Applying cluster analysis to the platonic question

by

H. THESLEFF
K. LOIMARANTA
Université de Turku - FINLAND

0. The purpose of this paper is to present some preliminary results of a two years research project started late in 1980 at the University of Helsinki by H. Thesleff (philologist, Helsinki) and K. Loimaranta (statistician, Turku).

1.- The 'Platonic Question' and stylometry.

Statistics of linguistic data used for classifying texts, here called 'stylometry' (adopting a term coined by Lutoslawski in 1896), have been associated with Platonic studies for more than a hundred years. Stylometry has been applied above all to the question of the relative chronology of the supposedly authentic texts contained in the Corpus Platonicum, i.e. roughly the half of altogether some 55 writings (counting the Republic and the Laws as single items and the Epistles as separate ones); to a less extent stylometry has been applied to questions of authenticity in the Corpus. Because various stylometrical methods had manifestly contributed to the separation of a 'late group' of Platonic texts (Laws and possibly Epinomis and the 7th Epistle, Philebus, Sophist-Politicus, and Timaeus-Critias), and they also seemed to indicate the existence of 'late group associates' (Parmenides, Phaedrus, Theaetetus) to be placed after the Republic, it has been customary since the 1920s to consider stylometry as potentially more reliable than other methods used for determining Platonic chronology. Although the separation of the 'early' and 'middle groups' of dialogues was effectuated mainly on other grounds, it is often thought that a cautious but extensive application of refined stylometry to all of the Platonic dialogues would produce an ever more precise picture of the order in which they were written. This is still the position taken by e.g. W.K.C. Guthrie in his important 'History of Greek Philosophy' (Vols. IV-V, 1975-78).

However, serious doubts as to the reliability of stylometry as hitherto practised have been expressed from various quarters, also by scholars who are not in principle averse to the method. The last to attempt a large-scale reassessment of the problem, L. Brandwood in his London thesis of 1958, compared the widely divergent results of different stylometrical studies and concluded that apart from the separation of the 'late group' and its 'associates' (and possibly the determining of the internal order within the 'late group'), very little has in fact been achieved by Platonic stylometry. In recent years Platonic stylometry has made even less progress, and a reluctance to take a definite position to the 'Platonic Question' is, for a variety of reasons, more widespread than before.

On the other hand, we are now provided with several new tools. There is Brandwood's new 'Word Index to Plato' (1976). The Platonic corpus and many comparable texts have been recorded on magnetic tape. And there is a constant development going on in computer technique and the science of statistics.

The Writers of the present paper are convinced that, in order to explain the shortcomings of conventional Platonic stylometry, and before a new approach to it is made, it is important to examine some of the premises commonly taken for granted in this kind of study. In particular, we should like here to call attention to the complications involved with the following three assumptions which are

commonly made or tacitly implied. A more detailed discussion of these complications will be published by H. Thesleff in a book called 'Studies in Platonic Chronology', forthcoming in 1982.

a) 'The Corpus Platonicum can be divided unambiguously into authentic and spurious writings'. It seems that several of the texts in the Corpus which are not unambiguously written by Plato himself, must have originated in his immediate environment, perhaps under his supervision, e.g. Alcibiades I, Hippias Major, or Ion. On the other hand Plato's so-called 'late style' includes mannerisms, such as the avoidance of hiatus and the preference of certain clausulae, which are easier to explain as adopted (from Isocrates ?) by pupils of Plato, than as adopted by Plato in his old age. So, in partial agreement with the position taken by S. Michaelson, A.Q. Morton and D.A. Gillies in their paper of 1977, though largely for different reasons, we find it highly questionable whether all of the so-called authentic works were entirely composed by Plato personally, and whether all of the so-called spuria are entirely unauthentic. And we find it methodically important to take also the latter texts into consideration. This is not usually done.

b) 'The authentic texts can be put into a single chronological chain'. Though theories of revision have not been popular with students of Plato, it seems today unavoidable to assume that some of the longer dialogues have been successively revised by Plato, or by his friends at the Academy during his lifetime. It is unclear, in principle, to what extent such revision may have affected linguistic details. But insofar as we do not know exactly what portions of a text have become later added, or later rewritten, and we do not know to what extent Plato and his circle were working on several texts at the same time, the assumption of a single chronological series is highly problematical.

c) 'The frequencies of the linguistic data to be studied are linearly dependent on chronology'. Few would today subscribe to Lutoslawski's 'Law of Affinity' according to which there exist linear trends of change in every author's linguistic usage, and these linguistic trends would make it possible to arrange all writings of any author in a chronological chain, by interpolation or extrapolation, if only two or some of these writings can be dated on other grounds. And critics of stylometry have very often noted that the assumption of linearity is particularly questionable in Plato's case, because he (and we would add : his circle) obviously varied the linguistic expressions according to artistic and other purposes, and depending on the audience addressed. Consequently, to see or to expect a direct correlation between changes in linguistic usage and a detailed chronological order of the texts studied, is mere illusion.

It seems to us, however, that if all of the writings of the Platonic corpus, and possibly different portions of text within these writings, could be put into approximately homogeneous groups according to criteria of linguistic affinity, some new light would possibly fall on the problems of authorship and/or chronology. Bearing in mind the complications referred to above, and by analyzing a sufficiently large body of linguistic data from the whole of the Corpus, we should perhaps be able to group together texts, or portions of texts, of approximately similar linguistic structure. Electronic data processing and modern multivariate statistical methods seem to facilitate such new approaches to

Platonic stylometry. In the first place, Cluster Analysis appears to provide a suitable way of approach.

2.- Cluster analysis.

The application of Cluster Analysis to a body of linguistic data requires a so-called Contingency Table based on frequency statistics of various common words and phrases. The table has, theoretically, the following general appearance :

word / text	1	2	. . .	c	Total
1. Euthyphro	n_{11}	n_{12}	. . .	n_{1c}	$n_{1.}$
2. Apology	n_{21}	n_{22}	. . .	n_{2c}	$n_{2.}$
- - - - -	- - - - - - - - - - - - -				- - - -
Sums	$n_{.1}$	$n_{.2}$	. . .	$n_{.c}$	$n_{..}$

The rows represent the different texts (writings, or portions of writings) to be considered, and the columns represent the different words (or phrases) counted. The numeric value of n_{ij} gives the frequency of the j:th word in the text number i; thus n_{11} is the frequency of the first word of the word-list in the first text (Euthyphro). The values of $n_{i.}$ and $n_{.j}$ are the marginal sums.

Our problem is now to gather to one group those rows where the distribution of the word-frequencies is as similar as possible. The row total is taken as a given constant measuring the size of the text.

Cluster analysis includes several methods based on different principles. The object of these methods is to divide the rows, or entities as they are usually called, of a contingency table into clusters. It is a common practice in cluster analysis today to use more than one method, especially if the application domain is a new one.

We believe that the so-called optimizing methods are best suited for dealing with the present problem. It is typical of these methods that a distance- or dissimilarity-measure is defined and that the number of clusters is given. The partitioning of the entities into clusters is performed so that the total of the within cluster distances is minimized. These methods are also called relocation methods, because the optimal solution is acquired by moving, 'relocating', the entities from one cluster to another. The Mixture Model method is related with the relocation methods, but behind it lies a more sophisticated probabilistic approach to the problem.

When we started our project we had at our disposal two clustering methods for a large computer : Method 1 : A standard relocation method with Euclidean distance-measure, the maximum number of variables (columns) being 20. Method 2 : A Mixture Model method for multinomial distribution. This method had been developed by Finnish insurance companies especially to treat contingency tables, i.e. materials like ours.

In the beginning of 1981, we programmed for a micro-computer (HP-85) two further programs for cluster analysis : Method 3 : A relocation method with χ^2–measure. Method 4 : A relocation method with G^2–measure. Both these methods can treat a contingency table with maximum 5200 cells. The number of clusters was maximated to 12. For Method 5, see below.

NOTE : The χ^2–measure is the standard measure to test the homogeneity of contingency tables. The G^2–measure is a similar measure, but it has some theoretical and practical advantages. In many cases the difference between the measures in unimportant, but if the cell-frequencies are small or the inhomogeneity is high the difference can be considerable. We have

$$\chi^2 = \Sigma \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \quad ; \quad G^2 = 2 \Sigma \log (n_{ij}/m_{ij})$$

where $m_{ij} = \frac{n_{i.}\ n_{.j}}{n_{..}}$

3.- Analyses performed.

The analyses were performed partly at the Computing Centre of the University of Helsinki, partly (by K. Loimaranta) with the aid of a desk-size computer (HP-85). We began with two bulks of material :

Material 1 : 114 relatively common formulae of reply (including negative ones not considered by von Arnim in 1912) occurring in 'elenchus' sections and the like (i.e. sections where the dialogue is characterized by comparatively brief questions or statements by the leader of the discussion, and brief replies). The frequencies were counted manually from Burnet's edition of Plato, with the aid of Brandwood's Word Index. Because of the restriction to elenchus sections, such writings as the Axiochus, Epistles and Timaeus were not considered.

Material 2 : 107 relatively common particles and particle combinations occurring in the entire Corpus (excluding the Definitions); for comparison, 6 writings of Xenophon were also studied. For the analyses, magnetic tapes of the texts of Plato and Xenophon were acquired from the Thesaurus Linguae Graecae (Irwine, California). The frequencies were counted separately in elenchus sections (as above), sections of more complex dialogue, and sections of continuous exposition; quotations were not considered.

Material 1. Methods 1 - 4 were applied. Results :

- With 2 clusters all four methods in a very similar manner separated the 'late group' and the 'late group associates' (including most of the Republic) from the rest (including Republic I). The place of Phaedo and Parmenides 1 (i.e. the first part of the dialogue) remained ambiguous.
- With 3 clusters all methods split the 'rest' but kept the 'late group' unchanged.
- A partition into 4 or more clusters gave satisfactory results only with method 4.

Comments : With method 1 the restriction to 20 formulae (though the most common ones were selected) obviously provided too limited a number of variables. With method 2 the program stopped after 3 clusters owing to too many zeros in the cluster sums. With method 3 the distribution to the clusters was too strongly guided by rare formulae. With method 4 there began to appear differences between the 'late group' and the 'late group associates', and a certain tendency to separate reported dialogues such as Euthydemus and to keep together the group Euthyphro, Hippias Major and Laches (which Thesleff considers semi-spurious) was recognizable. However, the material and the methods were apparently inadequate for determining finer distinctions. And the relative homogeneity, on this level, of the language of the Corpus, a part from the 'late group' and its associates (to which the main part of the Republic seems to belong), is notable.

An example of computer output with method 4 and 5 clusters :

Cluster 1 : R V R X Phdr Soph Polit Philb L I L X
Cluster 2 : HpMi Ion Crat AlcII Eryx Hippar Just Min Sis Theag
Cluster 3 : Charm Cri Euthphr Lach Gorg HpMa Lys Men R I AlcI Virt
Cluster 4 : Prot Euthd Symp Amat
Cluster 5 : Phdo R II Parm1 Parm2 Theaet

Material 2. Only methods a and 2 were applied. Results :

- Method 1 indicated that an analysis of the distribution of the 20 commonest particles (and combinations) does not suffice to distinguish Xenophon from Plato.
- Method 2 : From 6 clusters upward Xenophon was separated to one cluster.
- The other clusters distinguished roughly the 'late group', the 'late group associates', and the three types of exposition.
- The greatest variation occurred on the last-mentioned axis (elenchus - mixed dialogue - continuous exposition).

Comments : The negative result of the application of method 1 (cf. also Material 1) gives an additional warning not to trust conclusions as to authenticity or chronology drawn from a low number of linguistic data. With method 2, the different character of the language of Xenophon was obvious enough. Hence the relative homogeneity of the language of the Platonic corpus, on the particle level, should again be noted as a remarkable fact; and this gives some support to our view of the complications

of types (a) and (b) above. Another fact to be considered in subsequent studies is that there are on the whole greater differences between the different types of exposition separated here, than between the various writings. Yet the 'late group' and its 'associates' seem to stand out somehow.

'Two-way analyses'.

If a contingency table has, as in our case, a large number of columns, we can also apply cluster analysis to the latter. We then put together into one cluster those columns (e.g. particles) which show up a similar distribution among the rows (texts). Let us assume that we have a good division of columns into clusters and we wish to form row-clusters. We can expect to get more reliable results if we do not use the meagre frequencies in the original table, but use instead the bolder frequencies of the column-clusters. Vice versa clustering the columns, we can preferably make use of the sums of the row-clusters. We can do both these clusterings simultaneously, assuming that the measure used has a certain algebraic property. The G^2 has such a property.

This new method 5 was first and tentatively applied to the *elenchus* section of Material 2.
The computer output for 12 clusters was as follows :

Cluster 1 : Euthphr Lach HpMa
Cluster 2 : Marm2
Cluster 3 : Amat HpMi Min Just
Cluster 4 : Soph Polit L I
Cluster 5 : Cri Crat AlcI AlcII Hippar Charm Prot Gorg Men Ion R I
Cluster 6 : R III - X
Cluster 7 : Phdo Theaet Parm1 Symp Phdr
Cluster 8 : Lys Euthd
Cluster 9 : Theag
Cluster 10 : Virt Demod Eryx
Cluster 11 : Philb L X
Cluster 12 : Sis

Comments : Not here the separation of various odd pieces as singletons (the Axiochus was not included in this material), and the keeping together of the 'late group' (clusters 4 and 11), its 'associates' (clusters 7 and 6), and dialogues which otherwise seem to belong together somehow (clusters 1 and 8). The occurrence of two obvious 'spuria', Alcibiades II and Hipparchus, in cluster 5 indicates a short-coming of this method and at the same time, again, the relative homogeneity of the language of the Corpus (cf. complications (a) and (b), above). In any case this reasonably satisfactory division of the writings into 12 clusters is encouraging.

Thus we find it a promising task to develop further the use of modern statistical methods and the application of them to larger bodies of linguistic material in the Platonic corpus.