencies was 0 for each word and their standard deviation was 1; this is also known as the z-score normalization. The bag-of-words vector of normalized word frequencies is essential (Burrows, 2002; Evert *et al.*, 2017) to equalize the amount of information across the words included in the vector for authorship analysis; otherwise the information encoded in the top-ranked words substantially and unevenly influences the outcomes of authorship analysis experiments, as word frequencies follow the distribution described by Zipf's (1932) law.

Note that the z-score normalization is irrelevant to the Mahalanobis distances; that is, the result is the same regardless of whether normalization is applied or not.

3.4 Authorship verification based on likelihood ratios

In the likelihood ratio approach employed in the authorship verification experiments of this study, the similarity or difference between the two documents in comparison was assessed against two hypotheses, referred to as the same-author (H_{sa}) and the different-author (H_{da}) hypotheses. For that, each document was modelled as a feature vector of word frequencies, and the similarity or difference between the documents was quantified through the target distance measures introduced in Section 2. The measured similarity or difference is technically called a score.⁴ The likelihood of the score under either hypothesis was estimated from the probabilistic distribution of the same-author or the different-author scores. Thus, the likelihood ratio can be defined as in Equation (9), where f denotes a probability density function, x a y are the vectors of word frequencies (w_i^j , $i \in \{1 \cdots N\}$, $j \in \{x, y\}$) of the documents to be compared ($x = \{w_1^x, w_2^x \cdots w_N^x\}$ and $y = \{w_1^y, w_2^y \cdots w_N^y\}$), and $\Delta(x, y)$ is a distance function generating a score:

$$\begin{aligned} \text{likelihood ratio} &= \frac{f(\Delta(x, y)|H_{\text{sa}})}{f(\Delta(x, y)|H_{\text{da}})} \\ &= \frac{f(\Delta(\{w_1^x, w_2^x \cdots w_N^x\}, \{w_1^y, w_2^y \cdots w_N^y\})|H_{\text{sa}})}{f(\Delta(\{w_1^x, w_2^x \cdots w_N^x\}, \{w_1^y, w_2^y \cdots w_N^y\})|H_{\text{da}})} \end{aligned} \quad (9)$$

The probability density functions under H_{sa} and H_{da} need to be trained from a dataset of scores. As briefly mentioned in Section 3.1, the dataset of scores under H_{sa} consists of the scores obtained from the same-author comparisons (2,160 comparisons in total), and the dataset of scores under H_{da} consists of scores from the different-author comparisons (4,663,440 comparisons) that can be generated from the database. This process is called a score-to-likelihood ratio conversion or calibration,⁵ and this type of likelihood ratio estimation is known as a score-based likelihood ratio approach (Hepler *et al.*, 2012; Bolck *et al.*, 2015; Garton *et al.*, 2020).

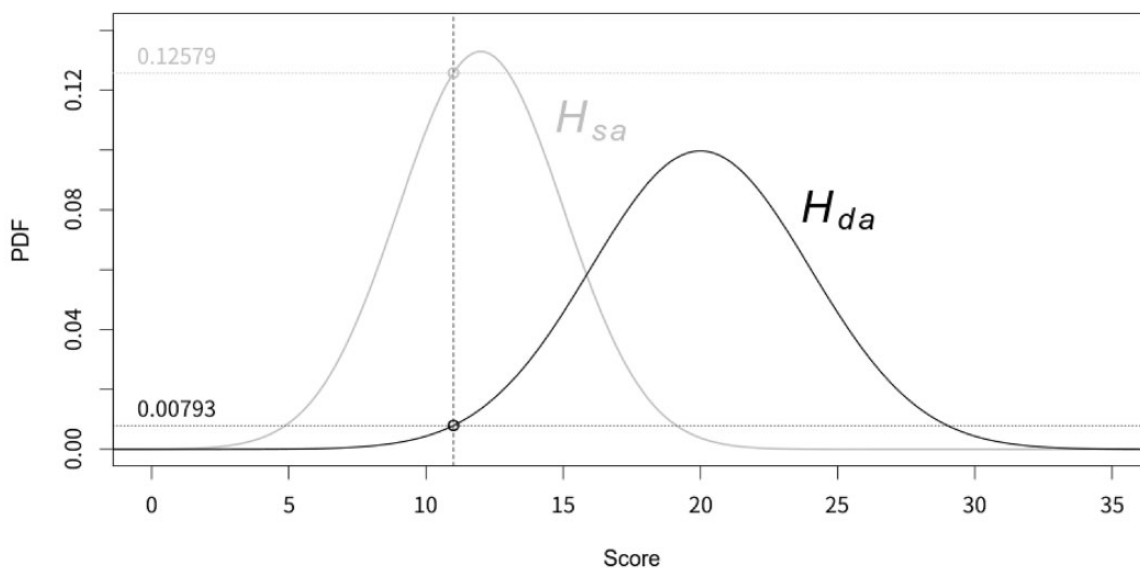


Fig. 4 Schematic illustration of a score-to-likelihood ratio conversion: x -axis=score; y -axis=PDF; grey curve=the model under H_{sa} (same-author model); black curve=the model under H_{da} (different-author model); PDF=probability distribution function. Adapted from Ishihara (2021)

Figure 4 illustrates the process of the score-to-likelihood ratio conversion with an example. Suppose that two documents are being compared, which are modelled as vectors by means of a distance (e.g. cosine distance) as a score. Further suppose that the calculated score is 10. For assessing the likelihood of this score value of 10 under the two hypotheses (H_{sa} vs. H_{da}), the within-author and the between-author variabilities, which are represented by the grey and black curves, respectively, were modelled with the same-author and different-author scores collected from the training database. Gaussian distribution is used for the models.

According to the models, the likelihood of a score of 10 under H_{sa} is 0.12579 and that of the same score value under H_{da} is 0.00793, with the likelihood ratio between them being 15.85330 ($\approx 0.12579/0.00793$). Setting a likelihood ratio of 1 as the threshold enables its use as a discriminatory function; a likelihood ratio higher than 1 denotes that the two documents in comparison are written by the same author, and a likelihood ratio lower than 1 denotes that the two documents are written by different authors. As such, a likelihood ratio of 15.8533062⁶ calculated for the example given in Fig. 4 suggests that the documents under comparison are written by the same author. If

the system is completely free from errors and perfectly calibrated, a likelihood ratio value being greater than 1 should be returned by the system for the comparisons of same-author documents, and mutatis mutandis, a likelihood ratio value being smaller than 1 for the comparisons of different-author documents.

As briefly mentioned, the same-author and different-author scores that are required for training a score-to-likelihood ratio conversion were obtained from 2,160 same-author and 4,663,440 different-author comparisons created from the entire database. The same two-author comparisons were also used as test data to investigate the discriminating accuracy of the system with the different distance measures (i.e. to determine how well the system can discriminate same-authored documents from different-authored documents). Ideally, the training data and testing data are separated, particularly if the database is small. However, judging from the size of the database, the effect on the experimental results arising from a non-partitioned database is relatively insignificant, if not negligible.

3.4.1 Distribution models

In the example score-to-likelihood ratio conversion given in Fig. 4, the distributions of same-author

(H_{sa}) and different-author (H_{da}) scores were modelled by a normal distribution. However, besides the normal distribution, three other models are used in this paper. They are the Weibull, log-normal and gamma distributions. This is because an initial observation of the distributional shapes of the same-author and different-author scores revealed that they do not always conform to normality, and that they often exhibit skewed distributions. The three functions are further used because they provide a better fit with the skewed distributions of the scores. The Weibull, gamma, and log-normal densities are similar in shape for the same coefficient of variation, but their differences become most significant in their tail behaviour (Tijms, 2003). Out of the three, the log-normal has the longest tail and the Weibull the shortest; thus, the log-normal model shows the most leptokurtic distribution. The best-fit model is selected by means of the Akaike (1974) information criterion for each experiment, separately for the same-source and different-source score distributions.

3.5 Evaluation of performance: equal error rate

The EER was used as a performance measure to show the discriminating power of the likelihood ratio-based authorship verification system. EER indicates the operating point at which the miss and false alarm rates are equal. The miss rate shows how many of the same-author comparisons the system failed to correctly assess as being from the same authors. The false alarm rate shows how many of the different-author comparisons the system failed to correctly assess as being from the different authors. The lower the EER is, the better.

The EER is graphically identifiable when the same-author and different-author likelihood ratios are plotted as a Tippett plot (Evelt and Buckleton, 1996). An example Tippett plot is given in Fig. 5, in which the same-author likelihood ratios are accumulatively plotted from the smallest to the highest (rising from the left) while the different-author likelihood ratios are from the highest to the smallest (rising from the right). The unit of the x -axis is the logarithm with base 10. The intersection of the two curves, which is indicated by the horizontal dashed line, is the point at which the miss and false alarm rates are equal (i.e. the EER). For the example given in Fig. 5, the EER

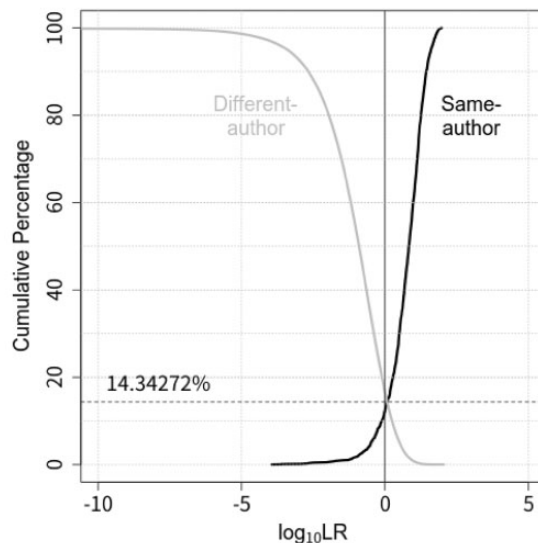


Fig. 5 Example Tippett plot in which the EER is identified as the intersection of the curves for the same-author (black) and different-author (grey) likelihood ratios. The dashed horizontal line that aligns with the intersection is the EER (14.34272%). The Tippett plot was generated for the likelihood ratios derived from Mahalanobis distance with within-author covariance for the documents of 1,400 words with 140 features

is 14.34272%. The `eer()` function of the `daivdav/ROC` library⁷ was used to calculate the EER.

As explained in Section 3.1, the entire database was used both for training a score-to-likelihood ratio conversion model and for testing the performance of the authorship verification systems with different distance measures. Although the database is large (thus, it would be less problematic), there is a possibility of overestimating likelihood ratios due to the use of the non-partitioned database. To confirm that the effect arising from the use of the entire database is insignificant, the likelihood ratios of some experiments were recalculated with a cross-validation approach. The EER values of the likelihood ratios with and without the cross-validated approach are given in Appendix I for comparison. In short, the EER values are virtually the same; the average difference is 0.07989%.

A tangible area to which the findings of the current study are applicable is the area of forensic science. As such, the derived likelihood ratios were also evaluated using the log-likelihood ratio cost (CLLR) (Brümmer

and du Preez, 2006), which is the standard metric used in forensic science for assessing the quality of likelihood ratios. For details on the assessment based on the CLLR, see Appendix II.

4 Results and Discussion

4.1 Overall performance

A series of experiments was carried out under a combination of conditions: for each experiment, we required (1) four distance measures for scores (Burrows's Delta, cosine, and Mahalanobis with two different covariance matrices), (2) different dimensions of the feature vector (N_{MFW}), and (3) different document lengths (700, 1,400, and 2,100 words). The EER values of the experiments were separately plotted for the four distance measures in Fig. 6 as a function of the N_{MFW} . The different panels of Fig. 6 denote the different document lengths.

Conforming to the results of previous studies (Evert *et al.*, 2017; Jannidis *et al.*, 2015; Ishihara, 2021), the cosine distance consistently performed better than Burrows's Delta regardless of the word length and number of most frequent words. As for the Mahalanobis distance, the selection of the covariance matrix determines its performance. The performance of the Mahalanobis distance with the between-author covariance matrix considerably underperformed the other distance measures, and this result echoes that of Evert *et al.* (2017) and Jannidis *et al.* (2015). However, exactly how the covariance matrix was adapted in their studies is not clear. Contrary to the between-author covariance matrix, the Mahalanobis distance with the within-author covariance matrix performed far better, but this outcome relative to the cosine distance is bound by word length and the order of the vector.

With a good amount of data (e.g. 1,400 and 2,100 words), the Mahalanobis distance with the within-author covariance matrix also performed better than the cosine distance, irrespective of the number of most frequent words. With documents consisting of 700 words, the cosine distance showed better discriminability up to the $N_{MFW} \approx 220$, after which the Mahalanobis distance with the within-author covariance matrix began performing better as the performance continued to improve with the increase in N_{MFW} .

Each distance measure exhibited unique trends in performance as a function of the number of the most frequent words. The Mahalanobis distance with the between-author covariance matrix hit the performance ceiling with a very low N_{MFW} (20), after which performance deteriorated at an alarming rate. With Burrows's Delta and the cosine distance, the performance of the system improved as the number of the N_{MFW} increased until a given N_{MFW} point (it is indicated by a large symbol in Fig. 6); then, the performance gradually deteriorated as the N_{MFW} increased further. The N_{MFW} at which the best performance was achieved was higher for the cosine distance than for Burrows's Delta. As far as the current data are concerned, the Mahalanobis distance with the within-author covariance matrix did not display any saturation in performance, even with the maximum N_{MFW} . The observation that the best-performing N_{MFW} is higher for the cosine distance than for Burrows's Delta, and that it is even higher for the Mahalanobis distance with the within-author covariance matrix than the cosine distance, indicates that better-performing distance measures make better use of the individuating information encoded in higher-order components of the feature vectors.

The substantial difference in performance between the between-author covariance matrix and the within-author covariance matrix demonstrated for the Mahalanobis distance in Fig. 6 is symptomatic of an important characteristic of textual data. That is, it can be inferred that the within-author and the between-author variabilities in word frequencies are different to the extent that the former cannot be estimated from the latter even with a large amount of between-author data. The extent of the relationship between the two sources of variability may be due to extrinsic reasons, such as the difference in the genre of text and its specific choice of words. However, this point warrants further research. Most importantly, the findings presented here empirically substantiate Argamon's (2008) argument that the rotated delta is theoretically better founded. The key point is that the covariance matrix not only needs to be selected based on the appropriate underlying assumption, but also needs to be trained by a good amount of data.

The current study was fortunately able to exploit a large amount of data being available from the database (i.e. texts collected from as many as 2,160 different

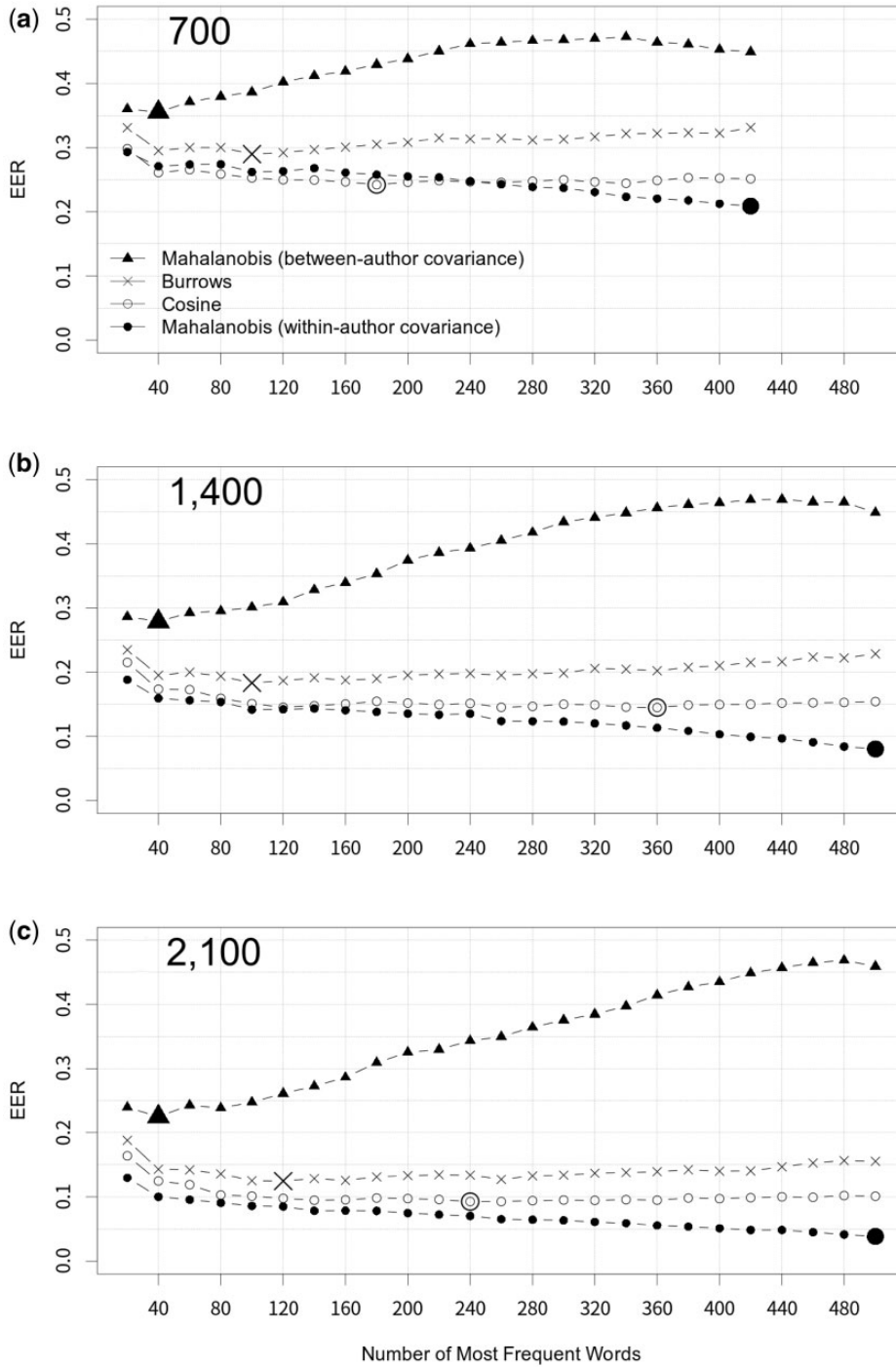


Fig. 6 EER values (y -axis) are plotted as a function of the number of N_{MFW} (x -axis) separately for Burrows's Delta, cosine, and Mahalanobis with within-author covariance and with between-author covariance matrix: (a) 700 words; (b) 1,400 words; and (c) 2,100 words. Large symbols=best EER values

authors). It is very likely that the superior performance of the Mahalanobis distance with the within-author covariance matrix is attributable to the sheer scale of data, which enabled an accurate estimate of within-author variability. Thus, it is prudent to investigate how the discriminability accuracy of the system will fluctuate as a function of the amount of data available for obtaining the within-author covariance matrix.

4.2 Amount of data for estimating the within-author covariance matrix

It goes without saying that availability of a larger dataset allows for more accurate estimates of the within-author covariance matrix. This consequently contributes to superior discrimination performance of the system with the Mahalanobis distance. In this section, the experiments that were conducted with the within-author covariance matrix for the system built on the Mahalanobis distance were repeated by changing the amount of data available for estimating the within-author covariance matrix. It is also expected that the amount of data required for the accurate estimate of this matrix depends on the size of N_{MFW} , as more features require more data for the model to be accurately built (Silverman, 1986).

As such, a series of experiments was carried out for each N_{MFW} of 100, 200, 300, 400, and 500 by increasing the number of samples that are used for estimating the within-author covariance matrix. For this, a given portion of the database was systematically selected and thereafter increased by 80 samples. The portion was chosen from the beginning of the database for one experiment, and the same experiment was repeated with the portion starting from the end of the database. That is, there were two experiments for the same experimental setting. As expressed by Equation (8), the within-author covariance matrix is the mean of the pooled covariance matrices from each of the maximum of 2,160 authors in the database. As such, the maximum number of samples was 2,160.

In this way of carrying out experiments, the samples used for estimating the covariance matrix are independent only up to 1,080 samples, after which more samples start to overlap for the estimate of the covariance as the sample increases; thus, the result is identical with a sample number of 2,160. Nevertheless, it

should be possible to obtain a ballpark figure of how the quantity of samples for estimating the within-author variability is relevant to the performance of the system. The initial portion of the database used to start the experiments was altered according to the number of N_{MFW} : 160, 240, 320, 480, and 560 samples, respectively. Documents of 2,100 words were used in the experiments.

The mean EER values were plotted as a function of the number of samples in Fig. 7; this was done separately for the different N_{MFW} values (100, 200, 300, 400, 500). Note that although the mean EER values of two experiments are used in the plot, the EER values of the two experiments are very similar to each other, even for sample numbers up to 1,080.

Figure 7 exhibits a clear descending trend in EER as the amount of data for an estimate of the covariance matrix increases. Figure 7 also shows how the degree of the descending trend differs—albeit systematically—across the different numbers for N_{MFW} . As anticipated, it is clear that the amount of data is a key factor for determining the performance of the system developed on the Mahalanobis distance with the within-author covariance matrix. When the number of the features in the vector is small (e.g. $N_{\text{MFW}} = 100$), the performance of the system starts settling down with less data—yet, it also continues to improve at a small but steady rate with more samples. Taking the best performance of the system with the cosine distance as the reference (EER = 9.25925%), the higher the order of the vector, the more samples the system requires to achieve a better performance than the reference performance. Describing this from the viewpoint of performance deterioration, it can be observed from Fig. 7 that the more features that are included in the vector, the more sensitive the system is to insufficiency of data; that is, the performance of the system deteriorates in a greater magnitude for larger N_{MFW} as the amount of available samples decreases. As a result, the system with the N_{MFW} of 500 (which performs the best with 2,160 samples) significantly underperforms in comparison to the system with the N_{MFW} of 100, when only a limited amount of data is available for the covariance matrix. This is a typical example of the phenomenon known as the ‘curse of dimensionality’ (Bellman, 1961, p. 97). As is well known, one factor that strongly influences the performance of pattern classifiers is the intrinsic dimensionality of the data

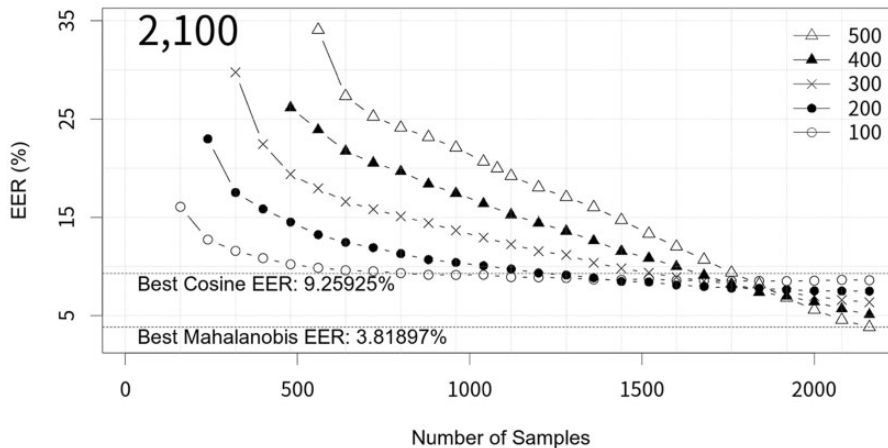


Fig. 7 Mean EER values are plotted against the number of samples separately for the different N_{MFW} (100, 200, 300, 400, and 500). The EER values of the best-performing systems with the cosine and Mahalanobis distances are included as a reference

(i.e. the ratio of the number of training samples to the feature-set dimensionality). In other words, a sufficiently large training-sample size relative to the size of the feature vectors is advisable to prevent suboptimal performance (Foley, 1972). Overall, the results presented in Fig. 7 show that, for the optimal outcome of the system built on the Mahalanobis distance with the adapted within-author covariance matrix, the N_{MFW} needs to be chosen according to the amount of available data for estimating the covariance matrix.

5 Conclusions

Since Burrows (2002), several distance measures have been developed and tested as possible improvements of Burrows's Delta for authorship analysis (Hoover, 2004b; Argamon, 2008; Eder, 2015; Smith and Aldridge, 2011). In particular, Argamon (2008) argued that the rotated delta was a theoretically better-founded distance measure reflecting the nature of textual data; thus, it was expected to perform better than other distance measures. However, Argamon's argument in this regard failed to be empirically warranted, in that, against expectation, the rotated delta considerably underperformed against other distance measures, including the original Burrows's Delta (Jannidis *et al.*, 2015; Evert *et al.*, 2017). The present study revisited the rotated delta, which is based on the

Mahalanobis distance, paying special attention to the estimation of the covariance matrix.

The results of a series of likelihood ratio-based authorship verification experiments using a large number of documents differing in word length (2,160 same-author and 4,663,440 different-author comparisons in total) demonstrated that the Mahalanobis distance works better than the cosine distance if the covariance matrix representing the within-author variability (not the between-author variability) is estimated using a good amount of training data. The results also showed that superior performance relative to the cosine distance is bound by word length and/or the order of the vector. The result of a follow-up experiment also showed that ample data are essential for making accurate estimates of the within-author covariance matrix, and to ensure that the Mahalanobis distance performs better than the cosine distance. Further shown, the higher the order of the vector, the more data are required for training for comparable results.

In conclusion, the current study empirically verified Argamon's argument for the rotated delta, but also portrayed the disposition of the Mahalanobis distance for authorship analysis; as such, the covariance matrix representing the within-author variability (not the between-author variability) needs to be both adapted and trained by a good amount of data. The quantitative results of the current study also infer that

the two sources of variabilities—notably within- and between-author variabilities—are independent of each other to the extent that the latter cannot accurately approximate the former. However, this inference should warrant attention for further research to confirm the relationship between them.

Acknowledgements

The author would like to thank the reviewer for their insightful and helpful comments. The author also would like to thank Frantz Clermont for his feedback on an earlier draft of this article. The author declares that he has no conflict of interest. Any opinions expressed in the final version of this paper are those of the author. This research did not receive any specific grant.

Appendix I

The EER was also calculated for the likelihood ratios estimated with a leave-one-out cross-validation method. In the cross-validation method, for calculating a likelihood ratio for each comparison, the documents attributable to the authors of the comparison were not included to train the score-to-likelihood ratio conversion model. That is, when you are estimating, for example, a likelihood

ratio of the comparison between Author *A* and Author *B*, the documents attributable to Authors *A* and *B* are not included for training the score-to-likelihood ratio conversion model. The likelihood ratios were calculated only for the documents of 1,400 words with the feature numbers (N_{MFW}) of 160, 320, and 480. This is because of the substantially demanding nature of cross-validation computations. The EER values with the cross-validation are presented in Table A1 together with the EER values without the cross-validation; i.e. the EER values given in Section 4.1.

Appendix II

In authorship verification, which is the target of the current study, likelihood ratio is commonly used as a discriminatory function to categorically determine whether a pair of texts were written by the same author or different authors. In such a case, EER, which assesses the accuracy of categorical decisions, is an appropriate metric. In forensic science, however, the likelihood ratio is used to quantify the strength of evidence, and the magnitude of likelihood ratio (not only whether the likelihood ratio correctly supports a hypothesis) needs to be taken into consideration for

Table A1. EER values obtained with and without cross-validation (Cross-valid.) calculation for Burrows's Delta, cosine, and Mahalanobis with within-author covariance and with between-author covariance matrix

Burrows's Delta				Cosine			
N_{MFW}	160	320	480	N_{MFW}	160	320	480
Cross-valid.	EER	EER	EER	Cross-valid.	EER	EER	EER
No	18.76753	20.60185	22.21909	No	15.04629	14.90810	15.28045
Yes	18.69727	20.55555	22.02333	Yes	14.93745	14.93410	15.30822
Mahalanobis with between-author covariance				Mahalanobis with within-author covariance			
N_{MFW}	160	320	480	N_{MFW}	160	320	480
Cross-valid.	EER	EER	EER	Cross-valid.	EER	EER	EER
No	33.97633	44.12037	46.50027	No	14.07407	12.03703	8.39727
Yes	33.87448	44.02777	46.66349	Yes	14.02777	11.97487	8.37962

It is clear from the comparisons between the EERs with and without cross-validation that the performance is virtually the same with or without a cross-validation. The values are for the documents of 1,400 words when the number of features (N_{MFW}) are 160, 320, and 480: Yes=with cross-validation; No=without cross-validation.

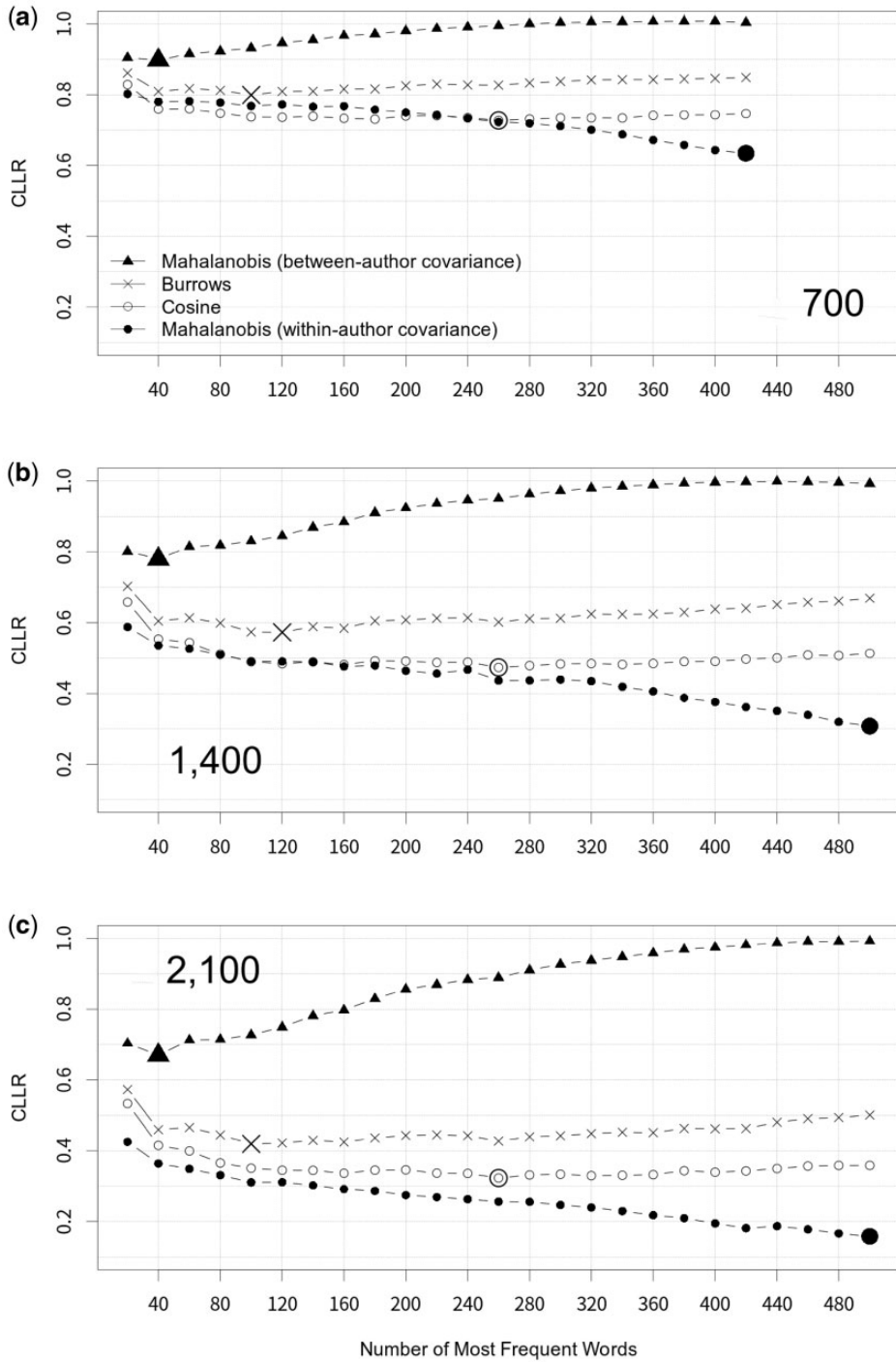


Fig. A1 Log-likelihood ratio cost (CLLR) values (y -axis) are plotted as a function of the number of N_{MFW} (x -axis) separately for Burrows’s Delta, cosine, and Mahalanobis with within-author covariance and with between-author covariance matrix: (a) 700 words; (b) 1,400 words; and (c) 2,100 words. Large symbols=best CLLR values

assessment. The log-likelihood ratio cost (CLLR) (Brümmer and du Preez, 2006) is a commonly used metric for assessing likelihood ratio-based forensic source detection systems. The mathematical explication of CLLR is beyond the scope of the present study; those who are interested, refer to Appendix C.2 of Morrison *et al.* (2021) and Section 2.4 of Morrison (2011). In brief, the CLLR assesses each likelihood ratio value, and assigns a penalty or cost. The likelihood ratio that supports a contrary-to-fact hypothesis attracts a large cost; the stronger the support for a contrary-to-fact hypothesis is, the greater the cost. The closer to 0, the better. If an CLLR is smaller than 1, it means that the system provides some useful information to the trier of fact.

Having the application of the current study in forensic science in mind, the CLLR values were calculated for the experiments, and they are separately plotted for the four distance measures in Fig. A1 as a function of the N_{MFW} . The different panels of Fig. A1 denote the different document lengths.

The CLLR values observable from Fig. A1 display very similar trends with the EER values observable from Fig. 6c. For all the experiments with different feature numbers and document lengths, the likelihood ratios derived with Burrows's Delta, cosine and Mahalanobis with within-author covariance were assessed with $CLLR < 1$. That is, the systems with these distance measures provide useful information. Only the Mahalanobis distance with between-author covariance failed to achieve $CLLR < 1$ for the document length of 700 words with 300 or more features, in which case the system must be discarded for casework.

References

- AbdulRazaq, A. A. and Mustafa, T. K. (2014). Burrows-Delta method fitness for Arabic text authorship stylometric detection. *International Journal of Computer Science and Mobile Computing*, 3(6): 69–78.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–23.
- Ali, T., Spreeuwers, L., Veldhuis, R. and Meuwly, D. (2015). Sampling variability in forensic likelihood-ratio computation: a simulation study. *Science & Justice*, 55(6): 499–508.
- Altamimi, A., Alotaibi, S. and Alruban, A. (2019). Surveying the development of authorship identification of text messages. *International Journal of Intelligent Computing Research*, 10(1): 953–66.
- Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2): 131–47.
- Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton: Princeton University Press.
- Benoit, K., Watanabe, K., Wang, H. et al. (2018). *quanteda*: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30): 774–76.
- Bolck, A., Ni, H. F. and Lopatka, M. (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: Applied to forensic MDMA comparison. *Law, Probability and Risk*, 14(3): 243–66.
- Brümmer, N. and du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20(2–3): 230–75.
- Burrows, J. F. (1987a). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press; New York: Oxford University Press.
- Burrows, J. F. (1987b). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2(2): 61–70.
- Burrows, J. F. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–87.
- Chowdhury, G. G. (2010). *Introduction to Modern Information Retrieval*, 3rd edn. London: Facet.
- Craig, H. (1999a). Contrast and change in the idiolects of Ben Jonson characters. *Computers and the Humanities*, 33(3): 221–40.
- Craig, H. (1999b). Jonsonian chronology and the styles of *A Tale of a Tub*. In Butler, M. (ed), *Re-Presenting Ben Jonson*. London: Palgrave Macmillan, pp. 210–32.
- Diederich, J., Kindermann, J., Leopold, E. and Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, 19(1–2): 109–23.

- Eder, M.** (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, **30**(2): 167–82.
- Evert, S., Proisl, T., Jannidis, F. et al.** (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, **32**(suppl_2): ii4–ii16.
- Evvett, I. W. and Buckleton, J. S.** (1996). Statistical analysis of STR data. In Carracedo, A., Brinkmann, B. and Bär, W. (eds), *Advances in Forensic Haemogenetics*. Santiago de Compostela: Springer, pp. 79–86.
- Foley, D.** (1972). Considerations of sample and feature size. *IEEE Transactions on Information Theory*, **18**(5): 618–26.
- Frigui, H. and Nasraoui, O.** (2004). Simultaneous clustering and dynamic keyword weighting for text documents. In Berry, M. W. (ed.), *Survey of Text Mining*. New York: Springer, pp. 45–72.
- Garton, N., Ommen, D., Niemi, J. and Carriquiry, A.** (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. <https://arxiv.org/abs/2002.09470> (accessed 20 July 2020).
- Halvani, O., Winter, C. and Graner, L.** (2017). Authorship verification based on compression-models. <http://arxiv.org/abs/1706.00516> (accessed 25 June 2020).
- Hansen, J. H. L. and Hasan, T.** (2015). Speaker recognition by machines and humans: a tutorial review. *IEEE Signal Processing Magazine*, **32**(6): 74–99.
- He, R. and McAuley, J.** (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, Montréal, Québec, Canada, 11–15 April 2016.
- Hepler, A. B., Saunders, C. P., Davis, L. J. and Buscaglia, J.** (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, **219**(1–3): 129–40.
- Hoover, D. L.** (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, **18**(4): 341–60.
- Hoover, D. L.** (2004a). Delta prime? *Literary and Linguistic Computing*, **19**(4): 477–95.
- Hoover, D. L.** (2004b). Testing Burrows's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75.
- Hoover, D. L.** (2017). The microanalysis of style variation. *Digital Scholarship in the Humanities*, **32**(suppl_2): ii17–ii30.
- Hoover, D. L. and Corns, T. N.** (2004). The authorship of the postscript to 'An Answer to a Booke Entitled, An Humble Remonstrance'. *Milton Quarterly*, **38**(2): 59–75.
- Ishihara, S.** (2017). Strength of forensic text comparison evidence from stylometric features: a multivariate likelihood ratio-based analysis. *The International Journal of Speech, Language and the Law*, **24**(1): 67–98.
- Ishihara, S.** (2021). Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. *Forensic Science International*, **327**:110980.
- Jannidis, F., Pielström, S., Schöch, C. and Vitt, T.** (2015). Improving Burrows' Delta. An empirical evaluation of text distance measures. In *Proceedings of Digital Humanities 2015*, Sydney, Australia.
- Juola, P.** (2008). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.
- Kestemont, M., Tschuggnall, M., Stamatatos, E. et al.** (2018). Overview of the author identification task at PAN-2018: Cross-domain authorship attribution and style change detection. In Cappellato, L., Ferro, N., Nie, J.-Y. and Soulier, L. (eds), *Proceedings of the CLEF 2018 Evaluation Labs*, Avignon, France, 10–14 September 2018.
- Koppel, M., Schler, J. and Bonchek-Dokow, E.** (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, **8**: 1261–76.
- Kreiman, J., Keating, P. and Vesselinova, N.** (2017). Acoustic similarities among voices. Part 2: Male speakers. *The Journal of the Acoustical Society of America*, **142**: 2519.
- Lee, Y., Keating, P. and Kreiman, J.** (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, **146**(3): 1568–79.
- Leegwater, A. J., Meuwly, D., Sjerps, M., Vergeer, P. and Alberink, I.** (2017). Performance study of a score-based likelihood ratio system for forensic fingerprint comparison. *Journal of Forensic Sciences*, **62**(3): 626–40.
- McKenna, W. and Antonia, A.** (1996). 'A few simple words' of interior monologue in *Ulysses*: Reconfiguring the evidence. *Literary and Linguistic Computing*, **11**(2): 55–66.
- McLachlan, G. J.** (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken: John Wiley & Sons.
- McMenamin, G. R.** (2002). *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton: CRC press.
- Morrison, G. S.** (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, **51**(3): 91–98.

- Morrison, G. S. and Enzinger, E.** (2018). Score based procedures for the calculation of forensic likelihood ratios—Scores should take account of both similarity and typicality. *Science & Justice*, **58**(1): 47–58.
- Morrison, G. S., Enzinger, E., Hughes, V. et al.** (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, **61**(3): 299–309.
- Morrison, G. S., Enzinger, E. and Zhang, C.** (2018). Forensic speech science. In Freckelton, I. and Selby, H. (eds), *Expert Evidence*. Sydney, Australia: Thomson Reuters.
- Neumann, C. and Ausdemore, M.** (2020). Defence against the modern arts: The curse of statistics-Part II: 'Score-based likelihood ratios'. *Law, Probability and Risk*, **19**(1): 21–42.
- Pawitan, Y.** (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.
- Reynolds, D. A., Quatieri, T. F. and Dunn, R. B.** (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, **10**(1–3): 19–41.
- Robertson, B., Vignaux, G. A. and Berger, C. E. H.** (2016). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, 2nd edn. Chichester: John Wiley and Sons, Inc.
- Rocha, A., Scheirer, W. J., Forstall, C. W. et al.** (2017). Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, **12**(1): 5–33.
- Rybicki, J. and Eder, M.** (2011). Deeper Delta across genres and languages: Do we really need the most frequent words? *Literary and Linguistic Computing*, **26**(3): 315–21.
- Sichel, H. S.** (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**(351a): 542–47.
- Silverman, B. W.** (1986). *Density Estimation for Statistics and Data Analysis*. London; New York: Chapman & Hall.
- Smith, P. W. H. and Aldridge, W.** (2011). Improving authorship attribution: Optimizing Burrows' Delta method. *Journal of Quantitative Linguistics*, **18**(1): 63–88.
- Stamatatos, E.** (2006). Authorship attribution based on feature set subsampling ensembles. *International Journal on Artificial Intelligence Tools*, **15**(5): 823–38.
- Stamatatos, E., Rangel, F., Tschuggnall, M. et al.** (2018). Overview of PAN 2018: Author identification, author profiling, and author obfuscation. In Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J.-Y., Soulier, L., San, E., Cappellato, J. and Ferro, N. (eds), *Proceedings of the 9th International Conference of the Cross-Language Evaluation Forum for European Languages*, Avignon, France, 10–14 September 2018.
- Stamou, C.** (2007). Stylochometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, **23**(2): 181–99.
- Stewart, L. L.** (2003). Charles Brockden Brown: Quantitative analysis and literary interpretation. *Literary and Linguistic Computing*, **18**(2): 129–38.
- Porgeirsson, H.** (2018). How similar are Heimskringla and Egils saga? An application of Burrows' delta to Icelandic texts. *European Journal of Scandinavian Studies*, **48**(1): 1–18.
- Tijms, H. C.** (2003). *A First Course in Stochastic Models*. New York: Wiley.
- Zipf, G. K.** (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge: Harvard University Press.

Notes

- 1 Before [Burrows \(2002\)](#), authorship attribution relied on multivariate statistical comparisons, which were superior in comparing a very small number of documents. However, such a multivariate statistical approach is not flexible in comparing a large number of documents and ranking them according to the similarity to the target document. Burrows's Delta was originally devised as a simple measure capable of identifying the most likely candidate from a large group of potential candidates. Considering also the fact that a text is usually represented by a large number of features (i.e. the dimension of a feature vector is inevitably high), it appears that the emergence of Burrows's Delta or the wide use of distance measures in authorship attribution is necessary; nevertheless, it is not clear whether Burrows was aware of the discrepancy between the statistical assumption of Burrows's Delta and the distributional nature of textual data described in this study.
- 2 See <http://bit.ly/1OjFRhJ>.
- 3 See <http://jmcauley.ucsd.edu/data/amazon/>.
- 4 This way of estimating likelihood ratios from similarity scores is called the 'similarity-only-score-based' method ([Morrison and Enzinger, 2018](#); [Ishihara, 2021](#)). The use of similarity-only-score-based methods has been criticized by some researchers ([Morrison and Enzinger,](#)

- 2018; Neumann and Ausdemore, 2020) for estimating forensic likelihood ratios (i.e. the strength of evidence) in forensic science as the score only assesses how similar a pair of objects is without considering the typicality of them. However, the current study is not a forensic study and uses the likelihood ratio as a discriminatory function, not as the strength of evidence.
- 5 Several statistical models and algorithms can be used for calibration, including the one based on kernel density estimation (Ali *et al.*, 2015), logistic regression (Morrison *et al.*, 2018), or pool adjacent violators (Brümmer and du Preez, 2006). To the best of our knowledge, the model used in the current study, which is based on the distribution of training scores, is a common calibration method for converting similarity-only scores to likelihood ratios (Bolck *et al.*, 2015; Leegwater *et al.*, 2017; Ishihara, 2021). The efficacy of the particular method used in the current study was demonstrated in Ishihara (2021) for linguistic textual data.
 - 6 Precisely speaking, a likelihood ratio of, for example, 15.85330, tells us that the degree of similarity/difference (expressed by the score) observed in the two documents would be around 16 times more likely observed if the two documents had been written by the same author than by different authors.
 - 7 See <https://rdr.io/github/davidavdav/ROC/>.