LMD konf. darbai, 209–216
1997, Vilnius

A semiparametric model for count data clusterization with application to medical data

M. Radavičius, J. Sušinskas (MII)

1. INTRODUCTION

Let $(K_1, N_1), \ldots, (K_r, N_r)$ be independent discrete observations each having a probability distribution that depends on an unobservable random variable (rv) ϑ_j, $j = 1, \ldots, r$. Suppose the rv's $\vartheta_1, \ldots, \vartheta_r$ are themselves independent and identically distributed (iid) and

$$\mathbf{P}\{\vartheta_j = \theta_i\} = p_i, \quad i = 1, \ldots, m, \quad \sum_1^m p_i = 1.$$

The problem is to estimate the parameters $\theta_1, \ldots, \theta_m$ and the prior probabilities $p_1, \ldots, p_m$ and to classify the observations (to divide into non-overlapping classes) according to a conjectural value of the unobservable rv's ϑ_j. This problem is called *classification without teaching* or *clusterization* problem. Clusterization problem for data of such a kind frequently occurs in social and medical studies. A typical example is, given r subpopulations, to classify them according to a certain feature on a basis of some agregated characteristic K_j (total number of jobless days and accidents, number of infected/recovered/new-born individuals, ect.) and the size N_j of j-th subpopulation, $j = 1, \ldots, r$.

Let us give more explicit description of the model. Denote

$$\Gamma_i(n) \stackrel{\text{def}}{=} \mathbf{P}\{N_j = n \mid \vartheta_j = \theta_i\}, \quad n = 1, 2, \ldots, \quad i = 1, \ldots, m.$$

We assume that the conditional disribution of K_j given N_j and ϑ_j is that of Poisson with the parametr $\theta_j N_j$, i.e.

$$\mathbf{P}\{K_j = k \mid N_j = n, \vartheta_j = \theta_i\} = \Pi_k(\theta_i n), \quad i = 1, \ldots, m,$$

where $\Pi_k(\lambda) = \lambda^k e^{-\lambda}/k!$, $k = 0, 1, \ldots$. Since rv's $\vartheta_1, \ldots, \vartheta_r$ are unobservable, rv's $(K_1, N_1), \ldots, (K_r, N_r)$ are iid sample from a finite mixture distribution

$$\begin{aligned} f(k, n) &\stackrel{\text{def}}{=} \mathbf{P}\{K_j = k, N_j = n\} \\ &= \sum_{i=1}^m p_i \Pi_k(\theta_i n)\Gamma_i(n), \quad k = 0, 1, \ldots, \ n = 1, 2, \ldots. \end{aligned} \tag{1}$$

The number of clusters m, the prameters $\theta_1, \ldots, \theta_m$, and the probability distributions $\Gamma_1(\cdot), \ldots, \Gamma_m(\cdot)$ are supposed to be unknown.

Poisson mixture models naturally arise in applications [3, 4, 7, 9]. A peculiarity of model (1) is that the mixture components include both the parametric $\Pi_k(\theta_i n)$ and the nonparametric $\Gamma_i(\cdot)$ parts. Since the definition of the clusters is based on the parameters $\theta_1, \ldots, \theta_m$ and we do not impose any condition on the probability distributions $\Gamma_1, \ldots, \Gamma_m$ the latter can be treated as infinite-dimensional nuisance parameters. The models of this type are called *semiparametric* [7].

In the paper we present a method for estimating the unknown parameters m, p_i, θ_i, $\Gamma_i(\cdot)$, $i = 1, \ldots, m$, and clustering the data. The method is based on the EM algorithm and nonparametric estimation theory, and follows main points of papers [5,11]. It also may be viewed as an approximation to the nonparametric maximum likelihood estimator (MLE) of the mixture distribution [2, 6, 7, 10]. We refer to the review paper of Bohning [2] for a description of other algorithms for calculating nonparametric MLE and further references. It should be noted that all of them are rather computer-intensive.

In section 2, necessary notation is introduced and some theoretical background is given. Section 3 is devoted to the description of the sequential estimation method. In the last section an application of the method to real count data is presented.

2. THEORETICAL BACKGROUND

The model under consideration is a special case of the following general framework. Let $X_1, \ldots, X_r$ be independent observations where X_j has a distribution density $\psi(\cdot, Y_j)$ with respect to σ-finite measure μ, $j = 1, \ldots, r$. Suppose the parameters $Y_1, \ldots, Y_r$ are themselves iid rv's with a common distribution G. If $Y_1, \ldots, Y_r$ are unobservable, $X_1, \ldots, X_r$ is iid sample from a mixture density

$$\psi(x, G) \stackrel{\text{def}}{=} \int \psi(x, y) G(dy). \tag{2}$$

Model (1) is obtained from (2) by the following substitutions: μ is a counting measure on $\mathbb{N}_0 \times \mathbb{N}$, $\mathbb{N} \stackrel{\text{def}}{=} \{1, 2, \ldots\}$, $\mathbb{N}_0 \stackrel{\text{def}}{=} \mathbb{N} \cup \{0\}$,

$$X_j = (K_j, N_j), \quad Y_j = (\vartheta_j, N_j), \qquad j = 1, \ldots, r, \tag{3}$$

$$f(k, n) = f(x) = \psi(x, G), \qquad G(\{\theta_i, t\}) = p_i \Gamma_i(t), \quad i = 1, \ldots, m, \tag{4}$$

$$\psi(x, y) = \psi(k, n, \theta, t) \stackrel{\text{def}}{=} \Pi_k(\theta n) \chi_{\{n\}}(t), \tag{5}$$

where $x = (k, n) \in \mathbb{N}_0 \times \mathbb{N}$, $y = (\theta, t) \in [a, b] \times \mathbb{N}$, $0 < a < b$, and $\chi_A(\cdot)$ denotes an indicator function of the set A.

For the sake of convenience, we assign a weight w_j to each observation X_j, $j = 1, \ldots, r$. We assume $w_j \equiv 1/r$, if the clusterization problem of the subpopulations is considered, and $w_j = N_j / \sum_l N_l$, when clustering the individuals. This enables one to treat both cases in an unified way.

Bayes classification rule. For brevity, put

$$f_{(i)}(k,n) = f_{(i)}(x|G) \stackrel{\text{def}}{=} \mathbf{P}\{K_j = k, N_j = n | \vartheta_j = \theta_i\}\} = \Pi_k(\theta_i n)\Gamma_i(n) \\ = \psi\big(x, \chi_{\{\theta_i\}}(\cdot)\, G(\{\theta_i\}, \cdot)\big)/p_i, \qquad i = 1, \ldots, m \quad (j = 1, \ldots, r). \tag{6}$$

According to the Bayes formula, posterior probabilities $\pi_i(x)$ that, given $X_j = x$, the cluster number of the j-th observation X_j equals i (i.e. $\vartheta_j = \theta_i$), is given by

$$\pi_i(x) = \pi_i(x|G) \stackrel{\text{def}}{=} \mathbf{P}\{\vartheta_j = \theta_i | X_j = x\} = p_i f_{(i)}(x)/f(x), \tag{7}$$

$i = 1, \ldots, m$, $j = 1, \ldots, r$. The minimal classification error is obtained by Bayes classification rule: assign the observation X_j to the i^*-th cluster if

$$i^* = \arg\max_{1 \leqslant i \leqslant m}\{\pi_i(X_j)\} = \arg\max_{1 \leqslant i \leqslant m}\{p_i f_{(i)}(X_j)\}, \quad j = 1, \ldots, r.$$

Thus, the clusterization problem reduces to that of estimation of the posterior probabilities. We estimate $\pi_i(\cdot)$ by the maximum likelihood method.

Nonparametric maximum likelihood estimator (NMLE). Let $\mathcal{G}$ be the family of all probability distributions on $[a, b] \times \mathbb{N}$ endowed with the vague topology (see, e.g. [10, p. 149]). Define

$$\mathcal{G}_m = \{G \in \mathcal{G}: \ |\operatorname{supp}(G(\cdot, \mathbb{N})| \leqslant m\}, \quad m \in \mathbb{N}.$$

$|A|$ denotes the number of elements of the set A. Note that both $\mathcal{G}$ and $\mathcal{G}_m$ are relatively compact. Let us write log-likelihood function for data (3) and mixture model (2)

$$L(G) \stackrel{\text{def}}{=} \sum_{j=1}^{r} \ln(\psi(X_j, G))\, w_j, \quad G \in \mathcal{G}.$$

Definition 1. A probability distribution $\hat{G} \in \mathcal{G}$ is called NMLE iff

$$L(\hat{G}) = \max_{G \in \mathcal{G}} L(G). \tag{8}$$

If the set $\mathcal{G}$ in (8) is replaced by a close proper subset $\bar{\mathcal{G}}$ of $\mathcal{G}$, then $\hat{G} \in \bar{\mathcal{G}}$ is called restricted NMLE.

Denote by $\hat{G}_m$ restricted NMLE for $\bar{\mathcal{G}} = \mathcal{G}_m$.

Definition 2. Mixture model (2) is identifiable on $\mathcal{G}$ ($\mathcal{G}_m$) iff $\psi(\cdot, G_1) \equiv \psi(\cdot, G_2) \bmod(\mu)$, $G_1, G_2 \in \mathcal{G}$ ($\mathcal{G}_m$), implies $G_1 = G_2$.

For nonidentifiable models it is impossible to estimate $G \in \mathcal{G}$ ($\mathcal{G}_m$) consistently ($r \to \infty$).

PROPOSITION 1. *The model defined by* (1), (2), (4), *and* (5) *is identifiable. The NMLE* $\hat{G}$ *for* $G \in \mathcal{G}$ *is strongly consistent and the same holds for* $\hat{G}_m$ *provided* $G \in \mathcal{G}_m$.

Since Poisson mixtures are identifiable, so are the mixtures with kernel function (5). Consequently, from [10, theorem 5.5] we conclude that $\hat{G}$ is strongly consistent. Moreover, it follows from [7] that $\hat{G}$ is a discrete distribution with the support consisting of r or fewer points. Hence $\hat{G} = \hat{G}_r$.

Strong consistency of restricted NMLE for rather general mixture models have been prooved in [6]. Unfortunatelly, conditions imposed on a mixture kernel in [6] are not satisfied in our case. However, a special form of our model (1)–(5) enables us to overcome this obstacle.

Remark. It is worth noting that consistency result as stated in Proposition 1 usually is not of great value for practice. The point is that rather frequently N_j, $j = 1, \ldots, r$, are of the same order or even much greater than r, say, the latter being fixed. Thus, asymptotic results of a different kind are needed. We reffer to an example of real data in the last section.

The EM algorithm. For computing NMLE $\hat{G}$ we apply the EM algorithm [1, 2, 4, 9]. The EM algorithm is an iterative procedure which, given an initial value of the parameter, calculates a new improved value that increases the log-likelihood function. The parameter values obtained converge to a stationary point. If the initial value is close enough to the MLE the EM algorithm converges to the MLE. Each iteration of the EM algorithm cosists of two steps: expectation (E) and maximization (M). In the E-step, the posterior probabilities $\pi_i(x \mid \hat{G}^{(0)})$ for a current parameter value $\hat{G}^{(0)}$ are computed. In the M-step a new value $\hat{G}^{(1)}$ of G maximizing the conditional expectation of log-likelihood for the complete data (X_j, ϑ_j), $j = 1, \ldots, r$, given the incomplete data X_j, $j = 1, \ldots, r$ (see (3)),

$$L(G \mid G^{(0)}) \stackrel{\text{def}}{=} \sum_{j=1}^{r} \sum_{i=1}^{m} \ln[f_{(i)}(X_j|G)]^m \pi_i(X_j|G^{(0)})\, w_j, \quad G \in \mathcal{G},$$

is found (see (6) and (7)). The parameter value $\hat{G}^{(1)}$ obtained is a current value in the next iteration of the EM algorithm. The process is repeated untill convergence. Suppose $\hat{G}^{(0)} \in \mathcal{G}_m$. Then solution of the maximization problem

$$L(G \mid G^{(0)}) \longrightarrow \max_{G \in \mathcal{G}_m}$$

is given by

$$\begin{aligned} &\operatorname{supp}\big(\hat{G}^{(1)}(\cdot, \mathbb{N})\big) = \{\hat{\theta}_1^{(1)}, \ldots, \hat{\theta}_m^{(1)}\}, \\ &\hat{\theta}_i^{(1)} = \sum_{j=1}^{r} K_j\, \pi_i\big(X_j \mid \hat{G}^{(0)}\big)\, w_j \Big/ \sum_{j=1}^{r} N_j\, \pi_i\big(X_j \mid \hat{G}^{(0)}\big)\, w_j, \\ &\hat{G}^{(1)}\big(\{\hat{\theta}_i^{(1)}, n\}\big) = \sum_{j=1}^{r} \pi_i\big(X_j \mid \hat{G}^{(0)}\big)\, \chi_{\{N_j\}}(n)\, w_j, \qquad i = 1, \ldots, m. \end{aligned} \tag{9}$$

As pointed out above, choosing good initial values are very important for successful performance of the EM algorithm. Since $\Pi_k(\lambda)$ attains its maximum value when $k = [\lambda]$ ([λ] denotes an integer part of the real number λ), it is natural to hope that the unknown parameters θ_i, $i = 1, \ldots, m$, are close to modes of the rv $\eta \stackrel{\text{def}}{=} K_1/N_1$. Although this is not the case in general, the modes of η, as we will see in the next section, can serve as a basis for a sequential procedure for estimation and cluster separation.

3. SEQUENTIAL ESTIMATING METHOD

Suppose we have found $\hat{G}_l \in \mathcal{G}_l$, $l \geqslant 1$, and want to increase the number of components by one. In other words, we are seeking for restricted NMLE $\hat{G}_{l+1} \in \mathcal{G}_{l+1}$. To apply the EM algorithm (9) we need initial estimates for θ_{l+1} and $G(\{\theta_{l+1}\}, \cdot)$.

Initial values for the EM algorithm. Let us first describe a probability 'density' of η and its nonparametric estimator. For $a \leqslant t \leqslant b$ set

$$g(t) \stackrel{\text{def}}{=} \sum_{n=1}^{\infty} n\, f([tn], n) = \sum_{n=1}^{\infty} n \sum_{i=1}^{m} p_i \Pi_{[tn]}(\theta_i n)\Gamma_i(n) \approx \mathbf{P}\{K_j/N_j \in dt\}/dt, \quad (10)$$

($j = 1, \ldots, r$). A nonparametric estimate $\hat{g}$ of g is obtained by exploiting standard smoothing technique (see, e.g. [5,9] and references therein), although the kernel $\Pi_k(\lambda)$ is not standard:

$$\hat{g}(t) \stackrel{\text{def}}{=} \sum_{j=1}^{r} \Big[H\Pi_{[tHN_j]}(HK_j)(1 - \chi_{\{0\}}(K_j)) + \chi_{\{0\}}(K_j))\chi_{[0,1]}(tN_j) \Big] N_j w_j,$$

H being a smoothing parameter. Preliminary practical investigatios aprove the choice $H = c_0 r^{2/5} + 1$, $c_0 \in [0.2, 2]$.

Let us introduce an auxiliary model f_l with l components and an additional uniform component $f_{(0)}$. Set

$$f_l(k, n) \stackrel{\text{def}}{=} \sum_{i=1}^{l} p_i \Big(\Pi_k(\theta_i n) + \frac{p_0}{n(1-p_0)(b-a)} \Big) \Gamma_i(n) = \sum_{i=0}^{l} p_i f_{(i)}(k, n), \quad (11)$$

$$f_{(0)}(k, n) \stackrel{\text{def}}{=} \frac{1}{n(1-p_0)(b-a)} \sum_{i=1}^{l} p_i \Gamma_i(n) = \frac{G([a, b], \{n\})}{n(1-p_0)(b-a)}, \qquad \sum_{i=0}^{l} p_i = 1.$$

The component $f_{(0)}$ can be thought of as a 'noise' or 'background' cluster. When estimating the mixture distribution G sequentially, it represents a part of the data under consideration ill fitted by the current model $\psi(\cdot, \hat{G}_l)$ with l components. Thus, the model obtained is more statistically and computationally stable. Furthermore, our model is well defined for $l = 0$ also.

Similarly as in (10), we obtain probability 'density' of η corresponding to model (11)

$$g_l(t) = g_l(t|G_l) \stackrel{\text{def}}{=} \sum_{n=1}^{\infty} n \sum_{i=1}^{l} p_i \Pi_{[tn]}(\theta_i n)\Gamma_i(n) + p_0/(b-a), \qquad a \leqslant t \leqslant b.$$

Substitution of restricted NMLE $\hat{G}_l$ of G for auxiliary model (11) into this expression gives an estimate $\hat{g}_l(\cdot) \stackrel{\text{def}}{=} g_l(\cdot|\hat{G}_l)$ for g_l. The initial value $\hat{\theta}_{l+1}^{(0)}$ for $\hat{\theta}_{l+1}$ is calculated by (successively) applying one of the following formulas (see [5] for more detailed description)

$$\hat{\theta}_{l+1}^{(0)} = \arg\max_{a \leqslant t \leqslant b} \left\{\hat{g}(t)/\hat{g}_l(t)\right\}, \tag{12}$$

$$\hat{\theta}_{l+1}^{(0)} = \arg\max_{a \leqslant t \leqslant b} \left\{\hat{g}(t)^{1/2} - \hat{g}_l(t)^{1/2}\right\}, \tag{13}$$

$$\hat{\theta}_{l+1}^{(0)} = \arg\max_{a \leqslant t \leqslant b} \left\{\hat{g}(t) - \hat{g}_l(t)\right\}. \tag{14}$$

The distribution $\hat{G}_{l+1}^{(0)}$ is evaluated by making use of k-means type method. Namely,

$$\hat{G}_{l+1}^{(0)}(\{\theta\}, \cdot) = \gamma \hat{G}_l(\{\theta\}, \cdot), \qquad \theta \in \operatorname{supp}\left(\hat{G}_l(\cdot, \mathbb{N})\right),$$
$$\hat{G}_{l+1}^{(0)}(\{\hat{\theta}_{l+1}^{(0)}, n\}) = \gamma \sum_{j \in J} w_j,$$

where γ is a normalizing factor, J is the subset of indeces $j \in \{1, \ldots, r\}$ such that $N_j = n$ and $\hat{\theta}_{l+1}^{(0)}$ is the closest point to K_j/N_j among $\{\hat{\theta}_{l+1}^{(0)}\} \cup \operatorname{supp}\left(\hat{G}_l(\cdot, \mathbb{N})\right)$.

Stopping rule. Since it is presupposed that m is much less than r, a stopping criteria for the successive increase of the number of components l in the current estimate $\psi(\cdot, \hat{G}_l)$ of mixture model (1)–(5) is needed. Interesting suggestions and results relative to this problem can be found in [6–8, 11]. However, they are too complex for handling or not directly applicable to our setting. Our criteria is based on a measure of closeness between two different values of the parameter θ. Define

$$\Delta(t_1, t_2) \stackrel{\text{def}}{=} (t_1 - t_2)^2/(t_1/M_1 + t_2/M_2)$$

where $t_1 \neq t_2$, $t_1, t_2 \in \operatorname{supp}\left(\hat{G}_{l+1}(\cdot, \mathbb{N})\right)$, and $M_i \stackrel{\text{def}}{=} \sum_{j=1}^{r} N_j \hat{G}_{l+1}(\{t_i, N_j\})\, w_j$, $i = 1, 2$. The corresponding components are joined if $\Delta(\hat{\theta}_i, \hat{\theta}_{i'}) < C$, $C = 1$, say. The parameters of the resulting component are calculated from that of being joined in a natural way. The sequential estimating procedure stops if all three initial estimation methods (12)–(14) fails to find a new value of the parameter θ that would not be joined, after the application of the EM algorithm, with the previous values $\operatorname{supp}\left(\hat{G}_l(\cdot, \mathbb{N})\right)$.

4. APPLICATIONS

The method proposed has been applied to real data. The data we deal with consist of medical observations of degeneracy frequency of a certain type among new-born children in Lithuania during 1994. N_j (K_j) signifies the total number of new-born children (respectively, the number of degenerate among them) in the j-th district during the period, $j = 1, \ldots, r$, the number of the districts $r = 46$. The values of N_j (K_j) range from 27 to 7128 (respectively, from 0 to 227). The parameter values of four clusters found are presented below:

Nr	1	2	3	4
p_i	0.02174	0.59833	0.23182	0.14811
θ_i	0.00163	0.01890	0.03042	0.04872

The performance of the procedure was also tested on simulated data. The preliminary results are encouraging. Detailed description of these results as well as extended inerpretation and discussion of applications to real data will be published elsewhere.

REFERENCES

[1] S. A. Aivazyan et al., *Applied Statistics. Classification and Reduction of Dimesionality*, Finansy i Statistika, Moscow, 1989 (in Russian).

[2] D. Bohning, A review of reliable maximum likelihood algorithms for semiparametric mixture models, *J. Statist. Plann. Inference*, **47** (1995), 5–28.

[3] M. A. J. van Dujin and U. Bockenholt, Mixture models for the analysis of repeated count data, *Appl. Statist.*, **44** (4) (1995), 473–485.

[4] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*, Chapman and Hall, New York, 1981.

[5] G. Jakimauskas and J. Sushinskas, Computational aspects of statistical analysis of Gaussian mixture combining EM algorithm with non-parametric estimation (one-dimensional case), Technical Report 96-6 (1996).

[6] B. G. Leroux, Consistent estimation of a mixing distribution, *Ann. Statist.*, **20** (3) (1992), 1350–1360.

[7] B. G. Lindsay and M. L. Lesperance, A review of semiparametric mixture models, *J. Statist. Plann. Inference*, **47** (1995), 29–39.

[8] B. G. Lindsay and K. Roeder, Residual diagnostics for mixture models, *J. Amer. Statist. Assoc.*, **88** (1993), 221–228.

[9] G. J. McLacklan and K. E. Basford, *Mixture Models. Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.

[10] J. Pfanzagl, Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures, *J. Statist. Plann. Inference*, **19** (1988), 137–158.

[11] R. Rudzkis and M. Radavičius, Statistical estimation of a mixture of Gaussian distributions, *Acta Applicandae Mathematicae*, **38** (1995), 37–54.

Semiparametrinis modelis diskrečių duomenų klasterizavimui ir jo taikymas medicininiams duomenims

M. Radavičius, J. Sušinskas

Darbe pasiūlyti Puasono semiparametriniai mišiniai sudaro plačią klasę statistinių modelių, aprašančių diskrečių duomenų klasifikavimo (klasterizacijos) uždavinį. Klasifikavimo taisyklė remiasi maišymo skirstinio neparametriniu maksimalaus tikėtinumo įverčiu. Aptariamos jo savybės ir pateikiama iteratyvi jo apskaičiavimo procedūra, kuri pagrįsta neparametrinių metodų ir EM algoritmo deriniu. Minėta procedūra pritaikyta realiems duomenims: Lietuvos rajonų klasifikavimui pagal 1994 metų naujagimių išsigimimų tikimybę.