

A New Procedure for Author Attribution

Roy Felton

4 Pitcairn Pl., New Windsor, Auckland 1007, New Zealand

KEYWORDS: stylometry, contrasts, binary

AFFILIATION: Business Computing and Economics Department, Manukau Institute of Technology, Auckland, New Zealand.

E-MAIL: rfelton@manukau.ac.nz

FAX NUMBER: 64(09)2730701

PHONE NUMBER: 64(09)8282100

The Bayesian approach of Mosteller and Wallace¹ to the question of the authorship of the Federalist Papers can be applied to the situation of disputed writing(s) being assigned to one or other of a limited number of known authors, each of whom has a known style. Another quite distinct situation frequently occurs in the case of classical and other ancient writings. Here there is often an undisputed corpus of texts (by a known or unknown author or group of authors) with another text or texts (hereafter treated in the singular only), which has been traditionally linked to this body but its authorship has been disputed on critical grounds. The problem is to determine whether this disputed text is close enough in style to that of the undisputed texts in order for it to belong to this group and hence presumably have the same author or group of authors. (Whether the corpus itself exhibits a unity of style between texts as well as the more usual assumption of unity of style within each text is unimportant.)

This problem can be handled by using contrasts on a random effects linear model. Either means or proportions (binary variables) can be used. The technique has a multivariate variant. The linear model may be generalised to include any number of factors such as text, genre and subject matter effects. Factors could be fixed or random, crossed or nested. Text is to be regarded as a random effect, but genre and subject matter as fixed. Genre and subject matter could be crossed or perhaps nested, whereas text and genre would be nested. At least two texts of undisputed provenance (not necessarily of the same length) are required. Usually contrasts (a linear combination of means with a zero sum for the coefficients of the means) are used with a fixed effects model to test hypotheses. However, in the above situation with ancient writings, the number of potential texts by an author may be conceived of as unlimited and hence a

random effect is appropriate. Since only some texts of this author are extant or indeed ever written, a contrast comparing only the extant ones against the disputed one is all that a researcher would probably be interested in. In the case of a random effect, its associated measure of variability (which is always made to be non negative) may be estimated from all the texts and used in the denominator of the test statistic ensuring that a more conservative test results. This would help to counter the well documented fact that frequently a stylometric test will reject an undisputed text². By making allowance for other sources of variation as well as that due to sampling, confounding of effects will be reduced resulting in a statistically superior methodology. The use of contrasts on a random effects model appears to be unique in stylometric research and possibly elsewhere as well.

A Bayesian type approach may be utilised to integrate out parameters, the estimates of which might be based on too little data (usually because the number of undisputed texts is small) to give very precise estimates. Autocorrelation, both within and between texts, may be incorporated into the linear model. Critical values can be obtained by simulation, which enables the case of discrete variables to be catered for in a straightforward manner as well as various types of autocorrelation. During simulation the effect of changing blocksize (number of words of running text that each observation is based on) but not text length was examined, noting that power was substantially increased when instead of 100-word blocks, 25-word blocks were used.

Words may be examined individually also (block-size of 1 word) giving rise to binary variables. These may be handled in an analogous way to count, derived or other variables. In this case it is more appropriate to estimate the text effect variability using only the undisputed writings.

The above approach of using contrasts on a random effects model was developed to solve a simply formulated problem drawn from the Christian Scriptures. There has been a somewhat persistent stream of critical thought this century which has sought to find an underlying source for a significant amount of the narrative portions of John's gospel. One proposal for a source for the synoptic-like miracle stories in this gospel, that has reasonably precisely defined limits, is Fortna's Gospel of Signs (hereafter FGOS).³ This reconstruction of the somewhat shadowy and elusive so-called signs source seems to enjoy a certain amount of support and its Greek text is available enabling a detailed statistical examination to be carried out.

John's gospel was divided into three equal sized main sections, each of 30 blocks of 100 words in

each. The three sections were F (derived from FGOS), N (derived from the narrative portions of the gospel) and D (derived from the discourse portions of the gospel). Each block began at the beginning of a sentence, which for this investigation was defined as a string of words terminated by fullstops, semicolons or question marks. Sentences that contained identifiable quotations from the Hebrew Scriptures were omitted from the count. Twenty-four simple-to-measure variables were used to measure style. Variables 1-10 were the number of words in each block of 1 letter, 2 letters up to 10 letters long. Variable 11 was the number of definite articles beginning with a tau. Variable 12 was the number of other words beginning with a tau. Variable 13 was the number of definite articles. Variable 14 was the number of indefinite and relative pronouns. Variable 15 was the mean word position of the first noun in a sentence (or part sentence). Variable 16 was the number of sentences (or part sentences). Variables 17-24 were the number of V-V, V-N, V-O, N-V, N-N, N-O, O-V and O-N transitions where V = verb, N = noun and O = other.

Variables 1, 3, 8, 11 and 13 were initially removed since they were highly autocorrelated although they were restored later when autocorrelation was built into the model. A modification of the usual pooled t test statistic was used to test the hypothesis whether the style of F differed significantly from the other two making allowance for the fact that these two themselves might differ significantly from each other. Variables 3, 7, 14, 16, and 23 were judged to be significant at the 5% level.

Because, if the N and D portions of the gospel were written at the same time, the estimate of the variability of the text effect would in fact be measuring that of the genre effect since the text effect would be non-existent, it was decided to include in the study the first Johannine letter (referred to subsequently as J) which was divided into 19 blocks of 100 words. It is almost certain J was written later than N. Hence N and J were used to estimate the variability of text effect which because of confounding included genre effect also. This variability was then used in the modified t test statistic to see if F and N (both belonging to the narrative genre) were significantly different. Since, if anything, the variability was inflated by genre this approach would be conservative. Variables 8, 15 and probably 16 were significant at the 5% level. The estimate of the inter-text variability was based on only two observations, meaning it would not be very precise. Since the critical value for the test statistic was a function of its true value it was decided to integrate this parameter out which resulted in no change in variables that were judged to be significant. However when further the test statistic was conditioned on its observed

value, variable 8 and probably 10 were now significant at the 5% level.

Variables 1-14 and 17-24 may also be regarded as binary - that is each word for variables 1-14 or pair of words for variables 17-24 may be coded as 1 or 0 depending on whether the particular property is present or not. The above stages of analysis were carried out using a binary random effects model with very similar results.

This investigation concluded that the balance of probability was that the style of F differed from that of N and D and indeed J by more than could be expected by chance even when an effort to take genre into account was made. However the decision was not as clear cut as one might have wished. If more undisputed texts (say 8) were available, the result could well have been quite significant. Also if more sophisticated (linguistically speaking) variables were utilised, a more decisive result could very well eventuate. Because a statistically significant test statistic was obtained with some of the 24 variables chosen, under the Baconian rule of the greater force of one counter-example than that of several supporting examples, it must be concluded that earlier, less statistically sophisticated stylometric investigations of John's gospel, which demonstrated its unity, have been superseded by this study.

Because it appears that numerous scholars champion either the unity of or some form of source theory for John's gospel it seems appropriate that the results of this investigation, while consistently supporting sources in a statistically significant sense, do so to a measured degree only, suggesting that New Testament scholars have not been completely misled in their more intuitive and less statistically sophisticated evaluation of the style of John's gospel. This investigation does not establish whether all the signs source was reproduced in the gospel or not or indeed whether that part which was, has been done so in a faithful manner. What has been demonstrated is that Fortna's reconstruction has a style sufficiently different from other narrative in the gospel to suggest a different origin.

References

1. Mosteller, F. and D. L. Wallace. *Inference and Disputed Authorship: The Federalist*. Reading, Massachusetts: Addison-Wesley, 1964.
2. Felton, T. R. *Stylometry - An Example*. Conference Proceedings. Operations Research Society of New Zealand / New Zealand Statistical Association Annual Conference (1994): 350-3.
3. Fortna, R. T. *The Gospel of Signs: A Reconstruction of the Narrative Source Underlying the Fourth Gospel*. Cambridge: Cambridge University Press, 1970.