ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/39728891

Quantitative Analysis of Literary Style

Article *in* The American Statistician · August 2002

DOI: 10.1198/000313002100 · Source: OAI

CITATIONS
49

READS
45

2 authors, including:

Nicolas W Hengartner
Los Alamos National Laboratory

134 PUBLICATIONS **2,251** CITATIONS

SEE PROFILE

All content following this page was uploaded by Nicolas W Hengartner on 02 September 2015.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

QUANTITATIVE ANALYSIS OF LITERARY STYLES

ROGER D. PENG AND NICOLAS HENGARTNER

Abstract. In this paper we use canonical discriminant analysis and principal component analysis to analyze literary styles and distinguish authorship. The use of these techniques in conjunction with statistical graphics reduces the need for imposing unrealistic distributional assumptions on the data while still retaining the ability to separate works into authorship groups. We use counts of function words as our units of analysis and find this approach to be effective.

1. Introduction

The analysis of literary style is a field with many opportunities for statisticians. The increasing amount of data available allows for the testing and application of various statistical techniques. Since the problems involved in analyzing literary style are inherently multivariate, classical techniques such as canonical discriminant analysis and principal component analysis can be used. Although it is impossible to identify all of the different factors that go into the writing of a piece of literature, one goal is to find a small set of variables which accurately summarize an author's writing style. Using these variables, we might be able to determine a "fingerprint" for an author that would be useful for discrimination.

Naturally, we are faced with all of the same challenges that any other multivariate analysis problem poses. We need tools to help us visualize and make sense of the data. The purpose of this paper is to evaluate the ability of a technique such as canonical discriminant analysis (CDA) to aid in detecting differences in authorship. CDA helps by providing useful plots and by identifying possible words for discrimination. Although constructing tests with explicit probability distributions is appealing, such tests usually impose restrictions on the data which put

Date: December 8, 2000.

Key words and phrases. authorship, literary style, canonical discriminant analysis, principal component analysis.

the subsequent conclusions into serious doubt. We try to avoid making distributional assumptions about the data as well as making statements about probability or optimality. We feel that a convincing case can be made for the use of CDA and graphics.

2. Previous Work

There has been much research covering different aspects of this field. For a comprehensive review we direct the reader to Holmes (1985). In 1901 T. C. Mendenhall (see Williams, 1975) analyzed the word length distributions of Shakespeare and Bacon in one of the earliest studies of its kind. Williams (1975) noted that Mendenhall was mistaken in some of his conclusions and that there was little evidence for or against the theory that some works written by Shakespeare could have been written by Bacon. Brinegar (1963) also used word length distributions to determine if Mark Twain had written the *Quintus Curtius Snodgrass* (QCS) letters. He used χ^2 tests and two-sample t-tests on the counts of 2, 3, and 4 letter words to check the agreement of the QCS letters with Twain's known writings. Holmes (1985) noted that these tests are based on the dubious assumption that an author's writings are random samples from a fixed frequency distribution of word lengths.

There is no general agreement on the unit of analysis that should be used in authorship studies. Williams (1940) analyzed the sentence lengths of works written by Chesterton, Wells, and Shaw. He noticed that the log of the number of words per sentence appeared to follow a normal distribution. After initially using criteria such as word length and sentence length, Mosteller and Wallace (1963) focused on using function word counts to discriminate between the works of Hamilton and Madison in the *Federalist Papers*. Specifically, they applied Bayes' theorem to the problem as opposed to taking a classical approach to discrimination. Särndal (1967) used word counts to determine the probabilities of Type I and Type II error when distinguishing between two authors. Although it is certainly desirable to quantify the probabilities of error, Särndal makes numerous arbitrary assumptions concerning the distributions of the word counts.

Holmes (1992), in an example of the use of a standard multivariate analysis technique, used hierarchical cluster analysis to detect changes in authorship in Mormon scripture. He used various measures of vocabulary richness to conduct his

analysis. Thisted and Efron (1987) also used the idea of vocabulary richness to determine the possibility of Shakespearean authorship of a newly discovered poem. They based their analysis of the poem on the rate of discovery of new words given the number of distinct words previously observed in the Shakespearean canon.

Each of the studies mentioned above, in particular the earlier ones, were concerned with finding ways to test authorship by computing test statistics and limiting distributions. We do not focus on the testing framework in this paper. Rather, we emphasize the role of graphics and descriptive statistics in analyzing literary style. In fact, one of the earliest examples of the use of graphics in this situation is in Mendenhall's "characteristic curves" showing the distribution of word lengths. While the average word length can be used to compare authors, the shape of the curve was found to be more helpful for discrimination (Williams, 1956).

One advantage to our approach is the ability to analyze multiple authors. Much work has been done in the area of distinguishing between two authors or perhaps two groups of authors. The idea of two possible authors or two groups naturally leads to the framerwork of hypothesis testing (assuming that a given document can only have one author). In a two author situation, one could be discriminating between James Madison and Alexander Hamilton, as in Mosteller and Wallace (1963). In a two group situation, one could be discriminating between 18th century authors and 19th century writers. However, given a block of text of unknown authorship, the choice of two groups may not be so obvious. It might be advantageous to examine multiple groups or authors.

It is important to note that in the cases mentioned above, the central goal was *not* to make a definitive statement of authorship. Rather, the interest of these studies is essentially methodological; to determine the effectiveness of statistical techniques on certain real problems. It seems to be the consensus that true questions of authorship should be left to historians and literary scholars. At best, statistical analysis can lead the scholar in a particular direction or rule out obvious cases.

3. Data

The nine authors selected for this study were Jane Austen, Willa Cather, Charles Dickens, Arthur Conan Doyle, Rudyard Kipling, Jack London, Christopher Marlowe, John Milton, and William Shakespeare. The data were obtained from Internet

websites such as Project Gutenberg and Etext. Multiple works for each selected author were downloaded in text format. Each of the authors' collected works were divided into blocks of 1700 words. Word counts were compiled for every block and tabulated. The number 1700 was chosen after some trial and error. It produced a fair number of "observations" for each author while maintaining the variability in the word counts. Furthermore, it was assumed that a larger block size would decrease the amount of block to block dependence in word counts. Mosteller and Wallace (1964) found in their study that words generally exhibited very little block to block dependence and that serious dependence only occurred in about 20% of the words used. In some instances (i.e. at the ends of documents), there were not enough words to complete a full block of 1700 words. These blocks were discarded. The total number of blocks collected for each author are listed in Table 1.

Author	Dates Lived	# of Blocks
Marlowe	1564-1593	56
Shakespeare	1564-1616	179
Milton	1608-1674	56
Austen	1775-1817	437
Dickens	1812-1870	598
Doyle	1859-1930	552
Kipling	1865-1936	157
Cather	1873-1947	237
London	1876-1916	299

TABLE 1. Total number of blocks collected for each author. Each block contains 1700 words.

We chose 69 words from the subset of the Miller-Newman-Friedman list of function words used by Mosteller and Wallace (1963). These words were chosen because of their relatively frequent use in the works being examined. It was desired to avoid having too many words for which authors had a large number of zero counts.

4. Canonical Discriminant Analysis

The main purpose of this section is to show how canonical discriminant analysis (CDA) can be used to aid in visualizing and quantifying the differences in the writing styles between the authors. CDA is needed to reduce the dimensionality of the original dataset. In that sense, the techinque is similar to principal component analysis. The goal is to summarize the data in a form that can be plotted in two dimensions.

Suppose X is our data matrix of word counts whose columns are centered around their respective means. X is an $n \times p$ matrix, where n is the total number of observations for all the authors being tested and p is the number of variables (i.e. different word types). Let G be the group matrix. G is a $n \times g$ matrix where g is the number of groups we are examining (i.e. the number of authors). Each row of G consists of a single 1 and the rest of the elements are 0's. The "1" represents the group membership for the observation. We can denote the sample total covariance matrix by $C_T = \frac{1}{n}X^TX$. Let $P = G(G^TG)^{-1}G^T$. Then we can denote the between groups covariance matrix as $C_B = \frac{1}{n}X^TPX$. Generally speaking we want to find linear combinations of the variables which maximize the between groups variance subject to a constraint on the total variance. Thus we want to maximze $\beta^TC_B\beta$ over all β satisfying $\beta^TC_T\beta = 1$. This defines a generalized eigenvalue problem and the eigenvectors are the solutions to the CDA problem. The computed β_i's are called the *discriminant functions*. For a given CDA problem, there are $\min(p, g-1)$ non-trivial discriminant functions and the same number of non-zero eigenvalues.

4.1. Variance Proportions.

We can compute an initial measure of the power of each discriminant function to separate the data into groups. The ith canonical correlation coefficient is defined as

$$r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

where λ_i is the eigenvalue for the ith discriminant function. The square of the canonical correlation coefficient can be interpreted as the proportion of variation in the discriminant function explained by the groups (Klecka, 1980). Thus, discriminant functions with r_i^2 values close to 1 are more effective for separating observations into groups. Geometrically, r_i^2 is a measure of the amount of space separating the

p-dimensional clouds of observations corresponding to each author in the ith dimension. Low r_i^2 values corresponds to the clouds being very close to each other and thus, less distinguishable. Unfortunately, in most of the groupings we tried the different authors' clouds of observations were generally close together and were not separated by much space. Although one could see the separation of the groups on the plot of the canonical vectors, the values of the r_i^2 were relatively low.

4.2. **Canonical Vector Plots.** Suppose β_1 is the first discriminant function. Recall that there are $\min(p, g-1)$ non-trivial discriminant functions in each analysis, where p is the number of variables and g is the number groups. If X is the $n \times p$ data matrix, then $y_1 = X\beta_1$ is the first canonical vector. Although it is only feasible to plot two canonical vectors at a time, the first two vectors are usually sufficient for observing separation between the groups.

4.3. **Loadings.** Sometimes it is useful to identify each canonical vector with a specific variable or perhaps a small subset of the total variables. If B is the matrix of discriminant functions, the columns of which are $\beta_1, \ldots, \beta_p$, the loadings are the correlations between the columns of X and XB. We can then identify each canonical vector with the variables which have the highest correlations.

5. Results

In the first example the canonical vectors (CV) are very effective at separating out the three authors, Austen, London, and Shakespeare. One can see on the canonical vector plot (CVP) in Figure 1 that the authors' blocks of text separate readily into three non-overlapping groups. One can see immediately that looking at the individual CV's is misleading. In either direction the CV's show poor separation of the groups. This poor separation is also indicated by the low values of the individual variance proportions: $r_1^2 = 0.48$ and $r_2^2 = 0.47$. While the individual CV's do not have much separation power, their effectiveness is greatly increased by using them together. Here, the second CV does the work of separating Shakespeare from London while the first vector separates Austen from the other two. The words with the largest loadings for each CV are shown in Table 2. The words are a mix of different parts of speech in a mix of tenses.

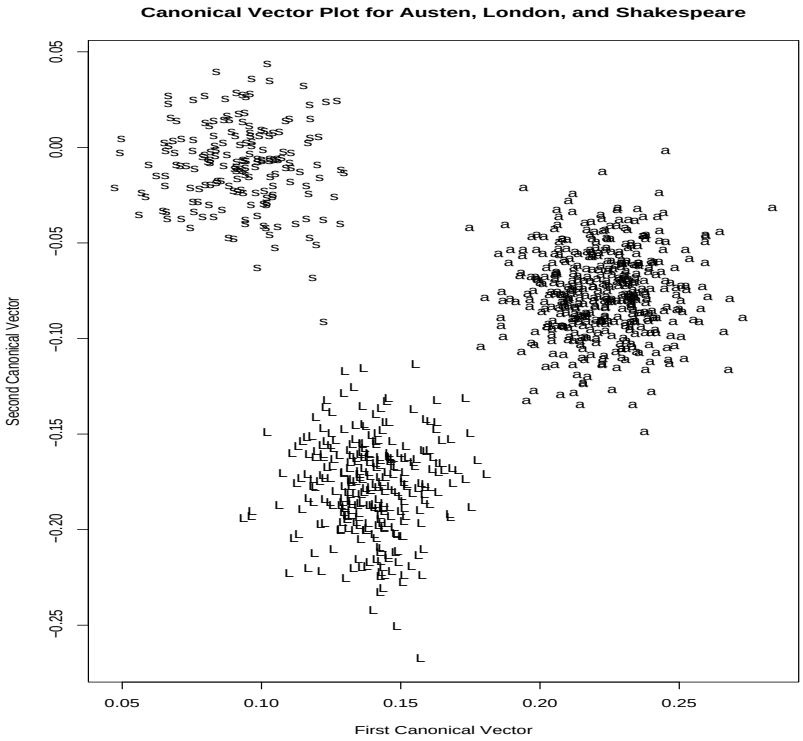

FIGURE 1. Canonical vector plot for Austen (a), London (L), and Shakespeare (s). Each point represents a block of 1700 words.

CV	Word	Loading
1	to	0.77
	her	0.72
	any	0.63
	been	0.62
2	the	−0.80
	was	−0.70

TABLE 2. Words with the largest loadings for the first two canonical vectors (Austen, London, and Shakespeare).

In the second example a larger grouping is used: Austen, Cather, Dickens, Doyle, Marlowe, Milton, and Shakespeare. At first glance the CVP (Figure 2) appears disappointing. However, some significant features do stand out. The cloud in the

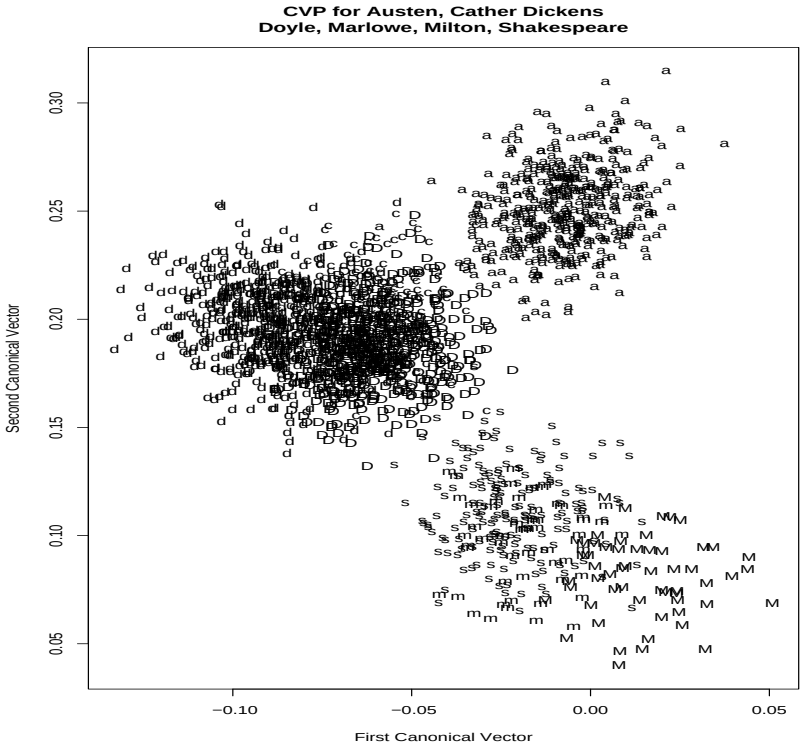

FIGURE 2. Canonical vector plot for Austen (a), Cather (c), Dickens (D), Doyle (d), Marlowe (m), Milton (M), and Shakespeare (s).

upper right corner of the plot consists solely of Austen blocks. In the middle-left portion Cather, Dickens, and Doyle completely overlap each other. Finally, in the lower right Marlowe, Milton, and Shakespeare partially overlap each other.

Overall, thre are three distinct clouds — Austen by herself, the group containing Cather, Dickens, and Doyle, and the group containing Marlowe, Milton, and Shakespeare. While the CV's are capable of separating the authors belonging to different centuries, they have difficulty distinguishing between authors who wrote in similar time periods. The three "time groups" are separated primarily in the direction of the second CV. The words with significant loadings in this case are *upon*, *the*, and *a* for the first CV, and *was*, *had*, *been*, *her*, and *any* for the second CV. Compared to the previous example there are some new words. However, *her*, *any*, *been*, and *was* are recurring.

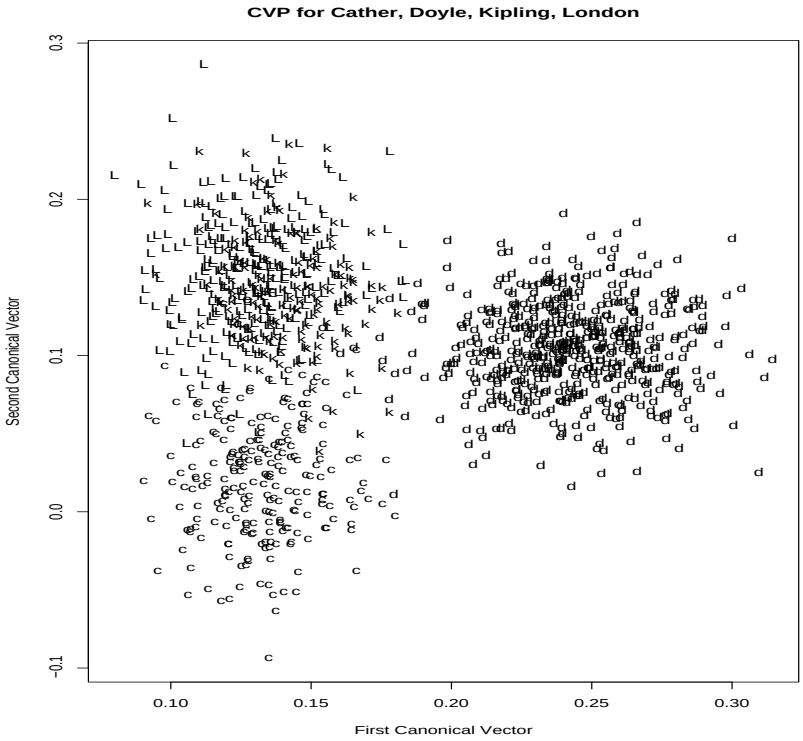

FIGURE 3. Canonical vector plot for Cather (c), Doyle (d), Kipling (k), and London (L).

Now four authors who lived in similar time periods will be examined: Cather, Doyle, Kipling, and London. One might expect there to be poor separation between the authors here. However the CVP in Figure 3 shows something slightly different. Doyle is isolated on the right hand side of the plot while Cather, Kipling, and London are bunched together on the left hand side. It appears that the first CV does the work of separating Doyle out from the others. The second vector makes an effort to separate the other three authors but little is achieved. The loadings give a hint as to which words are responsible for separating Doyle out from the other three authors. The words receiving the highest loadings for the first CV are *which* (0.80), *upon* (0.79), and *have* (0.63). Corroborating evidence can be obtained by looking at the mean counts for each author for these three words. They are shown in Table 3. Here there are significant differences in the mean word counts between

Author	*which*	*upon*	*have*
Doyle	12.3	8.2	12.7
Cather	3.3	1.4	6.6
London	3.0	2.3	5.0
Kipling	2.4	1.1	6.5

TABLE 3. Mean counts for *which*, *upon*, and *have* (Doyle, Cather, London, and Kipling).

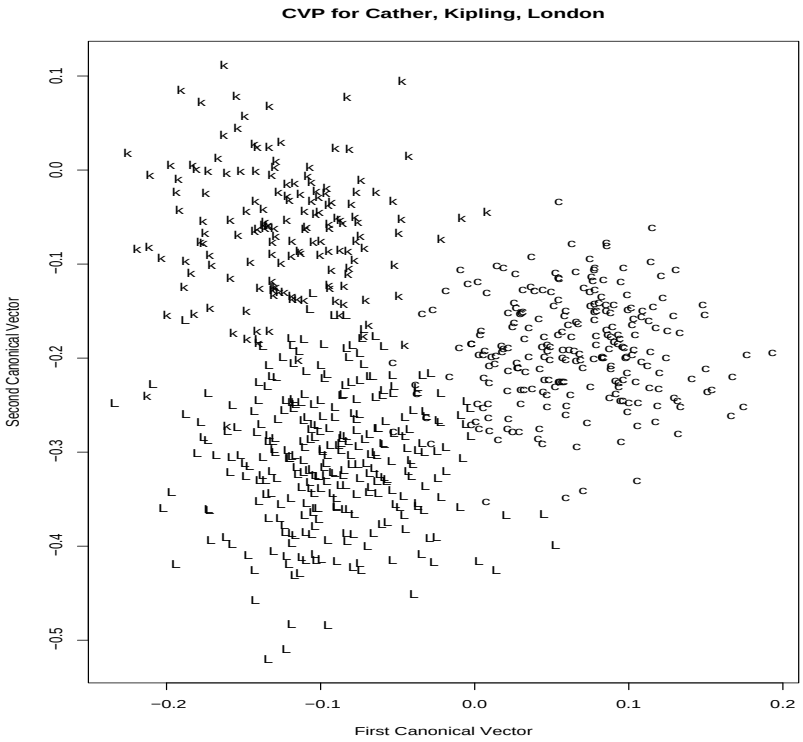

FIGURE 4. Canonical vector plot for Cather (c), Kipling (k), and London (L).

Doyle and the other three authors. Whereas the means for Cather, London, and Kipling are similar, the means for Doyle are at least twice as large.

When Cather, Kipling, and London are taken by themselves the result is shown in Figure 4. Compare the result here with Figure 1, where the separation was readily apparent. Although there are visible groupings in Figure 4, the borders between

the clouds are diffuse and not well-defined. The mean word counts in Table 4 are for the words with relatively large loadings for the first and second CV's. We see that the differences are not so obvious. Cather tends to use the words *her*, *had*,

Author	*her*	*had*	*to*	*was*	*of*	*may*
Cather	25.7	17.5	44.4	26.2	35.9	0.5
Kipling	5.3	8.0	36.9	16.8	35.8	1.4
London	8.9	14.1	37.4	33.5	50.5	0.4

TABLE 4. Mean counts for *her*, *had*, *to*, *was*, *of*, and *may* (Cather, Kipling, and London).

and *to* more often than Kipling or London. However, for *was*, *of*, and *may*, there do not seem to be any clear differences in usage. Looking at word means is useful in the more obvious case with Doyle. In the Cather-Kipling-London case, the CVP provides a potentially more useful tool for discrimination.

Now we will take a closer look at Marlowe's data. Compared to some of the other authors, his blocks show a high degree of variability within themselves. This is surprising considering that he had a relatively short writing career. In order to assess the variability we applied principal component analysis (PCA) to the Marlowe data. A plot of the first two principal components (PC) is shown in Figure 5. There appear to be two groups of outlying blocks. First is the sequence 13-15 (from *The Jew of Malta*), then we have a cluster with 41-42 (*Tamburlaine the Great, Part I*), and 54-56 (*Tamburlaine the Great, Part II*). It is interesting to note that the outlying blocks occur in sequences. It is possible that these sequences represent large scale changes in style within certain works. The nature of those changes, of course, is impossible to ascertain by just looking at word counts.

Principal component analysis was run on the other authors and the only one who had any peculiar features was Kipling. However, in his case, the suspicious blocks of text were not found in the plot of the first few PC's. Rather, they were apparent in the plot of the last two PC's. In the PC plot in figure 6 it appears that blocks 30-33 are in a corner by themselves. The fact that we see this on the plot of the last two PC's indicates that these blocks might be violating the correlation structure of the data (Jollife, 1986). Therefore, these blocks might not be obvious

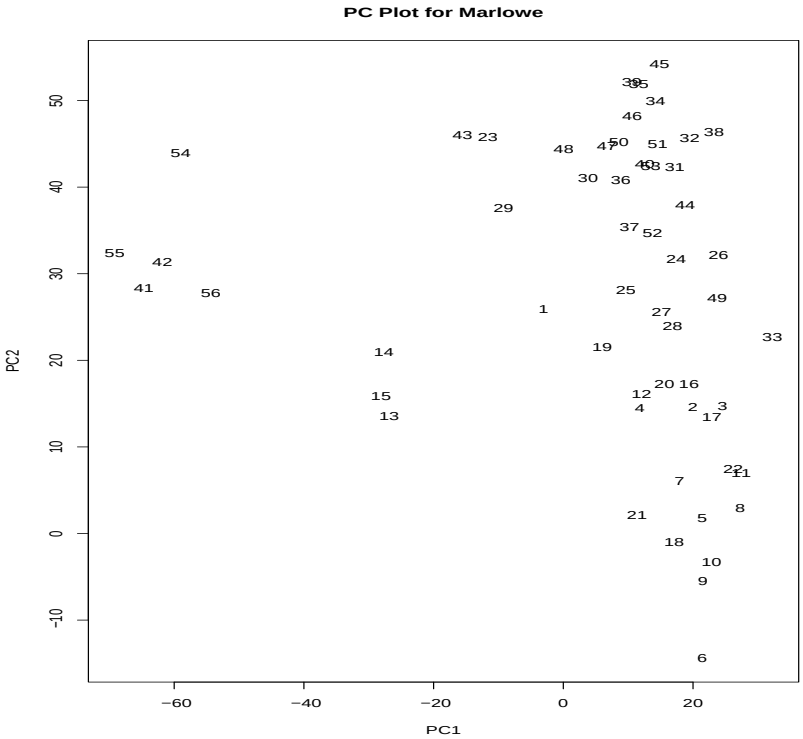

FIGURE 5. Plot of the first two principal components for the blocks of text collected for Marlowe. One can see that blocks 13-15, 41-42, and 54-56 are inflating the variance of the entire collection.

as outliers in the examination of the individual variables. Blocks 30-33 are all from the collection *The Writings in Prose and Verse of Rudyard Kipling.* Since this work is a collection of short works, it is possible that Kipling's style is more variable with respect to his word counts than in the longer works.

6. Summary

In this paper we have used canonical discriminant analysis and principal component analysis to aid in the analysis of literary style. The techniques have shown to be useful by providing interesting plots and by helping to determine possibly useful words for discrimination problems. Furthermore, CDA is useful when multiple authors are being examined simultaneously. In the field of statistical literary analysis, CDA is an effective alternative to fitting probability distributions, which require placing arbitrary assumptions on the data.

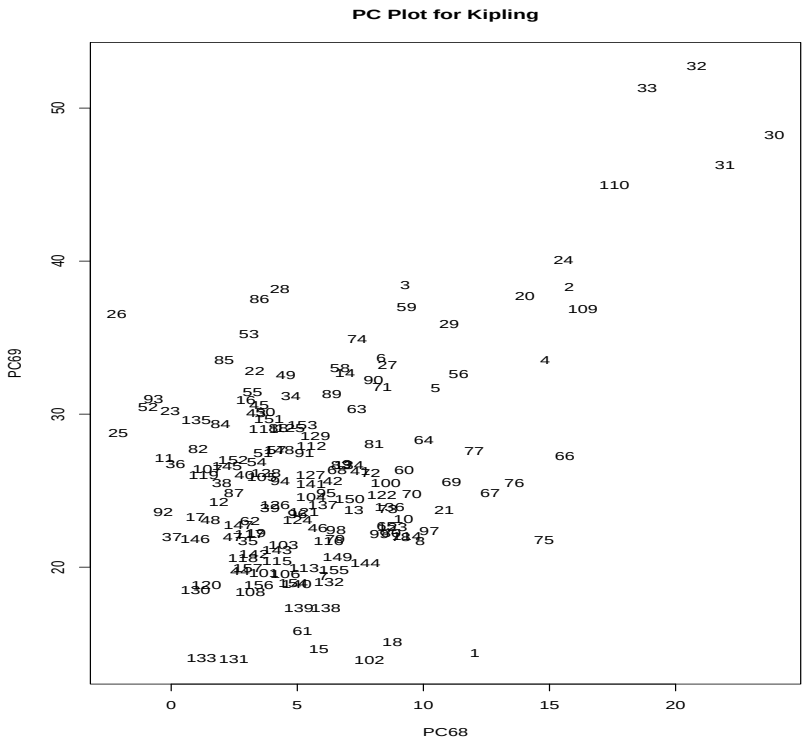

FIGURE 6. Plot of the last two principal components for Kipling. Blocks 30-33 appear in a grouping by themselves and could be violating the correlation structure of the data.

It should be noted that CDA will not be optimal in any way because of the inequality between the covariance matrices of the groups (Lachenbruch, 1975). Although, the method seemed to work well in many situations, there are cases where the discrimination is poor. One possible reason for this lack of power is a difference in covariance matrices.

REFERENCES

[1] Brinegar, C. S. (1963) Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association*. **58**, 85-96.

[2] Holmes, D. I. (1985) The Analysis of Literary Style — A Review. *J. R. Statist. Soc. A*. **148**, Part 4, 328-341.

[3] Holmes, D. I. (1992) A Stylometric Analysis of Mormon Scripture and Related Texts. *J. R. Statist. Soc. A*. **155**, Issue 1, 91-120.

[4] Jolliffe, I. T. (1986) *Principal Component Analysis*. Springer, New York.

[5] Klecka, W. R. (1980) *Discriminant Analysis*. Sage Publications, California.

[6] Lachenbruch, P. A. (1975) *Discriminant Analysis*. Hafner Press, New York.

[7] Mosteller, F. and Wallace, D. L. (1963) Inference in an Authorship Problem. *Journal of the American Statistical Association*. **58**, 275-309.

[8] Mosteller, F. and Wallace, D. L. (1964) *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, Massachusetts.

[9] Särndal, C.-E. (1967) On Deciding Cases of Disputed Authorship. *Applied Statistics*. **16**, 251-268.

[10] Thisted, R. and Efron, B. (1987) Did Shakespeare Write a Newly-discovered Poem? *Biometrika*. **74** 3, 445-455.

[11] Williams, C. B. (1940) A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style. *Biometrika*. **31** 3/4, 356-361.

[12] Williams, C. B. (1956). A Note on an Early Statistical Study of Literary Style. *Biometrika*. **43**, 248-256.

[13] Williams, C. B. (1975) Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon. *Biometrika*. **62** 1, 207-212.

8130 Math Sciences, UCLA Deptartment of Statistics, Los Angeles CA 90095

Yale University Department of Statistics, 24 Hillhouse Ave, New Haven CT 06511

View publication stats