Document Author Classification using Generalized Discriminant Analysis

Todd K. Moon, Peg Howland, Jacob H. Gunther
Utah State University

Abstract

Classification by document authorship based on statistical analysis — stylometry — is considered here by using feature vectors obtained from counts of *all* words in the intersecting sets of the training data. This differs from some previous stylometry, which used only selected "noncontextual" words with the highest counts, and also from conventional text search techniques, where noncontextual words are frequently left out when the term-by-document matrices are formed. The dimensionality of the resulting vector is reduced using a generalized discriminant analysis (GDA). The method is tested on three sets of documents which have been previously subjected to statistical analysis. Results show that the method is successful at identifying author differences and at classifying unknown authorship, consistent with previous techniques.

Keywords: author identification; LDA/GSVD; stylometry.

1 Introduction and Background

It has been suggested (see, e.g., [1, 2]) that authors leave tell-tale footprints in their writings indicative of authorship, which can be revealed by an appropriate statistical analysis. Following [1], we refer to such methods of authorship study as stylometry, or stylometric analysis. Stylometry is based on the assumptions that authors unconsciously use some word patterns in a manner more or less consistent across documents and across time and that, because the use of these words is unconscious, even imitators can be distinguished from the authors they would imitate. Extensive testing of stylometric analysis on works by various authors has provided at least partial validation of the underlying assumptions. For example, Sir Walter Scott showed little statistical variation in his style, even over a career interrupted by five strokes [1, Chapter 10]. And in a series of statistical tests, the author Robert Heinlein's signature uniquely showed through even when he was writing as two different narrators in *The Number of the Beast* [3, p. 106].

Statistical analysis of documents goes back to Augustus de Morgan in 1851 [4, p. 282], [1, p. 166], who proposed that word length statistics might be used to determine the authorship of the Pauline epistles. Since that initial proposal (not actually carried out by de Morgan), the Bible has been subjected to extensive statistical scrutinies, many of them reaching conflicting conclusions. Stylometry was also employed as early as 1901 to explore the authorship of Shakespeare [5]. Since then, it has been employed in a variety of literary studies (see, e.g., [6, 7, 8]), including twelve of *The Federalist* papers which were of uncertain authorship [9] (which we re-examine here), and an unfinished novel by Jane Austen (which we also re-examine here). Information theoretic techniques have also recently been used [10].

Stylometry is usually based on "noncontextual words," words which do not convey the primary meaning of the text, but which act in the background of the text to provide structure and flow. Noncontextual words are at least plausible, since an author may address a variety of topics, so particular distinguishing words are not necessarily revealing of authorship. As stated in [11]:

> The noncontextual words which have been most successful in discriminating among authors are the filler words of the language such as prepositions and conjunctions, and sometimes adjectives and adverbs. Authors differ in their rates of usage of these filler words.

(However, statistical analysis based on author vocabulary size *vs.* document length — the "vocabulary richness" — has also been explored [12].) In noncontextual word studies, a restricted set of "most common" words is selected [1], and documents are represented by word counts, or ratios of word counts to document length. As a variation, sets of ratios of counts of noncontextual word patterns to other word patterns are also employed [3]. However, it has largely been a matter of investigator choice which words are selected as noncontextual, opening the stylometric analysis to criticisms of nonobjectivity.

In this work, we examine *all* of the words in the intersection of the documents in question. This results in a higher dimensional space than has been conventional. The dimensionality is handled, however, using a generalized discriminant analysis (GDA) computed using the generalized singular value decomposition (GSVD).

The use of term-document indexing and latent semantic indexing (via the SVD) for document search and classification is by now very widespread (see, e.g., [13, 14]). In most instances, however, “uninformative” words — noncontextual words such as “the,” “a”, “that,” “and,” and the like — are not included in the term-by-document matrices, since they provide little information by which to distinguish documents by content. However, from the point of view of stylometry, such words are precisely those of interest since they allow for the possibility of author classification.

In this paper we present the idea of the GDA for author identification. The method is validated experimentally by performing some author identification tests on three sets of documents that have been classified by other stylometric analyses.

2 Problem Statement

The basic problem we address here is this: Given a set of documents alleged to have been written by one author, and another set of documents alleged to have been written by another author, determine if this is a valid allegation. A variation on this theme is as follows: given a document whose author is (assumed to be) a member of a finite class of authors, detect the author. In this paper, we simply consider binary comparisons, reserving more general classification problems for future research.

In addition to the stated goals, the idea is to make the classification based on the *style* of the authors’ writings, as opposed to the content of the document. A single author may address a multitude of subjects, but so may other authors, and the presence of absence of any particular word or set of words may be more indicative of topic than of author. A careful author will also resort to dictionary use, so even vocabulary richness is suspect as an indicator of authorship.

Let k denote the number of alleged authors (“classes”) of a series of training documents $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$, and let $n_i, i = 1, 2, \ldots, k$ be the number of documents attributed to author i, with $\sum_{i=1}^k n_i = n$. Let $\mathcal{W}_i$ be the set of words in $\mathcal{D}_i$. In this work, documents are compared on the basis of words they have in common. Let $\mathcal{W} = \cap_{i=1}^n \mathcal{W}_i$ be the set of words common to all documents, with $|\mathcal{W}| = d$. Denote by $\mathcal{D}_j \cap \mathcal{W}$ the subset of the document $\mathcal{D}_j$ with words in $\mathcal{W}$.

Loosely, the count vector $\mathbf{m}_j$ (a column vector) for the document $\mathcal{D}_j$ is formed from the words in $\mathcal{D}_j \cap \mathcal{W}$, with rows of $\mathbf{m}_j$ indexed by the words in $\mathcal{W}$. That is, $\mathbf{m}_j$ is a column of a term-by-document matrix, but *without* the usual set of “insignificant” stop words removed. The rationale for using the intersecting words is to make the method somewhat more context independent; the goal here is to separate authors on the basis of their writing *style*, not the basis of the topic they are writing about.

More precisely, it may be necessary to stem the words in the documents, treating noun pluralizations and verb tenses in a consistent way across documents, then form the count vectors from the list of stemmed words. (In some of the tests done, stemming has not been a significant issue in the words in the intersection.) Let $\tilde{\mathbf{m}}_j = \mathbf{m}_j / N_j$, where N_j is a normalization constant to be discussed below. We say that the vector $\tilde{\mathbf{m}}_j$ is derived from the document $\mathcal{D}_j$.

Let $A_i \in \mathbb{R}^{d \times n_i}$ be the matrix formed by stacking the n_i vectors $\tilde{\mathbf{m}}_j$ associated with author i, and let A be the $d \times n$ matrix

$$A = \begin{bmatrix} A_1 & A_2 & \cdots & A_k \end{bmatrix} \overset{\triangle}{=} \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix}.$$

Let $\mathcal{N}_i, i = 1, 2, \ldots, k$ be the set of column indices of A associated with author i.

With this data structure in mind, we can pose two problems:

- Given a vector $\tilde{\mathbf{v}} \in \mathbb{R}^d$ derived from a document $\mathcal{D}$ written by one of the authors, determine the author. This is a pattern recognition problem.
- Somewhat more problematically, given a set of documents $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$ alleged to have been written by a set of authors, determine if this is a valid allegation.

3 LDA/GDA for small sample size problems [15, 16]

Let $\mathbf{c}_i, i = 1, 2, \ldots, k$ denote the centroid of class i, $\mathbf{c}_i = \frac{1}{n_i} \sum_{j \in \mathcal{N}_i} \mathbf{a}_j$, and let $\mathbf{c} = \frac{1}{n} \sum_{j=1}^n \mathbf{a}_j$ denote the overall centroid. The within-cluster, between-cluster, and mixture scatter matrices are defined as [17, 18]

$$S_w = \sum_{i=1}^k \sum_{j \in \mathcal{N}_i} (\mathbf{a}_j - \mathbf{c}_i)(\mathbf{a}_j - \mathbf{c}_i)^T,$$

$$S_b = \sum_{i=1}^k \sum_{j \in \mathcal{N}_i} (\mathbf{c}_i - \mathbf{c})(\mathbf{c}_i - \mathbf{c})^T$$

$$= \sum_{i=1}^k n_i (\mathbf{c}_i - \mathbf{c})(\mathbf{c}_i - \mathbf{c})^T$$

and

$$S_m = \sum_{j=1}^n (\mathbf{a}_j - \mathbf{c})(\mathbf{a}_j - \mathbf{c}).$$

Applying a linear transformation $G^T \in \mathbb{R}^{l \times d}$ to the data matrix A to produce $\tilde{A} = G^T A$ yields scatter matrices of the transformed data

$$G^T S_w G, \qquad G^T S_b G, \qquad G^T S_m G,$$

respectively. To reduce the operational complexity, it is desirable to choose l such that $l \ll d$. A reasonable goal is to find a transformation G^T which produces small (measured with respect to some norm) within-cluster scatter $G^T S_w G$ while producing large between-cluster scatter and while reducing the dimension of the transformed data.

It is analytically attractive to use the trace as a measure of scatter, that is,

$$\operatorname{tr}(S_w) = \sum_{i=1}^{k} \sum_{j \in \mathcal{N}_i} \|\mathbf{a}_j - \mathbf{c}_i\|_2^2$$

and

$$\operatorname{tr}(S_b) = \sum_{i=1}^{k} \sum_{j \in \mathcal{N}_i} \|\mathbf{c}_i - \mathbf{c}\|_2^2.$$

We desire to compute G to minimize $\operatorname{tr}(G^T S_w G)$ while simultaneously maximizing $\operatorname{tr}(G^T S_b G)$. This joint optimization is approximated by maximizing

$$J(G) = \operatorname{tr}((G^T S_w G)^{-1} G^T S_b G).$$

This criterion cannot be applied, however, when S_w is singular, as occurs when $d > n$, which is typical for document processing. We use the generalized singular value decomposition (GSVD) in this case. We will do this using a factored representation.

Define the matrices $H_w \in \mathbb{R}^{d \times n}$, $H_m \in \mathbb{R}^{d \times n}$ and $H_b \in \mathbb{R}^{d \times k}$, by

(3.1)

$$H_w = \begin{bmatrix} A_1 - \mathbf{c}_1 \mathbf{e}_1^T & A_2 - \mathbf{c}_2 \mathbf{e}_2 & \cdots & A_k - \mathbf{c}_k \mathbf{e}_k^T \end{bmatrix}$$

$$H_m = \begin{bmatrix} (\mathbf{c}_1 - \mathbf{c}) \mathbf{e}_1^T & (\mathbf{c}_2 - \mathbf{c}) \mathbf{e}_2^T & \cdots & (\mathbf{c}_k - \mathbf{c}) \mathbf{e}_k^T \end{bmatrix}$$

and

(3.2)

$$H_b = \begin{bmatrix} \sqrt{n_1}(\mathbf{c}_1 - \mathbf{c}) & \sqrt{n_2}(\mathbf{c}_2 - \mathbf{c}) & \cdots & \sqrt{n_k}(\mathbf{c}_k - \mathbf{c}) \end{bmatrix},$$

where $\mathbf{e}_i = (1, 1, \ldots, 1) \in \mathbb{R}^{n_i \times 1}$. Then

$$S_w = H_w H_w^T \qquad S_b = H_b H_b^T$$

and

$$S_m = H_m H_m^T.$$

From classical discriminant analysis [17], it is known that when S_w is nonsingular, the columns of G maximizing $J(G)$ are the eigenvectors of $S_w^{-1} S_b$ corresponding to the l largest eigenvalues; the columns of G are thus the eigenvectors $\mathbf{x}_i$ in

$$S_w^{-1} S_b \mathbf{x}_i = \lambda_i \mathbf{x}_i \tag{3.3}$$

and the maximum value achieved is $J(G) = \lambda_1 + \lambda_2 + \cdots + \lambda_l$. This straightforward solution must be modified, however, when S_w is singular.

To treat the singular S_w case, express (3.3) as

$$\beta_i^2 S_b \mathbf{x}_i = \alpha_i^2 S_w \mathbf{x}_i \tag{3.4}$$

with $\lambda_i = \alpha_i^2 / \beta_i^2$. This can be expressed in factored form as

$$\beta_i^2 H_b H_b^T \mathbf{x}_i = \alpha_i^2 H_w H_w^T \mathbf{x}_i.$$

This is now in a form amenable to solution using the GSVD.

The GSVD of the matrix pair (H_b^T, H_w^T) finds a $k \times k$ orthogonal matrix U, a $n \times n$ orthogonal matrix V, a $d \times d$ matrix X of the form

$$X = Q \begin{bmatrix} R^{-1} W & \\ & I_{d-t} \end{bmatrix}$$

and matrices

$$\Sigma_b = \begin{bmatrix} I_r & & \\ & D_{b,s} & \\ & & 0_{k-r-s \times t-r-s} \end{bmatrix} \in \mathbb{R}^{k \times t}$$

$$\Sigma_w = \begin{bmatrix} 0_{n-t+r \times r} & & \\ & D_{w,s} & \\ & & I_{t-r-s} \end{bmatrix}$$

satisfying

$$U H_b^T X = \begin{bmatrix} \Sigma_b & 0_{k \times (d-t)} \end{bmatrix}$$

and

$$V H_w^T X = \begin{bmatrix} \Sigma_w & 0_{t \times (d-t)} . \end{bmatrix}$$

Here,

$$t = \operatorname{rank} \begin{bmatrix} H_b^T \\ H_w^T \end{bmatrix},$$

$$r = t - \operatorname{rank}(H_w^T)$$

and

$$s = \operatorname{rank}(H_b^T) + \operatorname{rank}(H_w^T) - 1$$

and the matrices $D_{b,s}$ and $D_{w,s}$ are (not generally the same) diagonal $s \times s$ matrices, and the 0 and and I matrices are 0 and identity matrices of the indicated dimensions. It is straightforward to show that the assignments

$$\begin{aligned} i = 1, 2, \ldots, r &: \alpha_i = 1,\ \beta_i = 0 \\ i = r+1, \ldots, r+s &: \alpha_i = [D_{b,s}]_{i,i}, \\ & \quad \beta_i = [D_{w,s}]_{i,i} \\ i = r+s+1, \ldots, t &: \alpha_i = 0,\ \beta_i = 1 \\ i = t+1, t+2, \ldots, d &: \alpha_i \text{ and } \beta_i \text{ arbitrary} \end{aligned}$$

results in a solution to (3.4). The columns of X are the generalized singular vectors for the matrix pair (H_b^T, H_w^T). The dimension-reducing transformation G is obtained by taking the first l columns of X.

The generalized discriminant analysis allows the dimension of the data to be reduced from d to l. This frequently results in improved performance because dimensions which are primarily noiselike are not used in the decision. Changing the dimension raises the question of how big l should be. One of the particular advantages of the approach employed here is that the dimension can be theoretically (as opposed to empirically) determined. It is known [19] that if $l = \mathrm{rank}(H_b)$ then no information will be lost from among the clusters. From a practical point of view, setting $l = k - 1$ (where k is the number of classes) avoids the need to compute $\mathrm{rank}(H_b)$ and provides essentially equivalent results, since $\mathrm{rank}(H_b) \leq k - 1$, and including extra columns "will have approximately no effect on cluster preservation." [15, p. 280]. This was validated in the experiments below, where for binary classification ($k = 2$), setting $l = 1$ gave equivalent results to $l = 4$.

The LDA/GSVD algorithm is summarized in Algorithm 1. It follows the construction of the Paige and Saunders [20] proof, but only computes the necessary part of the GSVD. The most expensive step of LDA/GSVD is the complete orthogonal decomposition of the composite H matrix in Step 2. When $\max(k, n) \ll d$, the SVD of $H = [H_b^T, H_w^T] \in \mathbb{R}^{(k+n)\times d}$ can be computed by first computing the reduced QR decomposition $H^T = Q_H R_H$, and then computing the SVD of $R_H \in \mathbb{R}^{(k+n)\times(k+n)}$ as

$$R_H = Z \begin{pmatrix} \Sigma_H & 0 \\ 0 & 0 \end{pmatrix} P^T.$$

This gives

$$H = R_H^T Q_H^T = P \begin{pmatrix} \Sigma_H & 0 \\ 0 & 0 \end{pmatrix} Z^T Q_H^T,$$

where the columns of $Q_H Z \in \mathbb{R}^{d\times(k+n)}$ are orthonormal. There exists orthogonal $Q \in \mathbb{R}^{d\times d}$ whose first $k + n$ columns are $Q_H Z$. Hence

$$H = P \begin{pmatrix} \Sigma_H & 0 \\ 0 & 0 \end{pmatrix} Q^T,$$

where there are now $d - t$ zero columns to the right of Σ_H. Since $R_H \in \mathbb{R}^{(k+n)\times(k+n)}$ is a much smaller matrix than $H \in \mathbb{R}^{(k+n)\times d}$, the required memory is substantially reduced. In addition, the computational complexity of the algorithm is reduced to $\mathcal{O}(mn^2) + \mathcal{O}(n^3)$ [21], since this step is the dominating part.

4 Some Experimental Methods and Results

In the classification, centroids are computed for each author class $\mathbf{c}_i$, using either all the data ("testing on the training data") or in a cross-validation or "leave-one-out" mode. Classification is done using nearest neighbor measurements with Euclidean distance.

To perform an initial validation of the GDA author identification technique, we have re-examined some classification experiments that have been previously done. Results are comparable to those of the previous analysis.

4.1 Textual Analysis of *Sanditon*

Up until recently before her death in 1817, Jane Austen was working on a novel posthumously titled *Sanditon* [22, p. 20]. Before her death she completed a draft of twelve chapters (about 24,000 words). The novel was posthumously "completed" by various writers, with varying success. The version best known was published in 1975 [23], coathored by "Another Lady," who remains unknown. Whoever she was, she was a fan of Austen's and attempted to mimic her style. Of this version, it was said, it "received, as compared with [its] predecessors, a warm reception from the English critics." [24, p. 76]. Notwithstanding its literary appeal and the attempts at imitating the conscious habits of Austen, she failed in capturing the unconscious habits of detail: stylometric analysis has been able to distinguish between the different authors [1, Chapter 16].

4.1.1 Textual Processing

We obtained a computer-readable document from the Electronic Text Center at the University of Virginia Library [25]. HTML tags and punctuation were removed and all words were converted to lower case. The document was evidently obtained from OCR from scanned documents, so it was necessary to carefully spell-check the document, but English and contemporary spellings were retained. Stemming of the document (for pluralizations and verb tense) was not done in this experiment.

The documents were scanned by a program written in `Python` which divided the texts into chapters and by author, with Author 1 encompassing chapters 1 through 12 (25,720 total words; 3729 distinct words), and Author 2 chapters 13 through 30 (75,974 total words; 6967 distinct words). After intersecting the words of Author 1 and Author 2 (to establish a more context-free set of words), 2518 distinct words in common were retained. These were used to form the data vectors for the 30 documents comprised of the individual chapters. Word counts for each chapter were normalized by the total number of words in each chapter (before intersection).

4.1.2 Tests and Results

Three tests were performed on the data.

Test 1 Testing on the training data (**resubstitution**). Centroids for each of the two classes were obtained using all of the data for each class. Then each column vector was classified in turn using minimum Euclidean distance. This is referred to as the LDA method. Then an l-dimensional representation was obtained using the GSVD for the generalized discriminant analysis (GDA). The value $l = 1$ is sufficient for theoretical reasons. However, to confirm the theory experimentally, the value $l = 4$ was also chosen and the experiments re-performed. Each column vector was classified in turn, again using minimum Euclidean distance. This is the GDA method. Numerical results are summarized in Table 1.

Test 2 Testing on non-training data (**cross-validation** or leave-one-out training). For each column vector, centroids were obtained for each class leaving that column vector out. Then the column vector was classified using this data. This was repeated for a 4-dimensional representation.

In each of these tests, the recognition rates were improved by the generalized discriminant analysis. This raises the possibility that perhaps the algorithm itself introduces structure into the data, allowing patterns to be recognized where, in fact, there are no patterns present. As a check on this possibility (which we considered remote from a mathematical perspective, but wanted to eliminate regardless), a third test was performed.

Test 3 Randomized columns. The columns of the data matrix A were randomized, then the cross-validation test was performed on the resulting matrix. This test was performed for 30 trials, with a different randomization each trial. If the generalized discriminant analysis is the cause of the good recognition accuracy, then very good recognition results should result. However, as the data in Table 1 indicate, the probability of misclassification is near 0.5, with the generalized discriminant analysis being slightly better. (This indicates that the generalized discriminant analysis *does* actually introduce some structure to the problem, but not enough for completely misleading classifications.)

As can be seen, the algorithm provides strong classification capability for both the resubstitution and cross-validation method, but nearly 50% probability of error for the randomized column test. From this we conclude:

- The GDA on this data provides a means of distinguishing authorship; and
- There actually is a statistically significant difference between the authors, as measured by this technique.

4.2 Textual Analysis of *The Federalist* *The Federalist* consists of a series of 85 papers written around 1787 by Alexander Hamilton, James Madison, and John Jay in support of the U.S. Federal Constitution [26]. Of these papers, 51 are attributed unambiguously to Hamilton, 14 to Madison, 5 to Jay, and 3 to both Hamilton and Madison. The remaining twelve papers are of uncertain attribution, but are known to be by either Madison or Hamilton. In [9, 10], statistical techniques were used to determine that all twelve ambiguous papers were due to Madison. We will use this as an experiment to validate the GDA technique.

4.2.1 Data Preparation A machine-readable copy of *The Federalist* was obtained from the Gutenberg project [27]. This version had two variants of paper No. 70 by Hamilton; both were retained in the data set. Footnotes, punctuation, and capitalization were removed. The papers were read and counts for each author were obtained using a `Python` program: Hamilton had 113884 total words; Madison had 38709; Jay had 8374; Hamilton and Madison had 5613; and Hamilton or Madison had 23944. Each paper constitutes a document, so there are 52 Hamilton documents and 14 Madison documents in the training set. The word lists for Hamilton, Madison, and "Hamilton or Madison" were intersected (to establish a more content-free set of words), resulting in a word list of 1497 words.

4.2.2 Tests and Results The three tests described in section 4.1.2 were performed on this data, with results as shown in Table 2. (Tests 1 and 2 use $l = 4$; Test 3 was computed with 10 randomizations of the columns.) A fourth test was run: classifying the twelve papers of uncertain authorship. In all cases, the unknown documents classified as "Madison."

4.3 Textual analysis of *The Book of Mormon* *The Book of Mormon* is a document regarded by The Church of Jesus of Christ of Latter Day Saints as scripture on par with the Bible and, like the Bible, has been subjected to numerous stylistic analyses. By its own account, the book is the translation by a nineteenth century American, Joseph Smith, of a compilation of ancient records set in Mesoamerica in the period from 600 B.C. to approximately 400 A.D. *The Book of Mormon* was first published in 1830. The first portion of the book is attributed to a writer named Nephi, whose writings were placed verbatim into the compilation. Most of the remainder of the book is due to Mormon (hence the name of the book), who wrote a narrative to tie together the historical account, and interspersed this narrative with primary historical and ecclesiastical documents of other writers. *The Book of Mormon* is a historical narrative interspersed with homiletic discourse. The document is complicated from an analytical point view. For example, in some parts, there is dialogue taking place which was probably not transcribed first hand but is reported by persons present, which is then summarized and compiled and written down by Mormon, then finally translated by Joseph Smith. We avoided difficulty by using text that could be unambiguously attributed to the author Mormon.

Several different authors can be identified from *The Book of Mormon* text. In an initial stylometric analysis [11] (dubbed a "wordprint" by its authors), 24 distinct authors/speakers accounting for 91.9% of the text were identified (with other authors contributing the remainder). In [11], analysis was performed based on a small set of commonly occurring words. The conclusion they reached was that the 24 authors were statistically distinguishable, and were also distinguishable from other nineteenth century authors. Since that initial analysis, criticisms [28] and challenges [12] have been made. In the latter, vocabulary richness was employed as a measure, and a statistical distinction between authors could not be obtained, suggesting the authorship and fabrication by the "translator" Joseph Smith. (Interestingly, in [29], vocabulary richness was also unable to distinguish between authors in *The Federalist*, so it seems that vocabulary richness may not be an appropriate measure for some texts.) A careful analysis based on word count ratios in [3] compared only two major authors, Nephi and Mormon and concluded that there *is* a statistical difference between their writing styles. This analysis has not (to our knowledge) been challenged. We therefore use this "known" difference of authorship to test our GDA technique. As for the other tests in this paper, only a binary comparison was performed, between the Nephi subdocuments and the Mormon subdocuments.

The fact that the *Book of Mormon* is (alleged to be) a translated document, rather than written in a primary language, rather complicates and enriches its analysis. Some study on translated documents appears in [3], where some results on translation from German are reported. And many of the biblical stylometric analyses have been performed on translations (which may account in part for differing conclusions reached by the various studies). Certainly this is an area where considerably more investigation and validation are warranted before firm conclusions may be reached.

4.3.1 Data Preparation A machine-readable copy of the *The Book of Mormon* was obtained from The Gutenberg Project [30] (edition unknown). This was labeled to indicate authorship. For authors whose writings exceed 5000 words, the document was further marked so that textual portions of approximately 5000 words are provided; we call these "subdocuments." For the authors of interest in these experiments, Nephi and Mormon, there were 5 subdocuments of Nephi text and 16 subdocuments of Mormon text, with Nephi having 27474 words total (1828 unique) and Mormon having 98446 words total (3544 unique). The words in "And it came to pass" were not counted among the totals, since this phrase — as common as a punctuation marker and probably serving a similar purpose in the original ancient language — was eliminated from consideration. A `Python` program read the data in and provided word counts. The total number of words in the book is 267,239, with 5599 distinct words.

An intersection of all the words common to all 21 blocks of data was found, resulting in the 105 words shown in table 3. (From these words, observe that plural/tense stemming is unnecessary.) Ratios of word counts to total number of words in each document were then computed and used to form the training vectors of dimension 105.

4.3.2 Tests and Results Tests 1, 2 and 3 described in section 4.1.2 were performed on this data (again with $l = 4$, except that Test 3 was performed with 100 random column permutations), with the results summarized in Table 4. In this case a single misclassification occurs on document "Mormon 15" for both the full and reduced dimensionality.

5 Conclusions and Extensions

This paper has introduced the use of generalized discriminant analysis for author classification. The dimension-reducing transformation allows the mathematics to weight which elements are most significant for pattern recognition purposes, eliminating subjective decisions and bias.

By tests performed on three previously-analyzed documents, we have established that the method is successful at identifying author differences, and that the GDA is generally superior to simple nearest neighbor testing.

The capability to deal with high dimensional vectors also opens up for future study a variety of possibilities for elements of feature vectors, besides the word count ratios considered here.

- The *position* of the word in the sentence (e.g., "and" as the first word of the sentence)
- The position of the part of speech (nouns, verbs, gerunds, prepositions) as a function of the position in the sentence: (e.g., a gerund as the first word of the sentence, or appearing in the first quarter of the sentence).
- Adjacent word pairs (e.g., "hardly ever")
- Word pairs not necessarily adjacent (e.g., "since ... because" or "if ..." as compared with "if ... then").

Use of several of these possibilities could lead to very large feature vectors. However, the generalized discriminant analysis weights which features are most significant from a classification point of view, resulting in a much smaller, but still effective, dimensionality. It is anticipated that these extentions will provide for sensitive classification based on smaller text sizes.

These initial efforts also suggest a variety of other studies that can be performed, such as performance as a function of block length. This initial paper also presents results for LDA/GSVD for the author discrimination problem. Follow-on still underway will also address the author discrimination problem using more conventional techniques such as K-NN or support vector machines and compare the methods. Also, the underlying assumptions of stylometry beg for further validation, a validation which is possible because of large databases of texts available.

References

[1] A. Morton, *Literary Detection*. New York: Charles Scribner's Sons, 1978.

[2] A. Ellegard, *A Statistical Method for Determining Authorship*. Gothenburg, 1962.

[3] J. L. Hilton, "On Verifying Wordprint Studies: Book of Mormon Authorship," *Brigham Young University Studies*, 1990.

[4] R. Lord, "de Morgan and the Statistical Study of Literary Style," *Biometrica*, 1958.

[5] T. Mendenhall, "A Mechanical Solution of a Literary Problem," *Popular Science Monthly*, 1901.

[6] C. D. Chretien, "A Statistical Method for Determining Authorship: The Junius Letters," *Languages*, vol. 40, pp. 95–90, 1964.

[7] D. Wishart and S. V. Leach, "A Multivariate Analysis of Platonic Prose Rhythm," *Computer Studies in the Humanities and Verbal Behavior*, vol. 3, no. 2, pp. 109–125, 1972.

[8] C. S. Brinegar, "Mark Twain and the Quintis Curtis Snodgrass Letters: A Statistical Test of Authorship," *Journal of the Americal Statistical Association*, vol. 53, p. 85, 1963.

[9] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison Wesley, 1964.

[10] P. Hanus and J. Hagenauer, "Information Theory Helps Historians," *IEEE Information Theory Society Newsletter*, vol. 55, p. 8, Sept. 2005.

[11] W. A. Larsen, A. C. Rencher, and T. Layton, "Who Wrote the Book of Mormon?," *Brigham Young University Studies*, 1980.

[12] D. Holmes, "A Stylometric Analysis of Mormon Scriptures and Related Texts," *Journal of the Royal Statistical Society, A*, 1992.

[13] M. W. Berry, Z. Drmac, and E. R. Jessup, "Matrices, vector spaces, and information retrieval," *SIAM Review*, vol. 41, no. 2, pp. 335–362, 1999.

[14] M. W. Berry, S. T. Dumas, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Review*, vol. 37, pp. 573–595, Dec 1995.

[15] P. Howland, J. Wang, and H. Park, "Solving the small sample size problem in face recognition using generalized discriminant analysis," *Pattern Recognition*, 2006.

[16] P. Howland, M. Jeon, and H. Park, "Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 1, pp. 165–179, 2003.

[17] K. Fudunaga, *Introduction to Statistical Pattern Recogntition*. New York: Academic Press, 1990.

[18] S. Theodoridis and K. Koutrombas, *Pattern Recognition*. New York: Academic Press, 1999.

[19] P. Howland and H. Park, "Equivalence of several two-stage methods for linear discriminant analysis," in *Proceedings of the Fourth SIAM International Conference on Data Mining*, pp. 69–77, 2004.

[20] C. Paige and M. Saunders, "Towards a generalized singular value decomposition," *SIAM J. Numer. Anal.*, vol. 18, pp. 398–405, 1981.

[21] G. Golub and C. V. Loan, *Matrix Computations*. Johns Hopkins University Press, 3rd ed., 1996.

[22] P. Poplawski, *A Jane Austen Encyclopedia*. London: Aldwych Press, 1998.

[23] J. Austen and A. Lady, *Sanditon*. London: Peter Davies, 1975.

[24] D. Hopkinson, "Completions," in *The Jane Austen Companion* (J. D. Grey, ed.), Macmillan, 1986.

[25] "*Sanditon* (machine readable)." `http://etext.lib.virginia.edu/toc/modeng/public/AusSndt.html`.

[26] A. Hamilton, J. Madison, and J. Jay, "The federalist," in *American State Papers* (R. M. Hutchins, ed.), vol. 43 of *Great Books of the Western World*, pp. 29–266, Encyclopedia Britannica, chicago ed., 1952.

[27] "*The Federalist* (machine readable)." `http://www..gutenberg.org/etext/18`.

[28] D. J. Croft, "Book of Mormon 'Wordprints' Reexamined," *Sunstone*, vol. 6, pp. 15–22, Mar. 1981.

[29] D. I. Holmes and D. Forsyth, "The Federalist Revisited: New Directions in Authorship Attribution," *Literary and Linguistic Computing*, vol. 10, p. 111, 1995.

[30] "*The Book of Mormon* (machine readable)." `http://www..gutenberg.org/etext/17`.

	1: resubstitution	2: cross-validation	3: randomized columns
LDA	6.7	10	49.5
GDA	0.0	6.7	46.0

Table 1: Classification results for *Sanditon* experiments, $l = 1$ or $l = 4$. Numbers show percent misclassification.

	1: resubstitution	2: cross-validation	3: randomized columns
LDA	18.2	18.2	58.2
GDA	0.0	7.6	42.3

Table 2: Classification results for *The Federalist* experiments. Numbers show percent misclassification.

a	according	after	again	against	all	also	among	an	are
as	at	away	be	because	been	before	behold	being	brethren
but	by	cause	children	come	concerning	could	day	did	do
done	down	even	for	forth	from	god	great	had	have
he	heard	him	himself	his	i	if	in	into	know
land	lord	man	manner	many	men	might	more	much	name
nephi	no	not	now	of	on	one	or	out	over
own	people	power	said	saying	shall	should	that	the	their
them	themselves	there	these	they	this	those	thus	time	together
until	unto	up	upon	was	we	were	when	which	who
will	with	words	would	yea					

Table 3: Words common to segments of approximately 5000 words in the writings of Nephi and Mormon

	1: resubstitution	2: cross-validation	3: randomized columns
LDA	4.8	4.8	45.8
GDA	0.0	4.8	41.24

Table 4: Classification results for *Book of Mormon* experiments. Numbers show percent misclassification.

Given a data matrix $A \in \mathbb{R}^{d \times n}$ with k clusters, this algorithm computes the columns of the matrix $G \in \mathbb{R}^{d \times (k-1)}$, which preserves the cluster structure in the reduced dimensional space, and it also computes the $k - 1$ dimensional representation Y of A.

1. Compute $H_b \in \mathbb{R}^{m \times k}$ and $H_w \in \mathbb{R}^{m \times n}$ from A according to (3.2) and (3.1), respectively.
2. Compute the complete orthogonal decomposition of $H = (H_b, H_w)^T \in \mathbb{R}^{(k+n) \times m}$, which is
$$P^T H Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}.$$
3. Let $t = \mathrm{rank}(H)$.
4. Compute W from the SVD of $P(1:k, 1:t)$, which is $U^T P(1:k, 1:t)W = \Sigma_b$.
5. Compute the first $k - 1$ columns of
$$X = Q \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix},$$
and assign them to G.
6. $Y = G^T A$

Algorithm 1: LDA/GSVD