A Mixture Model and K-Means Cluster Approaches to Fish Production in Commercial Quantity in Nigeria.

S.S. Abdulkadir, Ph.D.

Department of Statistics and Operations Research, Federal University of Technology, Yola, Nigeria.

E-mail: saidusauta@yahoo.com

ABSTRACT

The paper uses the methods of mixture model and K- means cluster to determine the number of states that can be grouped in production of fish in large quantity for foreign market and other for local consumption. The estimation of parameters was done using maximum likelihood by means of the EM algorithm for the finite mixture model. The data used was extracted from the Annual Abstract of Statistics of National Bureau of Statistics. The results show that only one state produces an average of 72,628.89 metric tons of fish that can be considered for export purpose, while others produce below average of 72,628.89 metric tons for internal consumption. The result obtained from K-means cluster is unstable therefore the finite mixture model outperform the K- means cluster.

(Keywords: mixture model, K-mean, cluster, components)

INTRODUCTION

Mixture modeling is a widely applied data analysis technique to identify unobserved heterogeneity in a population. The model arises in a natural way as marginal model with respect to an unobserved indicator variables Z_{ij} (McLachlan and Peal, 2000; Kuhnert and Bohning, 2006; etc.) indicating if observation i belongs to cluster j. It has been noted in literature that different form of heterogeneity can be reproduced by the mixture distribution, thus, making it easy for researchers in modeling.

According to Leish (2004) mixture models can be applied in areas ranging from biology, economics and marketing. He said mixture models can be applied to data where observations originate from various groups, where group affiliations are not known, and can provide approximations for multi-modal distributions. Wedel and Kamakura (2001) says that mixture models replace more traditional cluster analysis and cluster–wise regression techniques as state of the art.

Another traditional and most popular partition method is k-means cluster, first proposed by MacQueen (1967). The first step in this approach is to partition items into k clusters, and the second is to proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. This second step is repeated until no more reassignment takes place. The k must be predetermined as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends on the user to identify this in advance. One of the major problems of this approach is stability; however, it can be checked by rerun the algorithm with a new initial partition.

THE DATA SET ON FISH PRODUCTION IN NIGERIA

There are 37 states including Abuja, Federal Capital Territory in Nigeria, where fish production activities are taking place. Although the quantity of fish production varies between the states, some states produce in large quantity that can induce foreign market to generate more hard currency for the country.

The major area where Nigeria derives almost 75% foreign earning is petroleum and it is expected that one day the oil will finish because it is a deposit in a reservoir. According to CBN, (2005) fish production is regarded as fast growth rate in agriculture; it contributed to Gross Domestic Product in 2001 N76.6 million to N162.61 million in 2005. Olagunju *et al.*, (2007) asserted that Nigeria is one of the largest importers of fish in the developing world, importing some 600,000 metric tons annually.

Thus, development of fish production will serve as a supplement to earning foreign exchange. The question now is how do we determine the states that can produce in large quantity for foreign market?

In this paper we assume variation in average fish production between the states and the amount produce is also assume to have a probability density function, therefore we are led to hierarchical modeling, that is, non-parametric mixture model approach described above. The whole states and Abuja are expected to fall into two components (production for domestic consumption and export) Here we are faced with the problem of identifying the mean production for each component and the corresponding proportion of the overall production. This situation is considered as hidden (or latent) structure, since the component to which each state or Abuja belongs remains unobserved.

In this paper, the author considers the application of mixture model and k-means cluster to the fish production for foreign market. The paper discusses method of estimation of parameters in mixture model and classification approaches in mixture model and k-means cluster. The paper also represents data to illustrate the mixture model and k-means cluster results and offers the author's conclusions.

MODELS

In this paper two methods of clustering data are employed; namely Finite mixture model and K-means cluster.

Finite Mixture Model

A finite mixture model is a mixture model with a specified number of mixture components. Let K be the number of components or clusters. If K = 1, then this is homogeneous case, or if k= n the number of entities, then this situation is most general form. Typically, estimated number of components or clusters turns out to be less than number of entities. The mixture model for variable X is given by:

$$f(x)=\sum \alpha_{jf}\, f_j(x_i \mid \theta_j) \qquad \textbf{(1)}$$

where α_j is the mixture weight of the jth component and $\sum_{j=1}^{K}\alpha_j = 1$, $\theta_j = (\mu_j, \sigma_j)$ is the component parameters and $f(x_i \mid \mu_j, \sigma_j)$ = $\frac{1}{\sqrt{2\pi}\sigma_j}\exp(-\frac{1}{2}\left(\frac{(x_i-\mu_j)}{\sigma_j}\right)^2$ is the jth parameter density called the mixture kernel. The model is for uni-variate normal mixture model that is considered in paper. For multivariate case readers are referred to Everitt and Hand (1981), McLachlan and Peel (2000) among others.

Estimation of Parameters in Finite Mixture Model

Maximum Likelihood Estimates (MLEs) of parameters of mixture models can be obtained by using Expectation-Maximization algorithm (EM) developed by Dempster *et al.* (1977). It is often the method of choice for estimating the parameters of the model using unlabeled data.. The algorithm iterates between two steps.

"E"-step: calculate the expectation of the log-likelihood over all possible assignments of data points to source.

"M"- step: maximize the expectation by differentiating w.r.t the current parameters.

Consider a given set of data ($x_1, x_2, \ldots, x_n$), one seeks the values of the parameters that maximize the log-likelihood:

$$L((\theta) = \prod_{i=1}^{n}\sum_{j=1}^{K}\alpha_j f_j(x_i \mid \theta_j) \qquad \textbf{(2)}$$

For detail of EM-algorithm one is referred to McLachlan and Peel (2000).

Determination Number of Components in Mixture Models

The number of components is generally unknown and has to be estimated.

Let the MLEs obtained in Equation 2 be denoted by $\hat{\alpha}_j, \hat{\theta}_j$, j=1,….,K

The probability that observation i with measurement x_i belongs to mixture component j can be computed as:

$$P(j|x_i) = \frac{\alpha_j f_j(x_i|\theta_j)}{\sum_{k=1}^{K} \alpha_k f_k(x_i|\theta_k)}$$

$$= \frac{\alpha_j \frac{1}{\sqrt{2\pi}\sigma_j} \ell^{-\frac{1}{2\sigma_j^2}(x_i-\mu_j)^2}}{\sum_{k=1}^{K} \alpha_k \frac{1}{\sqrt{2\pi}\sigma_k} \ell^{-\frac{1}{2\sigma_j^2}(x_i-\mu_j)^2}} \quad \textbf{(3)}$$

By Bayes' theorem, the probability of membership in component j is computed by replacing α_j and θ_j with their maximum likelihood estimates.

Cluster was formed on the mixture models by associating each point x_i with component j having the highest probability of membership according to Equation 3.

As a guideline for selecting the number of component non-parametric maximum likelihood criteria (NPLME) descending from the gradient function (Bohning, 2003) which is defined as:

$$d(\theta, p) = \frac{1}{k} \sum_{i=1}^{k} \frac{f(x_i|\theta)}{f(x_i|p)} \quad \textbf{(4)}$$

and derived from the directional derivative was employed in this paper.

The implementation of estimation of parameters in the finite mixture model, classification of observations into components and the likelihood performance criteria were done using package for Computer Assisted Analysis of Mixture (C.A.MAN), Bohning *et al.* (1998),released by the author.

K-means Algorithm for Clustering

This approach involves three steps:

(i) Partition the items into k initial clusters

(ii) Proceed through the list of items; assigning an item to the cluster whose centroid (mean) is nearest. This is done using Euclidean distance with standardized or unstandardized observations. Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item

(iii) Repeat Step (ii) until no more reassignments take place.

The SPSS version 16 was used in the implementation of the clustering.

RESULTS AND DISCUSSION

The results obtained using finite mixture model and K- means cluster are displayed in Table 1 and Table 2, respectively. Since the objective of this paper was to determine a set of states including Abuja that can produce fish in abundant quantity for export, regarded as component one, and the remaining states as component two that will produce for internal consumption.

The mean production of fish (metric tons) obtained is similar for the two methods of classifications for the components and clusters. Also the number of states in each cluster in the k-means approach is similar to the mixing proportions if converted. The conversion is done by multiplying the mixing proportion by 37(the number of states including Abuja). For instance, in Table 1 the mean production is 6,152.68 metric tons for the first component and 58,148.86 metric tons for the second component while in Table 2 the values are 5,937.71 metric tons and 56,932.69 metric tons for the two clusters respectively.

The K – means cluster grouped 33 states into the first cluster and 4 states into the second cluster. Similarly, the mixing proportions in Table 1 corresponding to components are 0.8981 and 0.1019, respectively. If multiplied by 37(number of states including Abuja) result into approximately 33 and 4 for the two components. The finite mixture model used the sample data to determine 4 support points (i.e. 4 components) without predetermined number of components. This is not possible using K- means cluster.

Table 1: The Results of Mixture Model to Determine Number of Components.

Number of Components (K)	Component mean (μ_j)	Mixing proportion (α_j)	Log-likelihood
1	11,450.68	1	
2	6,152.966 58,148.86	.8981 .1019	-399.6462
3	2,673.717 21,778.11 63,563.09	.7179 .2010 .0811	-394.2888
4	123.56 14,624.62 58,127.82 72,628.89	.4480 .4891 .0365 .0264	-413.0498

Table 2: Result of K- Means Approach to Determine Number of Clusters.

	Initial Cluster	Final Cluster
K =2	1 2	1 2
Mean production of fish	123.56 72,628.89	5,937.71 56,932.69
No. of states in each cluster		33 4
K = 3	1 2 3	1 2 3
Mean production of fish	123.56 37,000 72,600	2,384.45 21,400 63,600
No. of states in each cluster		26 8 3
K =4	1 2 3 4	1 2 3 4
Mean production of fish	123.56 23,000 72,600 50,900	2,025.36 18,200 69,900 44,000
No. of states in each cluster		25 8 2 2

Then going by the log-likelihood (-413.0498) the appropriate number of components is four (4).Their corresponding mean productions (in metric tones) and mixing proportion shown in bracket are 123.56 (.4480), 14,624.62 (0.4891), 58,127.82 (.0364), and 72,628.89 (0.264), respectively. The same result in Table 2 also shows a similar trend up to cluster 3, but the mean of the fourth cluster is lower than what obtained in the third cluster. This might be due to stability usually occurs in cluster.

CONCLUSION

The results obtained from K-means cluster is unstable, therefore the finite mixture model outperform the k- means cluster in the classification problems. Thus, results obtained under mixture model are used to make the following conclusion. That the states that produce an average 72,628.89 metric tons should be considered to produce for foreign exchange.

There is only one state that can satisfy this condition going by the mixing proportion. Others state that produce less than average of 72,628.89 metric tons should produce for local market. The prioritization of fish production will bring more foreign exchange, employment opportunity and reduction poverty; and attract school leavers to the fishing industry.

REFERENCES

1. Central Bank of Nigeria. 2008. *Annual Report of Statement of Account*.CBN Publication: Lagos, Nigeria.

2. Dempster, A.P. N.M. Laid, and D.B. Rubin. 1977. “Maximum Likelihood from Incomplete Data via

the EM Algorithm (with discussion)". *Journal of the Royal Statistical Society*. B 39:1- 38.

3. Everitt, B.S. and D.J. Hand. 1981. *Finite Mixture Distribution-Monographs on Applied Probability and Statistics*. Chapman and Hall: New York, NY.

4. Dempster, A., N.M. Laird, N.M, and D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of Royal Statistics Society*.B39:1 – 38.

5. MacQueen, J.B. 1967. "Some Methods for Classification and Analysis of Multivariate Observations". *Proceedings of 5th Berkely Symposium on Mathematical Statistics and Probability*.1:281 -297. University of California Press: Berkeley, CA.

6. McLachlan, G. and D. Peel. 2000. *Finite Mixture Model*. John Wiley & Sons, Inc.: New York, NY.

7. Kuhnert and Bohning. 2006. "Equivalence of Truncated Count Mixture Distributions and Mixture of Truncated Count Distributions". *Biometric*, 62:1207 - 1215

8. Bohning, D. 2003. "The EM Algorithm with Gradient Function Update for Discrete Mixtures with Known (Fixed) Number of Components". *Statistics and Computing*.13:257-265.

9. Bohning, D., E. Dietz, and P. Schlattmann. 1998. "Recent Developments in C.A.MAN (Computer Assisted Analysis of Mixtures). *Biometrics*. 52:525–536.

10. Olagunju, F.I., I.O. Adesiyan, and A.A. Ezekiel. 2007. "Economic Viability of Cat Fish Production in Oyo State". *Nigeria Journal of Hum. Ecol*. 21(2).

11. Wedel, M. and W.A. Kamakura. 2001.*Market Segmentation-Conceptual and Methodological Foundations*. Kluwer Academic Publisher: New York, NY.

SUGGESTED CITATION

Abdulkadir, S.S. 2012. "A Mixture Model and K-Means Cluster Approaches to Fish Production in Commercial Quantity in Nigeria". *Pacific Journal of Science and Technology*. 13(1):384-388.

Pacific Journal of Science and Technology