Proceedings of the 8th WSEAS Int. Conference on Automatic Control, Modeling and Simulation, Prague, Czech Republic, March 12-14, 2006 (pp217-225)

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 (ioulia@aueb.gr)*

[(2)] *Laboratory of Archaeometry, Dept of Mediterranean Studies, University of the Aegean, 1 Demokratias Ave, Rhodes 85100, Greece (liritzis@rhodes.aegean.gr)*

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 $\mathbf{x}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$ 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 , $k=1,2,\ldots,g$ are the weights with $\sum w_k = 1$. Weights w_k represent the probabilities that a case $\mathbf{x}_j$ belongs to the kth 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 $(k=1,\ldots,g)$. $\boldsymbol{\mu}_k$ 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 $k=1,\ldots,g$. 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.