A Study on Authorial Consistency and Attribution : A Canonical Discriminant Analysis Approach

M. Bagavandas and G. Manimannan

Department of Statistics

Madras Christian College, Chennai-INDIA.

Email :mbdas49@yahoo.com

Abstract

This study makes an to attempt to use statistical technique to establish authorial consistency in writing styles of three Tamil scholars and to attribute authorship to unattributed articles written in the same period using linguistic variables. The eighteen linguistic variables are eleven morphological variables, four habitual words and three function words. Canonical discriminant analysis is used to establish consistency and also for attribution.

Keywords : authorial consistency, attribution, canonical discriminant analysis.

Résumé

Cette étude essaie d'utiliser la technique statistique pour établir l'homogénéité de authorial dans l'écriture de styles de trois lettrés Tamouls et attribuer la paternité à unattributed les articles écrits dans la période pareille utilisant des variables linguistiques. Les dix-huit variables linguistiques sont onze variables morphologiques, quatre mots habituels et trois mots grammaticaux. L'analyse canoniale de discriminant est utilisée pour établir l'homogénéité et aussi pour l'attribution.

Mots-clés : l'homogénéité de authorial, l'attribution, l'analyse de discriminant canoniale.

1. Introduction

In the modern era, authorship attribution is considered as a statistical classification and pattern recognition problem. Author attribution problems based on statistical analysis were not well established until Mosteller and Wallace's study on the Federalist papers (Mosteller and Wallace, 1964). These two American statisticians applied bayesian statistical analysis to function words to attribute authorship of the Federalist Papers. This most popular and successful study is still considered as the front-runner for all the modern statistical analysis based authorship studies. Holmes (1994) provides a comprehensive review on stylometric authorship attribution studies.

Initially, the best-known authorship studies have concentrated on univariate statistical analysis of the stylistic features extracted from a limited body of texts. Parameters of the statistical distributions of stylistic features have been used in attribution problems as characteristics of an author (Holmes, 1985). The introduction of modern computing facilities and easily available machine readable texts have given enough opportunities for stylometricians to employ multivariate techniques in attributional studies for analysing high dimensional data. All the multivariate statistical techniques, which need computer assistance, have met with

successful applications in stylometric attributional studies and their applications have given a new status to these methods within humanities scholarship (Roger Peng and Hengartner, 2002).

Holmes (1992) has used hierarchical clustering techniques to detect changes in the authorship of Mormon Scriptures. Burrow's uses of principal component analysis on a wide variety of authors and genres have established stylometry as a reliable attributional method (Burrows, 1987) and this method has been widely used. Good examples of the use of principal component analysis in authorship attribution can also be found in Holmes (1992) and Tweedie *et.al* (1998). Cluster analysis in conjunction with principal component analysis has been employed by many stylometricians to solve some important attribution problems (Holmes, 1992 and Mannion and Dixon, 1997). Canonical discriminant analysis has found application in discriminate registers and styles in the Modern Greek language (Fazlican and Patton, 2004) and the Gospel of St. Luke (Mealand, 1995). Correspondence analysis has been used by Dixon and Mannion (1998) to analyse Goldsmiths essays ; the same technique has been used by Mealand (1997) to study the Gospels. Bagavandas and Manimannan (2004) used factor analysis to quantify the writing styles of three scholars of Tamil language.

1.1 Stylistic features for authorship attribution

Attribution will be successful if proper stylistic features are used as discriminators. There is no general agreement on the stylistic features that should be used in attribution studies. Selection of these features depends on the type of problem involved. In general, when choosing such stylistic features, one must use something that has large variations across authors and relatively little variation among an author's own work. Initially, lexical variables have predominated in attribution studies, yet this decade has seen the applications of syntactic and semantic features (Baayen, 1996). In the present day attributional studies, commonly occurring context-free function words have been used. The rationale for using these words is that writers do not think about the way they use these words that these words straightaway flow from the mind and hence the usage of these words does vary too much from author to author (Roger Peng and Hengartner, 2002).

2. Data and Methods

The present study deals with the literary works of three contemporary Tamil scholars, namely, Mahakavi Bharathiar (MB), Subramanya Iyer (SI) and T.V.Kalyanasundaram (TVK). During pre-independence period, these three scholars have written a number of articles on India's freedom movement in the magazine called *India*. Initially, all the three scholars had written their articles by attributing their names in this magazine. Because of the opposition of the then British regime, the articles written on the same topic then appeared in the same magazine but without the authors' name. Ilasai Maniyan (1975) has compiled all these attributed and unattributed articles and brought out a book called *Bharathi's Dharisanam*. These scholars, in this book, also made a statement that the writing styles of these un-attributed articles indicate that Mahakavi Bharathiar may be the plausible author of these articles. This pioneering study makes an attempt to verify his statement using statistical techniques. For attributional studies the articles should belong to the same genre and time. For this quantitative study all the attributed and unattributed articles written on India's freedom movement published in 1906 in the magazine *India* are considered. Accordingly, there are nineteen articles of MB, seven of SI, and six of TVK and twenty-three unattributed articles. Eighteen stylistic features

considered for this study and they are eleven morphological variables, four habitual words and three function words (Table 1).

Abbreviations	Variables Name
P_NOUN	Occurrence of Noun
P_INT	Occurrence of Intensifiers
P_INF	Occurrence Infinites
P_PRO	Occurrence of Pronoun
P_NUME	Occurrence of Numerals
P_TWO	Occurrence of Two letter Words
P_THRE	Occurrence of Three letter Words
P_FOUR	Occurrence of Four letter Words
P_VOWE	Occurrence of Vowels
P_VERB	Occurrence of Verb
P_SYLLA	Occurrence of Syllable
P_POST	Occurrence of Postpositions
P_CLITIC	Occurrence of Clitics
P_CASE	Occurrence of Case Markers
P_ADVERB	Occurrence of Adverb
P_CONJUN	Occurrence of Conjunctions
TENSES	Type of Tenses
VOICES	Type of Voices

Table1. Lists of Morphology Variables of this study

For a comparative analysis the frequency counts of the stylistic features must be normalized to the text length in an article. In this study since each sentence is considered as a sample, to normalize the stylistic features, the raw frequency counts of each stylistic feature are divided by the number of words in each sentence and then multiplied by hundred to express it in percentage. Eighteen stylistic features are identified from each sentence ; only two features voice and tense are in frequencies but not in percentages.

If we represent each article by mean values of p stylistic features and if we have n such articles then we have a data matrix of size $n \times p$ for each author. Thus the entire information is converted as a data matrix and these data matrices form the basis for this quantitative study. The data matrix of MB is of size 19x18, of SI is of size 7x18 and of TVK is of size 6x18. The main objectives of this study are to establish authorial consistency and to identify the authorship of twenty-three un-attributed articles by analysing these data matrices using canonical discriminant analysis.

3. Canonical Discriminant Analysis

Canonical discriminant analysis is a multivariate method used for demonstrating the significance and nature of the differences between two or more pre-defined groups of objects, where data are available for several variables measured on each object (R. A. Johnson and D.W. Wichern, 2001). The main objectives of this analysis are (i) to find a set of discriminant functions with power in decreasing order of discrimination between groups identified a priori, (ii) to test whether the means of these groups along that axis are significantly different, and to attempt to assign individual objects of unknown origin to the given groups. The relationship between the groups can be assessed visually by means of a scatter plot in which the positions of individuals or the group means or both are plotted on axes known as canonical axes (discriminant functions), which depend on the original observations and are chosen by the analysis to represent the differences between groups. The main assumptions of canonical discriminant analysis are that there are multiple groups that can be unambiguously defined in advance and all the individuals of unknown origin can be assigned to only one group. This

analysis measures the distance between the means of any two samples taken at a time using Mahalanobis distance and determines whether this distance is significantly different from zero using Hotelling T^2 – statistic.

Canonical discriminant functions are obtained as multiple axes that separate sets of groups, assuming that there are more variables than groups, and (m-1) discriminant functions when there are m number of groups. Since the groups to be differentiated in our research are known, canonical discriminant analysis is a very powerful tool in determining distances between the groups. The main difference between Fisher's discriminant function and the canonical discriminant function is that Fisher's discriminant function connects pairs of centroids whereas canonical discriminant function summarizes the major axes between groups. This analysis is necessary to reduce the dimensionality of the original data set so that it can be plotted in two dimensions (Lanchenbruch, 1975). Klecka (1980) provides an introduction to canonical discriminant analysis.

This study proposes to use canonical discriminant analysis for establishing authorial consistency of the known articles and for attributing authorship for the twenty-three un-attributed articles. In performing this analysis, it is assumed that an author's collected works are forming a stable population and that all these populations have the same covariance structure. This analysis extracts discriminant function in such a way that the between-authors variability is maximized compared to within-authors variability and hence it can separate between-authors effects from within-authors effects. Given a classification variable, such as authors and several quantifiable stylistic features, this analysis derives canonical discriminant functions (linear combinations of quantifiable stylistic features) that have the highest possible multiple correlation with the classification variable such as authors and also summarizes between- authors variation in much the same way that the principal component analysis summarizes total variation. This analysis facilitates differentiation of authors by taking into account the inter relations of the stylistic features and the dependent variable. An important property of canonical variables is that they are un-correlated even though the underlying quantitative variables may be highly correlated (R. A. Johnson and D.W. Wichern, 2001).

4. Analysis of Morphology Data

The present study proposes to make use of canonical discriminant analysis to check the consistency of the writing style of each author and also to attribute authorship to the twenty-three unattributed articles. This analysis is employed to determine whether it is correct to state that a writer maintains his/her style of writing in all his/her texts written on a specific topic in a specified period of time. For example, whether it is possible for us to state that the three authors namely MB, SI and TVK had maintained their respective writing styles in all their articles of this study. Also canonical discriminant analysis is used to identify authorship for twenty-three unattributed articles. Authorial consistency and authorship attribution are made using eighteen stylistic features of this study.

In this analysis, the three authors, namely Mahakavi Bharathiar (MB), Subramaniya Iyer (SI) and T. V. Kalyanasundaranar (TVK), are designated as author 1, author 2 and author 3 respectively. As there are three authors to be differentiated, we get two canonical discriminant functions. The first canonical discriminant function accounts for 97 percent of the between authors variance. The second canonical discriminant function accounts for the remaining 3 percent of total between authors variance. Each canonical discriminant function is a linear combination of the eighteen stylistic features and is orthogonal to the other. The significant canonical correlation between authors and the first canonical discriminant function (r = 0.993)

and authors and second canonical discriminant function (r=0.789) indicating that both the canonical discriminant functions can explain the differentiation of the authors (*Table 2*).

It is a known fact that a canonical discriminant function is said to be a good discriminant function if it has more between-authors variability than within author's variability. In fact the coefficients of discriminant functions are chosen in such a way that the ratio of the between-authors sum of squares to within-authors sum of squares is as large as possible. The Wilk's lambda criterion measures this ratio and this lambda ranges from zero to unity. Values closer to zero are associated with functions that have much variability between authors and little variability within author and here smaller lambda values are associated with good discriminant functions For this study, the Wilk's lambda value is 0.006 (*Table 2*) and its observed significance level is zero. The smaller lambda value indicates that the two canonical discriminant functions are good discriminant functions.

Also the Wilk's lambda associated with function 2 after function 1 has been removed is 0.381. The significance associated with the second function is 0.284, indicating that this function does contribute in small scale to author differences. The mean values of the scores of the two discriminant functions indicate that author 1 has positive values for both the functions (6.339, 0.187), author 2 has a negative value for first function and a positive value for the second function (-11.364, 1.469), and author 3 has negative values for both the functions (-6.816,-2.306). *Figure 1* is the territorial map for the three authors on the two functions.

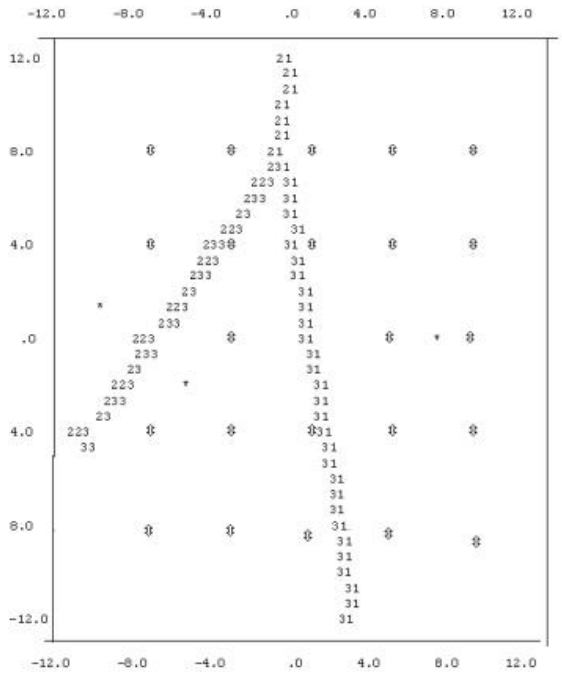

Figure 1. Territorial Map

- Indicates a group centroid

Canonical loadings provide correlation between stylistic features and the canonical discriminant function. Thus canonical loadings reflect the variance that stylistic features share with a canonical discriminant function and can be interpreted as amassing the relative contribution of each stylistic feature to each discriminant function. The canonical loadings of the first canonical discriminant function indicate that the function is dominated by the stylistic features like clitics, interjection, postpositions, case markers, conjunctions, four letter words, noun, adverb, verb and two letter words. The negative region of this function is associated with the loadings of the first seven stylistic features and positive region is associated with the loadings of the last three stylistic features. The negative region indicates that the small values of this function are associated with the usage of these stylistic features and the large values with less usage of these features. Similarly the features like voices, infinite, syllables, pronoun, word starting with vowels, numerals, three-letter words and tenses dominate the second canonical discriminant function. The regions of the canonical loadings of features like voices, pronoun, vowels and three letter words are positive and the regions of the loadings of the rest are negative (Table.2).

Variable Name	Functions	
	1	2
P_CLITIC	-.398*	-.238
P_POST	-.334*	-.006
P_INT	-.309*	-.283
P_CASE	-.256*	.050
P_CONJUN	-.170*	.040
P_NOUN	-.099*	.043
P_FOUR	-.091*	-.065
P_ADVERB	.052*	.041
P_VERB	.035*	-.002
P_TWO	.018*	.007
VOICES	.072	.512*
P_INF	-.056	-.453*
P_PRO	.012	.184*
P_SYLLA	.117	-.156*
P_VOWE	-.065	.151*
P_NUME	-.044	-.104*
P_THRE	-.001	.073*
TENSES	-.021	-.034*
Eigenvalue	67.111	1.644
Percentage Variance	97.6	2.4
Cumulative Percentage	97.6	100.0
Canonical Correlation	.993	.789
Wilk's Lambda	.006	.378
Chi-square	106.464	19.931
Degrees of Freedom	36	17
Significance	.000	.278

Table 2. Results of Canonical discriminat analysis

The classification matrix (*Table 3*) provides the summary of the classification results of this study. The percentages of cases classified correctly are often considered as an index of the effectiveness of the derived discriminant functions. The diagonal elements of this matrix are the number of cases classified correctly into groups and the non-diagonal elements are the misclassified cases. The articles of all three authors are classified into three groups correctly. The overall percentages of cases classified correctly are 100 percent. This result indicates that all the nineteen articles of author1, seven articles of author 2 and six articles of author 3 are correctly classified into three different groups. This result establishes that the percentages of occurrences of different stylistics features in all the nineteen articles of MB are the same and

hence it can be concluded that MB had maintained the same style in writing all the nineteen articles and it is true also in the cases of the other two scholars. Thus the consistency of the writing styles of these three scholars is established. Also all the twenty-three unattributed articles are attributed to author1. This shows that Mahakavi Bharathiar might have written all these articles. This important result endorses the statement of Ilasai Maniyan (1975).

Classification Results[a]

		GRO	Predicted Group Membership 1.00	2.00	3.00	Total
Original	Count	1.00	19	0	0	19
		2.00	0	7	0	7
		3.00	0	0	6	6
		Ungrouped cases	23	0	0	23
	%	1.00	100.0	.0	.0	100.0
		2.00	.0	100.0	.0	100.0
		3.00	.0	.0	100.0	100.0
		Ungrouped cases	100.0	.0	.0	100.0

a. 100.0% of original grouped cases correctly classified.

Table.3 Classification Results.

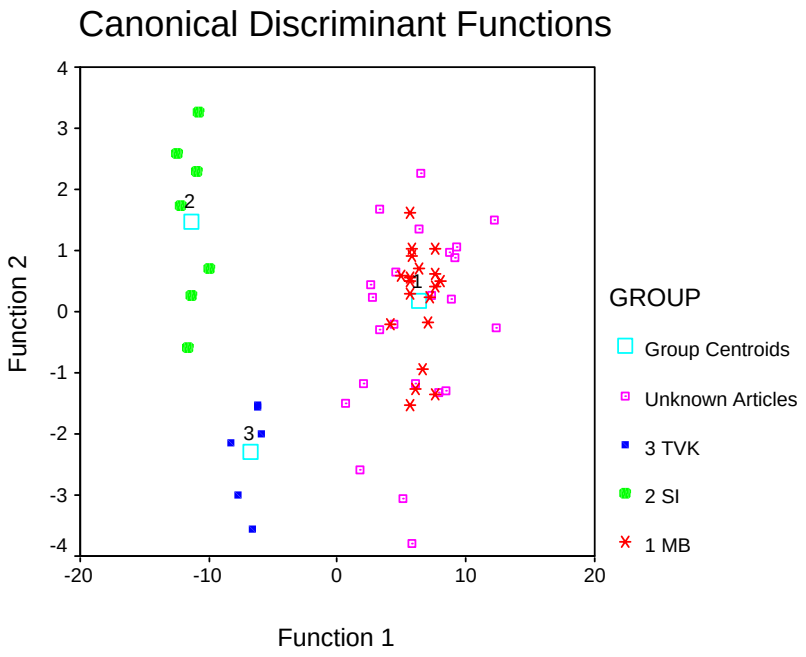

Figure 2. Canonical Discriminant Analysis

From the canonical discriminant function plot (*Figure 2*

the positive side of the second canonical discriminant function because of the positive loadings of features like voices, pronoun, vowel beginning words and three letter words.

5. Distinct and Consistent Writing Styles of Three Scholars

It is established that the three scholars have consistent but distinct writing styles. That is, the scholar Bharathi had used the same writing style to write all nineteen articles and hence all his nineteen articles are considered as one article by pooling all the articles together. The writing style of Bharathi is quantified by the averaging the values the eighteen stylistic features from this pooled article. The writing styles of other two scholars are quantified in the similar manner.

All these three scholars had used, on an average, one pronoun, one two-letter word and two three -letter words in a sentence of ten words. Also the smaller percentages of occurrence of features like intensifier, infinity and adverb indicate that these three authors had used these three features very rarely (Table 4).

The percentages of occurrences of stylistic features like noun, post-position, Clitics, case makers and conjunctions differentiate these three authors statistically from one another. This result indicates that Bharathiar is identified as the least user of these features whereas Kalyanasundranar is identified as the maximum user and Subramaniya Iyer is identified as the medium user of the same stylistic features.

Stylistic Features	Abbreviations	Mean values		
		MB	TVK	SI
Noun	P_Noun	34.260	45.47	41.80
Intensifier	P_Int	00.050	06.07	06.10
Infinitive	P_Inf	00.580	01.33	03.60
Pronoun	P_Pro	07.720	07.73	06.60
Tense	Tense	01.710	01.77	01.40
Numeral	P_Nume	03.990	05.27	05.30
Two-Letter Word	P_Two	10.610	09.79	09.50
Three-Letter Word	P_Thre	19.240	20.18	18.30
Four-Letter word	P_Four	20.460	25.66	25.90
Vowels	P_Vowe	27.740	33.36	28.80
Verb	P_Verb	23.430	21.40	23.60
Voice	Voices	02.250	01.55	02.10
Syllable	P_Sylla	151.10	119.7	163.0
Post position	P_Post	13.430	35.26	31.30
Clitics	P_Clitic	14.140	34.25	33.40
Case marker	P_Case	38.650	69.95	63.00
Adverb	P_Adverb	04.390	02.26	02.80
Conjunction	P_Conjun	22.650	42.15	35.50

Table 4. Mean values of the eighteen stylistic features.

The scholar MB had used passive voice sentences in past tense to narrate India's Freedom Movement. In a sentence of ten words, he had used, on the average, one clitic, one pronoun, two verbs, three words starting with vowels, four case makers and four nouns. The verbs and words starting with vowels are either two-letter or three-letter or four-letter words.

The scholar SI had made use of passive voice sentences in present tense. There will be six case markers, four nouns, three conjunctions and three postpositions and one pronoun in a sentence. The verb and word starting with vowel will be four-letter words with two or three syllables.

The author TVK had used active voice sentences in past tense to describe India's Freedom Movement. In these sentences of ten words, on the average, three postpositions, three Clitics,

three words starting with vowels, four conjunctions, four nouns and six case markers are accommodated.

6. Conclusion

This study has employed canonical discriminant analysis to establish the consistency of the writing styles of three scholars of Tamil language, namely, Mahakavi Bharathiar, Subramaniya Iyer and T.V. Kalyanasundranar and also to attribute authorship of one these three scholars to twenty-three unattributed articles using eighteen stylistic features.

This analysis has thus achieved a considerable degree of economy in its presentation of results. This study started with a data matrix with eighteen stylistic features and three authors. The eighteen stylistic features have been reduced to two canonical discriminant functions, whose scores best separate the three authors. This has the advantage that a bivarite scatter-plot of the two sets of discriminant function scores gives a visual picture of the discriminant analysis as in *Figure 2*.

Significant stylistic diversity was observed between the authors. It can be seen that the articles of each of the three authors have consistent writing styles and are well separated. Distinct and consistent writing styles of these three scholars are quantified. Also all the twenty-three un-attributed articles are grouped with the articles of Mahakavi Bharathiar.

References

Baayen, R. H., van Halteren, H. and Tweedie, F. J. (1996). Outside the Cave of Shadows : Using syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing,* Vol, 11 : 110-20.

Bagavandas, M. and Manimannan.G. (2004). Quantification Stylistic Traits : A Statistical Approach. *Proceedings of the 7th International Conference on Statistical Textual Data Analysis, Louvain-la-Neuve, Belgium*, Vol.1 : 71-78.

Burrows, J.F. (1987). *Computation into Criticism : A Study of Jane Austen's Novels and an experiment in Method*. Oxford, Clarendon Press, 1987.

Dixon, P. and Mannion, D. (1998). Goldsmith and the British Magazine. *Literary and Linguistic Computing*, Vol.13, No.1 : 37-50.

Fazlican and Patton, J.M. (2004). Change of Writing Style with Time. *Computers and the Humanities*, Vol.38 : 61-82.

Holmes, D. I. (1985). The Analysis of Literary Style : A Review. *Journal of the Royal Statistical Society, Series A,* 148 : 328-341.

Holmes, D. I. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society, Series A,* 155 : 91-120.

Holmes, D. I. (1994) Authorship Attribution. *Computer and the Humanities*, Vol.28 (2) : 87-106.

Illasai Manian (1975). *Bharathi Dharisanamm,* Maraimalai Adigal Padhipakam, Madras.

Johnson, R. A. and Wichern, D.W. (1982) *Applied Multivariate Statistical Analysis.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

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

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

Mannion, D. and Dixon, P. (1997). Authorship Attribution : The Case of Oliver Goldsmith. *The Statistician* 46 : 1-18.

Mealand. D. L. (1995). Correspondence Analysis of Luke. *Literary and Linguistic Computing*, Vol,10 : 171-82.

Mealand. D. I. (1997). Measuring Genre Differencies in Mark with Correspondence Analysis. *Literary and Linguistic Computing*. Vol.(12/4) : 225-247.

Mosteller, F. and Wallace, D. L. (1964). Inference in An Authorship Problem. *Journal of the American Statistical Association*, Vol. 58 : 275-309.

Roger Peng, D. and Hengartner, N.W. (2002). Quantitative Analysis of Literary Styles. *The American Statistician*, Vol.56 : 175-185.

Tweedie, F. J., Holmes, D. I. And Corns, T.N. (1998). The Provenance of De Doctrina Christina, attributed to John Milton : a statistical investigation. *Literary and Linguistic Computing*, Vol.(13/2) : 56-68.