INFORMATICA, 2002, Vol. 13, No. 2, 209–226

Comparison of Poisson Mixture Models for Count Data Clusterization

Jurgis SUŠINSKAS, Marijus RADAVIČIUS
Institute of Mathematics and Informatics
Akademijos 4, 2600 Vilnius, Lithuania
e-mail: jur@ktl.mii.lt, mrad@ktl.mii.lt

Received: March 2002

Abstract. Five methods for count data clusterization based on Poisson mixture models are described. Two of them are parametric, the others are semi-parametric. The methods emlploy the plug-in Bayes classification rule. Their performance is investigated by making use of computer simulation and compared mainly by the clusterization error rate. We also apply the clusterization procedures to real count data and discuss the results.

Key words: count data, clusterization, nonparametric Poisson mixtures, plug-in Bayes classification rule, maximum likelihood estimator, classification error rate.

1. Introduction

In the paper, we consider the clusterization problem of count data of the following type. A pair of random variables (K_j, N_j) is observed, where K_j is the number of cases (individuals) having a feature A, say, in the jth subsample (subpopulation) Ω_j and N_j is the total number of its elements, $j = 1, \dots, r$. Rather freaquently occur such data in medical, biological, social, and economical studies. Typical examples are as follows: the number of infected individuals among tested in various regions or medical institutions, the number of recaptured among the captured animals of some species in various inhabitats, the number of fatal cases in road-accidents in various towns, the defective portion of the whole production at various plants, etc. Our aim is to classify subpopulations $\Omega_j, j = 1, \dots, r$ according to the spread of the feature A in them.

The Bayes approach leads to different mixture-models of classification depending on the way we treat the totals $N_j, j = 1, \dots, r$. In the present paper we compare the performance of three mixture-models. The first model assumes that the observations $N_j, j = 1, \dots, r$, are non-informative. They are regarded as either nonrandom (incidental or selected in advance) parameters or independent and identically distributed random variables. In the second and third models the distribution of N_j depends on the unknown class number and, in the latter model, it is supposed that this distribution belongs to a parametric distribution family. As a compromise between flexibility and computational simplicity, the family of discretized normal distributions is chosen. No parametric structure is imposed on the distribution of N_j in the second model, and thus the model has a nonparametric part. As noted in (Sušinskas and Radavičius, 1998), when r is small in comparison to the totals $N_j, j = 1, \dots, r$, the maximum likelihood estimator of this case is significantly biased. The bias is reduced by smoothing the nonparametric part. Two smoothing procedures based on the EM algorithm are considered in the paper. The first one was proposed in (Sušinskas and Radavičius, 1998).

The Poisson mixture models for count data classification and the two smoothing procedures are compared by means of computer simulation and their impact on clusterization quality is evaluated mainly by the clusterization error rate. We also apply them to real data and discuss their performance. For a wider discussion of the applications of Poisson mixtures, we refer to (Van Dujin and Bockenholt, 1995; Everitt and Hand, 1981; Lindsay and Lesperance, 1995; McLacklan and Basford, 1988).

In the next section detailed description of the models is given. Section 3 is designated to the theoretical background of cluster analysis based on mixture-models and some computational aspects. In particular, the smoothing procedures of the nonparametric part are outlined and clusterization procedures to be compared are presented. Section 4 contains the results of the computer experiment and cluster analysis of real data. In the last section, some conclusions are drawn.

2. Mixture-Models for Clusterization

We assume that the following assumptions hold:

(a) The subsamples $\Omega_j, j = 1, \dots, m$ are sampled from m classes (populations) which differ from one another in probability of having the feature A. Let Z_j be an unobservable class (population) number of the jth subsample Ω_j. Under the condition that the class number $Z_j = i$ and the total $N_j = n$, the random variable K_j has the binomial distribution with the parameters N_j and θ_i, where $j = 1, \dots, r,\ \theta_i \in (0, 1),\ i = 1, \dots, m.$

(b) The random vectors $(K_j, N_j, Z_j),\ j = 1, \dots, r,$ are independent and identically distributed (i.i.d.).

(c) The probabilities $\{\theta_i\}$ are small: $\theta_i \leqslant \rho_0 \ll 1,\ i = 1, \dots, m.$

Denote by $\Gamma(\cdot \mid i)$ the conditional distribution of the total N_j given that the class number $Z_j = i$, i.e.,

$$\Gamma(n \mid i) \overset{\text{def}}{=} \Pr\{N_j = n \mid Z_j = i\}, \quad n \in \mathbf{N}, \quad i = 1, \dots, m, \tag{1}$$

where $\mathbf{N}$ denotes the set of all positive integers.

The Bayes approach and assumptions (a), (b) and (c) lead to the following probabilistic model

$$\begin{aligned} f(k, n) &\overset{\text{def}}{=} \Pr\{K_j = k, N_j = n\} \\ &= f_S\left(k, n \mid \lambda_S(m)\right) \overset{\text{def}}{=} \sum_{i=1}^{m} p_i \Pi_k(\theta_i n) \Gamma(n \mid i), \end{aligned} \tag{2}$$

where $k \in \mathbf{N}_0 \overset{\text{def}}{=} \mathbf{N} \cup \{0\}$, $n \in \mathbf{N}$. Here p_i is a prior probability of the ith class (population),

$$p_i \overset{\text{def}}{=} \Pr\{Z_j = i\} > 0, \quad i = 1, \dots, m, \quad \sum_{i=1}^{m} p_i = 1,$$

the functions

$$\Pi_k(t) = t^k \mathrm{e}^{-t}/k!, \quad k \in \mathbf{N}_0,$$

are the Poisson probabilities obtained from the Poisson approximation to the binomial distribution (justified by (c)),

$$\Pi_k(\theta_i n) \approx \Pr\{K_j = k \mid N_j = n, Z_j = i\} = C_n^k \theta_i^k (1 - \theta_i)^{n-k},$$

$k = 0, \dots, n,\ n \in \mathbf{N}$, the conditional distribution $\Gamma(\cdot \mid i)$ is introduced in (1), $\lambda_S(m) = \{p_i, \theta_i, \Gamma(\cdot \mid i), i = 1, \dots, m\}$ is a set of the unknown parameters of the model.

One can see from (2) that in the model proposed the random variables $\{N_j\}$ may carry some information about unobserved class numbers $\{Z_j\}$. If this is the case, we say that $\{N_j\}$ are *informative*. The following condition makes $\{N_j\}$ *non-informative*.

(d) The conditional distribution $\Gamma(\cdot \mid i)$ of the total N_j, given that the class number $Z_j = i$, is independent of i, i.e.,

$$\Gamma(\cdot) = \Gamma(\cdot \mid i),\ i = 1, \dots, m. \tag{3}$$

Then we have

$$\begin{aligned} f(k, n) &= \Gamma(n) f(k \mid n), \quad k \in \mathbf{N}_0,\ n \in \mathbf{N}, & (4)\\ f(k \mid n) &\overset{\text{def}}{=} \Pr\{K_j = k \mid N_j = n\} \\ &= f_N(k \mid n, \lambda_N(m)) \overset{\text{def}}{=} \sum_{i=1}^{m} p_i \Pi_k(\theta_i n), & (5) \end{aligned}$$

where $\lambda_N(m) = \{p_i, \theta_i, i = 1, \dots, m\}$ is a set of unknown parameters of the non-informative model (model (N) for short). Note that the distribution Γ is not included in the list of the unknown parameters as it is not involved in (5) and thereby is redundant for the clusterization (for details see the next section). Since in this case $\{N_j\}_{j=1}^r$ and $\{Z_j\}_{j=1}^r$ are mutually indepenent (see (3) and condition (b)), the totals $\{N_j\}_{j=1}^r$ can be treated as (incidental) model parameters (in a similar way as independent variables in regression models).

In mixture-model (2) we do not impose any parametric structure on the probability distributions $\Gamma(\cdot \mid i), i = 1, \dots, m$. Thus the model is *semiparametric*, i.e., includes both the parametric $p_i \Pi_k(\theta_i n)$ and the nonparametric $\Gamma(n \mid i)$ parts. Therefore it is refered to as model (S).

Assuming some parametric form for $\Gamma(\cdot \mid i)$ we obtain some kind of intermediate model between (2) and (4, 5). As a quite reasonable compromise between model flexibility and computational simplicity, we choose the family of discretized normal distributions.

(d') The conditional distribution $\Gamma(\cdot \mid i)$ is given by the following equality:

$$\Gamma(n \mid i) = \Gamma_P(n \mid a_i, \sigma_i) \overset{\text{def}}{=} \Pr\{\xi_i \in [n-1, n) \mid \xi_i \geqslant 0\}, \quad n \in \mathbf{N}, \tag{6}$$

where ξ_i is a normal random variable with the parameters $a_i > 0, \sigma_i > 0,\ i = 1, \ldots, m.$

A model resulting from (d') takes the form

$$f(k,n) = f_P\left(k, n \mid \lambda_P(m)\right) \overset{\text{def}}{=} \sum_{i=1}^{m} p_i \Pi_k(\theta_i n) \Gamma(n \mid a_i, \sigma_i), \tag{7}$$

where $k \in \mathbf{N}_0,\ n \in \mathbf{N}$, and in the sequel is referred to as *parametric* (model (P)). In this model the unknown parameters are listed in $\lambda_P(m) = \{\, p_i,\ \theta_i,\ a_i,\ \sigma_i,\ i = 1, \ldots, m\}$.

The three models obtained, namely, non-informative (4, 5), parametric (7), and semi-parametric (2), represent mixtures of Poisson distributions. We stress that the first two models are special cases of model (S) and are obtained from it by assuming (3) and (6), respectively, valid. Further examples and applications of Poisson mixtures can be found in (Van Dujin and Bockenholt, 1995; Everitt and Hand, 1981; Lindsay and Lesperance, 1995; McLacklan and Basford, 1988).

3. Theoretical Background

3.1. *Bayes Classification Rule (BCR)*

For brevity, set

$$f_{(i)}(k,n) \overset{\text{def}}{=} \Pr\{K_j = k, N_j = n \mid Z_j = i\} = \Pi_k(\theta_i n)\Gamma(n \mid i), \tag{8}$$
$$i = 1, \ldots, m \quad (j = 1, \ldots, r).$$

According to the Bayes formula, the posterior probability

$$\pi_i(k,n) \overset{\text{def}}{=} \Pr\{Z_j = i \mid\ K_j = k, N_j = n\}, \quad j = 1, \ldots, r,$$

i.e., the conditional probability that the class number Z_j of the jth observation (K_j, N_j) equals i, under the condition $(K_j, N_j) = (k, n)$, is given by

$$\pi_i(k,n) = \pi_i(k, n \mid \lambda_S(m)) = p_i f_{(i)}(k,n)/f(k,n), \quad i = 1, \ldots, m. \tag{9}$$

For parametric model (7), the unknown parameters $\lambda_S(m)$ should be replaced by $\lambda_P(m)$. If the non-informative model is assumed, we have by (4), (5), (8), and (9) that

$$\pi_i(k,n) = p_i\, \Pi_k(\theta_i n) \,/\, f(k \mid n), \quad (k,n) \in \mathbf{N}_0 \times \mathbf{N},\ i = 1, \ldots, m. \tag{10}$$

Hence the posterior probabilities are independent of $\Gamma(\cdot)$, and we conclude that $\pi_i(\cdot,\cdot) = \pi_i(\cdot,\cdot \mid \lambda_N(m))$, $i = 1,\dots,m$.

The minimal classification error is obtained by the Bayes classification rule (BCR) (Aivazyan *et al.*, 1989; Everitt and Hand, 1981; McLacklan and Basford, 1988): assign the observation (K_j, N_j) to the i^*th cluster if $i^* = d_B(K_j, N_j)$ where

$$d_B(K_j, N_j) \overset{\text{def}}{=} \arg\max_{1\leqslant i\leqslant m} \{\pi_i(K_j, N_j)\} = \arg\max_{1\leqslant i\leqslant m} \{p_i f_{(i)}(K_j, N_j)\} \tag{11}$$

is the Bayes decision function, $j = 1,\dots,r$. Thus, the clusterization problem reduces to estimation of the posterior probabilities. We estimate $\pi_i = \pi_i(\cdot,\cdot)$ by the *maximum likelihood* (ML) method.

3.2. *Maximum Likelihood Estimator (MLE)*

Let us consider model (S). First we will specify a range of the unknown parameters. Given $\rho \in (0, \rho_0/m)$ (recall that ρ_0 was introduced in condition (c)), set

$$\begin{aligned}\Lambda_S(m) \overset{\text{def}}{=} \Big\{\lambda_S(m) = \big(\theta_i, p_i, \Gamma(\cdot \mid i),\, i = 1,\dots,m\big)\colon \\ \theta_{i-1} + \rho \leqslant \theta_i < 1,\ p_i \geqslant \rho,\ i = 1,\dots,m\Big\},\end{aligned} \tag{12}$$

where $\theta_0 \overset{\text{def}}{=} 0$. The log-likelihood function takes the following form

$$L\big(\lambda_S(m)\big) = \sum_{j=1}^{r} \ln f\big(K_j, N_j \mid \lambda_S(m)\big) = \sum_{j=1}^{r} \ln\left(\sum_{i=1}^{m} p_i \Pi_{K_j}(\theta_i N_j)\Gamma(N_j \mid i)\right).$$

It follows from (Sušinskas and Radavičius, 1998) that the semiparametric MLE $\hat{\lambda}_S(m)$,

$$\hat{\lambda}_S(m) \overset{\text{def}}{=} \arg\max_{\lambda\in\Lambda_S(m)} L(\lambda), \tag{13}$$

is a consistent estimator of $\lambda_S(m)$ as $r \to \infty$. It is worth noting, however, that this result is not of great value for many applications since rather frequently N_j, $j = 1,\dots,r$, are of the same order or even much greater than r which sometimes is more natural to regard as being fixed. The problem of classifying districts of Lithuania according to the rate of still-borns, considered in the last section, is just a problem of this type. When all N_j, $j = 1,\dots,r$, are different the nonparametric MLE of $p_i\Gamma(\cdot \mid i)$, $i = 1,\dots,m$, for model (S) is determined by

$$\hat{p}_i\hat{\Gamma}(n \mid i) = \delta_{k\,i}\,\delta_{N_j\,n}\,/m,$$

where δ_{ki} is the Kronecker symbol and $k \overset{\text{def}}{=} \arg\max_{1\leqslant i\leqslant m}\{\Pi_{K_j}(\hat{\theta}_i N_j)\}$ is the maximum likelihood classification rule for classifying the observation K_j drawn from one of m Poisson populations with the parameters $\hat{\theta}_i N_j$, $i = 1,\dots,m$, respectively.

Thus, for strongly overlapping populations (classes) the semiparametric MLE of $\lambda_S(m)$ is significantly biased. The simulation results presented in the last section confirm this observation and give additional insight of the degree of bias. In order to reduce the bias we apply a smoothing technique (see subsection Smoothed EM algorithm).

Now let us turn to the parametric models (N) and (P). The parametric set $\Lambda_N(m)$ for model (N) is defined in the same way as (12) except that the nonparametric part $\Gamma(\cdot \mid i)$ is omitted. In this case, the probability distribution of the observation K_j depends on the (incidental) parameter N_j, and hence K_j, $j = 1, \dots, r$ are independent but not identically distributed random variables. For model (P), we suppose that

$$\Lambda_P(m) \overset{\text{def}}{=} \Big\{\lambda_P(m) = \big(\theta_i, p_i, a_i, \sigma_i,\ i = 1, \dots, m\big)\colon\ \theta_{i-1} + \rho \leqslant \theta_i < 1,\\ p_i \geqslant \rho,\ a_i \geqslant 0,\ \sigma_i \geqslant \rho_1,\ i = 1, \dots, m\Big\},$$

where $\theta_0 = 0$, $0 < \rho < \rho_0/m$, and $\rho_1 > 0$. Model (P) satisfies the usual regularity conditions for independent and identically distributed observations (see, e.g., (Ibragimov and Khasminskii, 1981)). Consequently, the MLE in this case is not only consistent, but also asymptotically efficient as $r \to \infty$. The same statement holds for independent nonidentically distributed observations satisfying model (N), provided $1 \leqslant N_j \leqslant C < \infty$, $j = 1, \dots, r$. Again, as mentioned above, to treat r as an asymptotic parameter and the parameters N_j, $j = 1, \dots, r$, bounded or the parametrs a_i, σ_i, $i = 1, \dots, m$, fixed, it is not natural for some (possible) applications.

3.3. *The EM Algorithm*

Let λ (Λ) denote any of the unknown parameters $\lambda_N(m)$, $\lambda_P(m)$, and $\lambda_S(m)$ (respectively, parameter sets $\Lambda_N(m)$, $\Lambda_P(m)$, and $\Lambda_S(m)$), the number of mixture components, m, being fixed. For computing the MLE $\hat{\lambda}$ of λ we apply the EM algorithm (Aivazyan *et al.*, 1989; Bohning, 1995; Everitt, Hand, 1981; McLacklan and Basford, 1988; Sušinskas and Radavičius, 1998). 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 consists of two steps: expectation (E) and maximization (M). In the E-step, the conditional expectation of the log-likelihood for the complete data (K_j, N_j, Z_j), $j = 1, \dots, r$, given the incomplete data (K_j, N_j), $j = 1, \dots, r$, is calculated with the current (initial) parameter value $\hat{\lambda}^{(0)}$ taken as a true value of the unknown parameter λ. In our case, this conditional expectation, denoted by $L(\lambda|\hat{\lambda}^{(0)})$, admits a simple expression in terms of the posterior probabilities $\pi_i(\cdot, \cdot|\hat{\lambda}^{(0)})$ (see (9)):

$$L(\lambda|\hat{\lambda}^{(0)}) \overset{\text{def}}{=} \sum_{j=1}^{r} \ln\left[f(K_j, N_j|\lambda)\right]\ \pi_i(K_j, N_j|\hat{\lambda}^{(0)}), \quad \lambda \in \Lambda.$$

In the M-step, a new value $\hat{\lambda}^{(1)}$ of λ maximizing $L(\lambda|\hat{\lambda}^{(0)})$ is found. The parameter value $\hat{\lambda}^{(1)}$ obtained is the current value in the next iteration of the EM algorithm. The process is repeated until the convergence.

Suppose $\hat{\lambda}^{(0)} \in \Lambda$. Then a solution of the maximization problem

$$L(\lambda|\lambda^{(0)}) \longrightarrow \max_{\lambda\in\Lambda}$$

is given by the following equations. The equations for the prior probabilities π_i and the parameters θ_i, $i = 1, \dots, m$, are the same for all the three models:

$$\hat{p}_i^{(1)} = \frac{1}{r}\sum_{j=1}^{r} \pi_i\Big(K_j, N_j|\hat{\lambda}^{(0)}\Big), \tag{14}$$

$$\hat{\theta}_i^{(1)} = \frac{\sum_{j=1}^{r} K_j\, \pi_i\big(K_j, N_j|\hat{\lambda}^{(0)}\big)}{\sum_{j=1}^{r} N_j\, \pi_i\big(K_j, N_j|\hat{\lambda}^{(0)}\big)}, \quad i = 1, \dots, m. \tag{15}$$

For the parameters a_i, σ_i, $i = 1, \dots, m$, we have

$$\hat{a}_i^{(1)} = \frac{1}{r\hat{p}_i^{(1)}}\sum_{j=1}^{r} N_j\, \pi_i\big(K_j, N_j|\hat{\lambda}^{(0)}\big), \tag{16}$$

$$\Big(\hat{\sigma}_i^{(1)}\Big)^2 = \frac{1}{r\hat{p}_i^{(1)}}\sum_{j=1}^{r} N_j^2\, \pi_i\big(K_j, N_j|\hat{\lambda}^{(0)}\big) - \Big(\hat{a}_i^{(1)}\Big)^2. \tag{17}$$

Finally, for model (S), the current estimate $\hat{\Gamma}_S^{(1)}$ of the nonparametric part Γ of the model is recalculated simply by taking normalized averages of the corresponding posterior probabilities:

$$\hat{\Gamma}_S^{(1)}(n \mid i) = \frac{1}{r\hat{p}_i^{(1)}}\sum_{j=1}^{r} \pi_i\Big(K_j, N_j|\hat{\lambda}^{(0)}\Big)\, \delta_{N_j n}, \quad i = 1, \dots, m. \tag{18}$$

3.4. *The Smoothed EM Algorithm (EMS)*

As noted above, the MLE of the unknown parameter λ_S of model (S) is significantly biased when r is small in comparison with $N_j, j = 1, \dots, r$, and clusters are strongly overlapping. To reduce this bias, we apply the smoothing technique to the nonparametric MLE $\hat{\Gamma}_S$ of the conditional distribution Γ. To be more precise, the improved estimator of λ_S, called a smoothed (semiparametric) MLE, is calculated iteratively by the EM algorithm with an additional smoothing step (S-step) for $\hat{\Gamma}_S$ at the beginning of each EM iteration.

Since the totals N_j, $j = 1, \dots, r$, are assumed to be large, we ignore their discrete character and treat them as continuous. Because of this, we apply the usual kernel smoothing (Nadaraya–Watson method for nonparametric regression) with Epanechnikov's kernel function $W(t) = 0.75\,(1 - t^2)\,\mathbf{1}(|t| < 1)$ and variable bandwidth $b = b(n)$, $n \in \mathbf{N}$, selected by the nearest neighbor method. Since the conditional variance of K_j/N_j, given N_j (i.e., accuracy of information about the parameter θ for jth observation), is proportional to $1/N_j$, $j = 1,\dots,r$, weights of the observations are assumed to be $w_j = N_j$, $j = 1,\dots,r$.

Two alternative procedures were implemented. The first one, (S1), for a given centre point N_c and the number of neighbors k, calculates the bandwidth $b(N_c) = b(N_c \mid k)$ by the formula $b(N_c) = |N^* - N_c|$, where N^* is the kth nearest neighbor to N_c among N_j, $j = 1,\dots,r$. Then

$$\hat{\Gamma}^{(1)}_{S1}(N_c \mid i) = \frac{1}{\overline{\overline{W}}} \sum_{j=1}^{r} \hat{\Gamma}^{(1)}(N_j \mid i)\, W\left(\frac{N_j - N_c}{b(N_c)}\right) w_j, \tag{19}$$

where $i = 1,\dots,m$ and

$$\overline{\overline{W}} = \sum_{j=1}^{r} W\left(\frac{N_j - N_c}{b(N_c)}\right) w_j. \tag{20}$$

The smoothing procedure of this type but with the weights $w_j \equiv 1$ was used in (Sušinskas and Radavičius, 1998).

The second, procedure (S2), finds the kth nearest neighbor of N_c among N_j, $j = 1,\dots,r$, for each cluster separately. This is performed in the following way. Let $(j_1,\dots,j_r)$ be a permutation of $(1,\dots,r)$ in increasing order of distances $|N_{j_l} - N_c|$, $l = 1,\dots,r$. Set

$$l_i^* \overset{\text{def}}{=} \min\left\{ l \geqslant 1\colon \sum_{s=1}^{l} \pi_i\Big(K_{j_s}, N_{j_s} | \hat{\lambda}^{(0)}\Big) \geqslant k \right\}$$

(although other reasonable definitions of l_i^* are possible naturally) and take $b_i(N_c) = b_i(N_c \mid k) = |N_{l_i^*} - N_c|$, $i = 1,\dots,m$. Then, just like in (19) and (20), we get

$$\hat{\Gamma}^{(1)}_{S2}(N_c \mid i) = \frac{1}{\overline{\overline{W}}_i} \sum_{j=1}^{r} \hat{\Gamma}^{(1)}(N_j)\, W\left(\frac{N_j - N_c}{b_i(N_c)}\right) w_j, \tag{21}$$

$$\overline{\overline{W_i}} = \sum_{j=1}^{r} W\left(\frac{N_j - N_c}{b_i(N_c)}\right) w_j, \quad i = 1,\dots,m. \tag{22}$$

Let us stress that $\hat{\Gamma}^{(1)}_{S1}$ and $\hat{\Gamma}^{(1)}_{S2}$ as well as the entire collection $\hat{\lambda}_S$ of parameters of the model calculated by the EMS algorithm depend on the smoothing parameter k, the number of the nearest neighbors. This parameter is taken to be of the form $k = 1 + [c_0 r^{\alpha}]$ for (S1) and $k = 1 + [c_1 (r/m)^{\alpha}]$ for (S2) ([x] is the integer part of the number x). Preliminary simulation results show that the choice $c_0 = 2.8,\ c_1 = 3.0$, and $\alpha = 0.33$ yields satisfactory results.

The nearest neighbor method is also used to bound from below the standard deviation estimates $\hat{\sigma}_i$, $i = 1, \ldots, m$, for procedure (P). To avoid the degeneracy problem, we set $\hat{\sigma}_i = c_i$ when $\hat{\sigma}_i$ is under c_i. Here $c_i = 2|N^* - \hat{a}_i|$, N^* is the kth nearest to $\hat{a}_i$ neighbor among N_j, $j = 1, \ldots, r$, and the number of neighbors k is the same as for procedure (S1) (see the previous paragraph).

4. Comparison of the Poisson Mixture-Models

4.1. *Clusterization Procedures*

The following five clusterization procedures were investigated. All of them are BCR's with the estimated parameters (estimated BCR's, for short) but differ from one another in the underlying mixture-model or the smoothing method involved in the EMS algorithm. By the estimated BCR, $\hat{d}_B$, we mean here BCR d_b (see (11)) in which "true" values of the corresponding unknown parameters λ are replaced by their MLE's $\hat{\lambda}$ based on a sample to be clusterized. To calculate the maximum likelihood estimates (MLe's) the EM (or EMS) algorithm is employed. Thus, the clusterization procedures under consideration differ from one another only in the parametrization of mixing distribution of the Poisson mixture and in the estimation method of this distribution.

The first three clusterization procedures correspond to the Poisson mixture-models introduced and we retain the same notation for them. Thus, procedure (S) is based on equations (2), (8), (9), (14), (15), (18); in procedure (N) calculations are performed using (4), (5), (10), (14), (15), and finally for procedure (P) formulas (7), (8), (9), (14)–(17) are applied. The last two clusterization procedures are based on model (S) (see formulas (2), (8), and (9)) fitted to data by iterative calculations using the EMS algorithm. Procedure (S1) exploits formulas (14), (15), (18), (19) and (20). Procedure (S2) differs from (S1) only in the definition of the k-nearest-neighbor which now depends on the cluster number i. This means that the last two formulas are replaced by (21) and (22).

To compare the clusterization procedures, we apply them to artificial data generated by computer and to real data.

4.2. *Computer Experiment*

The following Poisson mixture-models were used to generate artificial data. The number of classes (populations) $m = 2$,

$$\begin{aligned} p_1 &= 0.4; & p_2 &= 0.6; \\ \theta_1 &= 0.02; & \theta_2 &= 0.025;\ 0.0275;\ 0.03;\ 0.035;\ 0.04;\ 0.5; \end{aligned} \tag{23}$$

$$\begin{aligned} &\Gamma(1, n) = \phi(n \mid 500, 30), \\ &\Gamma(2, n) = q\,\phi(n \mid 500, 30) + (1 - q)\,\phi(n \mid 800, 30), \\ &q = 0;\ 1/6;\ 1/3;\ 1/2;\ 2/3;\ 5/6;\ 1. \end{aligned} \tag{24}$$

Here $\phi(n \mid a, \sigma) = \Pr\{n-1 < \eta \leqslant n\}/\Pr\{\eta > 0\}$ and η is a normal random variable with the mean a and the standard deviation σ.

Note that two extreme cases, $q = 0$ and $q = 1$, correspond to the well-separated and completely overlapping (with respect to N) parametric models, respectively. Thus, in the latter case we obtain non-informative model (N). The remaining values of q represent various intermediate situations. As the parameter θ_1 decreases closer and closer to θ_2 the classification problem becomes more and more complicated. The presented set of θ_1 values allows us to describe this phenomenon rather smoothly.

The models (N), (P), and the semiparametric model evaluated either by the EM algorithm (marked (S)) or by the two versions of the EMS (denoted (S1) and (S2), respectively) are compared by the clusterization error rate (CER) and estimating accuracy of the parameters p_1, θ_1 and θ_2 in a series of Monte–Carlo experiments.

Each Monte–Carlo experiment consists of the following steps.

Step 1. Generate a (complete) "teaching" sample $T_C \overset{\text{def}}{=} \{(K_j, N_j, Z_j),\ j = 1, \ldots, r\}$ of size $r = 200$ according to one of 42 possible mixtures of Poisson distributions with the collection of parameters $\lambda_S(2)$ presented in (23) and (24). Recall that Z_j stands for an (unobservable) class number of the jth observation.

Step 2. Using the incomplete (unclassified) "teaching" sample $T_I \overset{\text{def}}{=} \{(K_j, N_j),\ j = 1, \ldots, r\}$ obtained from T_C by dropping the class number Z_j, estimate the unknown parameters for each Poisson mixture-model by applying the corresponding procedure, (N), (P), (S), (S1) or (S2).

Step 3. Evaluate the deviations of the obtained estimates of the parameters p_1, θ_1, and θ_2 and CER for each clusteriztion procedure. CER of the clusterization procedure is calculated simply as a relative frequency of disagreements between decisions provided by this procedure and the true class (population) numbers $Z_j,\ j = 1, \ldots, r$, contained in the complete sample T_C.

Steps 1–3 were repeated $M = 100$ times and the overall performance of the clusterization methods (and the underlying models) are evaluated by average and standard deviation of the characteristics of interest.

The results are summarized in Table 1.

To save room, the averadges and standard deviations of the CER only are presented here. Since our goal is data clusterization, the accuracy of the parameter estimates is an auxiliary characteristic. Typically its behavior suits well with that of the CER.

Remark. It is well known that the convergence of the EM algorithm to the MLe depends on the starting point of the iteration process. Two collections of initial values of the parameters were used to start EM (EMS) iterations in order to get some insight of evaluating progress. If, for either collection, the parameter estimates obtained at the end of the iteration process are essentially the same, this fact is a fair indication that the MLe is actually found. The results in Table 1 correspond to the estimates with a greater likelihood.

In the first collection, the initial values for p_1, θ_1, and θ_2 are taken to be equal to the corresponding true values and $\Gamma(i, N_j) = 1/r,\ j = 1, \ldots, r,\ i = 1, 2$ (as if model (N) were valid). In the second collection, the initial values of the parameters θ_1, and θ_2 (the

Table 1

Monte Carlo estimates of the error rate of clusterization procedures

q	Theta	0.025		0.0275		0.03		0.035		0.04		0.05	
	Method	ave	std	ave	std	ave	std	ave	std	ave	std	ave	std
0	N	37.3	5.11	32.1	5.54	25.8	4.62	15.2	2.92	8.4	2.1	2.4	1.1
	P	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	S	32.8	3.64	26.8	3.38	21.3	3.44	12.8	2.71	6.9	1.85	1.8	0.87
	S1	11.8	6.84	6.2	3.87	3.9	2.5	1.8	1.24	1.0	0.78	0.3	0.4
	S2	3.5	6.05	1.6	2.25	1.1	1.02	0.8	0.69	0.5	0.55	0.2	0.35
1/6	N	36.9	5.09	31.9	5.53	25.5	4.51	15.1	2.89	8.6	2.09	2.5	1.1
	P	9.8	2.37	9.8	2.37	9.8	2.37	9.8	2.37	9.8	2.37	5.6	3.56
	S	33.0	3.53	26.6	3.29	21.0	2.99	12.7	2.34	7.3	1.78	2.1	0.95
	S1	20.9	7.2	14.6	4.25	11.3	3.05	7.2	1.93	4.2	1.52	1.5	0.79
	S2	15.3	6.88	11.5	4.25	9.1	2.5	6.1	1.87	3.8	1.4	1.3	0.79
1/3	N	36.5	4.78	31.5	5.19	25.4	4.81	15.1	2.88	8.7	2.1	2.7	1.14
	P	19.9	2.8	19.9	2.8	19.9	2.8	19.9	2.8	17.5	5.33	2.7	1.35
	S	33.6	3.46	27.2	3.29	21.4	3.01	13.2	2.41	7.8	1.87	2.3	1.01
	S1	27.1	6.83	20.5	4.37	16.2	3.46	10.1	2.42	6.1	1.8	2.0	0.94
	S2	23.3	5.94	18.2	3.56	14.9	3.09	9.4	2.37	5.7	1.74	2.0	0.98
1/2	N	36.6	4.73	31.2	5.73	25.5	5.25	15.2	2.69	8.7	2.09	2.8	1.15
	P	30.2	3.2	30.2	3.2	30.2	3.2	30.2	3.2	20.7	8.87	3.5	1.46
	S	33.9	3.49	27.5	3.25	21.8	3.1	13.8	2.45	8.3	1.9	2.5	1.08
	S1	30.8	6.11	24.7	4.08	20.1	3.5	12.7	2.71	7.7	1.97	2.5	1.11
	S2	29.4	5.67	23.6	4.1	19.1	4.2	11.9	2.53	7.4	1.82	2.4	1.08
2/3	N	36.5	4.9	31.1	6.24	25.6	5.7	15.4	2.73	9.0	2.1	2.9	1.14
	P	40.0	3.67	40.0	3.67	40.0	3.67	39.8	3.87	26.2	10.68	4.3	1.68
	S	34.4	3.53	28.0	3.3	22.3	3.19	14.3	2.64	8.7	2.02	2.7	1.09
	S1	34.2	5.44	28.4	5.55	22.6	4.2	14.2	3.1	8.8	2.17	3.0	0.99
	S2	33.5	5.66	27.6	5.91	22.1	4.47	13.6	2.92	8.5	2.06	2.8	0.94
5/6	N	37.0	5.77	32.2	6.45	26.3	6.12	15.5	2.76	9.3	2.04	3.1	1.17
	P	50.1	3.88	50.1	3.88	50.1	3.88	49.8	4.16	34.9	12.4	4.4	1.77
	S	34.8	3.57	28.6	3.4	22.8	3.06	14.8	2.65	9.1	1.95	3.0	1.15
	S1	37.3	5.94	31.4	5.44	25.6	4.87	15.8	3.55	9.8	2.21	3.2	1.07
	S2	37.5	6.3	31.3	6.61	25.1	5.85	15.2	3.45	9.5	2.29	3.1	0.99
1	N	36.9	6.0	32.6	7.02	26.6	6.23	15.9	2.95	9.8	1.99	3.3	1.16
	P	40.3	6.69	33.9	7.58	27.5	6.94	16.2	3.01	9.9	2.17	3.3	1.14
	S	35.1	3.6	29.1	3.4	23.4	3.14	15.3	2.76	9.8	2.11	3.3	1.17
	S1	39.7	6.42	34.0	6.48	27.7	5.55	17.3	3.36	10.6	2.35	3.5	1.15
	S2	39.6	6.76	33.3	6.18	26.9	5.86	16.5	3.38	10.2	2.27	3.3	1.14

prior probability p_1 and the mixing distribution Γ) are taken to be equal to the MLe (respectively, approximate MLe) based on the complete "teaching" data T_C. Namely,

$$\hat{\theta}_i^{(0)} = \frac{\sum_{j=1}^{r} K_j\, \delta_{iZ_j}}{\sum_{j=1}^{r} N_j\, \delta_{iZ_j}}, \qquad i = 1, 2, \tag{25}$$

$$\hat{p}_1^{(0)} = \hat{p}_{MLE} + \tau\,(1 - 2\,\hat{p}_{MLE}), \qquad \hat{p}_{MLE} \overset{\mathrm{def}}{=} \frac{1}{r}\sum_{j=1}^{r} \delta_{1Z_j}, \tag{26}$$

and

$$\hat{\Gamma}^{(0)}(i,n) = \frac{1}{r\hat{p}_i^{(0)}} \sum_{j=1}^{r} \big[(1-\tau)\delta_{iZ_j} + \tau(1-\delta_{iZ_j})\big]\,\delta_{nN_j}, \quad n \in \mathbf{N}, \tag{27}$$

where $\hat{p}_2^{(0)} = 1 - \hat{p}_1^{(0)}$ and $\tau > 0$ is a small quantity (e.g., $\tau = 0.1$).

Table 1 shows that parametric model (P) is too sensitive to model assumptions (non-robust). Procedures (S) and (S1) yield only a limited improvement in comparison to the method based on non-informative model (N), the second one being slightly better. Procedure (S2) outperforms (S) and (S1) and parametric clusterization methods (N) and (P) for misspecified (parametric) probability models (i.e., if $0 < q < 1$), and is quite competitive to the latter two in the case of an adequate probability model (recall that setting $q = 0$ or $q = 1$ yields probabilistic model (P); model (N) is adequate provided $q = 1$).

4.3. *Application to Real Data*

The real data investigated are taken from the database LIRECA maintained by the Lithuanian Human Genetic Centre. This database contains all registered cases of congenital anomalies among newborns in Lithuania since 1993.

The data we deal with consist of observations $\{(K_j, N_j),\ j = 1,\dots,r\}$, where N_j is the total number of newborns in the jth district of Lithuania during 1993–1997 and K_j is the number of newborns having a certain congenital anomaly. The number of districts (some of them are amalgamated) $r = 42$. The problem is to classify the districts according to the congenital anomaly rate. This problem was suggested to the first of the authors by Prof. V. Kučinskas of Vilnius University who also provided the data. We consider here the case where K_j is the number of stillborns.

The performance of all the five clusterization methods, (N), (P), (S), (S1), and (S2), is to be compared visually. It should be noted, however, that visual fitness of the clusters produced should be assessed with caution. It can be missleading in this case, since visual interpretation of common to us Gaussian mixtures considerably differs from that of Poisson (see, for instance, Fig. 18; the leftmost point of the 4th cluster and the rightmost point of the 5th cluster seem as "a mistake" of the classification procedure). The clusterization procedure based on non-informative model (N) tends to alocate observations with small totals N_j to the cluster with the greatest prior probability.

From the viewpoint of the applications, employment of different statistical methods for solving the same task follows an old statistical tradition. In our case, it means that, if all the five clusterization methods (or the major part of them) lead to (more or less) similar results, this fact is an added reason for their reliability.

Again, two collections of starting values of the parameters were used to start the EM (or EMS) iteration process (see Remark).

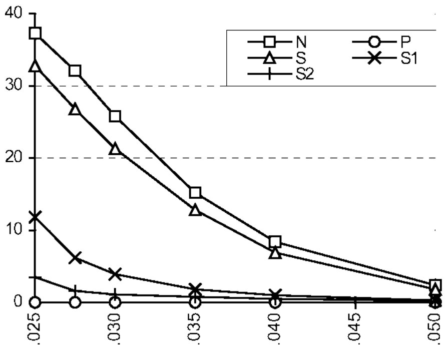

Fig. 1. BCR error rate ($q = 0$).

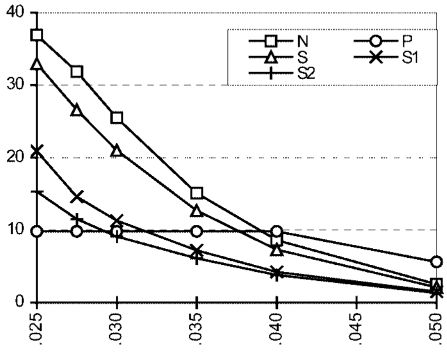

Fig. 2. BCR error rate ($q = 1/6$).

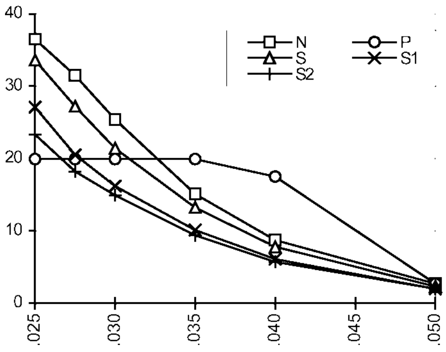

Fig. 3. BCR error rate ($q = 1/3$).

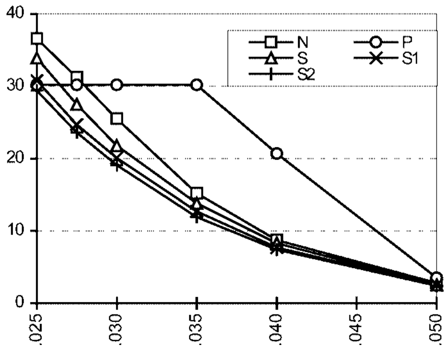

Fig. 4. BCR error rate ($q = 1/2$).

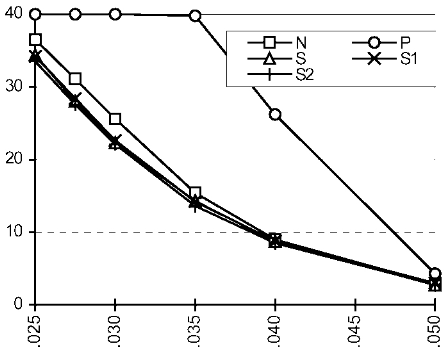

Fig. 5. BCR error rate ($q = 5/6$).

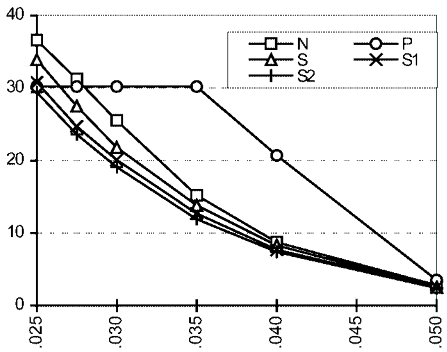

Fig. 6. BCR error rate ($q = 1$).

The initial values in the first collection were obtained by formulas (25)–(27). In these formulas, instead of the unobserved cluster numbers $\{Z_j,\ j = 1, \ldots, r\}$, cluster numbers resulting from the visual clusterization of the data were used (Fig. 7). Six initial clusters were distinguished.

The second collection of parameter values was obtained by an automatic cluster separation procedure described in (Sušinskas and Radavičius, 1998). In the sequential cluster

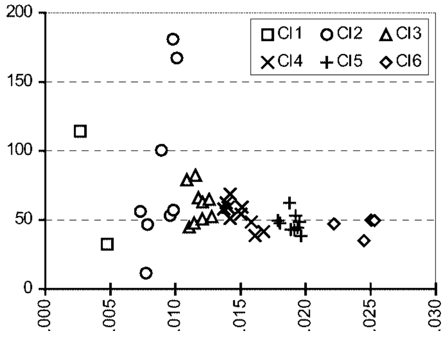

Fig. 7. Visual initial clusterization.

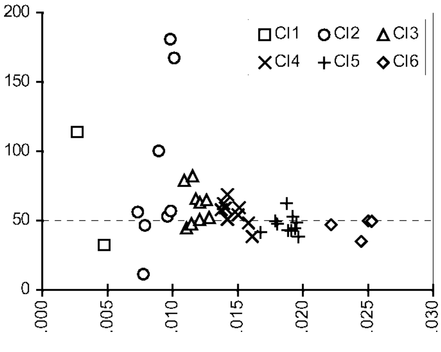

Fig. 8. Clusterization by method (S).

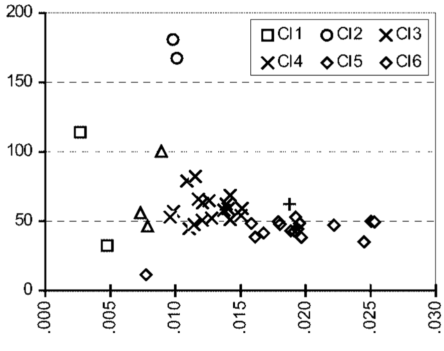

Fig. 9. Clusterization by method (S1).

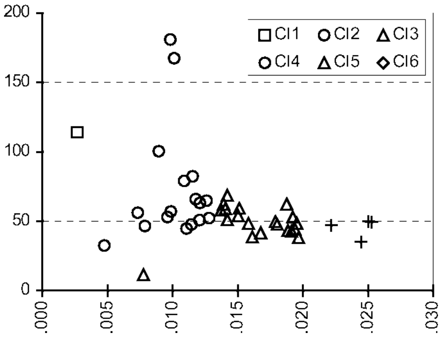

Fig. 10. Clusterization by method (S2).

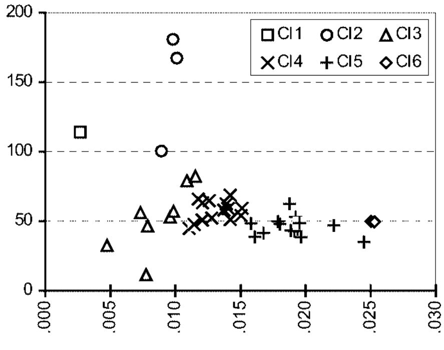

Fig. 11. Clusterization using model (P).

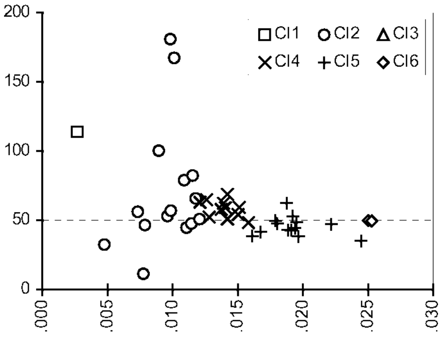

Fig. 12. Clusterization using model (N).

separating process we used the usual (i.e., without smoothing) EM algorithm trying to get the same number of clusters as in the first collection (Fig. 13).

We proceeded further in the same way as in the case of simulated data. The results are presented in Figs. 7–18. In general, they are similar to that in the case of simulated data.

The clusterization results produced by procedure (S) are the same as the initial ones (Figs. 7, 8, 13, and 14). While for the second initial clusterization this is natural, since the underlying models are the same and the only difference is in number of EM iterations,

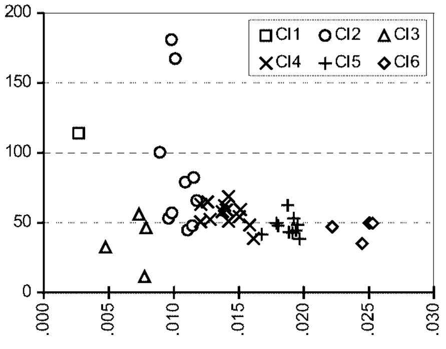

Fig. 13. Automatic initial clusterization.

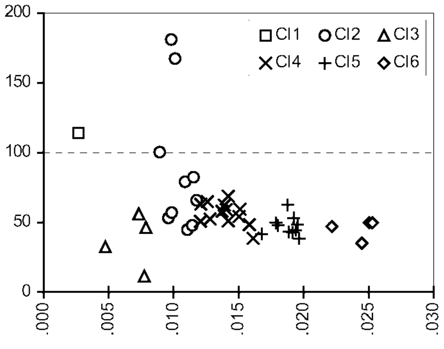

Fig. 14. Clusterization by method (S).

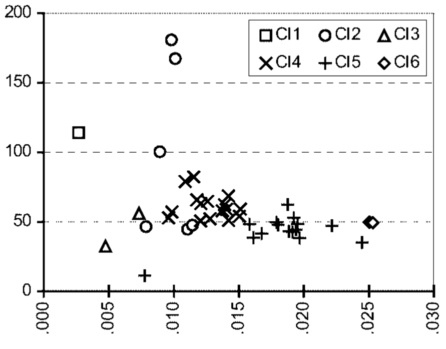

Fig. 15. Clusterization by method (S1).

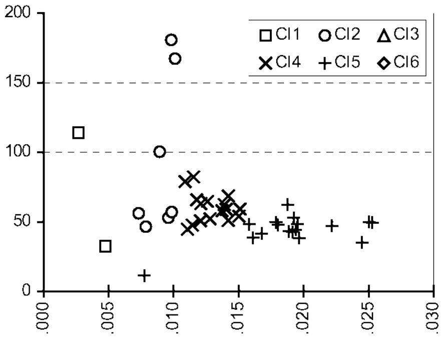

Fig. 16. Clusterization by method (S2).

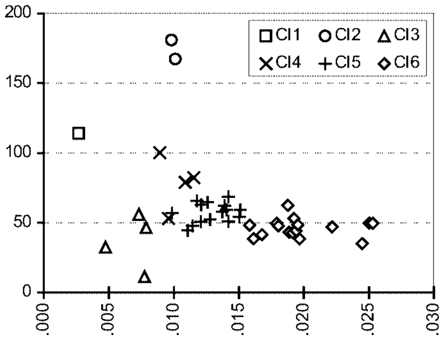

Fig. 17. Clusterization using model (P).

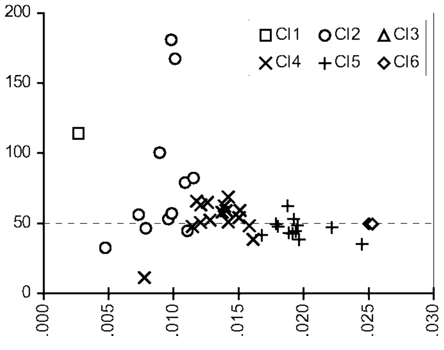

Fig. 18. Clusterization using model (N).

for the visual initial clusterization this fact indicates some extent of rigidity and hence biasedness of the method. Clusterization procedure (S) ignores the totals N_j and takes into account only ratios $K_j/N_j = 1, \ldots, r$, which is seen best when comparing with procedure (N). For the latter, the lesser the total N_j the greater impact of the prior probability upon the observation allocation. This is evident in broader allocation regions for massive clusters and narrower ones for small clusters as the total N_j decreases (see Figs. 8 and 12, 14 and 18).

Figs. 11 and 17 confirm the sensitivity of procedure (P) to the model assumptions. The picture is rather typical for the clusterization based on the bivariate Gaussian mixture model.

Procedure (S1) demonstrates irregular behavior assosiated with non-homogeneity of the clusters with respect to the totals. Clusters of high density of the totals in some places tend to “rob” elements from clusters with low density at these places even though these elements have quite different, from the typical elements of the former clusters, empirical rate K_j/N_j (Figs. 9 and 15). One can find some similarity of the partition produced by procedure (S1) with that of procedure (P).

The partitions of the data presented in Figs. 10 and 16 are, in a sense, intermediate between the alternative partitions considered. It seems that procedure (S2) yields the most reasonable clusterization results and is most flexible although the lefmost point of the 3rd cluster in Fig. 10 and the 5th cluster in Fig. 16 seems to be “suspicious”.

None of the procedures gives the same clusterization results for both starting partitions. One can interpret this fact as the lack of clear cluster structure in the data.

5. Conclusions

The simulation results show that the procedure (P) based on the parametric model is too sensitive to model assumptions (nonrobust). Procedures (S) and (S1) yield only a limited improvement in comparison to the method based on non-informative model (N), the second being slightly better. Procedure (S2) outperforms (S) and (S1) and parametric clusterization methods (N) and (P) for misspecified (parametric) probability models and is quite competitive with the latter two in the case of an adequite probability model.

The clusterization procedures under investigation were also applied to real data. The data consist of the number of stillborns among newborns in various districts of Lithuania during 1993–1997. In general, the results are similar to that in case of simulated data. Each of all the five clusterization procedures gives different results for two different starting partitions. This fact can be interpreted as lack of clear cluster structure in the data.

Acknowledgement. The authors are grateful to Prof. V. Kučinskas for the interesting problem and data provided.

References

Aivazyan S. A., V.M. Buchstaber, I.S. Yenyukov, L.D. Meshalkin (1989). *Applied Statistics. Classification and Reduction of Dimensionality*. Finansy i statistika, Moscow (in Russian).

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

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

Everitt, B.S., D.J. Hand (1981). *Finite Mixture Distributions*. Chapman and Hall, New York.

Ibragimov, I.A., R.Z. Khasminskii (1981). *Statistical Estimation: Asymptotic Theory.* Springer–Verlag, New York, Berlin.

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

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

Radavičius, M., J. Sušinskas (1998). Nonparametric Poisson mixtures and count data clusterization. In *Proceedings of 5th International Conference. Computer data analysis and modeling*, Vol. 2, pp. 49–54.

J. Sušinskas received the Ph.D. degree in mathematics from Minsk University in 1989. He is a reasercher at the Applied Statistics Department of the Institute of Mathematics and Informatics. His research interests include time series analysis, clusterization and classification, applications in medicine, biology, and economics.

M. Radavičius received the Ph.D. degree in mathematics from Steklov Mathematical Institute, Sankt-Petersburg division of Russian Academy of Science, in 1982. He is a senior reasercher at the Applied Statistics Department of the Institute of Mathematics and Informatics and an associate professor at Gediminas Technical University. His research interests are: efficient nonparametric and adaptive estimation, classification and cluster analysis, dimensionality reduction, application of statistical methods in medicine, biology, and social sciences.

Puasono mišinių modelių, skirtų diskrečių duomenų klasterizacijai, palyginimas

Jurgis SUŠINSKAS, Marijus RADAVIČIUS

Darbe aprašyti penki diskrečių duomenų klasterizavimo metodai, kurie remiasi Puasono skirstinių mišinių modeliu. Du iš jų yra parametriniai, o likusieji pagrįsti semiparametriniu modeliu. Klasterizavimui taikoma Bajeso klasifikavimo taisyklė su nežinomų parametrų reikšmėmis, pakeistomis jų didžiausio tikėtinumo įverčiais (’plug-in’ taisyklė). Aptariamų metodų veikimas ištirtas naudojant kompiuterinį modeliavimą, ir jie palyginti tarpusavyje pagal klasifikavimo klaidos dažnumą. Metodai taip pat yra taikomi realių medicininių duomenų klasterizacijai ir aptariami gauti rezultatai.