Ewa Genge
Uniwersytet Ekonomiczny w Katowicach

THE MULTINOMIAL MIXTURE MODEL – THE ANALYSIS OF STUDENTS' ATTITUDE TO THE SILESIA REGION

Introduction

Many statistical models involve mixture distributions in some way or other. In mixture distributions a population made up of *u* subgroups, mixed at random in proportion to the relative group sizes is considered. The interest lies in some random variable X which is heterogeneous across and homogeneous within the subgroups. Due to heterogeneity, X has a different probability distribution in each group, usually assumed to arise from the same parametric family, however, with the vector of parameter $\mathbf{\Theta}_s$ differing across the groups (*s*).

An overview of mixture models is given in Titterington et al. [1985] or McLachlan and Peel [2000, p. 81-116]. The most popular are multivariate normal mixture models (Gaussian mixture models). They are used in a lot of different areas such as astronomy, biology, economic, marketing or medicine [see i.e. Fraley and Raftery 2002, p. 611-631; Wedel and DeSarbo 1995, p. 21-55; Witek 2010a, p. 615-624; 2010b, p. 63-72]. Since the mixture of multinomial distributions is applied in the empirical part of this article we present the definition of this kind of mixture below.

1. The multinomial mixture model – definition

The data of n objects described by categorical variables $l_1,\ldots,l_m$ is considered. The data can be represented by the vector of objects $\mathbf{x}_i = (x_{ijh}; j = 1,\ldots,m; h = 1,\ldots,l_j; i = 1,\ldots,n)$ where $x_{ijh} = 1$ if the object i

belongs to the category h of the variable j. The total number of categories is given by $l = \sum_{j=1}^{m} l_j$, then the data is defined by the n by m matrix.

In the multinomial mixture model it is assumed that each observation $\mathbf{x}_i$ arises independently from a mixture of multivariate multinomial distributions defined by:

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

where:

f_s – density function of component *s*,

$\mathbf{x}_i$ – the vector of objects,

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

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

τ_s – the prior probability of component *s*;

$(\tau_s \geq 0 \wedge \sum_{s=1}^{u} \tau_s = 1), \boldsymbol{\Theta}_\mathbf{s} \neq \boldsymbol{\Theta}_l \forall s \neq l.$

The *s*th component of the mixture can be given as:

$$f_s(\mathbf{x}_i|\boldsymbol{\Theta}_s) = \prod_{j=1}^{m} \prod_{h=1}^{l_j} (\Theta_{sjh})^{x_{ijh}}, \tag{2}$$

where $\boldsymbol{\Theta}_s = (\Theta_{sjh}; j = 1,\ldots,m; h = 1,\ldots,l_j)$ and (2) formula is a product of m conditionally independent multinomial distributions of parameters $\boldsymbol{\Theta}_{sj}$.

Banfield and Raftery [1993, p. 803-821] proposed to constrain the covariances in the mixture of multivariate normal distributions, which resulted in 14 Gussian mixture models. Similarly, Celeux and Govaert [2008] imposed some constraints on the parameters of the mixture of multinomial distributions ($\boldsymbol{\Theta}$) and received 5 multinomial models.

The basic idea of this proposition is to impose the vector of components on distributions parameters $\boldsymbol{\Theta}_{sj} = (\Theta_{sj1},\ldots,\Theta_{sjl_j})$ to take the form $(\beta_{sj},\ldots,\beta_{sj},\gamma_{sj},\beta_{sj},\ldots,\beta_{sj})$, with $\gamma_{sj} > \beta_{sj}$. Since $\sum_{h=1}^{l_j} \Theta_{sjh} = 1$, we have:

$$(l_j - 1)\beta_{sj} + \gamma_{sj} = 1, \tag{3}$$

$$\beta_{sj} = (1 - \gamma_{sj})/(l_j - 1). \tag{4}$$

The constraint $\gamma_{sj} > \beta_{sj}$ can be finally written as $\gamma_{sj} > 1/l_j$. Then the vector $\mathbf{\Theta}_{sj}$ can be split into the following parameters:

- $\mathbf{a}_{sj} = (a_{sj1}, \ldots, a_{sjl_j})$, where $a_{sjh} = 1$ if h is equal γ_{sj}, $a_{sjh} = 0$ otherwise,
- $\varepsilon_{sj} = 1 - \gamma_{sj}$ corresponds to the probability that the data $\mathbf{x}_i$ arising from the sth component, such that $x_{ijh(s,j)} \neq 1$.

In other words, the multinomial distribution associated with the jth variable of the sth component is reparameterized by a center $\mathbf{a}_{sj}$ and the dispersion parameter ε_{sj}, which allows a interpretation similar to the center and the variance matrix used for continuous data in the Gaussian mixture models.

The relationship between the initial and new distribution parameters can be written as:

$$\Theta_{sjh} = \begin{cases} 1 - \varepsilon_{sj} & if \quad h = h(s,j), \\ \varepsilon_{sj}/(l_j - 1) & if \quad h \neq h(s,j). \end{cases} \tag{5}$$

Equation (2) can be for $\mathbf{a}_s = (\mathbf{a}_{sj}, j = 1, \ldots, m)$ and $\varepsilon_s = (\varepsilon_{sj}, j = 1, \ldots, m)$ rewritten as:

$$f_s(\mathbf{x}_i | \mathbf{\Theta}_s) = \bar{f}_s(\mathbf{x}_i | \mathbf{a}_s, \varepsilon_s) = \prod_{j=1}^{m} \prod_{h=1}^{l_j} ((1 - \varepsilon_{sj})^{a_{sjh}} (\varepsilon_{sj}/(l_j - 1))^{1 - a_{sjh}})^{x_{ijh}}. \tag{6}$$

This model will be denoted as $[\varepsilon_{sj}]$, in the following. On the basis of (6), three other models can be deduced:

- $[\varepsilon_s]$ – the model where ε_{sj} is independent of the variable j,
- $[\varepsilon_j]$ – the model where ε_{sj} is independent of the sth component,
- $[\varepsilon]$ – the model where ε_{sj} is independent both of the variable j and the sth component.

The most general model will also be denoted as $[\varepsilon_{sjh}]$. The number of the parameters associated with each models is given in Table 1, where $\sigma = 0$ in the case of equal prior probabilities and $\sigma = u - 1$ when prior probabilities are different for each class.

Table 1

The number of parameters of the 5 multinomial models

Model	Number of parameters
$[\varepsilon]$	$\sigma + 1$
$[\varepsilon_j]$	$\sigma + m$
$[\varepsilon_s]$	$\sigma + u$
$[\varepsilon_{sj}]$	$\sigma + um$
$[\varepsilon_{sjh}]$	$\sigma + u\sum_{j=1}^{m}(l_j - 1)$

Source: Celeux, Govaert [2008, p. 35].

2. Parameter estimation and model selection

The parameters of the mixture of multinomial models are usually estimated by maximum likelihood using the Expectation-Maximization (EM) algorithm [Dempster et al. 1977, p. 1-38]. Each EM iteration consists of two steps – an E-step and an M-step. In the M-step (for the a posteriori probabilities, obtained in E-step) new parameters of maximum likelihood given by (7) are obtained:

$$L(\mathbf{x}_i|\boldsymbol{\Theta}_s, \pi_s, z_{is}) = \sum_{i=1}^{n}\sum_{s=1}^{u} z_{is} \log\left[\tau_s f_s(\mathbf{x}_i|\boldsymbol{\Theta}_s)\right], \tag{7}$$

where $z_{is} = 1$ if $\mathbf{x}_i$ belongs to group s or $z_{is} = 0$ otherwise. Maximum likelihood estimators for each of the five models presented in Table 1 are given below. We adopt the notation:

$$e_{sjh} = n_s - \sum_{i=1}^{n} z_{is} x_{ijh}, \tag{8}$$

and $h(s, j)$ for the value which minimizes the difference given in (8).

For convenience, we assume that $e_{sj} = e_{sjh(s,j)}$.

1. Model $[\varepsilon_{sjh}]$:

$$\Theta_{sjh} = 1 - e_{sjh} / n_s . \tag{9}$$

2. Model $[\varepsilon_{sj}]$:

$$\Theta_{sjh} = \begin{cases} 1 - e_{sj} / n_s & if \quad h = h(s,j), \\ e_{sj} /(n_s(l_j - 1)) & if \quad h \neq h(s,j). \end{cases} \tag{10}$$

3. Model $[\varepsilon_{s}]$:

$$\Theta_{sjh} = \begin{cases} 1 - (\sum_j e_{sj}) / n_s m & if \quad h = h(s,j), \\ (\sum_j e_{sj}) /(n_s m(l_j - 1)) & if \quad h \neq h(s,j). \end{cases} \tag{11}$$

4. Model $[\varepsilon_{j}]$:

$$\Theta_{sjh} = \begin{cases} 1 - (\sum_s e_{sj}) / n_s & if \quad h = h(s,j), \\ (\sum_s e_{sj}) /(n(l_j - 1)) & if \quad h \neq h(s,j). \end{cases} \tag{12}$$

5. Model $[\varepsilon]$:

$$\Theta_{sjh} = \begin{cases} 1 - (\sum_{j,s} e_{sj}) /(nm) & if \quad h = h(s,j), \\ (\sum_{j,s} e_{sj}) /(nm(l_j - 1)) & if \quad h \neq h(s,j). \end{cases} \tag{13}$$

The M steps for each of five models ($[\varepsilon_{sjh}]$, $[\varepsilon_{sj}]$, $[\varepsilon_{s}]$, $[\varepsilon_{j}]$, $[\varepsilon]$) could also be written using the new parameterization $\mathbf{a}_s$ and ε_s. Then it is assumed that:

$$a_{sjh} = \begin{cases} 1 & if \quad h = h(s,j), \\ 0 & if \quad h \neq h(s,j). \end{cases} \tag{14}$$

$$\varepsilon_{sj} = 1 - \Theta_{sjh}(s,j). \tag{15}$$

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 [see Wang 1994 for more details].

In order to select the optimal clustering model several measures have been proposed [see i.e. McLachlan and Peel 2000, p. 81-116]. Four information criteria are available in `mixtools` package of **R**: BIC (*Bayesian Information Criterion*), AIC (*Akaike Information Criterion*), ICL (*Integrated Completed Likelihood*) and CAIC (*Consistent Akaike Information Criterion*). The performance of some of these criteria was compared by Biernacki et al. [1999, p. 49-71] and Bozdogan [2000, p. 62-91]. In general, BIC was found to be consistent under correct specification of the component densities [Kass and Raftery 1995, p. 928--934; Keribin 2000, p. 49-66] and has given good results in a range of applications [i.e. Fraley and Raftery 2002, p. 611-631; Stanford and Raftery 2000, p. 601-609]. The criteria used in further analysis are defined:

$$AIC_s = 2\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) - 2v_s, \tag{16}$$

$$BIC_s = 2\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) - v_s \log(n), \tag{17}$$

$$ICL_s = 2\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) + \frac{v_s}{2}\log(n), \tag{18}$$

$$CAIC_s = 2\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) - v_s(\log(n)+1), \tag{19}$$

where: $\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.)$ – is the maximized loglikelihood for the model M_s, v_s is the number of parameters to be estimated in that model, n is the number of observations in the data.

The first term in criteria measures the goodness-of-fit, whereas the second term penalizes model complexity.

3. Example

In this example the data collected by the Marketing Department of University of Economics in Katowice in 2008 were analysed. The main goal of this sampling survey was to recognize students' attitudes to the Silesia region and its

promotion. The survey comprised different areas of the Silesia region: central, the Dabrowa Basin, south, north, south-west. The respondents studied at:

- the University of Economics in Katowice,
- the University of Economics in Katowice (Rybnik Centre),
- the University of Economics in Katowice (Bielsko Campus),
- the Katowice School of Economics (Katowice Piotrowice),
- the Katowice School of Finance and Banking,
- the Czestochowa University of Technology,
- the Czestochowa School of Linguistics,
- the Academy of Fine Arts in Katowice,
- the Higher School of Applied Sciences in Ruda Slaska.

Students were asked 12 questions about their background and their attitude to Silesia, its culture, tradition and promotion.

There were 627 polls collected. The main goal of the analysis was to find clusters with similar students' attitudes to our region. The mixture of multinomial distributions were applied. All computations in this paper were done in `mixtools` package of **R** and SPSS software. Some results of `mixtools` package of **R** are presented in Figure 1.

```
> x.new<-makemultdata(slask, cuts = 2)
> multmixmodel.sel(x.new$y, comps = c(1,2),
  epsilon = 1e-03)
number of iterations= 114
1 2 Winner
AIC -3244.819 -1764.462 2
BIC -3247.039 -1771.123 2
CAIC -3247.539 -1772.623 2
ICL -3247.039 -1770.603 2
Loglik -3243.819 -1761.462 2
```

Fig. 1. The results of `mixtools` package of R

The optimal number of the mixture components was chosen using four different information criteria. Figure 1 shows that the optimal number of components is 2 (for each of criterion). We estimated parameters of two components using EM algorithm. The mixture of multinomial distribution methodology outlined before yields two groups of students consisting of 255 and 372 students respectively.

The first group comprises students who feel a strong bond with Silesia. For question: "Do you feel ties with Silesia?", 58% chose answer "yes", 32% – "rather yes". There were no negative answer. Students are also rather intent on staying in Silesia: 61% of students are going to stay in Silesia, 34% have not decided yet and 5% are going to leave. The students in this group like Silesian traditions. The question "Do you like Silesian traditions?" elicited 37% "yes" answers and 46% "rather yes" answers. As far as the Polish Silesian dialect is concerned, the majority of students like it (38% "yes" answers and 28% "rather yes" answers). However, 33% of students do not like it too much (the percentage of students who chose answers: "neither yes, neither no"). High pollution is perceived as the main disadvantage of living in the Silesia area (64% "yes" and 28% "rather yes" answers). Nearly three-quarters of students polled believe in the improvement of the Silesia's image. However, as many as 75% of students did not observe any Silesia's promotion. There were different opinions concerning Silesia's promotion in our country: 38% think that the Silesia region should be promoted as a whole, 24% claim that the separate subregions should be promoted and 38% think that the separate subregions should be promoted but under the common logo of the Silesia region. Silesia is perceived as a region attractive for tourists by 42% of students, 26% think the opposite and 32% do not have any opinion. We can say that students of this group have a positive attitude towards Silesia. We can suppose that this kind of attitude and the sense of belonging to this region stem from students' background. 70% of students of this group were brought up here and their parents come from here, 21% of students have been living in Silesia for years, but their parents come from another part of Poland, only 8% of students polled came here just to study.

Quite a different attitude towards Silesia can be observed in the second group of students. The ties with Silesia are quite weak, i.e. only 39% of students feel strong ties with Silesia, 27% feel some kind of bond, 20% of the respondents feel no ties with Silesia, 13% haven't even thought about it. Only 46% of students have decided to stay here in the future, as many as 17% are intent on leaving and 37% haven't taken any decision on this issue yet. The students belonging to this group do not like Silesian traditions very much: 23% chose "yes" answers, 31% chose "rather yes" answers, 16% do not like the traditions at all. The last part of this group do not have any opinion (answer "neither yes, nor no"). The vast majority of this group do not like the Silesian dialect either. The question "Do you like the Silesian dialect?" elicited 30% "no" answers and 20% "rather no" answers. The positive attitude to the infrastructure development is almost at the same level in both groups. The air pollution in this region is also very negatively perceived in the second group of students. As far as the im-

provement of the image of the Silesia region is concerned, 5% less than in the first class believe that it is at all possible. Most of the students have not observed the new promotional campaign (64%), but there are also 12% of students who like it very much (16% have no opinion). There are also different opinions about the way of promoting the Silesia region, similarly to the first group. The vast majority of students (35%) think that the separate subregions should be promoted but under the common Silesian logo. A large part of this group perceives Silesia as unattractive for tourists (35%), 34.7% of students do not have any opinion. For 40% of the respondents, Silesia is as an industrial area, comprising an area of the former Katowice voivodship, for 28% of students Silesia is a region associated with the current area of this part of Poland. However, as many as 12% less than in the first group of students do perceive the Dabrowa Basin as a separate part of Silesia. We think that the reason of this split approach is that many people looking for a job came and settled down in this part of Silesia many years ago.

We think that the definitely skeptical attitude to the Silesia, its customs, dialects, tradition and different Silesian borders in this group is connected with students' and their parents' background. 59% of students and their parents come from Silesia, 29% of parents come from other regions of Poland and 12% of students came only to study here.

Conclusions

We have shown the use of the mixture models in the classification of students studying in different parts of Silesia. The mixture of multinomial models analysis yields two groups of students. The first group comprises students who feel strong ties with Silesia. The bond with Silesia in the second group of students is quite weak.

The mixture model analysis has confirmed that students' and their parents' background has the influence on those two different attitudes. The difference can be especially observed among students living/studying in the Dabrowa Basin. Administratively, they feel Silesian. They live in this region, but do not have the roots here, so they do not necessarily identify with everything that Silesia is connected with.

Literature

Banfield J.D., Raftery A.E. (1993): *Model-based Gaussian and Non-Gaussian Clustering.* "Biometrics", No. 49.

Biernacki C., Celeux G., Govaert G. (1999): *Choosing Models in Model-based Clustering and Discriminant Analysis*. "Journal of Statistical Computation and Simulation", No. 64.

Bozdogan H. (2000): *Akaike's Information Criterion and Recent Developments in Information Criterion*. "Journal of Mathematical Psychology", No. 44.

Celeux G., Govaert G. (2008): http://www.mixmod.org/IMG/pdf/statdoc_2_1_1.pdf.

Dempster A.P., Laird N.P., Rubin D.B. (1977): *Maximum Likelihood for Incomplete Data Via the EM Algorithm (with discussion)*. "Journal of the Royal Statistical Society", No. 39, ser. B.

Fraley C., Raftery A.E. (2002): *Model-based Clustering, Discriminant Analysis, and Density Estimation.* "Journal of the American Statistical Association", No. 97.

Kass R.E., Raftery A.E. (1995): *Bayes Factors*. "Journal of the American Statistical Association", No. 90.

Keribin C. (2000): Consistent Estimation of the Order of Mixture Models. "Sankhya Indian Journal Statistics", No. 62.

McLachlan G.J., Peel D. (2000): *Finite Mixture Models*. Wiley, New York.

Stanford D., Raftery A.E. (2000): *Principal Curve Clustering with Noise*. "IEEE Transactions on Pattern Analysis and Machine Intelligence", No. 22.

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

Wang P. (1994): *Mixed Regression Models for Discrete Data*, *PhD thesis.* University of British Columbia, Vancouver.

Wedel M., DeSarbo W.S. (1995): *A Mixture Likelihood Approach for Generalized Linear Models.* "Journal of Classification", No. 12.

Witek E. (2010a): *Analysis of Massive Emigration from Poland – the Model-based Clustering Approach.* Proceedings of the 32nd Annual Conference of the Gesellschaft für Klassifikation, Springer.

Witek E. (2010b): *Wykorzystanie mieszanek rozkładów w regresji*. W: *Współczesne problemy modelowania i prognozowania zjawisk społeczno-gospodarczych*. Red. J. Pociecha. Wydawnictwo UE, Kraków.

MIESZANKI ROZKŁADÓW WIELOMIANOWYCH – ANALIZA POSTAW STUDENTÓW WOBEC WOJEWÓDZTWA ŚLĄSKIEGO

Streszczenie

Mieszanki rozkładów są stosowane wówczas, gdy zbiór obserwacji charakteryzuje się nadmiernym rozproszeniem. W literaturze najczęściej są spotykane mieszanki rozkładów normalnych (*model-based clustering*). W referacie zostaną przedstawione mieszanki rozkładów wielomianowych oraz wyniki ich zastosowań do podziału studentów o podobnych postawach wobec województwa śląskiego (jego tradycji, kultury, możliwości rozwoju itd.).

Badania zostaną przeprowadzone za pomocą pakietu `mixtools` programu komputerowego **R**.