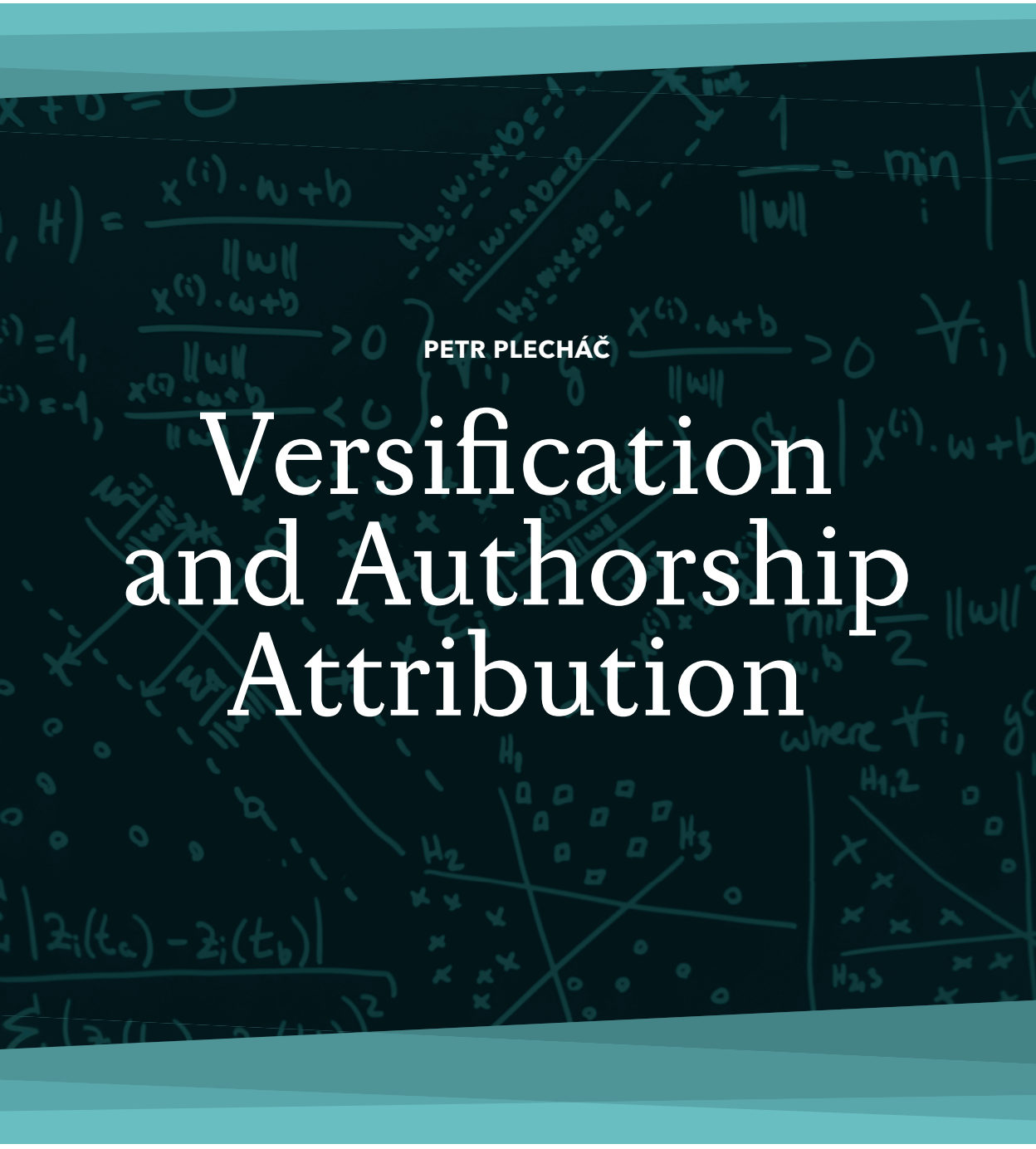
CZECH LITERATURE STUDIES



PETR PLECHÁČ

Versification and Authorship Attribution

INSTITUTE OF CZECH LITERATURE
KAROLINUM PRESS

This PDF includes a chapter from the following book:
Versification and Authorship Attribution

1 Quantitative Approaches to Authorship Attribution

Petr Plecháč
Institute of Czech Literature, Czech Academy of Sciences
e-mail: plechac@ucl.cas.cz

https://doi.org/10.14712/9788024648903.2

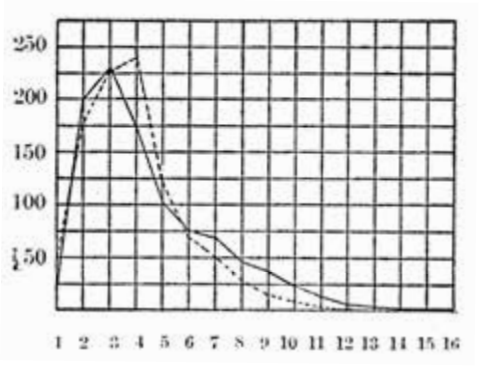
1 Quantitative Approaches to Authorship Attribution

1.1 Origins of Stylometry

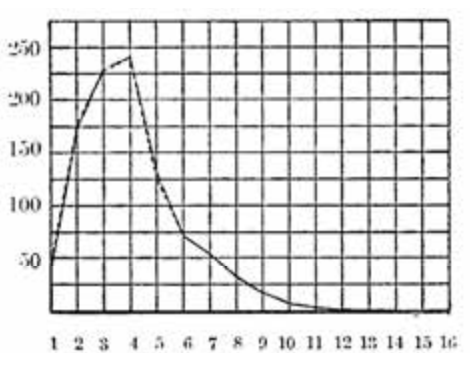
Many scholars (e.g. Holmes 1998; Juola 2006) trace the origins of stylometry to several passages in a letter written by the British mathematician Augustus De Morgan to Reverend W. Heald on August 18, 1851 (De Morgan 1851/1882). After considering how to distinguish the Pauline epistles actually written by St. Paul from those written by other author(s), De Morgan mused that the average word length measured by the number of characters might give some clue: "If St. Paul's epistles which begin with Παυλος gave 5.428 and the Hebrews gave 5.516, for instance, I should feel quite sure that the *Greek* of the Hebrews (passing no verdict on whether Paul wrote in Hebrew and another translated) was not from the pen of Paul" (De Morgan 1851/1882: 216; emphasis in the original). Later he complained: "If scholars knew the law of averages as well as mathematicians, it would be easy to raise a few hundred pounds to try this experiment on a grand scale" (De Morgan 1851/1882: 216).

In fact, it was not until the end of the 19th century that the American physicist Thomas Corwin Mendenhall raised the money for this experiment. In an initial article entitled "The Characteristic Curve of Composition" (1887), Mendenhall suggested ignoring averages and dealing with overall word length distribution instead. Eventually, thanks to the support of a benefactor, August Hemenway, he applied this method to a real-world case of disputed authorship. The results of that experiment were published in the article "A Mechanical Solution to a Literary Problem" (1901). There, Mendenhall compared the shape of a curve determined by the relative frequencies of words of different lengths in works ascribed to William Shakespeare with equivalent curves for works by Francis Bacon and Christopher Marlowe (FIG. 1.1). Based on the similarities and differences, he cautiously concluded that while Bacon had not written the works in question, there was strong evidence that Marlowe had (Mendenhall 1901: 104–105). The discrepancies between the curves for Shakespeare and Bacon were, however, later found to be due to the comparison of verse texts by the former with non-verse texts by the latter (see Williams 1975).

https://doi.org/10.14712/9788024648903.2



(a) Texts ascribed to Shakespeare (dashed line) and texts by Bacon (solid line).



(b) Texts ascribed to Shakespeare (dashed line) and texts by Marlowe (solid line almost covering dashed line).

FIG. 1.1: Relative frequencies (per thousand) of word lengths measured by number of characters; source: Mendenhall 1901: 104 (facsimile).

Independently of Mendenhall, the American mathematician William Benjamin Smith had also been employing quantitative methods in the 1880s. In his article "Curves of Pauline and Pseudo-Pauline Style", published under the pen name Conrad Mascol (1888a; 1888b), he, like De Morgan, considered the authorship of the Pauline epistles. In line with Mendenhall, he took the shape of the curves representing various textual features (e.g. the average number of words or prepositions per page) to be a criterion. On comparing the curves for epistles generally agreed to be written by St. Paul with those of doubtful authorship, Smith concluded that the author of the former had probably not written the latter. Significantly, he also stressed that the key consideration when selecting features should be their topic independence.[1] This principle, though now taken for granted, was not generally accepted until the mid-20th century, as we will see in Section 1.2.

A third pioneering work usually mentioned in this field is an article by Lucius Adelno Sherman (1888) that was probably also conceived independently of Mendenhall's studies.[2] It analysed the average sentence length measured by the number of

1 Smith wrote: "When we now ask, What are the elements of style to be considered? The answer must be: All such as are affected not at all, or apparently and comparatively very little, by the subject-matters of discourse" (Mascol 1888a: 456).

2 Grzybek (2014) notes, however, that Sherman may have been inspired by a response to Mendenhall's initial article that was published in an 1887 issue of *Science*. Its author observed: "There are other characteristics of writers equally susceptible of treatment by the statistical and graphical method, in

words in the work of novelists writing in English. Still Sherman did not highlight the possibility of using this metric for authorship recognition.

Outside of these studies, there is, however, another branch of stylometry which, although only sporadically recognised by scholars (Grzybek 2014 and Grieve 2005 rank among the exceptions), dates back some 100 years before Mendenhall's first article and more than 60 years before De Morgan's letter. This concerns the attributions of Shakespearean scholars based on the quantification of rhythm and rhyme.

One of the earliest examples of this approach can be found in a study by Edmond Malone (1787/1803) which proposed that none of the three parts of the play *Henry VI* had actually been written by Shakespeare. Malone's arguments were based, among other things, on attention to versification: he argued that there were far fewer rhymes and enjambments in the texts in question than in other works by Shakespeare.

Another instance can be seen in a comment by the scholar Henry Weber about the play *The Two Noble Kinsmen* (1812), which was first published in 1634 as a collaborative work by William Shakespeare and John Fletcher (see Section 4.1 for details). Weber worked out a scene-by-scene division of authorship between Shakespeare and Fletcher based on the frequencies of certain line endings among other factors:

> Taking an equal number of lines in the different parts which are attributed to Shakespeare and to Fletcher, the number of female, or double terminations in the former, is less than one to four; on the contrary, in the scenes attributed to Fletcher the number of double or triple terminations is nearly three times that of single ones. (Weber 1812: 166)

Decades later, James Spedding (1850) used the same metric to arrive at a theory of joint authorship by Shakespeare and Fletcher that he also applied to *Henry VIII*.

The real rise of versification-oriented stylometry did not come, however, until the 1870s and 1880s after the founding of the *New Shakspere Society*.[3] In the first volume of their *Transactions,* one Society member, John Kells Ingram (1874) suggested dividing unstressed blank verse endings into "light endings" and "weak endings"[4] and using

---

which their personal peculiarities differ more widely, and which are therefore more characteristic than the habitual selection and use of long or short words. For example: it seems to me that the length of the sentence is such a peculiarity" (Eddy 1887: 297).

3 Concerning its name, the Society's members maintained: "This spelling of our great Poet's name is taken from the only unquestionably genuine signatures of his that we possess, the three on his will, and the two on his Blackfriars conveyance and mortgage." (Furnivall 1874a: 6).

4 Ingram described these two forms as follows: "It is evident that amongst what have been called as a class weak endings, there are different degrees of weakness. […] There are *two* such degrees, which require to be discriminated, because on the words, which belong to one of these groups the voice can

the ratio of instances of the two to support Spedding's attribution of *Henry VIII*. Ingram himself called this method the "weak-ending test". Other members proposed (or adopted) and applied several other such verse tests designed to distinguish Shakespeare's works from those of other authors based on the prevalence of particular features. These included the "rhyme test" (for rhymed lines), the "stopt-line test" (for enjambment), the "middle-syllable test" (for extra-metrical syllables at the end of the first half-line) and the "caesura test" (for word breaks after the sixth syllable in alexandrines).[5]

Many of these attributions by New Shakspere Society members were later proven wrong owing to the simplistic nature of their methods or errors in their source data (Grieve 2005: 6). Even so, they are an important part of the history of stylometry and should not be neglected.

1.2 Searching for the "Golden Feature"

The works of George Kingsley Zipf seem to have inspired a new era in the development of 20th-century stylometry (see Koppel, Schler and Argamon 2009: 4–5). The formulation of Zipf's law (1932), which states that all natural language texts follow the same rank-frequency word distribution, likely encouraged scholars to rethink the possibilities for authorship attribution. This meant finding a similar textual feature that would remain stable across the works of one author while differing in those of other authors.

Of great influence in this period were the stylometric works of George Udny Yule, who initially proposed using sentence length measured by the number of words (Yule 1939). Unlike Sherman (see Section 1.1), Yule considered not only average values but also other distribution characteristics. These included the median, the $Q_{0,25}$ and $Q_{0,75}$ quartiles, the interquartile range and also—since sentence length generally tends to follow a positively skewed log-normal distribution—the decile $Q_{0,9}$.

Just a couple of years later, Yule's book *The Statistical Study of Literary Vocabulary* (1944) introduced a new metric designed to capture vocabulary richness. He defined that measure as follows:

---

to a certain small extent dwell, whilst the others are so essentially *proclitic* in their character […] that we are forced to run them, in pronunciation no less than in a sense, into the closest connection with the opening words of the succeeding line. The former may with convenience be called 'light endings', whilst to the latter may be appropriated the name (hitherto vaguely given to both groups jointly) of 'weak endings'" (Ingram 1874: 447; emphasis in original).

5 See Fleay 1874a, 1874b, 1874c, 1874d; Furnivall 1874b, 1874c.

$$K = \frac{10^4 \left[\left(\sum_{m=1}^{m_{\max}} m^2 V_m \right) - N \right]}{N^2} \tag{1.1}$$

where N is the text length measured by the number of tokens and V_m is the number of word types with a frequency of m.

Importantly, Yule did not take into account the entire vocabulary when he applied his metric to real attribution tasks. Instead, he confined his analysis to nouns alone. He explained this choice as follows:

> My object in limiting myself to nouns for the investigation into the vocabularies of Thomas à Kempis and Gerson was in part simply the limitation of material and the exclusion of words of little or no significance as regards style, such as prepositions, pronouns, etc. Of the three principal parts of speech, nouns, adjectives and verbs, I thought nouns would probably be the most significant or characteristic. (Yule 1944: 21).

In fact, it was fairly common for mid-20th-century scholars to assume that high-frequency function words had no authorial signal and, thus, could not contribute to authorship recognition (see Grieve 2005: 32–34). This assumption was wrong, however, as we will see in Section 1.3.

Many other simple features were proposed for authorship attribution purposes in this period. They included average word length measured by the number of syllables (Fucks 1952) and the frequency of loan words (Herdan 1956). None of them, however, turned out to be sufficiently robust, and when they were applied to attribution tasks other than those they were designed for, they usually failed (see Hoover 2003; Grieve 2005).

1.3 Multivariate Analyses

The most important contribution to 20th-century stylometry came from a publication by Frederick Mosteller and David L. Wallace (1964). In a groundbreaking study of the authorship of *The Federalist Papers*, the two revived a principle introduced by W. B. Smith (see Section 1.1) that remains widely accepted today. This held that as far as possible, the features used for authorship recognition should be topic independent. Rejecting the content-based word tests that dominated studies by their contemporaries, these scholars, thus, turned their gaze to the most common function

words and the frequencies of their variations (e.g. *while/whilst*). Crucially, their analysis was based not on the usual comparison of isolated values but rather on one of entire sets. This is where the turn from simple univariate methods to more sophisticated multidimensional analyses began. By the 1980s, it had led to the application of such statistical methods as multivariate variance analysis (Larsen, Rencher and Layton 1980) and principal component analysis (Burrows and Hassal 1988; Burrows 1989). Of all of these methods, however, the so-called Burrows' Delta would prove the most popular.

1.3.1 Burrows' Delta

The Delta was proposed by John F. Burrows (2002, 2003) as a simple measure of stylistic similarities between two texts. This metric was primarily designed to resolve cases where there was a text of unknown or doubtful authorship (target text: t_0) and a corpus of works produced by a finite set of candidate authors (candidate set: $T = \{t_1, t_2, t_3, \ldots, t_m\}$). The goal was to find the candidate whose texts showed the greatest similarity to the target text, i.e. the one whose texts had the lowest Delta value.

Like Mosteller and Wallace's analysis, Burrows' Delta relied on a set of high-frequency words. The most straightforward approach to such data would have been to plot their frequencies in both the target and the candidate texts and then compare the resulting curves just as Mendenhall had (Section 1.1). Such visual assessments tend, however, to be vague and unreliable. Instead, Burrows suggested an alternative: the discrepancy between the texts could be expressed as the mean value of the differences between the frequencies of specific words. This method was set out as follows:

(1) From the entire body of work (i.e. $t_0 \cup T$), select the n most common words w_1, w_2, w_3, …, w_n.

(2) Each text $t_i \in \{t_0, t_1, t_2, \ldots, t_m\}$ is represented as a vector $\mathbf{f}_i = (f(t_{i,1}), f(t_{i,2}), \ldots, f(t_{i,n}))$, where $f(t_{i,j})$ denotes the relative frequency of w_j in t_i.

(3) Word frequency tends to decrease sharply after the uppermost entries (Zipf's law). The differences in the prevalence of the most common words will, thus, generally be much larger than those between, say, the 50th and 100th most common words in any given body of texts. To make each word a marker of equal weight, the frequencies of individual words are transformed into *z*-scores: $\mathbf{t}_i = (z(t_{i,1}), z(t_{i,2}), \ldots, z(t_{i,n}))$.

$$z\left(t_{i,j}\right) = \frac{f\left(t_{i,j}\right) - \mu_j}{\sigma_j} \tag{1.2}$$

$$\mu_j = \frac{1}{m}\sum_{i=0}^{m} f\left(t_{i,j}\right) \tag{1.3}$$

$$\sigma_j = \sqrt{\frac{1}{m}\sum_{i=0}^{m}\left(f\left(t_{i,j}\right) - \mu_j\right)^2} \tag{1.4}$$

The *z*-score transforms the frequency distribution for each word across the corpus to give it a mean of 0 and a standard deviation of 1. (In very rough terms, this transformation contracts or extends the frequency ranges so that they are approximately the same for each word.)

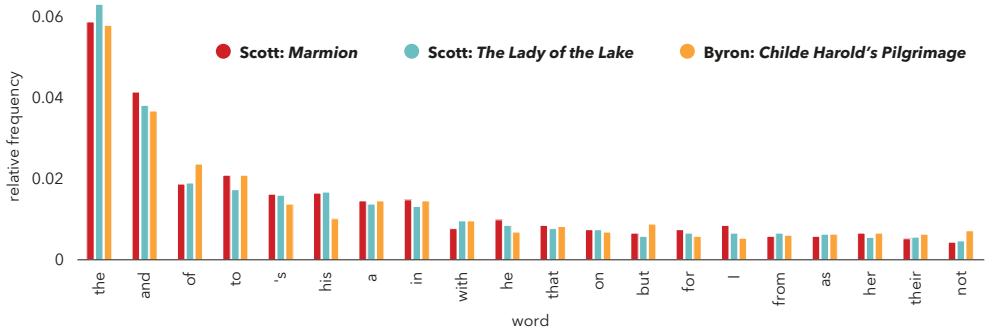
(4) The stylistic dissimilarity (Δ) between texts t_a and t_b is finally calculated as the arithmetic mean of the absolute values of the differences between the *z*-scores for individual words:

$$\Delta\left(t_a, t_b\right) = \frac{\sum_{j=1}^{n}\left|z\left(t_{a,j}\right) - z\left(t_{b,j}\right)\right|}{n} \tag{1.5}$$

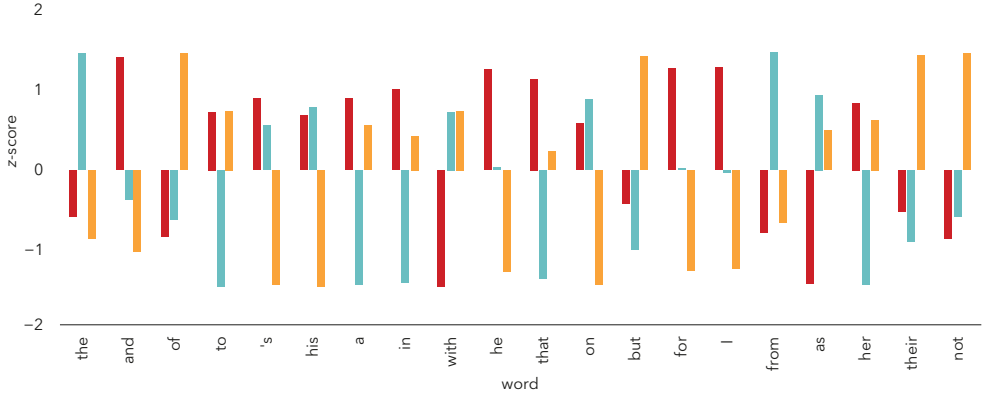
(5) The candidate whose text $t_a \in T$ yields the lowest value $\Delta(t_a, t_0)$ is considered the most likely author of the target text.

To illustrate this approach, we may consider a model situation where Walter Scott's *The Lady of the Lake* is the target text and the candidate set consists of *Marmion* by the same author and *Childe Harold's Pilgrimage* by George Gordon Byron. FIG. 1.2a shows the relative frequencies of the 20 most common words in these three poetic works. FIG. 1.2b presents the data transformed into *z*-scores. FIG. 1.2c gives the absolute values of the differences between the *z*-scores for works in the candidate set and works in the target text. The last two columns highlight the mean values (Δ).

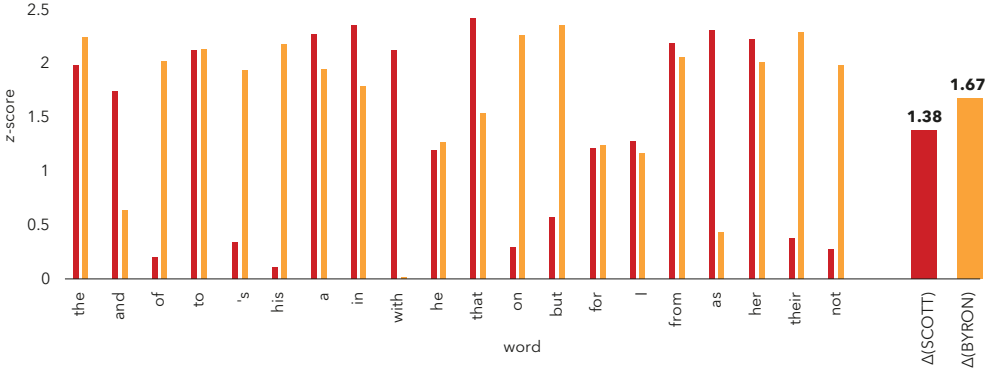
Thanks to the simple, intuitive and fairly accurate nature of the Delta measure, it was embraced soon after it was presented and became a popular authorship attribution method. Several modifications have since been proposed (e.g. Hoover 2004a, 2004b). From a contemporary perspective, however, the most important advance was arguably Shlomo Argamon's interpretation of the Delta's key principle.



(a) Relative frequencies of the 20 most common words in each text.



(b) Relative frequencies of the 20 most common words transformed into *z*-scores.



(c) Absolute values of the differences between the *z*-scores for each candidate and the target text; the final two columns show the mean values.

FIG. 1.2: Burrows' Delta for Walter Scott's *The Lady of the Lake* (target text), Walter Scott's *Marmion* and George Gordon Byron's *Childe Harold's Pilgrimage* (candidate set).

1.3.2 The Geometric Interpretation of Burrows' Delta and Its Modifications

Argamon (2008) pointed out that the Delta measure that Burrows had stumbled on by intuition was actually the equivalent of measuring the *Manhattan distance* between two vectors. As such, the entire method could be seen as an instance of nearest neighbour classification or a special case of the popular *k*-nearest neighbour classifier where $k = 1$.

Argamon proceeded from a simple consideration: Since the process was based on candidate ranking, there was no need to divide the sum of differences by the number of analysed words (n). After all, division by a constant would not affect the ranking. Once the denominator was dropped from formula 1.5, we obtain a simple summary of the absolute values of the *z*-score differences, i.e. the Manhattan distance (D_{M}; see FIG. 1.3):

$$\Delta\left(t_a, t_b\right) \propto D_{\mathrm{M}}\left(\mathrm{t}_a, \mathrm{t}_b\right) = \sum_{j=1}^{n} \left| z\left(t_{a,j}\right) - z\left(t_{b,j}\right) \right| \tag{1.6}$$

In the same article, Argamon also suggested a modification of Burrows' original method, or what he called the quadratic Delta (Δ_{Q}) based on the Euclidean distance (D_{E}) between the given vectors:

$$D_{\mathrm{E}}\left(\mathbf{t}_a, \mathbf{t}_b\right) = \sqrt{\sum_{i=1}^{n} \left(z\left(t_{a,i}\right) - z\left(t_{b,i}\right) \right)^2} \tag{1.7}$$

Just as dividing each distance by a constant did not affect the final ranking in Burrows' Delta, the same was true for extracting the root in the formula for the Euclidean distance (square root is a monotonically increasing function). The formula for Δ_{Q} was, thus, defined as the square of the Euclidean distance:

$$\Delta_{\mathrm{Q}}\left(t_a, t_b\right) = \sum_{j=1}^{n} \left(z\left(t_{a,j}\right) - z\left(t_{b,j}\right) \right)^2 \tag{1.8}$$

The cosine Delta ($\Delta_{\angle}$; Smith and Aldridge 2011) is another recent popular modification of Burrows' Delta. It is based on the cosine similarity of vectors, that is, the cosine of the angle θ between them:

$$\cos(\theta) = \frac{\sum_{j=1}^{n} z\left(t_{a,j}\right) z\left(t_{b,j}\right)}{\| \mathbf{t}_a \| \; \| \mathbf{t}_b \|} \tag{1.9}$$

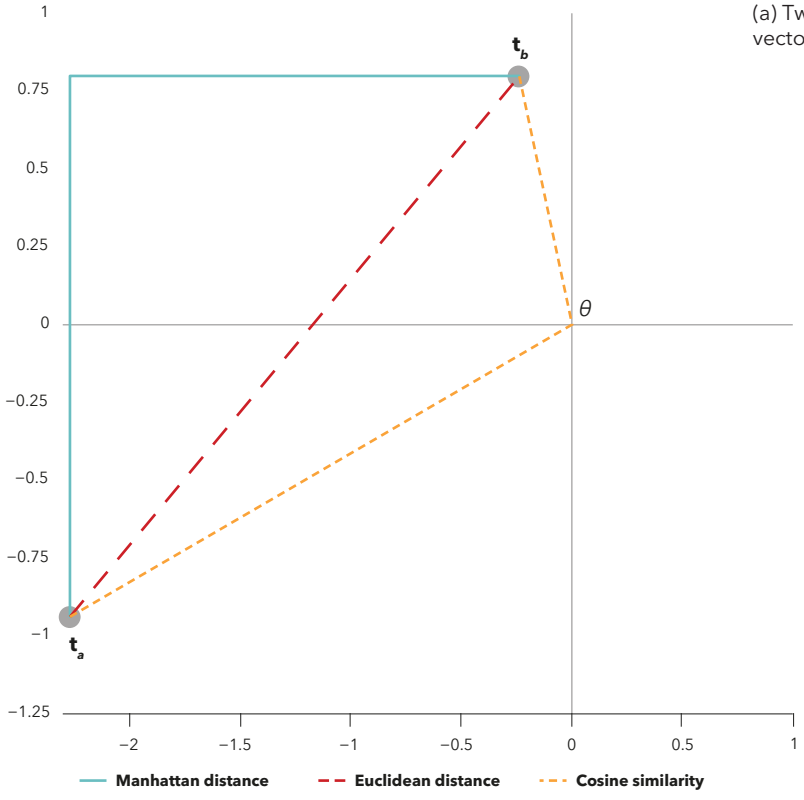


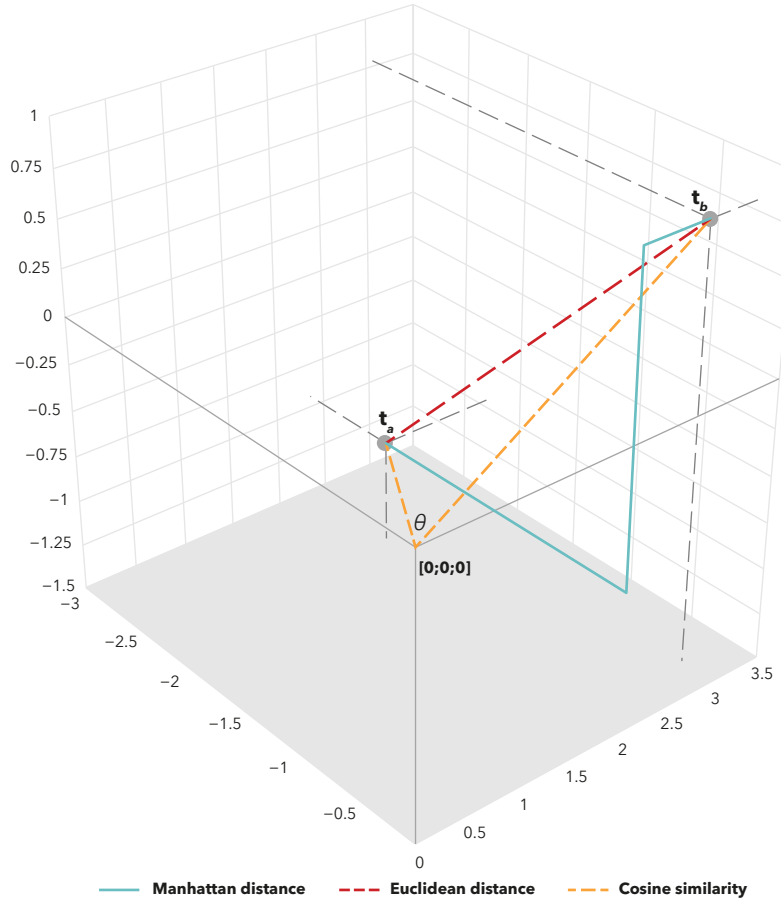
FIG. 1.3: Manhattan distance, Euclidean distance and cosine similarity of vectors $\mathbf{t}_a$ and $\mathbf{t}_b$.

Since $\cos(\theta) \in [-1, 1]$, the formula is modified so that—as with Burrows' Delta and the quadratic Delta—the greater the similarity between two texts, the lower the cosine Delta value and *vice versa*:

$$\Delta_{\angle}(t_a, t_b) = 1 - \cos(\theta) \tag{1.10}$$

Metrics from the Delta family have been tested across languages and text types with various settings for the number of the most common units (n) and with other features such as lemmata, character n-grams and word n-grams (see, e.g. Eder 2011; Rybicki and Eder 2011; Jannidis et al. 2015).

(b) Three-dimensional vector space.



1.4 Support-Vector Machines

Outside of the Delta, more sophisticated machine learning methods have gained increasing attention over the last decade or two. These include *random forest* (e.g. Tabata 2012), *naïve Bayes classifier* (e.g. Zhao and Zobel 2005) and above all *support-vector machine* (SVM) techniques (e.g. Diederich et al. 2003; Koppel and Schler 2004). The SVM technique is still probably the most popular in contemporary stylometry although deep-learning methods seem poised to overtake it (see, e.g. Savoy 2020). This section outlines the general principles behind SVM.

An SVM is a supervised learning technique, which means that its algorithm uses labelled training data to infer a classification function for new data. This key principle can be illustrated with a very simple example based on artificial data. Imagine

a target text t_0 and 20 samples from each of two candidates (author 1, author 2). All of the texts are represented by *z*-scores for the two most common words ("the" and "and").

During the first (learning) phase, the SVM is fed data from author 1 and author 2 (training data). These data are labelled according to author, and the SVM tries to find a function that correctly separates them by their labels. This is done using a hyperplane—a subspace with one dimension fewer than the original vector space. In our example with its two-dimensional data, this means a one-dimensional space, i.e. a line. During the second phase (classification), the hyperplane inferred from the training data is used to classify the target text.

FIG. 1.4a shows that if the data are linearly separable, then an infinite number of potential hyperplanes can separate them correctly. Some of these may attribute the target text to author 1 while others may attribute it to author 2. From all these possibilities, the SVM chooses the hyperplane that maximises the distance to the nearest vectors on each side (also known as the support vectors), as shown in FIG. 1.4b (this is the maximum-margin hyperplane). In this case, the SVM classifies the target text as the work of author 1.

Generally, for *n*-dimensional data, the task is formulated as follows: We are given the training data $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, …, $(\mathbf{x}_m, y_m)$ where the first member of each pair denotes the *n*-dimensional vector $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$ and the second member denotes one of two classes to which the vector belongs: $y_i \in \{-1, 1\}$. The goal is to find a normal vector $\mathbf{w}$ and a parameter b to define a hyperplane H

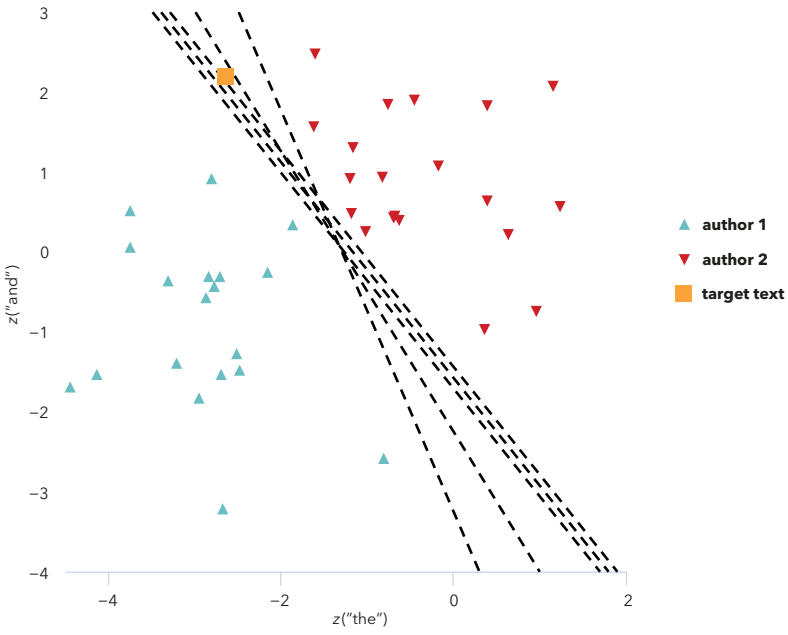
$$H : \mathbf{w} \cdot \mathbf{x} + b = 0 \tag{1.11}$$

that separates the vector space into two half-spaces so that each half-space contains only data of the same class and the distance to the nearest vector is maximised.

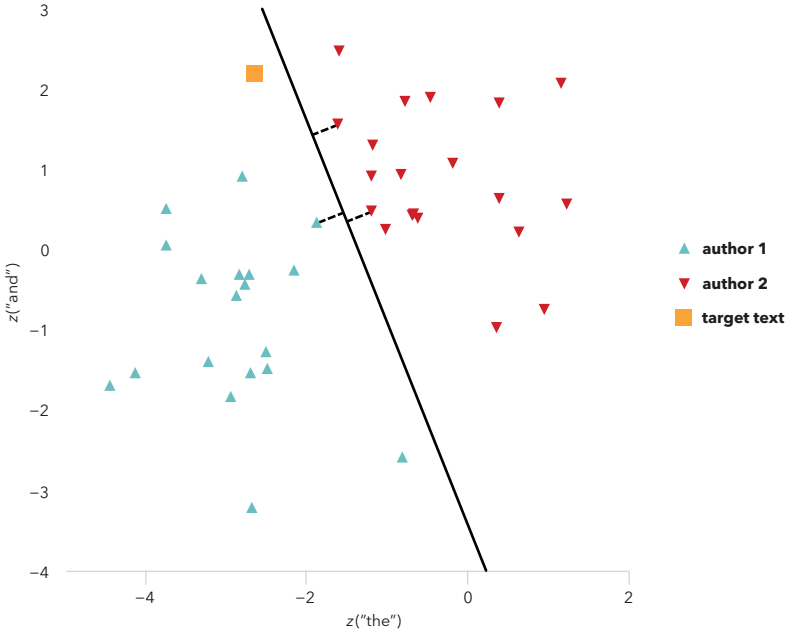
These requirements may be defined formally using the oriented distance d of the vectors $\mathbf{x}_i$ to hyperplane *H.* This will be positive for vectors in one half-space and negative for vectors in the other one:

$$d(\mathbf{x}_i, H) = \frac{\mathbf{x}_i \cdot \mathbf{w} + b}{\| \mathbf{w} \|} \tag{1.12}$$

As we have two classes $y_i \in \{-1, 1\}$, the requirement that each half-space contain vectors belonging to the same class may be formulated as:



(a) Various possible hyperplanes separating training data from author 1 and author 2.



(b) Maximum-margin hyperplane; dashed lines indicate distances to support vectors.

FIG. 1.4: A Support-Vector Machine (artificial data).

$$\begin{aligned} \forall i : y_i = 1, \quad & \frac{\mathbf{x}_i \cdot \mathbf{w} + b}{\| \mathbf{w} \|} > 0 \\ \forall i : y_i = -1, \quad & \frac{\mathbf{x}_i \cdot \mathbf{w} + b}{\| \mathbf{w} \|} < 0 \end{aligned} \tag{1.13}$$

This may be simplified as:

$$\forall i, \quad y_i \frac{\mathbf{x}_i \cdot \mathbf{w} + b}{\| \mathbf{w} \|} > 0 \tag{1.14}$$

Next, we require the maximum possible margin. We therefore try to maximise the Euclidean (non-oriented) distance of the nearest (support) vectors to hyperplane *H*. All these requirements may be expressed as:

$$\begin{aligned} & \max_{\mathbf{w},b} \min_i \left| \frac{\mathbf{x}_i \cdot \mathbf{w} + b}{\| \mathbf{w} \|} \right| \\ & \text{where } \forall i, \quad y_i \frac{\mathbf{x}_i \cdot \mathbf{w} + b}{\| \mathbf{w} \|} > 0 \end{aligned} \tag{1.15}$$

The number of solutions to this task remains infinite, however, since the direction of vector **w** is specified but its magnitude ||**w**|| is not. For practical reasons, the magnitude ||**w**|| should be inversely proportional to the Euclidean distance of the support vectors to hyperplane *H*:

$$\frac{1}{\| w \|} = \min_i \left| \frac{\mathbf{x}_i \cdot \mathbf{w} + b}{\| \mathbf{w} \|} \right| \tag{1.16}$$

This allows for the simplification of the support vector requirement as follows:

$$\left| \mathbf{x}_i \cdot \mathbf{w} + b \right| = 1 \tag{1.17}$$

For all of the vectors, the requirement is therefore:

$$\forall i, \quad y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) \geq 1 \tag{1.18}$$

This brings us to a basic statement of the optimisation problem for an SVM: If we are looking for a normal vector **w** and a parameter *b* to define the hyperplane *H* with

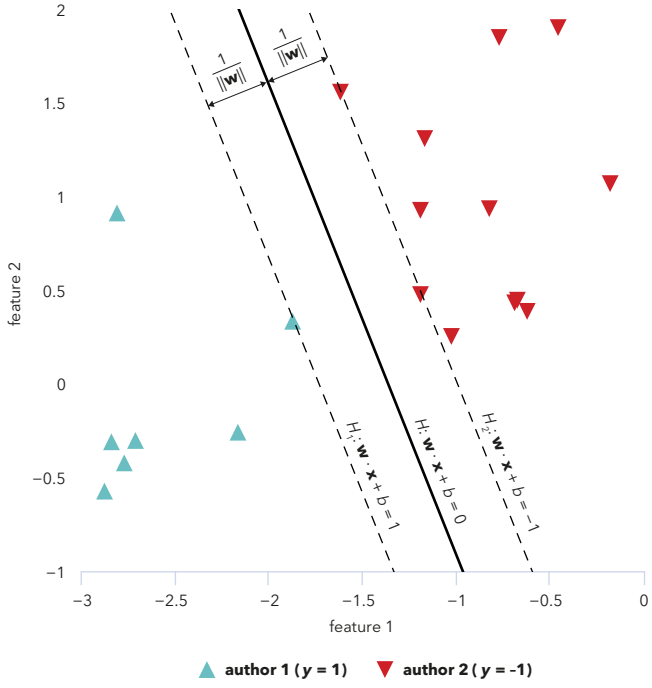


FIG. 1.5: A Support-Vector Machine.

the maximum possible margin, and if the width of that margin should be inversely proportional to the magnitude $\|\mathbf{w}\|$ (formula 1.16), then the solution is the minimal possible normal vector $\mathbf{w}$ which satisfies inequation 1.18 (see FIG. 1.5).

Again for practical reasons, it is not the magnitude $\|\mathbf{w}\|$ that we minimise but rather its square divided by two:

$$\min_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|^2$$
$$\text{where } \forall i, \quad y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) \geq 1 \tag{1.19}$$

This task is then solved by Lagrange multipliers (see, e.g. Abney 2007: 117–119).

The example above is the simplest instance of the classification of n-dimensional data. In practice, however, we are often faced with more complex issues. Those challenges include (1) linearly inseparable data and (2) the need for classification into more than two classes.

1.4.1 Linearly Inseparable Data

If there is no hyperplane that would correctly separate the classes, one of two approaches is usually employed: (1) the hyperplane condition is relaxed (the soft-margin SVM) or (2) we perform kernel transformation of the data into higher dimensions. Below I consider each of these techniques:

(1) *A soft-margin SVM* tends to be used with data with a fairly low noise level. This method relaxes the condition that each half-space must only contain vectors of the same class. Instead, a slack variable ξ is introduced to penalise vectors on the "wrong" side of the hyperplane. Here the goal is to find the hyperplane with the maximum margin and minimum "overlap" of vectors into the half-space of a different class.

For a vector $\mathbf{x}_i$ occurring in the half-space of a different class, ξ_i denotes the Euclidean distance $\mathbf{x}_i$ measured from the side of the margin defined by support vectors of its own class (H_{y_i}) and normalised by the margin width (see FIG. 1.6).

For these vectors, thus:

$$\xi_i = \frac{\left| \frac{\mathbf{w} \cdot \mathbf{x}_i + b - y_i}{\|\mathbf{w}\|} \right|}{\frac{1}{\|\mathbf{w}\|}} \qquad \xi_i = \left| \mathbf{w} \cdot \mathbf{x}_i + b - y_i \right| \tag{1.20}$$

For other vectors $\xi_i = 0$.

The optimisation problem (formula 1.19) is therefore extended to:

$$\min_{\mathbf{w}, b, \xi_i} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i$$
$$\text{where } \forall i, \quad y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) \geq 1 - \xi_i \quad \text{and} \xi_i \geq 0 \tag{1.21}$$

where C is the penalty parameter of the model. This determines how much it will penalise misclassifications.

(2) In *kernel transformation*, noisy linearly inseparable *n*-dimensional data are transformed into an (*n*+*k*)-dimensional space. In this way, they eventually become linearly separable (the "kernel trick"). As an example, we may consider the

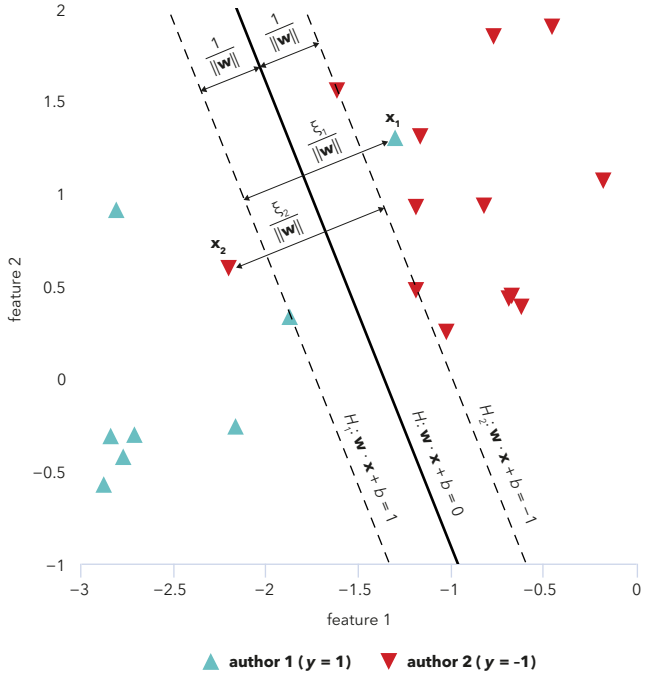


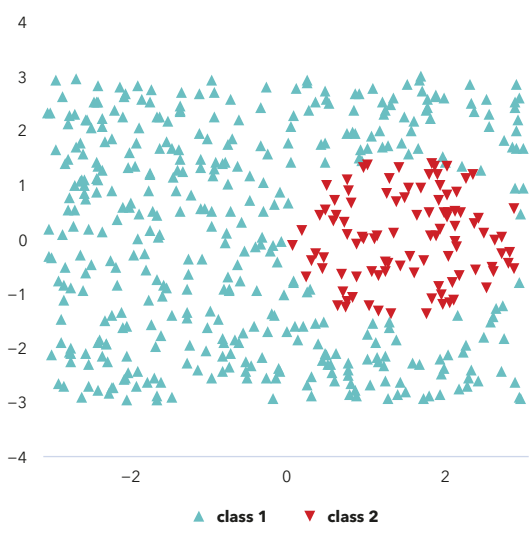
FIG. 1.6: A soft-margin SVM.

transformation of two-dimensional data (FIG. 1.7a) into a three-dimensional space (FIG. 1.7b) where each original vector $\mathbf{x} = (x_1, x_2)$ is converted into $\mathbf{x}' = (x_1, x_2, x_1^2 + x_2^2)$. Since linguistic data tend, however, to include quite a few instances per class and a very high number of dimensions, kernel transformation is not usually required.

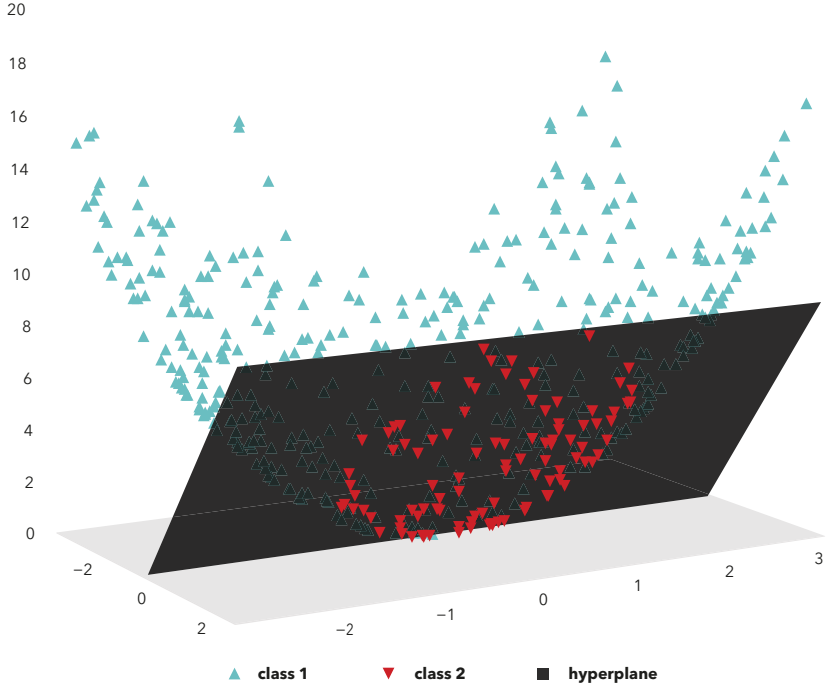
1.4.2 Multiclass Classification

As we have seen, an SVM is inherently a binary classifier. The most common way to perform multiclass classification is therefore to split the problem into multiple binary tasks. There are two ways that this can be done: the *one-vs.-rest* strategy and the *one-vs.-one* strategy.

(1) In the *one-vs.-rest* strategy, a classification function is constructed for each class in order to separate its data from the rest of the data (*k* classes, thus, produce *k* classification functions, i.e. *k* hyperplanes). If only one out of all of the *k* classification



(a) Original data.



(b) Transformed data.

FIG. 1.7: Kernel transformation of linearly non-separable two-dimensional data. Transformation function: $\Phi\left(x_1, x_2\right) = \left(x_1, x_2, x_1^2 + x_2^2\right)$.

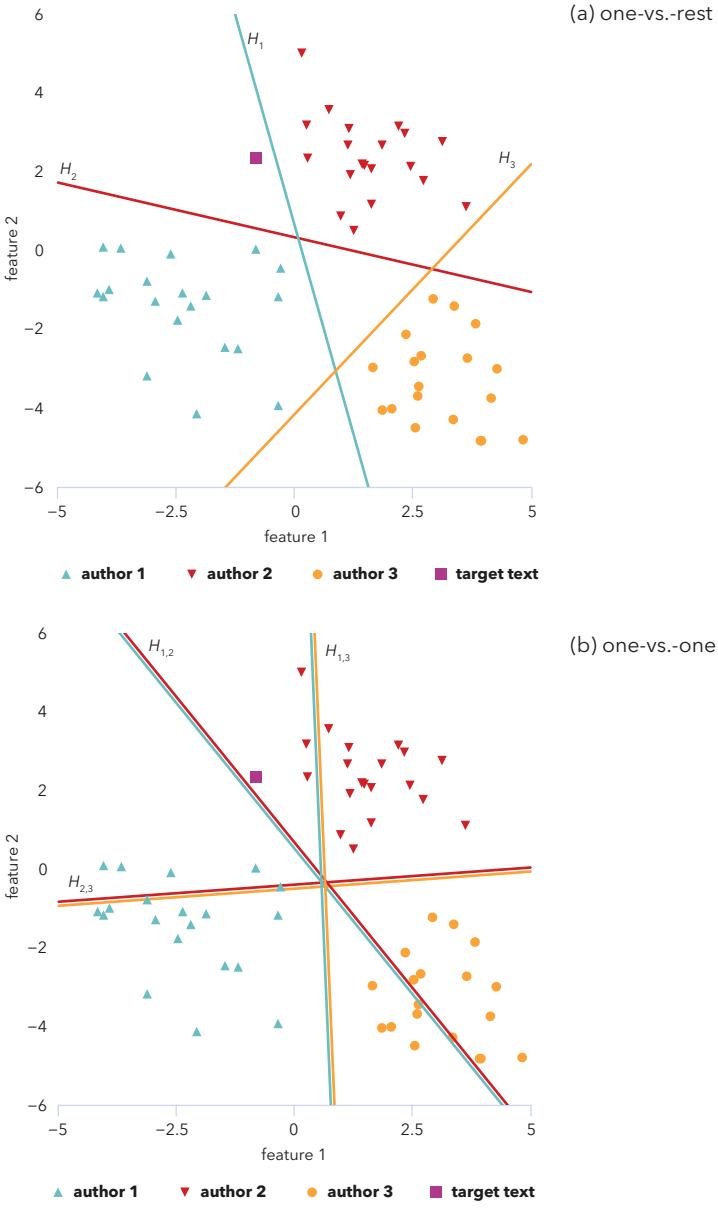


FIG. 1.8: Multiclass classification with an SVM using (a) the one-vs.-rest strategy and (b) the one-vs.-one strategy. In both cases, the target text is attributed to author 2. In case (a), hyperplane H_1 also classifies the text as author 1, but the distance to H_2 is greater. In case (b), the target is classified as author 2 by two hyperplanes ($H_{1,2}$; $H_{2,3}$) and only classified as author 1 by one hyperplane ($H_{1,3}$).

functions attributes the target text to a particular author and all the other functions ascribe it to the "rest" group, the target is simply classified as the work of that author. If many classification functions assign the target to a particular author, a decision is made based on which hyperplane is farther from the target vector. In the example given in FIG. 1.8a, the target text is, thus, attributed to author 2.

(2) In the *one-vs.-one* strategy, we construct a classification function for each pair of classes (k classes, thus, produce $\frac{k(k-1)}{2}$ classification functions). Each of these functions attributes the target text to a single author. The final verdict reflects the author selected by the most classifiers.

1.4.3 The Normal Vector as an Indicator of Feature Importance

A hyperplane constructed with an SVM has one particularly useful property: the coordinates of its normal vector can reveal the importance of particular features for the classification.

For simplicity's sake, we will remain in the two-dimensional vector space with its hyperplane (i.e. line) defined by the general equation $w_1x + w_2y + b = 0$. The normal vector $\mathbf{w} = (w_1, w_2)$ defines the slope of the line while parameter b is its vertical shift. And this slope also indicates the importance of each feature for the classification.

We can illustrate this with a real-world example. Consider a simple device placed deep in a forest that measures the shoulder height and speed of any animal passing by. Since we know that wolves and moose are the forest's only inhabitants, we want to train the device to tell them apart. Intuitively we might guess that height is a good discriminator (wolves are generally much smaller than moose) while speed is not as informative. Not only does speed vary greatly (an animal may be ambling along or running for its life), but the maximum speeds of wolves and moose also happen to be more or less the same (55 to 60 kilometres per hour). As FIG. 1.9 shows, using labelled training data for 50 wolves and 50 moose, we can distinguish reliably between the animals. As expected, the classification is done solely by height; speed is distributed more or less equally across the two animal populations, as can be seen in the histogram on the top of the chart. It is therefore completely useless as an indicator. This is also captured in the hyperplane's position parallel to the x-axis ($w_1 = 0$). In other words, we would achieve the very same level of precision if our data were one-dimensional (based on height only) and the animals were simply classified based on whether they were taller or shorter than 118.5 cm (midway between the height of the tallest wolf and that of the shortest moose).

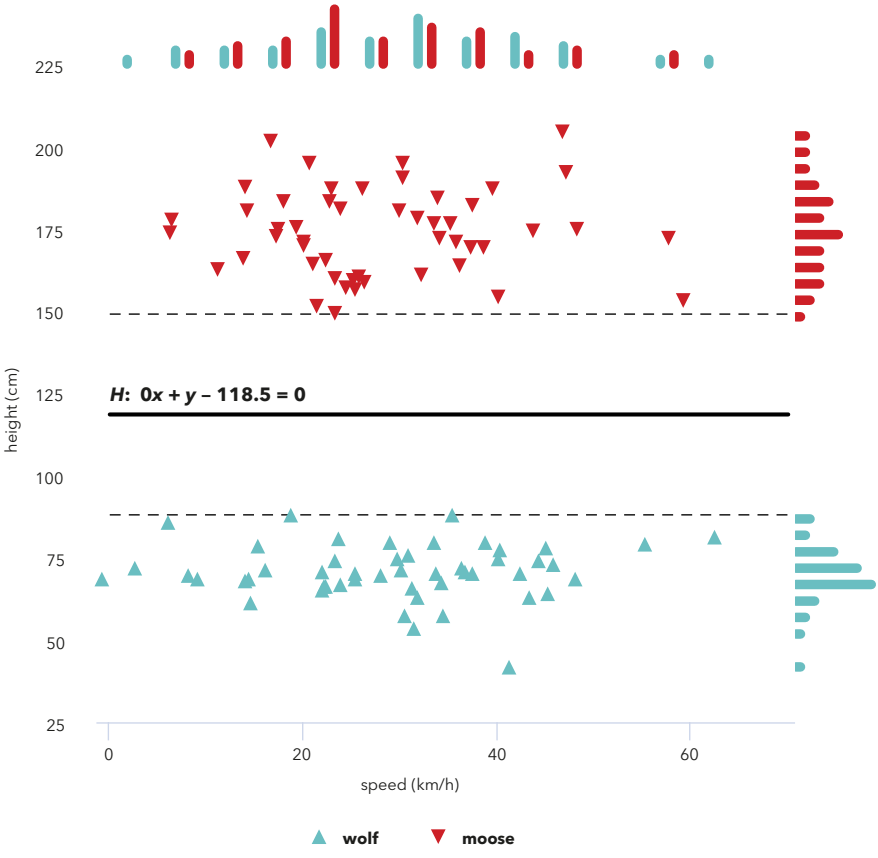
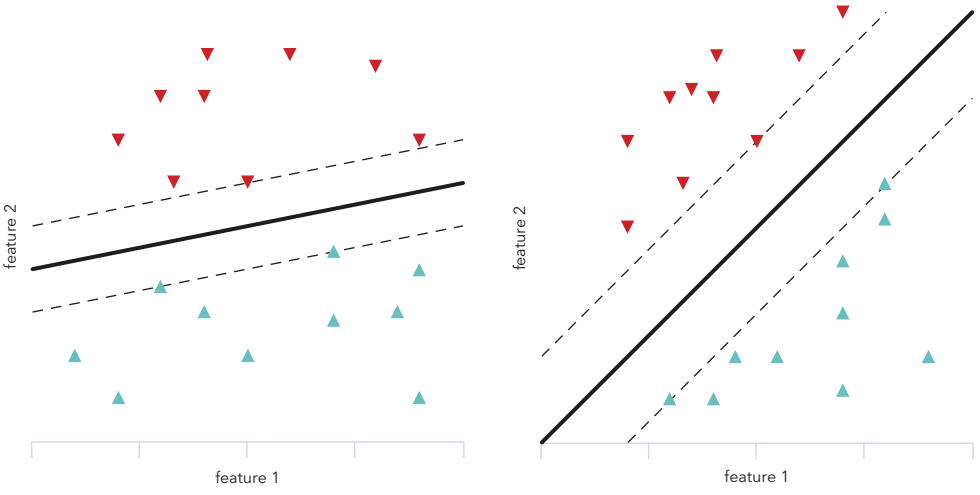


FIG. 1.9: Speed and height of wolves and moose. Artificial data.

Generally speaking, the greater the importance of a feature on the x-axis, the steeper the gradient of the hyperplane, and thus, the greater the $w_1 : w_2$ ratio. In FIG. 1.10a, we can see that feature 1 (x-axis) contributes somehow to the classification but its role is far less important than that of feature 2 (y-axis), i.e. $w_2 > w_1 > 0$. In FIG. 1.10b, both features contribute equally ($w_1 = w_2$). FIG. 1.10c captures the opposite situation to the one in FIG. 1.9: feature 2 has no importance and the classification is done entirely based on feature 1 ($w_2 = 0$).

We can use the same approach to interpret normal vectors of the hyperplane in spaces with more than two dimensions. This, however, only holds true for linear SVM. After kernel transformation (Section 1.4.1(2)), the relationship between a normal vector and particular features can no longer reasonably be interpreted.



(a) Feature 2 is more important than feature 1 ($w_1 < w_2$).

(b) Both features are equally important ($w_1 = w_2$).

FIG. 1.10: Feature importance. Normal vector of hyperplane: $\mathbf{w} = (w_1, w_2)$.

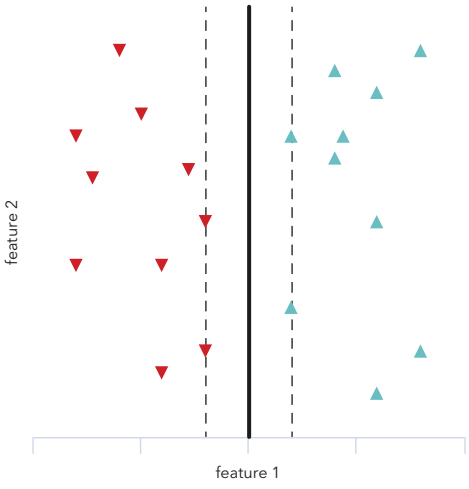
1.4.4 Validation

One crucial aspect of any machine learning model is its accuracy. There are several ways that accuracy can be estimated.

In the *holdout method*, we split the data into training and test sets. This split is usually done at random and at a ratio of 2:1. The training set is then used to train the model that will classify data from the test set. The share of correctly classified samples provides a general accuracy estimation.

In contrast, *k-fold cross-validation* can produce a better picture by dividing the data into k groups of equal size. Under this approach, one group is treated as the test set while the remaining $k-1$ groups are the training set. This is repeated for each group, which leads to k accuracy estimations. These results are then averaged to produce a single estimation.

When the data contain only a few samples from each class—a fairly common situation with linguistic data—*leave-one-out cross-validation* is the preferred method. In this case, the data consisting of n samples are split into $k = n$ groups. For each iteration, the model is tested on a single sample. The portion of correct classifications is used to estimate accuracy.



(c) Feature 2 has no importance; the classification is done solely based on feature 1 ($w_2 = 0$).

On its own, however, this accuracy estimation has only limited relevance. For a classifier to be useful, its accuracy must exceed the threshold (baseline) that could be reached by sheer guesswork. If, for example, a binary classifier has a 90%-accuracy rate for data where 90% of the samples belong to one class, it will hardly be useful in practice. A trivial classifier that always chose the most common class would achieve the very same level of accuracy. Outside of circumstances where this majority class baseline is most suitable (i.e. imbalanced datasets), the *random baseline* (RB) can help us determine the accuracy threshold. This tells us the most likely accuracy of a classifier that predicts the class at random:

$$\mathrm{RB} = \sum_{a=1}^{N} \left(\frac{n_a}{X} \right)^2 \tag{1.22}$$

where N denotes the number of classes, X is the number of samples and n_a is the number of samples in class a.

* * *

Bringing together all of these observations, we may sum up the main benefits and drawbacks of SVM compared to Delta measures as follows:

— While SVM models give different weight to each feature (see Section 1.4.3), in Delta metrics, all these features contribute equally to the classification. An SVM is, thus, theoretically more resistant to data noise. A good illustration can be seen in FIG. 1.10c where the SVM recognises that feature 2 is irrelevant to the classification. In contrast, Delta metrics would weigh both features equally. As such, Δ and Δ_Q would misclassify the lower support vector of the class on the right of the chart since its nearest neighbour is the other class's support vector.
— On the other hand, the SVM approach requires quite a large number of samples to carry out training. If only limited samples are available for some (or all) of the candidate authors, then we may still solve the task by using the less robust Delta measures.

1.5 Versification-Based Attribution

In the previous sections, we saw that stylometry employs a wide variety of both techniques and textual features. With the exception of early studies of Shakespeare (see Section 1.1), however, stylometry has not included features from the domain of versification. Yet despite this lack of interest from mainstream stylometry, versification features were taken up in the 20th century in the studies of verse experts associated with the so-called Russian school of metrics.

In the early 1920s, for example, Boris Tomashevsky used versification to prove that the ending which Dmitry Zuev claimed to have found to Pushkin's unfinished poem "The Mermaid" in 1889 was a forgery (Tomashevsky 1923/2008). Elsewhere verse rhythm and rhyme have been used to dispute the authenticity of alleged fragments of the tenth chapter of *Eugene Onegin* (Lotman and Lotman 1986), to challenge works newly added to Alexander Iliushin's edition of Gavriil Batenkov's poems (Shapir 1997, 1998; see Section 4.2 for details) and, above all, in the extensive work of Marina Tarlinskaja on Shakespeare and his contemporaries (Tarlinskaja 1987, 2014).

Because of the isolation of these versification-based approaches, however, a gulf has opened up between mainstream stylometry with its increasingly advanced methods and these studies, which have remained bound to the simple methods of descriptive statistics.

This can be illustrated with an example from Tarlinskaja's book *Shakespeare and the Versification of English Drama, 1561–1642* (2014), which deals with the authorship of the play *Henry VIII*.

Most scholars agree that *Henry VIII* was a collaborative text in which certain sections were written by John Fletcher (the "A" part) and the remainder were the work of

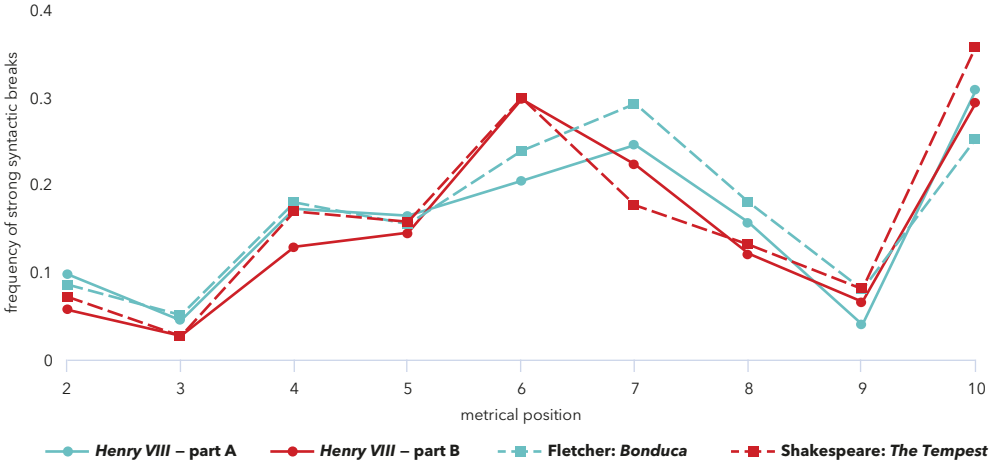


FIG. 1.11: Frequency of "strong syntactic breaks" after particular syllables (metrical positions) in Parts A and B of *Henry VIII*, Fletcher's *Bonduca* and Shakespeare's *The Tempest*. Source: Tarlinskaja 2014: table B.3.

William Shakespeare (the "B" part).[6] Tarlinskaja (2014: 140–149) sets out to support this hypothesis with versification-based evidence. She, thus, points out that the two parts have different distributions of "strong syntactic breaks".[7] She also measures the frequencies of these breaks not only in Parts A and B but also in two other plays from the same period: Fletcher's *Bonduca* and Shakespeare's *The Tempest*. She finds that within Part A, these breaks occur most frequently after the seventh syllable in a line (disregarding the line's final syllable) and that the same holds true for *Bonduca*. In contrast, in Part B and *The Tempest,* they are most common after the sixth syllable (see FIG. 1.11).

In the same way, Tarlinskaja compares the frequencies of monosyllabic words and enjambments (i.e. the lack of a "strong syntactic break") at the end of lines. Here too she discovers a significant similarity between Part A and *Bonduca* on the one hand and Part B and *The Tempest* on the other.

While these are strong and valid arguments, this analysis does not, in fact, differ substantially from Mendenhall's approach (cf. Section 1.1). Since his time, however, methods have emerged that are far more reliable and robust than the simple comparison of two measurements.

6 See also the attributions by Spedding and Ingram (Section 1.1).

7 "A strong syntactic break occurs, for example, at the juncture of sentences, or a sentence and a clause, […] between the author's [speech] and direct speech, […] or between a direct address and the rest of the utterance" (Tarlinskaja 2014: 24).

1.6 Summary

Authorship attribution, as we have seen, generally relies on the notion that authorship can be determined based on the *similarity* between the *numerical representation* of a target text and the *numerical representations* of the texts of candidate authors.

While 19th-century stylometry used simple quantifications such as word length (Mendenhall), in the years since, the field has turned to far more complex characteristics. At the same time, the understanding of *similarity* has evolved from the simple comparison of two isolated measures to multidimensional analyses and machine learning methods.

Various style markers have been taken into account for these purposes. They include the frequency of words, character *n*-grams, collocations and parts of speech, to name only a few. Nevertheless, a key aspect of the style of an important literary form—poetry—has almost completely been disregarded. While versification-based features are generally seen as author-specific, they have not been properly tested or used to attribute the authorship of poetic texts. The case for the stylometric study of versification features also has the following support:

— Most features measured in stylometry (e.g. words and *n*-grams) amount to what are known in statistics as "rare events", or more specifically, large numbers of rare events (LNRE; cf. Baayen 2001). Therefore, fairly large text samples are required. In practice, however, these are rarely available for authorship attribution studies with poetic texts. Usually only a small number of poems are concerned and not an entire collection. On the other hand, versification features are generally far more frequent. This means that they may be analysed even with significantly smaller samples.

— The vocabulary of a poetic text is not determined only by its author and genre/topic. It may also be affected by poetic metre. Forstall and Scheirer (2010), for example, found an association between metre and the frequencies of certain character *n*-grams.

— Some stylometrists have proposed combining different feature sets within a single analysis. One example might be most common words + character *n*-grams + word *n*-grams (cf. Mikros and Perifanos 2013; Eder 2011). These features are, however, already strongly correlated. Versification, on the other hand, tends to be almost entirely independent of these correlations. We may, thus, expect a combined analysis of lexicon and versification to be more powerful than one of lexicon alone.

In the following chapters, I seek to test the applicability of versification features to modern methods of authorship attribution. To begin, I explore this method with Czech, German and Spanish poetry. To the best of my knowledge, this approach has only ever been tested sporadically. Two studies, conducted with small samples of Latin poetry (Forstall, Jacobson and Scheirer 2011) and old Arabic poetry (Al-Falahi, Ramdani and Bellafkih 2017) respectively, both yielded rather unsatisfactory results. There are also some reports of research with Middle Dutch poetry (Kestemont and Haverals 2018) and Portuguese poetry (Mittmann, Pergher and dos Santos 2019). Most recently, versification features have been used with greater success to attribute the authorship of Latin poetry (Nagy 2021). Some of my own attempts to test versification-based features can also be found elsewhere (Plecháč, Bobenhausen and Hammerich 2018; Plecháč and Birnbaum 2019).