Multivariate Mixture of Normals with unknown number of components. An application to cluster Neolithic Ceramics from Aegean and Asia Minor

Ioulia Papageorgiou[1] and Ioannis Liritzis[2]

[1] *Department of Statistics, Athens University of Economics and Business, Patission 76, 10334 Athens, Greece*

[2] *Laboratory of Archaeometry, Dept of Mediterranean Studies, University of the Aegean, 1 Demokratias Ave, Rhodes 85100, Greece*

ABSTRACT

Multivariate techniques and especially cluster analysis have been commonly used in Archaeometry. Exploratory and model-based techniques of clustering have been applied in geochemical (continuous) data of archaeological artifacts for provenance studies. Model-based clustering techniques like classification maximum-likelihood and mixture maximum likelihood had been used in a lesser extent in this context and although they seem to be suitable for such data, they either present practical difficulties -like high dimensionality of the data- or their performance give no evidence to support that they prevail on the standard methods (Papageorgiou *et al.*, 2001). In this paper standard statistical methods (hierarchical clustering, principal components analysis) and the recently developed one of the multivariate mixture of normals with unknown number of components (see Dellaportas and Papageorgiou, 2005) in the category of the model–based ones, are applied and compared. The data set comprises of chemical compositions in 188 ceramic samples derived from the Aegean islands and surrounding areas.

KEYWORDS: CERAMIC COMPOSITIONS, CLUSTER ANALYSIS, MIXTURE MAXIMUM LIKELIHOOD, REVERSIBLE JUMB, OUTLIERS, XRF, AEGEAN.

1. INTRODUCTION

Provenance studies of the raw materials used during prehistoric lithic industry are of key importance in researchers on ancient man. During Palaeolithic, this provides basically information on the extension of the territory exploited by small groups of hunter-gatherers. In the Neolithic and Bronze Age provenance studies contribute to the knowledge of long-distance circulation and exchanges of raw materials and goods, hence on the *chaines operatoires* of lithic and clay artifacts. Indeed, reconstructing mobility strategies is a major goal of researchers interested in prehistoric hunter-gatherers and the use of geochemical source characterization of ceramics found at sites in a region offers a way to reconstruct the precurement range, or distance traveled to obtain resources of prehistoric groups.

Pottery, due to its remarkable storage properties was a vital item used in every day life food activities. Not only but aesthetic qualities was frequently used by ancient man. Ceramics is also one of the preferred materials in provenance studies. This is because of its mode of formation from characteristic clay sources the physico-chemical properties are most often different at a major, minor but mainly trace element level.

Early ceramic provenance studies were based on bulk physical properties, such as, typology, technology, etc, as well as on petrography. Although useful for sample description, these observations generally do not provide valuable criteria for provenance studies.

The impact on characterization studies was made during sixties when spectroscopic methods allowed the determination of elemental compositions from small-sized samples. Since then till today nearly all provenance studies are based on elementary composition. Among the destructive methods of analysis are electron microprobe (for about 10 major elements), neutron activation analysis (up to ~27 major to trace elements), ICP-MS/AES, with up to more than 50 elements determined, Optical Emission Spectroscopy, Atomic Absorption Spectroscopy, PIXE, and XRF, depending of instrumentation availability and allowance to sample in a destructive manner (Pollard & Heron, 1996).

However non-destructive analysis is progressively used employing X-ray fluorescence (Liritzis *et al*., 2002)

In this study the characterization of the presently analyzed ceramics was made with the application of standard statistical methods such as hierarchical Clustering Analysis and Principal Components Analysis as well as model based clustering of Multivariate Mixture of Normals with unknown number of components (see Dellaportas and Papageorgiou 2004).

Statistical Analysis and data reduction employing Multivariate techniques lead to a number of various variables that characterize a certain group of objects (ceramic in this context). The question is defining groups in the data set, based on their compositional proximity. Such a comparison would result in groupings of ceramics and the raw materials they derive from. Because of the nature of the data (a number of continuous variables) and the problem of identification of such distinct groups, cluster analysis is the most appropriate multivariate method to use and has been widely used in bibliography together with Principal Component Analysis (PCA).

In the next section we briefly describe mixture maximum likelihood, since the innovative technique employed in this work is directly linked with this methodology. In fact this approach tends to overcome the disadvantages of mixture likelihood. A brief also presentation and the idea beyond the novel methodology is given. An application is made as a case study on chemical element composition of ceramics derived from prehistoric settlements in the wide region of the Aegean and its results are presented in the section titled Statistical Analysis. Finally, the obtained groupings are discussed along with current archaeological evidence and statistical evaluation.

2. STATISTICAL MODEL –BASED METHODOLOGIES FOR CLUSTERING

Model-based statistical methodologies assume that the observations forming the data are generated from a distribution. Usually the distribution is normal and because the dimension of the data is higher than one, it is multivariate normal. The assumption of the normality is not essential, but quite common in such techniques. Mixture maximum likelihood approach assumes a mixture of multivariate normals regarding the data distribution. Let **x=(x**$_1$**,x**$_2$**,…,x**$_n$**)** denotes the data table with $\mathbf{x_j}$ to be a p-vector, representing the j observation (a ceramic sherd in this context) and p is the dimension of the data (number of variables that for each subject we have available measurements). Under this approach the data are coming from a population with density

$$w_1 f(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+w_2 f(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)+\cdots+w_g f(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) \qquad (2.1)$$

where $f(x;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ is the density of the multivariate normal distribution, the mixture is assumed to be finite consisted of g components with the same distributions, but different parameters and w_k , *k=1,2,…,g* are the weights with $\sum w_k = 1$. Weights w_k represent the probabilities that a case $\mathbf{x}_j$ belongs to the *k*th component.

The likelihood function for a sample of size *n*, will be

$$\prod_{i=1}^{n}\sum_{k=1}^{g} w_k f(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k).$$

Clustering the data will result after estimating the unknown parameters in the population, $\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k, w_k$ *(k=1,…,g)*. $\boldsymbol{\mu}_k$ is a vector of *k* parameters and $\boldsymbol{\Sigma}_k$ is a symmetric matrix, thus p×(p+1)/2 parameters for each of the g matrices. Following the estimation of the parameters and weights in mixture, clustering the data in groups (components) will be performed in the basis of w_k. More precisely an observation $\mathbf{x}_j$ is classified to the component with the largest weight w_k.

There are two disadvantages of different nature in this technique. A practical one, that induces problems in the progress of the method and a rather fundamental and methodological one. The practical problem arises from the fact of the large number of parameters needed to be estimated in contrast with the small number of observations available from excavation. Estimation suffers and a way to overcome this is to restrict ourselves imposing constrains in parameters among components to minimize the total number. The most usual constrain is matrices $\boldsymbol{\Sigma}_k$ to be equal across *k=1,…,g*. Expectation-Maximization Mixture (EMMIX) is a software discussed in McLachlan *et al.* (1999) that implement mixture maximum likelihood. The second problem is more fundamental. It is needed for the technique to have pre-fixed the number of the components in the mixture in order to work. This leads to the necessity of reliable statistical tests to define the number of component in the finite mixture as a separate problem with the estimation. The methodology of maximum likelihood has to be executed for a variety of values for g (the number of components) and at a later separate stage, tests like approximate weight of evidence AWE (Banfield and Raftery, 1993) and Bayesian information criterion BIC (Fraley and Raftery, 1999) suggest the most powerful *g* value. Unfortunately AWE, BIC and other similar tests are all approximation tests. As a result, they might even not provide the same suggested *g* value (Fraley and Raftery 1998).

In an attempt to deal with the problems of estimation and choice of the number of components in the finite mixture simultaneously, another approach, based in Bayesian inference was developed and presented in Dellaportas and Papageorgiou (2005). The assumption for the basis of the problem is the same: A finite mixture with unknown number of normal components. Making use of the powerful Bayesian technique of Reversible Jump Markov Chain Monte Carlo (RJMCMC) (Richardson and Green, 1997) that allows testing models with different number of unknown parameters, it is possible to estimate the parameters in a mixture of *k* components and compare with another mixture of *l* components with $l \neq k$.

Some applications of univariate normal mixtures that use reversible jump are presented in Nobile and Green (2000), Robert *et al.* (2000), Fernandez and Green (2002), Green and Richardson (2001), Bottolo *et al.* (2003). An extension to multivariate mixture is the novelty in the approach by Dellaportas and Papageorgiou (2005). The multivariate context is appropriate for the application in clustering and moreover there are no constrains in the form of variance-covariance matrices of the components.

The method works iteratively and each iteration include a stage of testing if a move (either split or merge) would be accepted or staying in the same number of components and in any case estimate the parameters of mixture based on the data. A more detailed technical description is presented in Dellaportas and Papageorgiou (2005).

3. CLUSTERING DATA FROM AEGEAN AND ASIA MINOR

The data set under study consists of 188 samples, deriving from eight archaeological excavation sites: Yali and Pergussa near Nissiros (Dodecanese), Kalithies cave in the island of Rhodes, Sarakinos Cave in Beotia, Cetral Greece, two settlements in Cyprus and Ulucak in Asia Miror near Smyrna. Most samples have an age overlap during late Neolithic and bronze age period and a question of interest is to provide statistical evidence in exchange of goods via a *chaines repertoire* model. Samples derive from well stratified archaeological sections dated by C-14 and represent characteristic typology provided by the excavator per case.

3.1 Sample Preparation

In all ceramic sherds the outer surface was discarded to avoid weathering implying leaching/ infiltration of ions, thus altering elemental composition. Solid pieces of ceramic pieces and soils were powdered (<90 μm), dried, and measured by a portable ED-X-Ray Flourescence analyzer (ED-XRF).

3.2 The Analyser

The EDXRF field portable analyzer Spectrace 9000 TN was used with a mercuric iodide (HgI_2) detector, which has a spectral resolution of about 260 eV FWHM at 5.9 keV, and three excitation sources of radioisotopes within the probe unit – Americium Am-241 (26.4 KeV K-line and 59.6 KeV L-lineV) measuring Ag, Cd, Sn, Ba, Sb; Cadmium Cd-109 (22.1 K-line, 87.9 K- & L-line KeV) measuring Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Sr, Zr, Mo, Hg, Pb, Rb, Th, U; and Iron Fe-55 (5.9 KeV K-line) measuring K, Ca, Ti, Cr. The system was calibrated on several standard clay and brick, and the application software Fine particle of soil application was used.

4. STATISTICAL ANALYSIS AND DISCUSSION

For the total number of samples (188) the analyzer described above provided us with measurements in nine elements: Ba, Fe, Rb, K, Ti, Mn, Sr, Zr, Ca. The set of nine elements was the common subset of elements for which measurements were available for all the parts of the data.

Thus the processed data set is a nine-dimensional (188×9) data matrix. An initial hierarchical clustering allows us to separate some very clear and compact groups that separate well from the remaining. Several hierarchical techniques, like complete Linkage, Average Linkage, Single Linkage, Ward's method have been applied and agreement among all was possible about this issue. The obvious distinct groups that all above techniques agree are: Samples from Ftelia, Mykonos, with codes MFC1-28, Kalithies, Rodos with codes KR1-10, samples from Cyprus, both origins (Koufovounos and Sotiras) CK and CS

The second stage of the analysis is to remove the above described clear groups of Ftelia, Kalithies, Cyprus and singletons/outliers like SARA-25 and continuing further analysis to the remaining samples. After this "peeling-off" procedure to the data we end up with 125 samples mainly consisted of Ulucak, Yali and Sarakinos.

For this remaining data set, both hierarchical clustering and model-based (with RJMCMC) techniques have been applied and results have been compared. The algorithm converged rather quickly and suggested a three-component mixture as the most powerful model. Figures 1 and 2 present the predictive density based on all iterations of RJMCMC. Predictive density of future data is a posterior inference statistic. Samples from predictive density can be generated by sampling one, or more, data points for each sampled points of parameter θ selected from the RJMCMC iterations. The density is plotted in all possible 2-dimensional projections of the first 5 principal components. The samples (points in the figures) are also plotted in the same images. Although the images are the projections of the density, they show that predictive density captures the data quite well.

With a detailed examination of the remaining samples and figures 1 and 2 in addition with the results of the analysis in the first stage, the following observation are made:

1) The pottery from the eight sites indicate a clear intra- and inter-site correlation. Robust clusters of the major sites are clearly seen. For example, the *Ftelia group, the Yali group, the Kalithies* and *Sarakinos caves,* the two *Cyprus settlements.*

2) There appears an intrasite distribution and occasionally extremely distant "outliers". This is the case with; Sotiras CS2, and Koufovouno Cyprus, CK2. In fact these two are; a combed bowl and a monochrome flask respectively, derived from different floors and phases too, in relation to the rest. In fact the ceramics from two settlements have sub-groups implying more than one clay source, and at the same time, clusters comprising from either site, indicating communication through pottery exchange. An anticipated result accounting for their proximity and same cultural phase.
Similarly, the two Ftelia MFC1, 8 of the most later part (c.4500-4700 BC) indicate a quite different origin of clay in relation to the rest.

3) Sarakinos cave group exhibits a greater spread around an apparent central nucleus, with an obvious "outlier" SARA25, and others falling within neighbor clusters, - e.g. SARA20 close to Kalithies Rhodes, SARA30 along the elongated distribution of Cyprus groups, and several RHO (Ulucak) (77, 86, 101, 83, 87, 75, 96) form separate distinct sub-groups within SARA main cluster. Though some Ulucak sherds (RHO39 EBA, RHO107 LN, RHO80 LC, RHO81 LN, RHO92 LN) of LN, EB and LC periods, overlap with some SARA (4 EBA, 29 EBA, 3 MN, 17 MN) from Early Bronze and Middle Neolithic, while SARA42 of LN I a-b period belongs to the same subgroup of RHO: 72 LC, 95 LN, 69 LC. This interesting pattern imply possible interactions (exchange of ceramics and/or sharing same clay source), enhanced from the fact they are of the same period i.e. Late Chalcolithic / Early Bronze Age (4000-2500 BC), Late Neolithic and Middle Neolithic. This finding needs further verification.

4) Two soil samples from local floor of Ulucak settlement (RHO60, 61) though form an expected group, was not used as a clay source and pottery production. In all techniques they both are quite distant from main Ulucak cluster (s).

5) Yali and Pergussa ceramics form distinct subgroups. Several RHO (Ulucak) ones (49, 93, 38, 102, as well as, those of 72, 95, 69, 78, 108) fall within Yali subgroups but form distinct clusters, and RHO-102 is close to Pergussa one- both of LN period- but on another tree-branch. Also RHO98 resembles Yali YALD3, both of LN period, too. Such interaction is possible during Late Chalcolithic (for Ulucak) and Late Neolithic (Greek Neolithic at Yali). The two sites are close to the Asia Minor coastline, Ulucak being c.15 km from Smyrna.

6) In Ulucak, a quite interesting observation is the apparent use of a particular clay source throughout the long period of successive cultural phases (Early Bronze, Late Chalcolithic, Late Neolithic, late Early Neolithic), evidenced from subgroups containing ceramic sherds from these periods.

The extremely interesting Ulucak- Yali-Pergoussa and Sarakinos-Ulucak interaction needs further verification.

5. CONCLUSION

The attempted characterization on a diversified nature (temporal, contemporary and geographical) of pottery samples, mainly to test the success of the novel grouping model-based method, has proved highly satisfactory. An additional advantage in contrast with the non-model based clustering techniques is that fully estimation of the parameters exist after model-based classification and it is possible to classify a new incoming sample to one of the existing groups (discriminant analysis).

The obtained results indicated several useful information regarding long distance trade exchange, usage of same clay source by successive cultural phases, interaction of settlements via sea routes. Some ‘outliers’ imply very different clay sources.

The standard methods in cluster analysis used here are distribution free which means they make no use of data distribution assumption. However, in the model-based techniques applied the resulting groups overly the multivariate normal. The recently introduced iterative methodology that is applied in this paper is a model-based technique with same philosophy as mixture maximum likelihood, under a different formulation (Bayesian) and improved in the sense of it is devoid of disadvantages that mixture maximum likelihood. Standard and model-based techniques were used in our application made as a case study on chemical element composition of ceramics derived from prehistoric settlements in the wide region of the Aegean.

It is a first time to our knowledge the endeavor to group prehistoric ceramic fabric derived from seemingly distant cultures in and across the Aegean. Questions posed by archaeologists often refer to the use of common clay sources, exchange trade routes, diachronical accessibility of same clay source.

Acknowledgements: We thank Prof. Sampson, Prof. Mantzourani and Prof. Ciliniroglou for providing samples and Ministry of Culture for granting permission.
The General Secretariat for Research & Technology for partial funding of the project and the Ministry of the Aegean for funding initial stages of the project.

REFERENCES

Banfield, J. D. and Raftery, A.E (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803-21.
Bottolo L., Consonni G., Dellaportas P. and Lijoi A. (2003) Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6, 25-47.

Dellaportas, P. and Papageorgiou Ioulia (2005) Multivariate of normals with unknown number of components (to appear in *Statistics and Computing*). **http://stat-athens.aueb.gr/~ptd/finmix.pdf**

Fernandez C. and Green P.J. (2002) Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Socisety, Series B*, 64, 805-826.

Fraley, C. and Raftery, A. E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis, *Computer Journal,* **41**, 578-88.

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28, 355-376.

Lindsay B. G. (1995) *Mixture models: Theory, Geometry and Applications*. Hayward: Institute of Mathematical Statistics.

Liritzis.I (2005) ULUCAK (Smyrna, Turkey): chemical analysis with clustering of ceramics and soils and obsidian hydration dating. Mediterranean Archaeology & Archaeometry, Vol.5, Special Issue (in press).

Liritzis.I, Drakonaki.S, Vafiadou.A, Sampson.A and Boutsika.T (2002) Destructive and non-destructive analysis of ceramics, artifacts and sediments of Neolithic ftelia (Mykonos) by portable EDXRF spectrometer: first results. In Sampson.A (ed.) (2002) *The Neolithic settlement at Ftelia, Mykonos.* Dept. of Mediterranean Studies, Univ. of the Aegean, Rhodes, Greece, Chapter 11, 251-271.

McLachlan G. J., Peel, D., Basford, K. E. and Adams, P. (1999). The EMMIX algorithm for the fitting of mixtures of normal and t-components, *Journal of Statistical software*, **4**, 2.

McLachlan G. J. and Basford K. E. (1988) *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

Nobile, A. and Green, P.J. (2000). Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, 87, 15-35.

Papageorgiou, Ioulia, Baxter M.J. & Cau M. A. (2001) Model-based clustering techniques in archaeological ceramic provenance studies. *Arhaeometry*, **43**, 4, 571-588.

Pollard.M and Heron.C (1996) *Archaeological Chemistry*. The Royal Society of Chemistry, London.

Richardson, S. and Green, P. J. (1997). On the Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Socisety, Series B*, 59, 731-792.

Robert C.P., Ryd¶en T. and Titterington D.M. (2000) Bayesian inference in hidden Markov models through the reversible jump MArkov chain Monte CArlo method. *Journal of the Royal Statistical Socisety, Series B*, 62, 57-76.