A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 235, 2010

Ewa Witek[*]

THE APPLICATION OF MIXTURE MODELS IN CLUSTERING OF THE EUROPEAN UNION COUNTRIES

Abstract. In finite mixture models, each component corresponds to a cluster. In the 1990's finite mixture models were extended by mixing standard linear regression models and generalized models (Wedel, Kamakura, 1995). An important area of application of mixture models and also of their extensions is in marketing segmentation, where finite mixture models replace more traditional cluster analysis and cluster-wise regression techniques. The article presents an application of mixture models in economic analysis, i.e. clustering of the EU countries.

Key words: Mixture models, EM algorithm, information criteria, EU countries, rootogram.

I. FINITE MIXTURE MODELS

Finite mixture models are a popular technique for modelling unobserved heterogeneity or approximating general distribution functions. They are used in a lot of different areas such as astronomy, biology, economics, marketing or medicine . An overview of mixture models is given in Titterington, Smith and Makov (1985) or McLachlan and Peel (2000).

The mixture is assumed to consist of s components where each component follows a parametric distribution. Each component has a weight assigned which indicates the *a-priori* probability for an observation to come from this component and the mixture distribution is given by the weighted sum over the u components. If the weights depend on other variables, these are referred to as concomitant variables.

The mixture model is given by

$$f(y|\mathbf{x},\boldsymbol{\omega},\boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s(\boldsymbol{\omega},\boldsymbol{\alpha}) f_s(y|\mathbf{x},\boldsymbol{\Theta}_s), \tag{1}$$

where:
f_s – density function of component s,

[*] Ph.D student, Department of Statistics, The Karol Adamiecki University of Economics, Katowice .

y – dependent variable,

$\mathbf{x}$ – vector of independent variables,

$\boldsymbol{\omega}$ ($\boldsymbol{\alpha}$) – the concomitant variables and their parameters,

$\boldsymbol{\Theta}_s$ – the component specific parameter vector for the density function f_s,

$\boldsymbol{\Theta}$ – the vector of all parameters for the mixture density function $f()$, $\boldsymbol{\Theta} = (\tau_s, \boldsymbol{\alpha}_s, \boldsymbol{\Theta}_s)$,

τ_s – the *prior* probability of component *s*, $(\tau_s(\boldsymbol{\omega}, \boldsymbol{\alpha}) \geq 0 \wedge \sum_{s=1}^{u} \tau_s(\boldsymbol{\omega}, \boldsymbol{\alpha}) = 1)$, $\boldsymbol{\Theta}_s \neq \boldsymbol{\Theta}_l \forall s \neq l$.

In marketing, choice behaviour is often modelled with a use of marketing mix variables such as price, promotion and display. Based on the assumption that groups of respondents vary in terms of different price or promotion elasticities, there are mixtures of regressions which are fitted both to model consumer heterogeneity and to segment the market. Socio-demographic variables such as age and gender have often been shown to be related to different market segments even though they generally do not perform well when used to *a-priori* segment the market. The relationship between the behavioural and the socio-demographic variables is then modeled through concomitant variables models where the group sizes (i.e. the weights of the mixture) depend on the socio-demographic variables.

We assume that the component specific densities are from the same parametric families. If f_s is from the exponential family of distributions and for each component a generalized linear model is fitted (GLM's McCullagh and Nelder 1989), these models are also called GLIMMIX (Wedel and DeSarbo 1995).

The posterior probability that observation $(\mathbf{x}, y)$ belongs to class *r* is given by

$$P(r|\mathbf{x}, y, \varphi) = \frac{\tau_r f(y|\mathbf{x}, \boldsymbol{\Theta}_r)}{\sum_{s=1}^{u} \tau_s f(y|\mathbf{x}, \boldsymbol{\Theta}_s)}. \tag{2}$$

II. PARAMETER ESTIMATION OF MIXTURE MODELS

The parameters of the mixture model are usually estimated by maximum likelihood using the Expectation-Maximization (EM) algorithm (Dempster et al. [1977]). The log-likelihood of a sample of *n* observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is given by

$$\log L = \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \mathbf{\Theta}) = \sum_{i=1}^{n} \log \left(\sum_{s=1}^{u} \log \tau_s f_s(y_i|\mathbf{x}_i, \mathbf{\Theta}_s) \right), \tag{3}$$

and can usually not be maximized directly. Each EM iteration consists of two steps – an E-step and an M-step:

– E step – estimation of the *posterior* class probabilities for each observation:

$$\hat{p}_{is} = P(s|\mathbf{x}_i, y_i, \mathbf{\Theta}_s), \tag{4}$$

using equation (2) and derivation of the prior class probabilities:

$$\hat{\tau}_s = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{is} \tag{5}$$

– M step – maximization of the log-likelihood for each component separately using the *posterior* probabilities as weights:

$$l(y_i|\mathbf{x}_i, \mathbf{\Theta}_s, \tau_s, \hat{p}_{is}) = \sum_{i=1}^{n} \sum_{s=1}^{u} \hat{p}_{is} \log[\tau_s f_s(y_i|\mathbf{x_i}, \mathbf{\Theta}_s)]. \tag{6}$$

The E and M steps are repeated until the likelihood improvement falls under a pre-specified threshold or a maximum number of iterations is reached. The EM algorithm does not to have to be used for mixture models only, but it rather provides a general framework for fitting models on incomplete data. Suppose we augment each observation $(\mathbf{x}_i, y_i)$ with an unobserved multinomial variable $\mathbf{z}_i = [z_{i1}, \ldots, z_{iu}]$, where $z_{is} = 1$ if $(\mathbf{x}_i, y_i)$ belongs to class s and $z_{is} = 0$ otherwise. The EM algorithm can be used to maximize the likelihood on the "complete data" $(\mathbf{x}_i, y_i, z_i)$; the z_i encode the missing class information.

One of the well known limitations of the EM algorithm is that convergence can be slow. There can also be numerical instabilities at the margin of parameter space, and if the component gets to contain only a few observations during the iterations, parameter estimation in the respective component may be problematic. As a result, numerous variations of the basic EM algorithm described above exist, most of them exploiting features of special cases for $f(y|\mathbf{x})$.

III. MODEL SELECTION

In order to select the optimal clustering model, several measures have been proposed (McLachlan and Peel [2000]). Three information criteria are available in *flexmix* package of **R**: BIC (*Bayesian Information Criterion*), AIC (*Akaike Information Criterion*) and ICL (*Integrated Completed Likelihood*). The criteria are defined as

$$BIC_s = -2\log p(\mathbf{x}, y \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) + v_s \log(n)\,, \tag{7}$$

$$AIC_s = -2\log p(\mathbf{x}, y \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) + 2v_s\,, \tag{8}$$

$$ICL_s = -2\log p(\mathbf{x}, y, \mathbf{z} \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) - \frac{v_s}{2}\log(n)\,, \tag{9}$$

where $\log p(\mathbf{x}, y, \mathbf{z} \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.)$ is the maximized loglikelihood for the model M_s, v_s is the number of parameters to be estimated in the mixture model and n is the sample size.

The fit of a mixture model to a given data set can only improve as more terms are added to a model. Hence, likelihood cannot be used directly in the assessment of models for cluster analysis. In the criteria mentioned above a term to the loglikelihood is added to penalize the complexity of the model. The first term in BIC, AIC and ICL criteria measures the goodness-of-fit, whereas the second term penalizes model complexity. One selects model that minimizes either AIC, BIC or ICL. Quantitatively, those criteria differ only by the factor by which v_s is multiplied. Qualitatively, the criteria provide a mathematical formulation of the principle of parsimony in model building, although for large data sets their behavior is rather different.

In general, BIC was found to be consistent under correct specification of the family of the component densities (Kass and Raftery 1995, Keribin 2000), whereas BIC selected too many components when one of the true, normally distributed components was substituted by a different distribution, such as the uniform distribution. AIC tends to select too many components even for a correctly specified mixture.

IV. IDENTIFIABILITY OF MIXTURE MODELS

The identifiability of many mixture models remains open question. Statistical models are in general represented by parameter vector $\boldsymbol{\Theta}$ which consists of the component weights and the component specific parameters determine a mixture distribution, i.e., there is a mapping from the parameter space to the model space. For identifiability this mapping has to be injective, i.e. for each model in the model space there is a unique parameter vector in the parameter space which is mapped to the model. Lack of identifiability can be a problem for model estimation or if the parameters are interpreted. A comprehensive overview of this topic is beyond the scope of this paper, however, users of mixture models should be aware of this problem:

- Relabeling of components: mixture models are only identifiable up to a permutation of the component labels. For EM-based approaches this only affects interpretation of results, but is no problem for parameter itself.
- Overfitting: if a component is empty or two or more components have the same parameters, the data generating process can be represented by a smaller model with fewer components. This kind of unidentifiability can be avoided by requiring that the prior weights τ_s are not equal to zero and that the components specific parameters are different.
- Generic unidentifiability: it has been shown that mixtures of univariate normal, gamma, exponential, Couchy and Poisson distributions are identifiable, while mixtures of discrete or continuous uniform distributions are not identifiable. A special case is the class of mixtures of binomial and multinomial distributions which are only identifiable if the number of components is limited with respect to, e.g. the number of observations per person (Titterington *et al.* 1985, Grün 2002).

V. EXAMPLE

The goal of this example is to find groups of EU countries and to estimate the regression parameters in each of them. All computations and graphics in this paper were done in flexmix and clustvarsel packages of R (version 2.7.2).

The data was sourced from AMECO database[1]. The following variables in the period 1999–2007 were used in the analysis: y – gross fixed capital formation, x_1 – foreign trade shares in the world, x_2 – balance of current transaction with the rest of the world, x_3 – gross domestic product, x_4 – gross public debt, E – the euro area country (binary variable).

[1] http://ec.europa.eu/economy_finance/db_indicators/db_indicators8646_en.htm

At the very beginning of our analysis we checked variable's evidence for being useful for clustering using the variable selection for model-based clustering (Raftery and Dean 2006). The variables selected by the variable selection procedure were (in order of selection) x_3, x_2, x_1.

The optimal number of clusters was chosen using information criteria. Figure 1 gives a plot of AIC, BIC and ICL criteria mentioned in section III of this article.

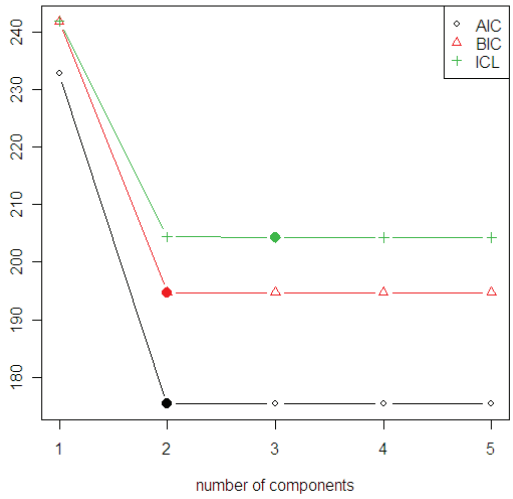

Figure 1. AIC, BIC, ICL values
Source: Own computations based on AMECO database.

In several applications, the BIC approximation to the Bayes factor has performed quite well (Fraley and Raftery [1998], [2002]), so we decided to choose the number of components according to this criterion.

The model which was chosen as the best is a finite mixture of two regression models with y – gross fixed capital formation as a dependent variable, x_1 – the trade shares in the world, x_2 – balance of current transaction with the rest of the world as independent variables and E – the euro area country as a concomitant variable. In further analysis we ran the test for significance of regression coefficients. For x_2 the coefficient of the second component was not significantly different from 0. The significant parameter estimates of both components, the estimated prior probabilities τ_s and n_s – the number of observations assigned to the corresponding clusters are presented in Table 1.

Table 1. The significant parameter estimates of two components in the mixture

Cluster	τ_s	n_s	Regression model
I	0,33	9	$y = 29{,}18x_1 + 0{,}12x_3 - 15{,}09$
II	0,67	18	$y = 4{,}24x_1 + 0{,}15x_3 + 1{,}93$

Source: Own research.

Histograms or rootograms of the posterior class probabilities can be used to visually assess the cluster structure (Tantrum, Murua, and Stuetzle, 2003). Rootograms are very similar to histograms, the only difference is that the height of the bars correspond to square roots of counts rather than the counts themselves, hence low counts are more visible and peaks are less emphasized.

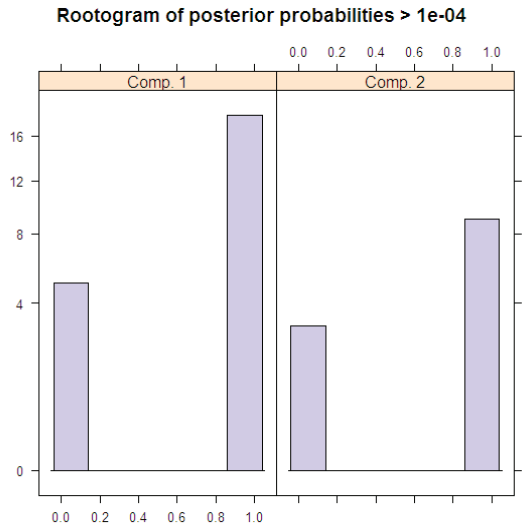

Figure 2. Rootogram
Source: Own computations based on of AMEC0 database.

A rootogram of the posterior probabilities of the observations is shown in Figure 2. It can be used for arbitrary mixture models and indicates how well the observations are clustered by mixture. For ease of interpretation the observations with *a-posteriori* probability less than $eps = 10^{-4}$ were omitted as otherwise the peak at zero dominated the plot. The posteriors of two components have modes at 0 and 1, indicating well-separated clusters (Leisch, 2004).

VI. CONCLUSIONS

We have shown the use of the mixture models in classification of EU countries. The analysis of mixture models yields two groups of countries. The first class is characterized by an average foreign trade share of 2.87%, an average GDP of € 788.22 bn € and gross fixed capital formation at the average level of € 164.2 bn. Those are mostly old Euro zone member states: Germany, Austria, the Netherlands, France, Ireland, Greece, Spain, Italy, Portugal. As far as the second class is concerned the average values are: $\bar{x}_1 = 0.75\%$, € $\bar{x}_3 = 178.69$ bn, € $\bar{y} = 33.73$ bn. This class remaining the rest of EU countries, not belonging to the Euro area yet, with the exception of Belgium, Luxemburg, Finland and new Euro area country- Slowenia. We could think that investment in the new European Union countries[2] will be much higher than in the first class of countries. However, different reports show that the greatest investment ever made by the EU through cohesion instruments (worth € 308 billion in 2004 prices) will be made in the period 2007–2013. 82% of the total amount will concentrate on the "convergence" objective, under which the poorest member states and regions are eligible. We expect that the gross fixed capital formation will be much higher for countries of the second class in 5 years time.

REFERENCES

Dempster A.P., Laird N.M., Rubin D.B. (1977), *Maximum likelihood for incomplete data via the EM algorithm (with discussion)*, „Journal of the Royal Statistical Society", ser. B, 39, 1–38.

Fraley C., Raftery A.E. (1998), *How many clusters? Which clustering method? Answers via model-based cluster analysis*, „The Computer Journal", 41, 577–588.

Fraley C., Raftery A.E. (2002), *Model-based clustering, discriminant analysis, and density estimation*, „Journal of the American Statistical Association", 97, 611–631.

Grün B., (2002), *Identifiezierbarkeit von multinomialen Mischmodellen*, Master,s thesis, Technische Univesität Wien, Vienna, Austria, Kurt Hornik and Friedrich Leisch, advisors.

Kass R.E., Raftery A.E. (1995), Bayes Factors, *Journal of the American Statistical Association*, 90, 928–934.

Keribin, C., Consistent estimation of the order of mixture models. Sankhya Indian J. Stat. v62. 49–66.

Leisch R., 2004 *Exploring the structure of mixture model components*, Compstat 2004 – Proceedings in Computational Statistics, s. 1405–1412.

Leisch R., 2004, *FlexMix: A general framework for finite mixture models and latent class regression in R*, "Journal of Statistical Software", 11(8), s. 1–18, http://www,jstatsoft,org/v11/i08/

McCullagh, P., Nelder J., 1989*, Generalized linear models,* 2 Ed. Chapman and Hall, New York, USA

McLachlan G.J., Peel D. (2000), *Finite mixture models*, Wiley, New York.

[2] 10 *countries* which joined the EU on 1 May 2004.

Raftery A.E., Dean N. (2006), *Variable selection for model-based clustering*, Journal of American Statistical Association, 101 (473), s. 168–179.

Schwarz G. (1978), *Estimating the dimension of a model*, „The Annals of Statistics", 6, 461–464.

Tantrum J. Murua A., Stuetzle W., 2003, Assesment and pruning of hierarchical model-based clustering, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining,* ACM Press, New York, NY, USA, s.197–205.

Titterington D.M.., Smith A.F., Makov U.E., 1985, *Statistical Analysis of Finite Mixture Distribution*, John Wiley & Sons, San Diego.

Wedel M., DeSarbo W., 1995, A mixture likelihood approach for generalized linear models, *Journal of Classification,* Springer, vol. 12(1), pages 21–55, March

Wedel M., Kamakura W.A., 2001, *Market Segmentation: Conceptual and Methodological Foundations,* Kluwer Academic Publishers Boston Dordrecht London.

Wedel, M., 2002, *Concomitant Variables in Finite Mixture Models*, "StatisticaNeerlandica", nr 55, s. 362–375

Ewa Witek

WYKORZYSTANIE MODELI MIESZANEK DO KLASYFIKACJI KRAJÓW UNII EUROPEJSKIEJ

Modele mieszanek, których składowe charakteryzowane są przez rozkłady prawdopodobieństw (tzw. rozkłady składowe mieszanki) już od dawna znajdują swoje zastosowanie w taksonomii. Wedel i Kamakura (1995) przedstawili pojęcie modelu mieszanek w szerszym ujęciu – rozkłady składowe określone są za pomocą funkcji regresji lub uogólnionych modeli liniowych (GLM). Modele te znajdują zastosowanie przede wszystkim w badaniach marketingowych. W artykule przedstawiono charakterystykę modeli mieszanek, sposobów estymacji jej parametrów, wyboru stosownej liczby składników mieszanki, a także przykład wykorzystania modeli mieszanek do klasyfikacji krajów Unii Europejskiej.