Interpreting Burrows's Delta: Geometric and Probabilistic Foundations

Shlomo Argamon
Department of Computer Science, Illinois Institute of Technology, Chicago

Abstract

While Burrows's intuitive and elegant 'Delta' measure for authorship attribution has proven to be extremely useful for authorship attribution, a theoretical understanding of its operation has remained somewhat obscure. In this article, I address this issue by introducing a geometric interpretation of Delta, which further allows us to interpret Delta as a probabilistic ranking principle. This interpretation gives us a better understanding of the method's fundamental assumptions and potential limitations, as well as leading to several well-founded variations and extensions.

Correspondence:
Shlomo Argamon, Linguistic Cognition Laboratory, Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616.
E-mail: argamon@iit.edu

1 Introduction

In his 2001 Busa Award lecture, John F. Burrows (2003) proposed a new measure for authorship attribution which he termed 'Delta', defined as:

> the mean of the absolute differences between the *z*-scores for a set of word-variables in a given text-group and the *z*-scores for the same set of word-variables in a target text. (Burrows, 2002)

The measure assumes some set of comparison texts is given, with respect to which *z*-scores are computed (based on the mean and standard deviation of word frequencies in the comparison corpus). The Delta measure is then computed between the target text and each of a set of candidate texts (generally comprising the comparison corpus), and the target is attributed to the author of the candidate text with the lowest Delta score.

A number of literary authorship studies (Burrows, 2002, 2003; Hoover, 2004a, 2004b, 2005) have shown the Delta measure to be exceptionally useful for authorship attribution studies, even with a large number of candidate authors (as long as genre is controlled for). However, while Delta is a powerful new tool in the arsenal of the computational stylist, *why* it works so well has remained somewhat obscure. As Hoover (2005) states:

> In spite of the fact that Burrows's Delta is simple and intuitively reasonable, it, like previous statistical authorship attribution techniques, and like Hoover's alterations, lacks any compelling theoretical justification.

The purpose of this article is to partially fill this lacuna by examining a geometric interpretation of Burrows's original Delta measure. As we will see, this interpretation allows Delta to be related to some well-understood notions in the theory of probability. This interpretation leads directly to several well-founded extensions of the Delta measure, as well as a better understanding of the method's assumptions and potential limitations. In an upcoming sequel, we will address the effectiveness of several different Delta variants derived from this interpretation,

Literary and Linguistic Computing, Vol. 23, No. 2, 2008.
doi:10.1093/llc/fqn003 Advance Access Published on 1 March 2008

by empirical tests on several authorship attribution tasks.

2 The Geometry of Delta

2.1 Delta as a nearest-neighbor classifier

We proceed by first examining the algebraic structure of Burrows's Delta metric, and then considering how it may be interpreted geometrically as a kind of 'distance measure', where the 'nearest' authorship candidate is chosen for attribution.

Call the set of n words of interest (usually the most frequent) for computing Delta $\{w_i\}$, defining $f_i(D)$ as w_i's frequency in document D, μ_i its mean frequency in the comparison corpus, and σ_i its standard deviation. Then, the z-score for w_i in document D is

$$z(f_i(D)) = \frac{f_i(D) - \mu_i}{\sigma_i}$$

Thus, we have the following mathematical expression for the Delta measure ('average absolute difference of the z-scores') between documents D and D':

$$\Delta(D, D') = \frac{1}{n}\sum_{i=1}^{n} |z(f_i(D)) - z(f_i(D'))|$$

This formula can be simplified algebraically as follows:

$$\begin{aligned}\Delta(D, D') &= \frac{1}{n}\sum_{i=1}^{n} |z(f_i(D)) - z(f_i(D'))| \\ &= \frac{1}{n}\sum_{i=1}^{n} \left| \frac{f_i(D) - \mu_i}{\sigma_i} - \frac{f_i(D') - \mu_i}{\sigma_i} \right| \\ &= \frac{1}{n}\sum_{i=1}^{n} \left| \frac{f_i(D) - f_i(D')}{\sigma_i} \right|\end{aligned}$$

This simplification shows that Delta does not actually depend on the mean frequencies in the comparison set, but may be viewed as a normalized difference measure between frequencies in D and in D'. We further note that averaging is just multiplication of the sum by a constant factor of $1/n$, which depends only on the number of words considered, and thus is irrelevant when using Delta as a ranking metric. In the remainder of this article, therefore, we will use the following simplified formula as equivalent to Burrows's Delta:

$$\Delta_B^{(n)}(D, D') = \sum_{i=1}^{n} \frac{1}{\sigma_i} \left| f_i(D) - f_i(D') \right|$$

i.e. the sum of the standard deviation-normalized absolute differences of the word frequencies. Note that n is the number of frequent words used, and the subscript B indicates equivalence to Burrows's original Delta formula (we will develop several variants later on).

This formulation shows clearly that Delta ranks authorship candidates D' by a kind of distance from the test text D, where each dimension of difference (word frequency) is scaled by a factor of $1/\sigma_i$ (i.e. small differences count more for dimensions with less 'spread'). Thus, Delta may be viewed as an axis-weighted form of 'nearest neighbor' classification (Wettschereck *et al.*, 1997), where a test document is classified the same as the known document at the smallest 'distance'.

2.2 Manhattan and Euclidean distance

A deeper understanding of Delta is obtained by considering each comparison of target text D and given text D' as a point in a multi-dimensional geometric space, where the differences between word frequencies give the point's coordinates in the space. Mathematically, we use lower-case deltas δ_i to indicate, for each word w_i, the difference between w_i's frequencies in the two texts: $\delta_i = f_i(D) - f_i(D')$. (We defer for now consideration of the effect of conversion to z-scores.) Every point in this n-dimensional *difference space* corresponds to one possible set of differences between target and given texts over all words of interest. The Delta measure assigns a numeric score to each such *difference point*, allowing them to be ranked for likelihood of similarity or difference of authorship. To make this more concrete Fig. 1a depicts the scores assigned by the simplest Delta function, where all standard deviations are 1, to points in a 2-dimensional space (i.e. one where only two words are used in the Delta calculation). The figure shows several *iso-Delta* diamonds, i.e. sets of points where Delta values are equal. In three dimensions the lines become planes,

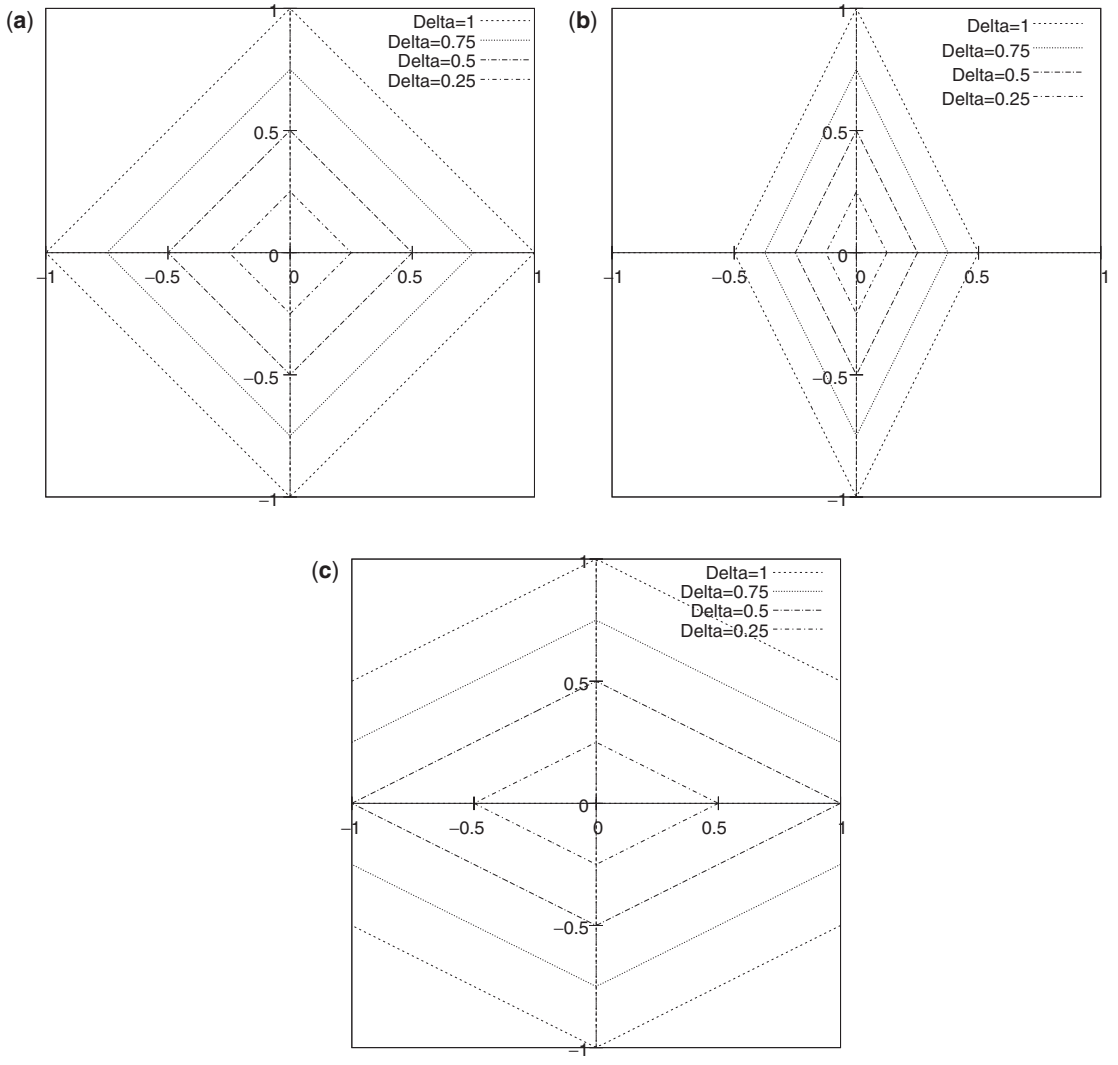

Fig. 1 Graphs showing the structure of the distance measure implied by Burrows's Delta in two dimensions ($\Delta_B^{(2)}$). Each graph shows a space assuming two words are used for computing the measure, where the coordinates of each point in the space correspond to the frequencies of these words in a possible text. The diamonds show the 'iso-Delta' lines, such that all points on each diamond have the same value for Delta. (**a**) The standard deviation is 1 for the frequencies of both words. (**b**) The standard deviation for the word on the x axis is 2, while the other is 1. (**c**) The standard deviation for the word on the x axis is 1/2, while the other is 1

and the diamonds octahedra; in higher dimensions it becomes rather more complex.

For comparison, Fig. 1b shows iso-Delta lines where $\sigma_1 = 1/2$ and $\sigma_2 = 1$, while Fig. 1c shows iso-Delta lines where $\sigma_1 = 2$ and $\sigma_2 = 1$. As these graphs clearly show, the effect of dividing by σ_i is to rescale the scoring in the direction of each of the axes, making differences in some directions more salient than in others.

Note that regardless of the rescaling of the axes, the distance of a given point in the difference space from the origin is computed as the sum of its distances along each of the axes from the origin. Such a distance measure has been termed

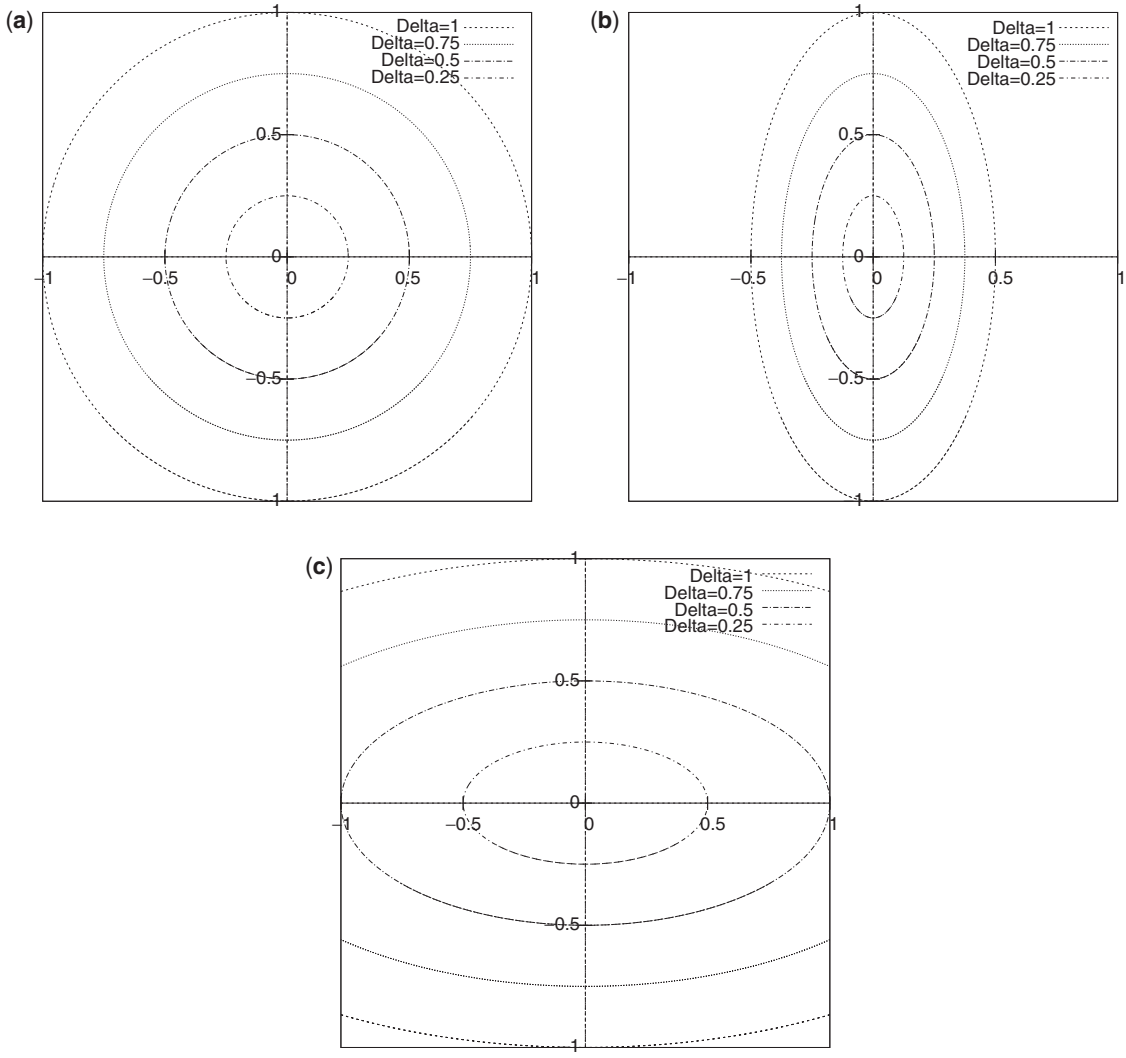

Fig. 2 Graphs showing the structure of the distance measure implied by axis-parallel quadratic Delta in two dimensions ($\Delta^{(2)}_{Q\perp}$). Each graph shows a space assuming two words are used for computing the measure, where the coordinates of each point in the space correspond to the frequencies of these words in a possible text. The ellipses show the 'iso-Delta' lines, such that all points on each ellipse have the same value for Delta. (**a**) The standard deviation is 1 for the frequencies of both words. (**b**) The standard deviation for the word on the x axis is 2, while the other is 1. (**c**) The standard deviation for the word on the x axis is 1/2, while the other is 1

'Manhattan distance' by analogy to measuring driving (or walking) distance in Manhattan (or any other grid of city streets) by summing distances in each of the principle directions. Another, possibly more natural, measure of distance would be the straight-line 'Euclidean' distance, in which the iso-Delta curves would be circles, as in Fig. 2a. Using Euclidean distance gives the distance formula (derived from the Pythagorean Theorem):

$$\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2} (f_i(D) - f_i(D'))^2}$$

Instead of the sum of absolute differences, we have the square-root of the sum of the squared differences.

Note that if the standard deviations (for rescaling) vary between the axes, the circular iso-Delta curves become axis-parallel ellipses, as shown in Fig. 2b for $\sigma_1 = 1/2$ and Fig. 2c for $\sigma_1 = 2$ (compare to Fig. 1). Again, note that since Delta is to be used as a ranking principle (and so we only care about relative values), we can simplify the formula by removing the square root, giving as our first variant the *quadratic Delta* formula:

$$\Delta^{(n)}_{Q\perp}(D, D') = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(f_i(D) - f_i(D'))^2$$

The 'Q' in the subscript denotes the use of a quadratic function in the sum, while the '⊥' symbol indicates that the scaling is in line with the perpendicular axes (we will return to this point in Section 4).

3 Delta as a Probabilistic Ranking Principle

3.1 Quadratic Delta

The quadratic Delta formula just introduced leads us to a deeper conception of why such a ranking principle for authorship candidates can make sense. Consider first the case of using just a single word frequency. In this case,

$$\Delta^{(1)}_{Q\perp}(D, D') = \frac{1}{\sigma_1^2}(f_1(D) - f_1(D'))^2,$$

where the author of the D' giving the smallest value will be attributed as D's author. It can be straightforwardly shown mathematically that an identical decision process is given by choosing the highest probability value given by a Gaussian distribution[1] with mean $f_1(D')$ and standard deviation σ_1, whose probability density function is given as:

$$G_{\langle f_1(D'),\sigma_1\rangle}(f_1(D)) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{\sigma_1}(f_1(D)-f_1(D'))^2} = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\Delta^{(1)}_{Q\perp}(D, D')}$$

This is because the normalizing factor $1/\sqrt{2\pi}\sigma_1$ is a constant (and so does not change candidate rankings), and exponentiating minus Delta just inverts candidate rankings (turning minimizing into maximizing), since exponentiation is a monotonic operator (its outputs have the same ordering as its inputs).

This is easily generalized, so that minimizing the n-dimensional quadratic Delta

$$\Delta^{(n)}_{Q\perp}(D, D') = \sum_i \frac{1}{\sigma_i^2}(f_i(D) - f_i(D'))^2$$

may be seen to be equivalent to maximizing a probability according to the n-dimensional Gaussian distribution[2]:

$$G_{\vec{\mathbf{f}}(D'),\vec{\sigma}}(\vec{\mathbf{f}}(D)) = \frac{1}{\sqrt{(2\pi)^n}\prod_i \sigma_i} e^{-\sum_i \frac{1}{i\sigma_i^2}(f_i(D)-f_i(D'))^2} = \frac{1}{\sqrt{(2\pi)^n}\prod_i \sigma_i} e^{-\Delta^{(n)}_{Q\perp}(D,D')}$$

What this equivalence means is that the use of quadratic Delta for choosing an authorship candidates amounts to choosing the highest-probability candidate, where the frequency of each indicator word w_i, in texts written by the author of D', is assumed to be randomly distributed (in the abstract n-dimensional word frequency space) with probabilities given by a Gaussian distribution with mean $f_i(D')$ and standard deviation σ_i. Thus, a main assumption of this method is that, for each candidate author, the author's candidate document D' is taken as the 'prototype' for all documents by that author, while the potential 'spread' of the word frequency distribution for other documents is fixed as the overall spread for that word in the entire background set. In addition, every indicator word's frequency is assumed to be statistically independent of every other indicator word's frequency. (This second assumption is indeed rather questionable, and is dealt with in more detail in Section 4.)

More specifically, in this case the distribution is taken to be a Gaussian distribution over n independent variables (the n indicator word frequencies) (Fig. 3). This distribution (also called the 'normal' distribution) is the one most commonly used for modeling probabilistic phenomena, for two main reasons. The first is the distribution's convenient mathematical structure (for example, the projection of an n-dimensional Gaussian distribution onto any

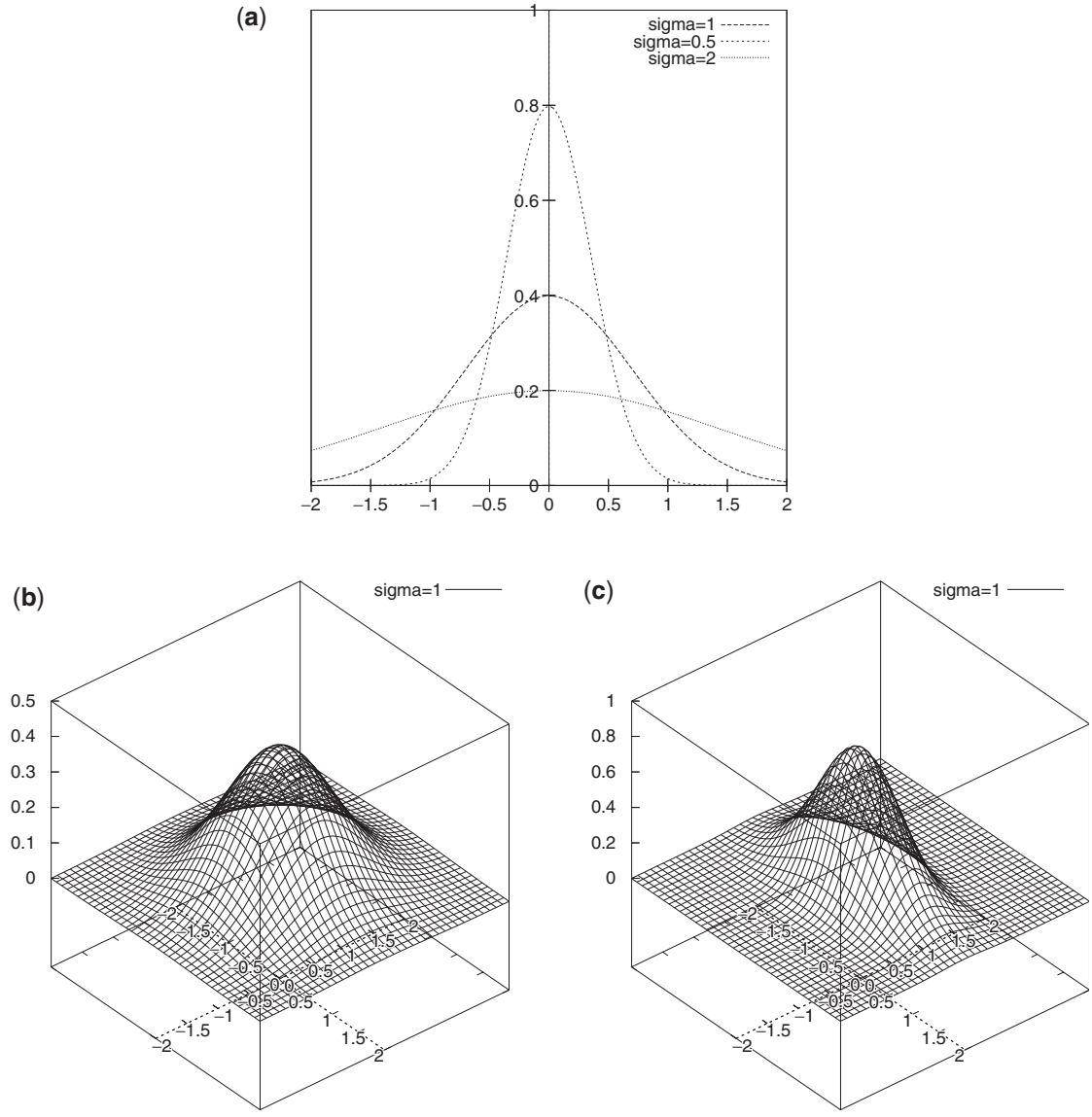

Fig. 3 Gaussian probability distributions. (**a**) 1-dimensional Gaussian distributions for several values of standard deviation σ. (**b**) 2-dimensional Gaussian distribution (probability shown along the vertical axis) where both standard deviations are 1. Note that horizontal cross-sections are circles. (**c**) 2-dimensional Gaussian distribution where one standard deviation is 1 and the other is 2. Note that horizontal cross-sections are ellipses

linear combination of a subset of variables is also a Gaussian distribution). The second is the Central Limit Theorem, which states that the distribution of the sum of k identically distributed independent random variables will approach a Gaussian distribution as k increases, regardless of the type of distribution of the individual variables (so long as the variance of the sum remains finite).

3.2 Linear Delta

While all of this discussion of the quadratic Delta variant is well and good, we may wonder what this

approach can say about Burrows's original Delta. In fact, it can be shown that attribution by Burrows's Delta function is also equivalent to a probabilistic attribution principle, just using a different underlying probability distribution function. Recall that Burrows's Delta between texts D and D' may be computed by the function:

$$\Delta_B^{(n)}(D, D') = \sum_{i=1}^{n} \frac{1}{\sigma_i} \left| f_i(D) - f_i(D') \right|$$

In the same manner as we showed that distance ranking of authorship candidates by quadratic Delta is equivalent to probability ranking using a Gaussian distribution, we can show that distance ranking by Burrows's Delta is equivalent to probability ranking using what is known as the Laplace distribution.[3] The Laplace distribution for one variable is given by the equation (Evans *et al.*, 2000, p. 117):

$$L_{\langle a,b \rangle}(x) = \frac{1}{2b}\ e^{-\frac{|x-a|}{b}}$$

This distribution has two parameters, a, indicating the median, and b indicating the amount of 'spread' in the distribution. Graphs of this density function (centered around 0) are given in Fig. 4a for several values of b.

The multivariate form (assuming independent variables) is given by the formula:

$$L_{\langle \vec{a}, \vec{b} \rangle}(\vec{x}) = \frac{1}{\prod_i 2b_i}\ e^{-\sum_i \frac{1}{b_i}|x_i - a_i|}$$

Graphs of the 2-dimensional Laplace density function with $b_1 = b_2 = 1$ is given in Fig. 4b, and for $b_2 = 2$ in Fig. 4c. The connection to Burrows's Delta is the same in form as the connection of the Gaussian density function to quadratic Delta, seen by substituting σ_i for b_i, $f_i(D')$ for a_i, and $f_i(D)$ for x_i:

$$\frac{1}{\prod_i 2\sigma_i}\ e^{-\sum_i \frac{1}{\sigma_i}|f_i(D) - f_i(D')|} = \frac{1}{\prod_i 2\sigma_i}\ e^{-\Delta_B^{(n)}(D, D')}$$

Note that the fraction in the front $(1/\prod_i 2\sigma_i)$ is a constant for all D and D', and so will not affect any ranking, and that exponentiating minus Delta just reverses the ranking (so we seek a maximum probability instead of a minimum distance).

There are several important differences between the Laplace distribution and the Gaussian distribution that are relevant to us. Most significantly, while the mean of the distribution is in fact equal to the parameter a, the standard deviation of the Laplace distribution is not b_i, but rather is given by $\sqrt{2}b_i$. If a_i and b_i are to be estimated from w_i's frequencies in the comparison corpus, the best (maximum likelihood) estimates are (Evans *et al.*, 2000, p. 120):

$$a_i = \text{median}(\langle f_i(D_1), f_i(D_2), \cdots, f_i(D_m) \rangle)$$

$$b_i = \frac{1}{n} \sum_{j=1}^{m} |f_i(D_j) - a_i|$$

The *median* of a set of numbers is that number such that half of the set are higher and half are lower.

The implications are crucial for properly establishing Burrows's Delta as a well-founded probabilistic ranking principle. As discussed earlier, with respect to the quadratic variation of Delta, the most straightforward interpretation is that parameters estimated from the comparison corpus are taken to describe the probability distribution of word frequencies for different authors (whose distributions just vary by 'location', i.e. $f_i(D')$). If so, then we would want to interpret Burrows's Delta as a principle for ranking authorship candidates based on assuming that:

- Word frequencies for a given author are randomly distributed according to a multivariate Laplace distribution;
- All authors have the same b parameters for such distributions, varying only by the distribution mean;
- Such parameters can be best estimated from a varied comparison corpus of texts.

Using z-scores, which divide by standard deviations σ_i in the comparison corpus, does not precisely allow such an interpretation, since the Laplace distribution does not divide by σ_i but rather by b_i. So to establish such an interpretation, instead of using z-scores the method should instead divide by

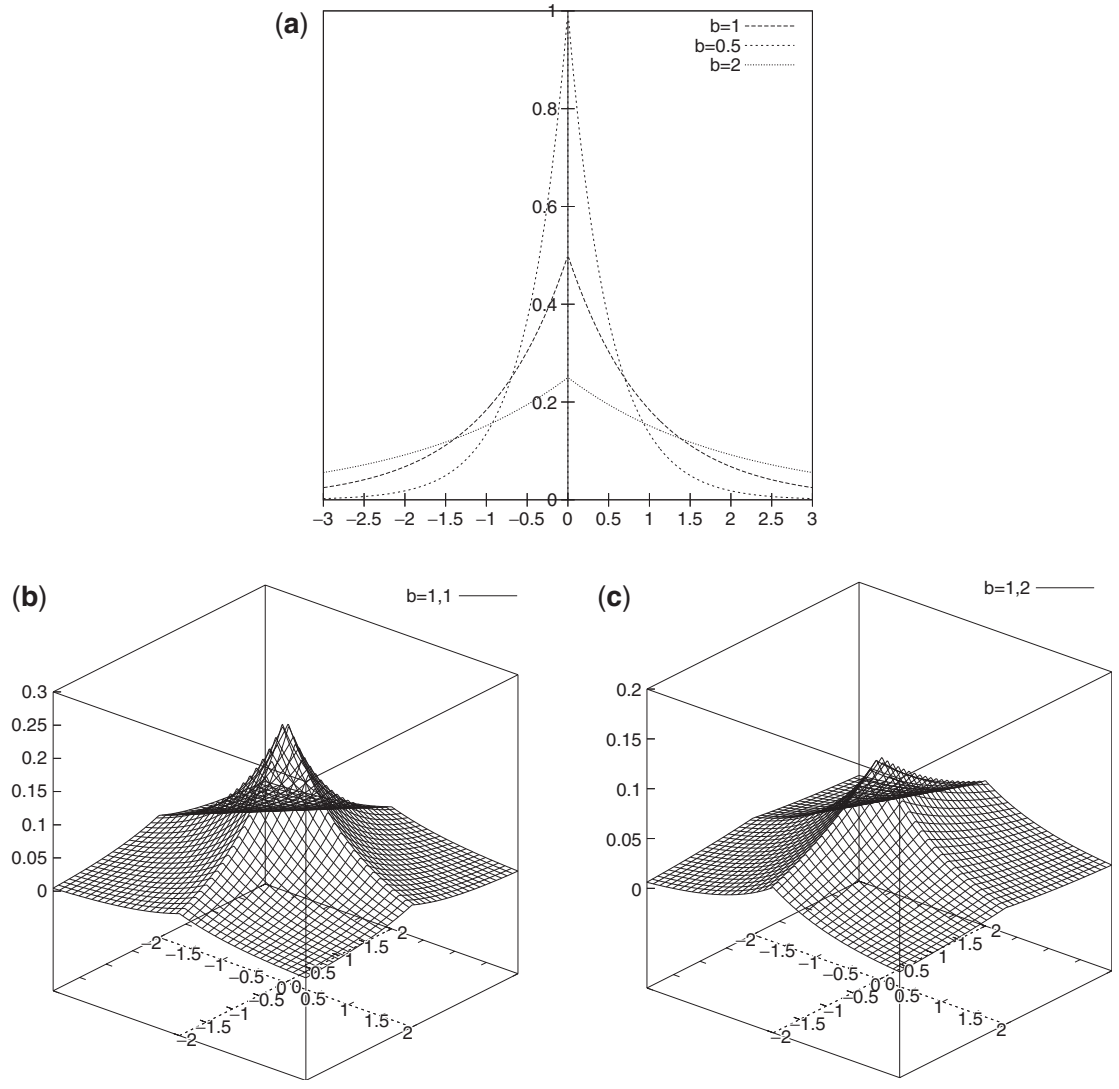

Fig. 4 Laplace probability distributions. (**a**) 1-dimensional Laplace distributions for several values of standard deviation σ. (**b**) 2-dimensional Laplace distribution (probability shown along the vertical axis) where both standard deviations are 1. Note that horizontal cross-sections are diamonds. (**c**) 2-dimensional Laplace distribution where one standard deviation is 1 and the other is 1/2. Note that horizontal cross-sections are elongated diamonds

b_i estimates as given above, i.e. the average absolute deviation of comparison text word frequencies from the median word frequency. This gives us the alternative *linear Delta* formulation:

$$\Delta_{L\perp}^{(n)}(D, D') = \sum_{i=1}^{n} \frac{1}{b_i} \left| f_i(D) - f_i(D') \right|$$

with b_i defined as above (average absolute deviation from the median). This function retains the form of Burrows's original Delta, but bases it more firmly as a probabilistic ranking principle. (It may still be the case that the difference in practice between Burrows's Delta and linear Delta may be negligible.)

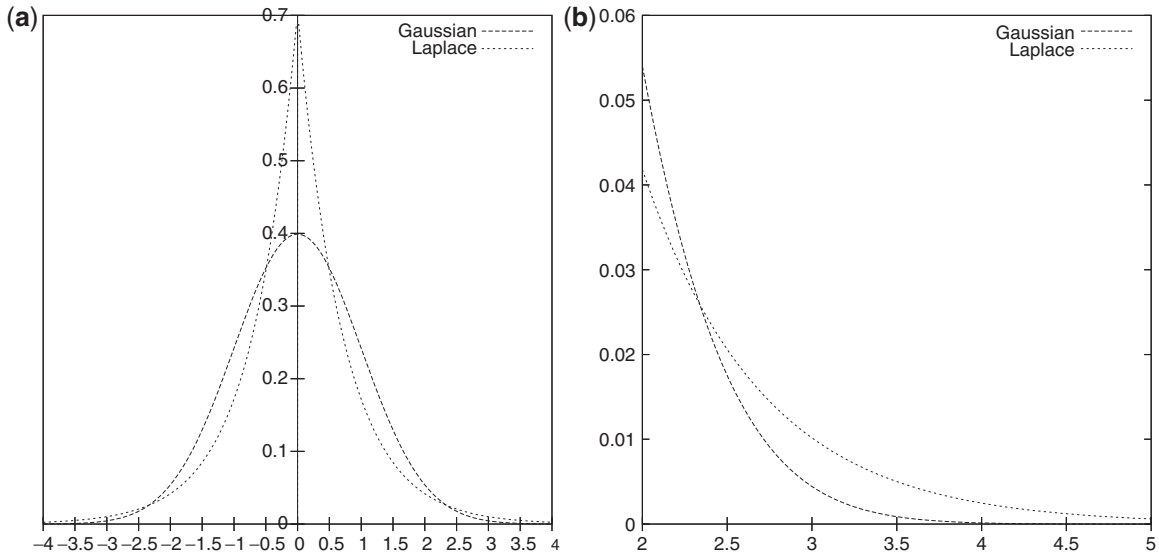

Fig. 5 Comparison of Gaussian and Laplace probability distributions, showing a Gaussian density function with $\mu = 0$ and $\sigma = 1$ and a Laplace density function with $a = 0$ and $b = \sqrt{2}$ (such that its standard deviation is also 1. (**a**) From $x = -4$ to $x = 4$. (**b**) Detail view of the range $x = 2$ to $x = 5$, showing the heavier tail of the Laplace distribution

3.3 Contrasting the Gaussian and Laplace distributions

Both the Gaussian and Laplace distributions have useful properties for our purposes; the choice between them for any given problem is largely empirical. We should also note that as both of them are unbounded, i.e. they give any number some nonzero probability of occurrence, neither is accurate for describing word frequencies, strictly speaking, as word frequencies cannot be less than 0 nor greater than 1. However, they may still provide useful models for word frequencies.

A useful way of understanding the difference between the Gaussian and Laplace distribution is by viewing them as *error laws*, that is, as describing the random distribution of errors in measurement about some true value. Keynes (1911) has shown that, assuming positive and negative errors to be equally likely, that:

- If the most probable true value is the *arithmetic mean* (i.e. average) of the noisy observations, then the Gaussian distribution is the most likely correct error law;
- If the most probable true value is the *median* of the noisy observations, then the Laplace distribution is the most likely correct error law.

The mean of a set of numbers is more strongly affected by faraway values ('outliers') than the median, hence the Laplace distribution is more stable (hence generally to be preferred) when such outliers are more common. Intuitively, we can see this also in comparison of the distributions' shapes (Fig. 5), where we see two main differences.

- The Laplace distribution is more 'peaked' than the Gaussian, that is, it gives more probability to values very close to the mean, whereas the Gaussian distribution allows more spread, and
- The Laplace distribution is more 'heavy-tailed' than the Gaussian, in that it also gives more probability to values very far from the mean.

Thus, the Gaussian distribution allows for more mid-range spread around the mean than the Laplace, while the Laplace allows for more faraway 'outliers' to occur. Thus, if we expect the texts of a given author to have similar frequencies for all common words,

with a medium amount of spread but almost no 'outliers', then a Gaussian distribution will be more appropriate. However, if we think that the frequencies will be very tightly gathered around the center, but that there is a greater (though still low) likelihood that some texts by the author will have a few highly atypical word frequencies, then the Laplace distribution may be a more appropriate choice.

We emphasize again that the choice of proper distribution for modeling word frequencies for authorship attribution is entirely an empirical one, which we will address in the sequel.

4 Interdependence among Word Frequencies

The understanding that Delta (in the various incarnations thus far presented here) corresponds to choosing the author that assigns highest probability to the target text (under certain assumptions) goes a long way to giving the method theoretical justification, as well as elucidating its underlying theoretical (in this case probabilistic) assumptions. One of those assumptions, however, is particularly troubling—the assumption that word frequencies are utterly independent of each other. This assumption is clearly false in general, and so raises a significant theoretical difficulty for the general use of the Delta method.[4] It is the purpose of this section to address this problem by examining how versions of Delta could be developed that relax the strong assumption of independence. Such methods will stand on a firmer theoretical foundation, and so possibly lead to more easily justified results (provided that empirical evidence also supports their efficacy).

4.1 The Gaussian distribution

We begin, for mathematical simplicity, with the Gaussian distribution. If two random variables (in this case, the frequencies of different frequent words in a random text) are not independent, then the have nonzero *covariance*, where the covariance σ_{ij} of the frequencies of words w_i and w_j is defined as:

$$\sigma_{ij} = \mathrm{Exp}_D[(f_i(D) - \mu_i)(f_j(D) - \mu_j)]$$

that is, the expected value (over all possible random texts D) of f_i's deviation from its mean times f_j's deviation from its mean. If the two words occur independently, then the covariance will be zero. Naturally, since we do not have access to all possible random texts, we must estimate the covariance from given data in the comparison corpus C, so in what follows we use the estimate:

$$\sigma_{ij} = \frac{1}{|C|}\sum_{D\in C}(f_i(D) - \mu_i)(f_j(D) - \mu_j)$$

It is easy to see that the variance of a single variable is simply its covariance with itself:

$$\sigma_i^2 = \frac{1}{|C|}\sum_{D\in C}(f_i(D) - \mu_i)^2$$

Mathematically, we arrange the covariances among all the variables in a *covariance matrix*, a square matrix defined as:

$$\mathbf{S} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

In the case that all the variables are independent, all elements of the matrix other than those on the diagonal are zero. When there are statistical dependencies, however, some (perhaps many or all) of the nondiagonal elements will be nonzero. Note that the diagonal elements (variances) are always positive, while covariances between different variables may be negative (i.e. when one variable is high, the other tends to be low). Correspondingly, we represent the relevant word frequencies in a given document D by a *vector* of all the frequencies:

$$\vec{\mathbf{f}}(D) = \begin{bmatrix} \vec{\mathbf{f}}_1(D) \\ \vec{\mathbf{f}}_2(D) \\ \vdots \\ \vdots \\ \vec{\mathbf{f}}_n(D) \end{bmatrix}$$

This representation allows us to use the powerful tools of *linear algebra* (Meyer, 2000) to compactly represent and manipulate complex expressions (a brief explanation of basic operations is given in the Appendix I).

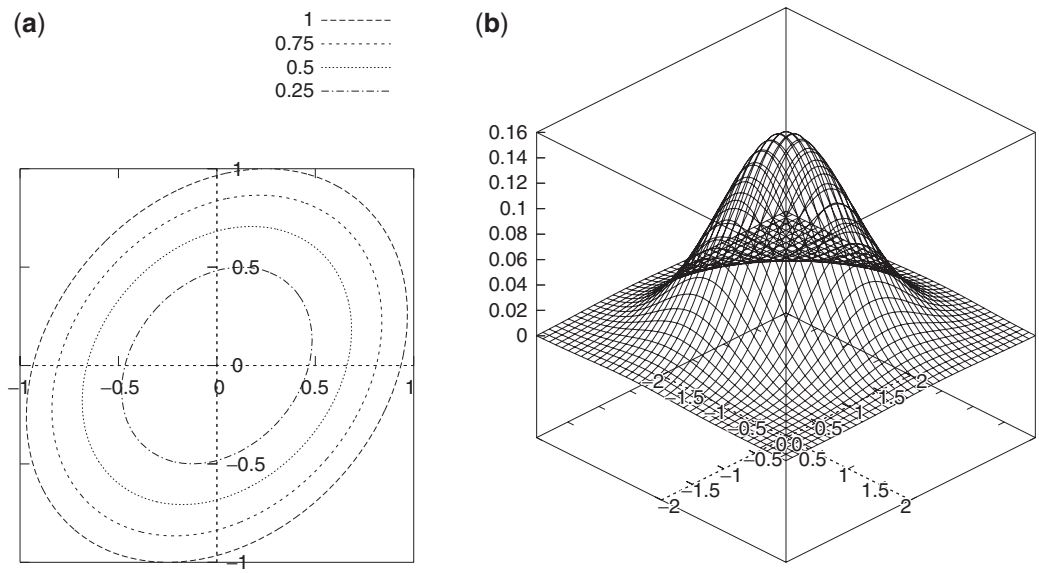

Fig. 6 Two-dimensional quadratic ranking functions with dependent variables; standard deviations $\sigma_1 = 1$ and $\sigma_2 = 2$ and covariance $\sigma_{12} = \sigma_{21} = 1/2$. (**a**) Iso-Delta curves; note that the ellipses are rotated 45° due to the covariance. (**b**) The Gaussian probability distribution

The matrix equivalent of the reciprocal of the variance $(1/\sigma^2)$ is the *inverse* of the covariance matrix, denoted $\mathbf{S}^{-1}$, and defined such that $\mathbf{S}^{-1}\,\mathbf{S} = \mathbf{I}$, where $\mathbf{I}$ is the *identity matrix* which has all ones on the diagonal and zeros elsewhere. To take a simple case in two dimensions, suppose that the covariance between words w_1 and w_2 is $\sigma_{12} = \sigma_{21} = 0.5$, with variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$. Then the covariance matrix $\mathbf{S}$ is

$$\mathbf{S} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$

and its inverse $\mathbf{S}^{-1}$ is

$$\mathbf{S}^{-1} = \begin{bmatrix} 1.143 & -0.286 \\ -0.286 & 1.071 \end{bmatrix}$$

The corresponding quadratic Delta function, allowing for dependent variables, is:

$$\begin{aligned}\Delta^{(2)}_{Q\not\parallel}(D, D') &= (\vec{\mathbf{f}}(D) - \vec{\mathbf{f}}(D'))^T \mathbf{S}^{-1} (\vec{\mathbf{f}}(D) - \vec{\mathbf{f}}(D')) \\ &= 1.143(f_1(D) - f_1(D'))^2 \\ &\quad - 0.286(f_1(D) - f_1(D'))(f_2(D) - f_2(D')) \\ &\quad - 0.286(f_2(D) - f_2(D'))(f_1(D) - f_1(D')) \\ &\quad + 1.071(f_2(D) - f_2(D'))^2\end{aligned}$$

(The subscript '$Q\not\parallel$' indicates a quadratic function—'Q' —that is nonaxis-parallel— '$\not\parallel$'.) Graphical depictions of this function are given in Fig. 6. As the figure shows, when variables covary, the iso-Delta ellipses (or in higher dimensions, the ellipsoids) are rotated relative to the axes. The direction and amount of rotation depends on the amount of covariance.

To generalize, when $\mathbf{S}^{-1}$ exists (see subsequently), the n-dimensional nonaxis parallel quadratic Delta function is given by

$$\begin{aligned}\Delta^{(n)}_{Q\not\parallel}(D, D') &= (\vec{\mathbf{f}}(D) - \vec{\mathbf{f}}(D'))^T \mathbf{S}^{-1} (\vec{\mathbf{f}}(D) - \vec{\mathbf{f}}(D')) \\ &= \sum\nolimits_i \sum\nolimits_j (f_j(D) - f_j(D'))(\mathbf{S}^{-1})_{ij} \\ &\quad \times (f_i(D) - f_i(D'))\end{aligned}$$

where $(\mathbf{S}^{-1})_{ij}$ denotes the i, jth element in the $(\mathbf{S}^{-1})$ matrix. Note that this choice of nonaxis-parallel Delta function is not arbitrary. Choosing a document D' to minimize this function directly corresponds to choosing one to maximize the probability given by the standard multivariate Gaussian

distribution with mean at $\vec{\mathbf{f}}(D')$ and covariance matrix **S**:

$$G_{\vec{\mathbf{f}}(D'),\mathbf{S}}(\vec{\mathbf{f}}(D)) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{S})}} \times e^{-(\vec{\mathbf{f}}(D)-\vec{\mathbf{f}}(D'))^T \mathbf{S}^{-1}(\vec{\mathbf{f}}(D)-\vec{\mathbf{f}}(D'))} = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{S})}} e^{-\Delta^{(n)}_{Q\_}(D,D')}$$

where det(**S**) denotes the *determinant* of matrix **S** (Meyer, 2000).

This is all well and good, but there is a fly in the ointment. The fact is that **S** does not always have an inverse; clearly such a situation will play havoc with the formulae presented above. In fact, if there are fewer texts in the comparison corpus than there are dimensions (i.e. relevant words) considered, we are *guaranteed* that **S** has no inverse! Roughly speaking, this is because **S** only has an inverse if the set of points used to estimate it (i.e. the comparison texts, viewed as points in an n-dimensional space) has thickness in every single direction in the n-dimensional space (technically, this is equivalent to the matrix having 'full rank'). For example, consider the case of two relevant words and two comparison texts (Fig. 7a). Note that while in each axis-parallel direction the set of two points has thickness, along the direction perpendicular to the line joining the two points the set perforce has zero thickness. That is, the two points define a line which has lower dimension (one) than the full space (two). In general, a set of m points can define a space of at most $m-1$ dimensions, and hence if the number of word variables n is greater than $m-1$, the covariance matrix will not be invertible.

However, all is not lost. As Fig. 7 shows for the low-dimensional case, it is still possible to use a Gaussian distribution as an authorship ranking principle by first rotating the space to align with the 'natural' axes defined by the estimated covariance matrix **S**, and then ranking according to a *lower-dimensional* Gaussian probability distribution. Mathematically, this is accomplished by *eigenvalue decomposition* of the matrix **S**. This decomposition factors **S** into a product of three matrices, as $\mathbf{S} = \mathbf{EDE}^T$ where **D** is a diagonal matrix (zero except on the diagonal) and **E** is a square matrix called the *eigenvector* matrix. Geometrically, **E** represents an n-dimensional rotation of the space around the origin such that after such rotation, the variables corresponding to the new axes are statistically independent. The values along the diagonal in **D** (the *eigenvalues*) are then the estimated variances of the distribution for each of those new composite variables. In the case that **S** is noninvertible, some of the eigenvalues will be zero, meaning that the distribution is flat in the corresponding directions. Let us call the number of nonzero eigenvalues (the lower number of 'thick' dimensions) k. By using just the k columns of **E** that correspond to nonzero eigenvalues, we can both *rotate* the space to accord with the 'natural' axes of the distribution, and *project* points into a k-dimensional space all of whose dimensions have some 'thickness'.

The idea, then, is to do an eigenvalue decomposition of **S** (using a standard software library) to get **E** and **D**. Then we remove columns of **E** that correspond to zero eigenvalues in **D** to get $\mathbf{E}_*$, and remove the zero eigenvalue columns and rows in **D** to get $\mathbf{D}_*$. Note that for any n-dimensional vector $\vec{\mathbf{x}}$, computing the vector $\vec{\mathbf{y}} = \mathbf{E}_*^T\vec{\mathbf{x}}$ rotates $\vec{\mathbf{x}}$ into the natural axes of the space and then projects it into the k-dimensional 'thick' space, transforming the n-dimensional vector $\vec{\mathbf{x}}$ to a k-dimensional vector $\vec{\mathbf{y}}$. In this new k-dimensional space, the elements d^*_{ii} along the diagonal of $\mathbf{D}_*$ are just the variances σ_i^2 of each new variable. Thus, in the new space, we can just use $\Delta^{(k)}_{Q\perp}$ as a ranking principle (equivalent to Gaussian distribution-based ranking), and get a probabilistic ranking method that takes into account dependence among the variables. We also note that in the case where **S** has full rank, i.e. no eigenvalue is zero, this method will give identical results to just directly using $\mathbf{S}^{-1}$ as above.

Thus, given matrices $\mathbf{D}_*$ and $\mathbf{E}_*$ derived from **S** by eigenvalue decomposition, we define the nonaxis-parallel quadratic Delta function as:

$$\Delta^{(n)}_{Q\_}(D, D') = (\vec{\mathbf{f}}(D) - \vec{\mathbf{f}}(D'))^T \mathbf{E}_* \mathbf{D}_*^{-1} \mathbf{E}_*^T \times (\vec{\mathbf{f}}(D) - \vec{\mathbf{f}}(D'))$$

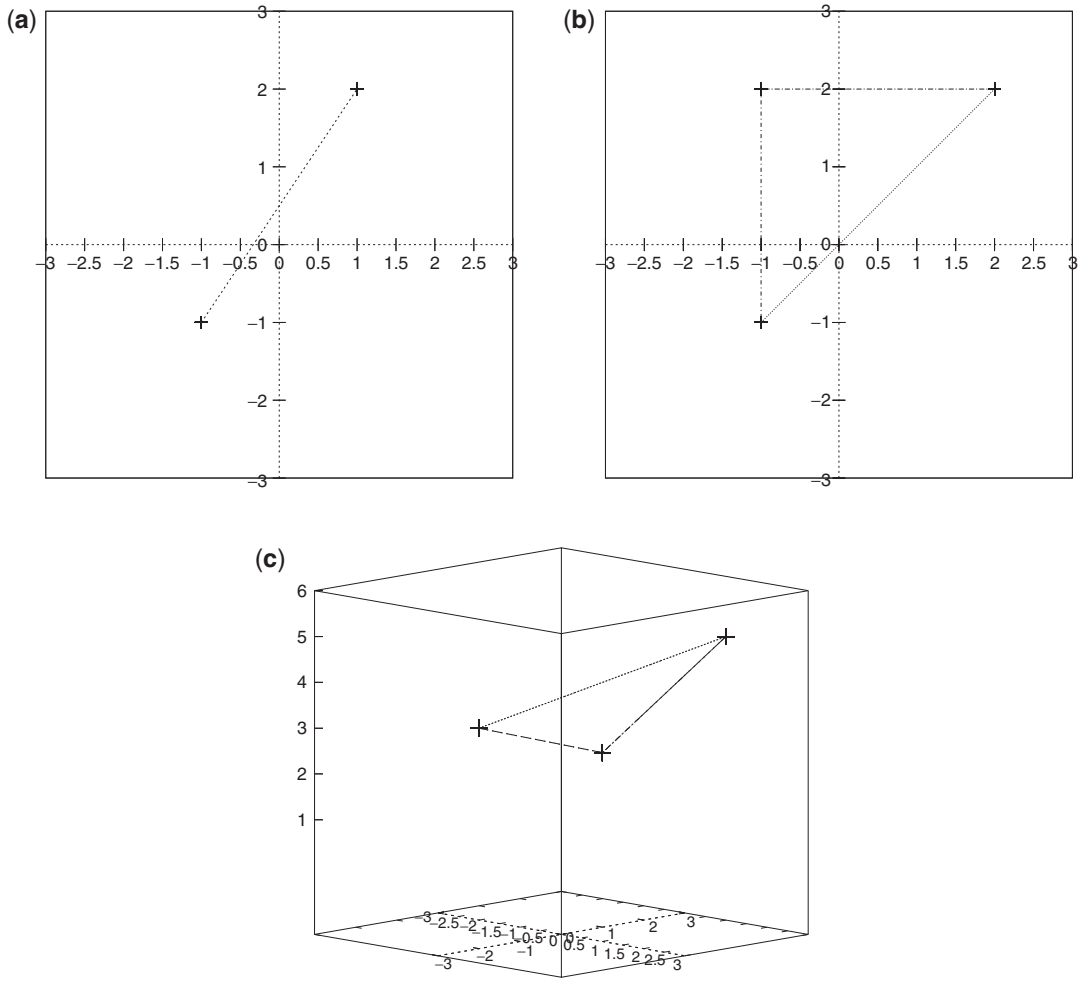

Fig. 7 Illustrations of subspaces spanned by small numbers of points (i.e. comparison texts). (**a**) Two points in a two dimensional space can give variation along both axes, but span a 'natural' space (the line segment connecting them) with no 'thickness', regardless of how the points are placed. Hence two points can only span a 1-dimensional subspace. (**b**) Three non collinear points in a 2-dimensional space give a triangle, which has 'thickness' in all directions, and so span a 2-dimensional subspace. (**c**) Even if embedded in three dimensions, such that the points have different locations on all three axes, they still only form a two-dimensional triangle regardless of where they are located, and so three points can only span a 2-dimensional space

with the equivalent Gaussian distribution:

$$G_{\vec{\mathbf{f}}(D'),\mathbf{S}}(\vec{\mathbf{f}}(D)) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{D}_*)}} \times e^{-(\vec{\mathbf{f}}(D)-\vec{\mathbf{f}}(D'))^T \mathbf{E}_* \mathbf{D}_*^{-1} \mathbf{E}_*^T (\vec{\mathbf{f}}(D)-\vec{\mathbf{f}}(D'))} = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{D}_*)}}\ e^{-\Delta_{Q-}^{(n)}(D,D')}$$

4.2 The Laplace distribution

Unfortunately, no such mathematically and computationally tractable treatment exists for multivariate Laplace distributions with correlated variables. Indeed, there is no single universally accepted multivariate generalization of the Laplace distribution. Some possibly applicable multivariate extensions

of the Laplace distribution do exist (Kotz *et al.*, 2001), but since such generalizations are neither unique nor straightforward, they are outside the scope of this article.

A heuristic (i.e. not probabilistically well-founded) extension of the Laplace distribution is based on using the eigenvalue decomposition trick presented above to transform document vectors onto an (empirically) independent set of variables. The idea is to take each document vector $\vec{\mathbf{f}}(D)$, and convert it to a (lower-dimensional) vector $\vec{\mathbf{g}}(D)$, defined by:

$$\vec{\mathbf{g}}(D) = \mathbf{E}_*^T \vec{\mathbf{f}}(D)$$

where $\mathbf{E}_*$ is the reduced-dimension eigenvector matrix computed from the comparison set, as described in Section 4.1 above. All document vectors (in the comparison and test sets) are transformed thusly, then the standard (independent variable) linear Delta function $\Delta_{L\perp}$ can be applied to the $\vec{\mathbf{g}}(D)$ vectors, as the variables are now (assumed to be) independent. Keep in mind that the parameters a_i and b_i for $\Delta_{L\perp}$ are to be calculated over the variables in $\vec{\mathbf{g}}(D)$, rather that $\vec{\mathbf{f}}(D)$.

Note that this method is only approximate, as the eigenvector method for transforming the vectors to independent variables is based on assuming that the true distribution is actually Gaussian. As noted earlier, an exact solution using any of the currently available multivariate generalizations of the Laplace distribution would be excessively complicated for our purposes here.

4.3 Relationship to PCA

The sort of eigenvector decomposition described in this section highlights some of the similarities and differences between the Delta method and the older method of using principal components analysis (PCA). Briefly, PCA uses eigenvector decomposition to project points in a high-dimensional space into a low-dimensional (usually 2-dimensional) space, while losing as little as possible information about the variability in the data (see Binongo and Smith, 1999 for a more detailed exposition). Applications of PCA for authorship attribution (Burrows, 1992; Binongo and Smith, 1999; Baayen *et al.*, 2002) vary in their details, but the basic idea is to obtain a 2-dimensional visualization of the relative positions of the n-dimensional representations of comparison documents and the target document, and then to see if (a) comparison documents by different authors are divided neatly into different regions of the space, and (b) whether the target can be clearly associated with one of those regions (corresponding to a particular attribution of its authorship).

Use of PCA can thus also be viewed as a form of nearest-neighbor classification in a transformed space, where the transformation is the rotation and projection to a particular 2-dimensional space (the *PC space*), defined by the two main principal components. Measuring distance between points in this space can be viewed as a probabilistic ranking principle as discussed earlier (Section 3), which assumes that:

- The probability distribution is Gaussian,
- Variances are equal in all directions (if the PC space is not scaled according to the eigenvalues), and
- Only distances in the PC space are significant–all others can be safely ignored.

These characteristics of the PCA method may account (singly or together) for the fact that Burrows's Delta often works better in practice than PCA, despite its strong assumption of word frequency independence. The Laplace distribution may be a better approximation of the true distribution of mean word frequencies than the Gaussian, accounting for differences in frequency variation among different words may by critical, and two dimensions may simply not be enough to capture the true structure of the space (though the success of linear discrimination methods (Diederich *et al.*, 2003; Baayen *et al.*, 2002) can argue against this last notion). Use of the (approximate) Laplace-based method given immediately above may, therefore, enable more accurate attribution by combining the benefits of PCA with those of Delta.

5 Discussion

We have shown how Burrows's Delta measure for authorship attribution may be viewed as an

approximation to ranking candidates by probability according to an estimated Laplace distribution. This view leads directly to some theoretically well-founded extensions and generalizations of the method based on using Gaussian distributions in place of Laplace distributions, as well as correcting for statistical correlations between the various word frequencies being used. The choice among these various methods, of course, is an empirical question, which we will address in the sequel to this article, by applying these methods to several previously examined authorship problems.

In addition to giving several justifiable variants to the method, this view of Delta also gives a clearer idea of its assumptions and likely limitations. The method clearly assumes that word frequencies for all authors are distributed with similar spreads, only differing in the central value (which is taken from a relevant comparison document). In cases where only one or two documents are available from a given author, this assumption is virtually unavoidable; however, when more documents are available, they should be used to adjust the estimates of the likely spread in frequencies of the various words under consideration. More significantly, this assumption appears to fundamentally limit use of the method when all the samples (from all authors) are of pretty much the same textual variety, otherwise we would expect the word frequency distributions over the comparison set to be a mixture of several disparate distributions, one for each genre found in the set, thus potentially biasing results depending on the variety of the test text.

Acknowledgements

Thanks to David Hoover, Mark Olsen, and the anonymous reviewers for their readings and helpful comments on earlier drafts of this article.

References

Baayen, H., van Halteren, H., Neijt, A., and Tweedie, F. (2002). An Experiment in Authorship Attribution. In Morin, A. and Sébillot, P. (eds), *JADT 2002: Journées internationales d'Analyse statistique des données textuelles*.

Binongo, J. N. G. and Smith, M. W. A. (1999). The application of principal components analysis to stylometry. *Literary and Linguistic Computing*, **14**: 445–65.

Burrows, J. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, **7**: 91–109.

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

Burrows, J. (2003). Questions of authorship: Attribution and beyond; a lecture delivered on the occasion of the Roberto Busa Award ACH-ALLC 2001, New York. *Computers and the Humanities*, **37**(1): 5–32.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, **19**(1): 109–23.

Evans, M., Hastings, N. A. J., and Peacock, J. B. (2000). *Statistical Distributions*. New York: Wiley.

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. (2005). Delta, Delta Prime, and modern American poetry: Authorship attribution theory and method. *Proceedings of the 2005 ALLC/ACH conference,* Victoria, BC.

Keynes, J. M. (1911). The principal averages and the laws of error which lead to them. *Journal of the Royal Statistical Society*, **74**: 322–8.

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Boston: Birkhäuser.

Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics.

Wettschereck, D., Aha, D. W., and Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, **11**(1–5): 273–314.

Notes

1 Named after mathematician Johann Carl Friedrich Gauss (1777–1855).

2 The arrow notation in $\vec{\mathbf{f}}$, and $\vec{\sigma}$ indicates that these quantities here are *vectors* containing the values of f, and σ for all words w_i; i.e. $\vec{\mathbf{f}}(D) = \langle f_1(D), f_2(D), \cdots, f_n(D)\rangle$. Also see Appendix I.

3 Named after mathematician and physicist Pierre-Simon Laplace (1749–1827).
4 Even assuming overwhelming empirical evidence to be adduced for Delta's efficacy, the underlying gap between the method's theoretical assumptions and the known facts about language could still raise serious doubts about its validity in novel cases, were the discrepancy not explained.
5 After its originator, French mathematician and philosopher René Descartes (1596–1650).

Appendix 1

1 Some Key Concepts in Linear Algebra

The mathematics of *linear algebra* studies analytical methods deriving from the problem of finding a solution to multiple algebraic equations in multiple unknowns simultaneously. The theory is also useful for geometric reasoning, based on Cartesian[5] principles of analytic geometry. This appendix perforce must be a rather superficial overview; an excellent comprehensive text is Meyer's (Meyer, 2000).

1.1 Matrices and vectors

The two key notions are the *vector* and the *matrix*. A vector represents a point in a multi-dimensional space as a list of n numbers, each number giving a position along one of n notional axes (on some arbitrary scale). A vector may also be viewed as an arrow from the origin to its position. Conventionally, we notate a vector as $\vec{x}$, and view it as a vertical column of its elements:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}$$

A fundamental operation between vectors is the *scalar product*, notated $\vec{x}^T\vec{y}$, defined to be the sum of the products of the corresponding elements of the vectors (the operation is only defined if both vectors have the same number of elements). The scalar product of a vector with itself is just the vector's length (in Euclidean terms) squared. The scalar product of two different vectors can be shown to be equal to the cosine of the angle between the vectors times the product of the two vectors' lengths. Thus, holding length equal, vectors that point in roughly the same direction will have a larger scalar product than vectors that point in different directions; the scalar product of vectors at right angles to one another is always 0.

A matrix, on the other hand, represents a *linear transformation* of the n-dimensional space into a (possibly different) m-dimensional space, and may be viewed as comprising an $n \times m$ array of numbers, as:

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}$$

A matrix is viewed as a transformation of the space, via the operation of *matrix multiplication*, where a matrix is used as a function to move points (i.e. vectors) in the space. In matrix multiplication, a new m-dimensional vector $\vec{z}$ is constructed from a given n-dimensional vector $\vec{x}$ by taking the scalar product of each *row* of the matrix (viewed as a vector) with $\vec{x}$ to form each element of the new vector $\vec{z}$. In mathematical terms:

$$\begin{array}{ccc} \vec{z} & = & \mathbf{A}\vec{x} \\ \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} & = & \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots A_{1n}x_n \\ A_{21}x_1 + A_{22}x_2 + \cdots A_{2n}x_n \\ \vdots \\ A_{m1}x_1 + A_{m2}x_2 + \cdots A_{mn}x_n \end{bmatrix} \end{array}$$

Multiplication of an $n \times m$ matrix $\mathbf{A}$ by an $k \times n$ matrix $\mathbf{B}$ is defined similarly to give a $k \times m$ matrix $\mathbf{C}$, notated $\mathbf{C} = \mathbf{AB}$, by defining each column of $\mathbf{C}$ to be the product of $\mathbf{A}$ with each corresponding column of $\mathbf{B}$.

If both dimensions of a matrix are equal ($m = n$), then the matrix is *square*, and it transforms points in n-dimensional space to other points in n-dimensional space. A special kind of square matrix is the *identity matrix*, notated $\mathbf{I}$, whose

elements are all zero except for those along the diagonal which are all one (the dimensionality is assumed known). It serves the same function in linear algebra as the number 1 does in regular algebra. Some square matrices $\mathbf{A}$ have an inverse $\mathbf{A}^{-1}$ defined such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, such a matrix is called *invertible*; other matrices are termed *singular*.

A square matrix can be used also to define an arbitrary quadratic function over points in the n-dimensional space, by generalizing the notion of scalar product:

$$\begin{aligned}\vec{x}^T\mathbf{A}\vec{x} = {} & x_1(A_{11}x_1 + A_{12}x_2 + \cdots A_{1n}x_n) + \\ & x_2(A_{21}x_1 + A_{22}x_2 + \cdots A_{2n}x_n) + \\ & \vdots \\ & x_n(A_{n1}x_1 + A_{n2}x_2 + \cdots A_{nn}x_n) \\ = {} & A_{11}x_1^2 + A_{12}x_2x_1 + \cdots A_{1n}x_nx_1 + \\ & A_{21}x_1x_2 + A_{22}x_2^2 + \cdots A_{2n}x_nx_2 + \\ & \vdots \\ & A_{n1}x_1x_n + A_{n2}x_2x_n + \cdots A_{nn}x_n^2\end{aligned}$$

1.2 Eigenvalue decomposition

If we view an $n \times n$ matrix $\mathbf{A}$ as a transformation on an n-dimensional vector space (via the matrix product, $\vec{\mathbf{y}} = \mathbf{A}\vec{\mathbf{x}}$), we may ask 'Are there any vectors whose direction does not change when multiplied by $\mathbf{A}$ (though its length might)?' That is, is there any vector $\vec{\mathbf{x}}$ such that $\mathbf{A}\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$ for some value of λ? This question leads to an extremely fruitful area of linear algebra, via the twin notions of the *eigenvector* and the *eigenvalue*. The essential definition is this:

> Given an $n \times n$ matrix $\mathbf{A}$, an n-dimensional vector $\vec{\mathbf{x}}$ with at least one nonzero element is an *eigenvector* of $\mathbf{A}$ with associated *eigenvalue* $\lambda \neq 0$, if $\mathbf{A}\vec{\mathbf{x}} = \lambda\vec{\mathbf{x}}$.

The eigenvectors of a matrix thus form, in a sense, the fundamental 'modes' of the matrix, viewed as a vector-space transformation. If $\mathbf{A}$ is invertible (i.e. if $\mathbf{A}^{-1}$ exists), then it has n nonzero eigenvalues (hence n eigenvectors); if it is non-invertible, then there will be fewer nonzero eigenvalues and associated eigenvectors.

Note that multiplying any eigenvector by a number (i.e. stretching it) will also give an eigenvector. Thus, we can standardize eigenvectors to all have unit length, such that for any eigenvector $\vec{\mathbf{e}}$, we have that $\vec{\mathbf{e}}^T\vec{\mathbf{e}} = 1$. To avoid confusion, we will call these *unit eigenvectors*.

Also note that any two eigenvectors $\vec{\mathbf{e}}_i$ and $\vec{\mathbf{e}}_j$ will be *orthogonal* to each other (that is, perpendicular), such that their dot-product is zero: $\vec{\mathbf{e}}_i^T\vec{\mathbf{e}}_j = 0$.

Thus, we may define the *eigenvector matrix* $\mathbf{E}(\mathbf{A})$ (or simply $\mathbf{E}$ when not ambiguous) for $n \times n$ matrix $\mathbf{A}$ with $m \leq n$ eigenvectors to be the $n \times m$ matrix whose columns are $\mathbf{A}$'s unit eigenvectors. If $\mathbf{A}$ is invertible (i.e. has n eigenvectors), then $\mathbf{E}(\mathbf{A})$ is a square matrix such that $\mathbf{E}(\mathbf{A})^T\,\mathbf{E}(\mathbf{A}) = \mathbf{I}$, that is, its transpose is its inverse (this follows directly from the two properties noted just above).

It can then be shown that $\mathbf{E}^T\mathbf{A}\mathbf{E} = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix (one all of whose elements are zero except along the diagonal), whose diagonal elements are the eigenvalues of $\mathbf{A}$.