Attempt to statistically identify writer for a text using a computer program

Bhaskar Ray

October 2019

Introduction

Let us consider two pieces of text

- My bounty is as boundless as the sea,
  My love as deep; the more I give to thee,
  The more I have, for both are infinite.
- Continuous as the stars that shine
  And twinkle on the milky way,
  They stretched in never-ending line
  Along the margin of a bay:

Suppose we are tasked with finding the piece by William Shakespeare. Anyone acquainted with English literature will indisputably vouch for the former.
The above recognition is based on our perception of the writing styles of different authors from past experiences. Is the same feasible using computers? In this article, I intend to discuss on a statistical methodology that will aid us in this task. Recognition based on previously known samples is a trending topic. Numerous methods for such recognition have been developed and incisively analyzed in the recent years. Here, I shall try to find a possible approach to deal with the problem at hand.

Our problem

We wish to identify the probable writer (of course well-known) for a given text using texts from various authors. This can be done if we can identify the difference in their writing styles mathematically.

Arriving at a solution

First, we shall frame the problem mathematically and then try to solve it using tools from both mathematics and statistics.

Framing the problem mathematically

The key to this is frequency distribution. We can consider the frequencies of the ordered pairs of consecutive letters. About the punctuation marks, we may opt to ignore them. We may also ignore the cases[1] of the letters For the ease of mathematical manipulations, we can construct a frequency vector for each text that we consider as:

$$\vec{v_f} = \{f_{aa}, f_{ab}, \ldots, f_{az}, f_{ba}, f_{bb}, \ldots, f_{bz}, f_{ca}, f_{cb}, \ldots\ldots, f_{zy}, f_{zz}\}$$

where $f_{\alpha\beta}$ is the frequency of occurrence of the ordered pair $\alpha\beta$ in the text. From basic combinatorics, we know that there can be $26 \times 26 = 676$ such ordered pairs. Hence, the vector, $\vec{v_f} \in \mathbb{R}^{676}$.

[1] Here case refers to whether the letter is in capital or not. A capital letter is said to be in upper case while a small letter is said to be in lower case

However, there is a serious drawback of this representation. The larger the text, the longer will v_f corresponding to it tend to be. For instance, suppose we have a piece of text. We construct another piece of text by copying the original text twice. Although we expect both of them to give us the same information, the vector corresponding to the latter piece of text is twice of that for the original piece. Since we are not interested in the length of the text, rather interested in the trend it follows, we can normalise the vector v_f to get a normalised frequency vector,

$$\vec{v} = \frac{\vec{v_f}}{||\vec{v_f}||_2}$$

We can illustrate how the normalised frequencies vary between authors by taking some text samples of William Shakespeare and some of William Wordsworth and presenting a spider plot corresponding to each author. The plots are presented in *Figure 1* and *Figure 2*.

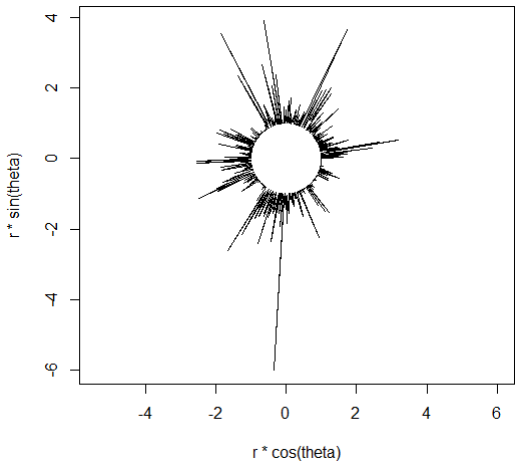

Figure 1: Spider plot for frequency distribution of Shakespeare's text

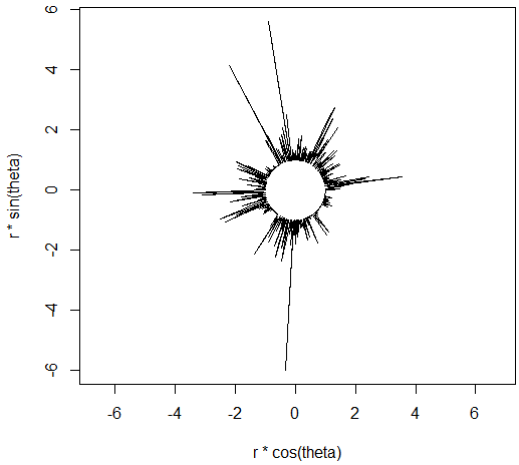

Figure 2: Spider plot for frequency distribution of Wordsworth's text

We would, now want to associate a cluster of points in $\mathbb{R}^{676}$ with each author. We shall conclude that a text is written by an author if the normalised frequency vector obtained from that text is sufficiently close to the cluster corresponding to that author.

Finding a representative vector for each author

We can choose the representative vector for an author to be the normalised mean of all the vectors obtained for that author. For the i^{th} author, let the vectors obtained from n different texts be

$\vec{v}_{i_1}, \vec{v}_{i_2}, \ldots, \vec{v}_{i_n}$. We compute the mean as $\vec{v}_{mean_i} = \frac{\sum_{j=1}^{n} \vec{v}_{i_j}}{n}$. Then, the representative vector for the author is given by

$$\vec{v}_i = \frac{\vec{v}_{mean_i}}{||\vec{v}_{mean_i}||_2}$$

Finding author for a given file

Having found representative vectors for the authors, we now have to classify a normalised frequency vector, $\vec{v}$ found for some file under one of these authors. We shall now discuss some possible approaches to do this.

L_1 distance

We shall compute the L_1 norm, $d_i = ||\vec{v} - \vec{v}_{mean_i}||_1$ for all i. The i^{th} author for which the value of d is minimum is predicted to be the author for the given text.

L_2 distance

This is similar to the above except that here, we compute the L_2 norm, $d_i = ||\vec{v} - \vec{v}_{mean_i}||_2$.

There is a serious drawback of the two above techniques. In these two techniques, we have put equal priority to each of the pairs of letters. From the above figures 1 and 2, we can observe that there are certain pairs of letters that contribute more in distinguishing between the writing style of those two authors as compared to others. We should, therefore, assign greater priority to them. This can be achieved by assigning some weight to each of the pairs. But, how do we assign the weights and on the basis of what? Our next technique will deal with these questions and try to find a satisfactory solution.

Weighting the pairs

Let us construct a weighted frequency vector, $\vec{w} = \begin{bmatrix} l_1 v_1 \\ l_2 v_2 \\ \vdots \\ l_{676} v_{676} \end{bmatrix}$ for each text, where the normalised frequency vector is $\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{676} \end{bmatrix}$, constructed as described above and $l_1, l_2, \ldots, l_{676}$ are the weights, such that $\sum_{i=1}^{676} l_i^2 = 1$.

We now need some condition(s) to decide the weights. We shall tend to assign constrained weights to the letter pairs such that the variance increases between authors and decreases for texts from the same author. This can be done in either of the two following ways:

- Considering the $\vec{w}$ vectors and working with the variance-covariance matrices associated with those in order to maximise the spread of the representative vectors corresponding to different authors and simultaneously minimise the spread of the vectors corresponding to the same author upto an optimal level.
- Let us consider a vector containing the weights, $\vec{l} = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_{676} \end{bmatrix}$. For each text, we shall consider the quantity, $\vec{l}^T \vec{v}$ and try to maximise its variance between authors and minimise the same between texts from the same author.

We observe that the second approach is simpler and a smart solution may be feasible. Indeed, maximising variance between different authors is something that we can achieve through a technique called ***Principal Component Analysis***. However, we need to simultaneously reduce variance between texts of the same

author. Hence, we need something more than that.
Since $\vec{l}$ is unknown, we need to find the variance of $\vec{l}^T\vec{v}$ in terms of quantities that can be obtained from raw data.

Fact 1. *For an n-dimensional random vector $\underset{\sim}{X}$, $Var(\vec{a}^T\underset{\sim}{X}) = \vec{a}^T Var(\underset{\sim}{X})\vec{a}$, where $\vec{a}$ is an n-dimensional constant vector and $Var(\underset{\sim}{X})$ is the variance-covariance matrix of $\underset{\sim}{X}$.*

Proof. Let, $a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$. Therefore,

$$\begin{aligned}
Var(\vec{a}^T\underset{\sim}{X}) &= Var(\sum_{i=1}^{n} a_i X_i) \\
&= \sum_{i=1}^{n} Var(a_i X_i) + \sum_{1\le j\le n; j\neq i} Cov(a_i X_i, a_j X_j) \\
&= \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{1\le j\le n; j\neq i} a_i a_j Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} a_i \left(a_i Var(X_i) + \sum_{1\le j\le n; j\neq i} a_j Cov(X_i, X_j) \right) \\
&= a^T \begin{bmatrix} a_1 Var(X_1) + \sum\limits_{1\le j\le n; j\neq 1} a_j Cov(X_1, X_j) \\ a_2 Var(X_2) + \sum\limits_{1\le j\le n; j\neq 2} a_j Cov(X_2, X_j) \\ \vdots \\ a_n Var(X_n) + \sum\limits_{1\le j\le n; j\neq n} a_j Cov(X_n, X_j) \end{bmatrix} \\
&= \vec{a}^T Var(\underset{\sim}{X})\vec{a}
\end{aligned}$$

□

Optimisation

As described above, let $\vec{v}_i$ be the representative vector for the i^{th} author and $\vec{v}_{i_j}$ be the vector corresponding to the j^{th} text of the i^{th} author. We want to maximise variance of $\vec{l}^T\vec{v}_i$'s and for each i, minimise variance of $\vec{l}^T\vec{v}_{i_j}$'s.
Let V be the variance-covariance matrix of the v_i's and for each i, let V_i be the variance-covariance matrix of the v_{i_j}'s. The matrix, $V_{intra} = \frac{1}{n}\sum_{i=1}^{n} V_i$ gives us an estimate of the variance between texts from the same author. So, we consider maximising $\vec{l}^T V \vec{l}$ and minimising $\vec{l}^T V_{intra}\vec{l}$ optimally. This is equivalent to solving

$$\max_{\vec{l}\in\mathbb{R}^{676}; ||\vec{l}||_2=1} \frac{\vec{l}^T V \vec{l}}{\vec{l}^T V_{intra}\vec{l}}$$

The following theorem will help us arrive at a solution.

Theorem 1. *If A and B are real symmetric matrices, then the maximum value of $\frac{x^T Ax}{x^T Bx}$ under the condition $||x||_2 = 1$ is given by the absolute value of the largest eigenvalue (in terms of magnitude) of $B^{-1}A$ and the vector which the maximum is attained is the eigenvector of $B^{-1}A$ corresponding to the largest eigenvalue.*

We know that a variance-covariance matrix is always real and symmetric. Hence, both V and V_{intra} are real,symmetric.
Thus, applying the above theorem, the required value of $\vec{l}$ is the eigenvector corresponding to the eigenvalue of largest magnitude corresponding to $V_{intra}^{-1}V$.

Computational optimisations

Working with two 676×676 matrices is computationally intensive. We can optimise this by ignoring pairs of letters for which the variance computed in the matrix V is below a certain level, say 0.00001, i.e., the inter-author variance is very low. This cut-off was selected on an ad-hoc basis, observing the general pattern in the variance-covariance matrices. We can justify our action by arguing that such a pair would not help us much in distinguishing between authors.
Another problem that we might encounter is that the system might ignore small, yet significant values. For this, when we normalise the vectors, we normalise them to length 100 instead of the standard procedure of normalising them to length 1. Thus, the variance cut-off becomes 0.1.
A major challenge is the existence of an inverse of the matrix V_{intra}. Due to the variance cut-off, the inverse exists in most practical cases. Yet, to tackle the adverse cases, we may make use of the *Moore-Penrose pseudo-inverse*.
To further reduce computations, we may opt to ignore those pairs corresponding to the least few values (in terms of magnitude) in the eigenvector obtained, which are less than a certain cut-off. Since the squares of the elements in the eigenvector add upto 1, the cut-off may be set as *the least few values whose sum of squares is less than, say, 0.01*. It was observed that the technique performs well even without this adjustment.

Predicting for given files

From the above procedures, we have representative vectors, $\vec{v_i}$'s for all the authors and the weight vector, $\vec{l}$. For a given file, we obtain the normalised frequency vector, $\vec{v}$. We then compute differences, d_i's as

$$\begin{aligned} d_i &= |\vec{l}^T \vec{v_i} - \vec{l}^T \vec{v}| \\ &= |\vec{l}^T (\vec{v_i} - \vec{v})| \end{aligned}$$

We search for the index, k for which the value d_k is minimum. The k^{th} author is expected to be the author of the given text.

Analysing the success of the technique

After we have arrived at a solution, we need to check if the solution is successful. To analyse this, we construct a confusion matrix. The confusion matrix is created from predictions made about texts from known authors. It looks like

	Author 1	Author 2	...	Author n
Author 1	a_{11}	a_{12}	...	a_{1n}
Author 2	a_{21}	a_{22}	...	a_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Author n	a_{n1}	a_{n2}	...	a_{nn}

In the confusion matrix so constructed, a_{ij} gives the number of texts which are actually written by the i^{th} author and were predicted to be written by the j^{th} author.

A case study

In a case study involving 6 authors, viz, Robert Frost, William Shakespeare, William Wordsworth, Charles Dickens, Samuel Taylor Coleridge and Percy Bysshe Shelley, a few texts corresponding to each author was used for learning. After all the computations, predictions were made on all those texts and also, on some more texts from these authors. The following confusion matrix was obtained.

	Robert Frost	William Shakespeare	William Wordsworth	Charles Dickens	Samuel Taylor Coleridge	Percy Bysshe Shelley
Robert Frost	7	1	0	2	2	0
William Shakespeare	4	13	2	1	1	0
William Wordsworth	0	0	8	0	1	0
Charles Dickens	0	0	0	10	0	0
Samuel Taylor Coleridge	0	0	0	0	7	0
Percy Bysshe Shelley	0	0	0	0	0	3

Figure 3: Confusion matrix for the case study

The overall success rate was 77.419%. Considering individual authors, the success rate was less in case of Frost and Shakespeare while the performance was much better in the other cases.

Conclusion

This entire technique is based on statistical manipulation of the available data. It is quite obvious that a 100% success rate is too hard to guarantee. The technique might be improved further considering different sampling techniques for more unbiased estimates and otherwise. However, for all practical purposes, we might consider the success rate to be good enough.

Computer program for implementation

I have written some programs to implement the idea. It has two parts, viz., a C++ program with multiple headers included and an R code. The C++ program reads and processes all the text files and converts them to a form which is feasible for statistical computations. The statistical computations are done using the R code. After all computations, for verifying author for a given text or producing a confusion matrix for all authors considered, the C++ program can be used.

Acknowledgements

I want to thank Dr. Arnab Chakraborty (Associate Professor, Applied Statistics Unit, Indian Statistical Institute, Kolkata) for his guidance on this project and his suggestions on a draft of this article.

Suggested reading

1. Applied Multivariate Statistical Analysis, by Wolfgang Karl Härdle and Léopold Simar.