Journal of Classification 17:51-65 (2000)
DOI: 10.1007/s003570000004

Robust Classification of Multinomial Observations with Possible Outliers

Yu. S. Kharin

Belarusian State University

E. E. Zhuk

Belarusian State University

Abstract: Cluster analysis problem of a mixture of multinomial random observations is considered with the presence of outliers in the sample. Robust decision rule based on the truncation principle is proposed and investigated.

Keywords: Cluster analysis; Multinomial observations; Outliers; Robust decision rule; Truncation principle.

1. Introduction

Many effective clustering algorithms are already developed with hypothetical model assumptions about observations: independence, absence of missing and outlying data, continuous and Gaussian probability distributions of sample elements, etc. (e.g., Bock 1989; Bock 1996; McLachlan 1992; Hartigan 1975; Anderberg 1973). But the hypothetical model assumptions are often violated in practice (Huber 1981; Hampel, Ronchetti, Rousseeuw, and Stahel 1986; Kharin 1996) and the classical clustering methods usually forfeit their optimality properties as a consequence (cf.,

---

The authors would like to thank three referees and Professor Phipps Arabie, Editor, for their constructive reviews and useful remarks. Also we thank INTAS for financial support (project INTAS-93-725-ext).

Authors' Addresses: Yu. S. Kharin and E. E. Zhuk, Department of Mathematical Modeling and Data Analysis, Belarusian State University, 4 Fr.Skariny av., 220050 Minsk, Republic of Belarus; email: kharin@fpm.bsu.unibel.by

Kharin 1996; Kharin and Zhuk 1993). In this situation, stable (robust) clustering procedures are necessary. The monograph of Kharin (1996) and some papers (e.g., Kharin and Zhuk 1993) were devoted to this problem of stability (robustness) in clustering of continuous multivariate observations.

But in some applications the observations can be integer-valued: empirical investigation of different medicament effects in medicine, detection of nonhomogeneity in contingency tables (e.g., Kendall and Stuart 1967; Lui 1996), processing of guestionnaire forms in sociology, etc. (e.g., Anderberg 1973; Abusev 1998). For this case we need new special clustering methods intended for processing of discrete data.

In the present paper the problem of cluster analysis of discrete (multinomial) random observations is investigated, assuming the presence of outliers.

2. Mathematical Model

Let a complete system of $k \geq 2$ random events $\{A_l\}_{l \in K}$, $K = \{1, \ldots, k\}$, be partition on a probability space $(\Omega, \mathbf{F}, \mathbf{P})$:

$$\Omega = \bigcup_{l \in K} A_l,\ A_l \in \mathbf{F};\ \ A_l \cap A_j = \emptyset,\ l \neq j \in K\,. \tag{1}$$

The observations are obtained by n series of experiments $\{E_t\}_{t=1}^{n}$. The series E_t consists of m_t independent experiments. Each experiment results in an event from $\{A_l\}_{l \in K}$. The results of the t-th series E_t are recorded by a random integer-valued k-vector ($t = 1,n$):

$$\mathbf{X}_t = (X_{t1}, \ldots, X_{tk})' \in \mathbf{N}_o^k,\quad \mathbf{N}_o = \{0,1,2,\ldots\}\,, \tag{2}$$

where "′" is the transposition symbol, X_{tl} is a random number (frequency) of the appearance of A_l in E_t. Note that

$$\sum_{l \in K} X_{tl} = m_t$$

is the total number of experiments in the t-th series.

An observation $\mathbf{X}_t$ belongs to one of $L \geq 2$ classes $\{\Omega_1, \ldots, \Omega_L\}$: $D_t^o \in S$, $S = \{1, \ldots, L\}$, is a random index of the class from which the observation $\mathbf{X}_t$ originated, and

$$\pi_i = \mathbf{P}\,\{D_t^o = i\} > 0,\quad i \in S\,, \tag{3}$$

are the prior class probabilities: $\pi_1 + \ldots + \pi_L = 1$. Given the fixed $D_t^o = i$, a random vector $\mathbf{X}_t \in \mathbf{N}_o^k$ has a conditional multinomial probability distribution (p.d.):

$$\mathbf{P}\{\mathbf{X}_t = \mathbf{x} \mid D_t^o = i\} = q(\mathbf{x};\theta_i^o,m_t) \tag{4}$$
$$= \frac{m_t!}{\prod_{l\in K} x_l!} \prod_{l\in K} (\theta_{il}^o)^{x_l},\ \mathbf{x} \in Q^k(m_t),\ i \in S\,;$$
$$Q^k(m) = \{\mathbf{x} = (x_1,\ldots,x_k)' \in \mathbf{N}_o^k:\ \sum_{l\in K} x_l = m\}\,,$$

where $\theta_i^o = (\theta_{i1}^o,\ldots,\theta_{ik}^o)'$ is the vector of conditional probabilities, which characterizes the class Ω_i;

$$\theta_{il}^o = \mathbf{P}\{A_l \mid D_t^o = i\} > 0 \tag{5}$$

is the conditional probability of the event A_l ($l \in K$) from the complete system of events (1) under fixed class index $D_t^o = i$ ($i \in S$): $\Sigma_{l\in K}\,\theta_{il}^o = 1$. It is assumed that all L vectors $\theta_1^o,\ldots,\theta_L^o$ are distinct.

The cluster analysis problem consists in the classification of the sample $\tilde{\mathbf{X}}^n = \{\mathbf{X}_t\}_{t=1}^n$ of size n, i.e., in the construction of a statistical estimator $\hat{\mathbf{D}} = \hat{\mathbf{D}}(\tilde{\mathbf{X}}^n) = (\hat{D}_1,\ldots,\hat{D}_n)' \in S^n$ for an unknown random classification vector $\mathbf{D}^o = (D_1^o,\ldots,D_n^o)' \in S^n$ with unknown class characteristics $\{\pi_i,\theta_i^o\}_{i\in S}$.

But in practice (e.g., McLachlan 1992; Kharin 1996) the sample $\mathbf{X}^n$ usually contains some outliers. In this situation an observation

$$\mathbf{X} = (X_1,\ldots,X_k)' \in \mathbf{N}_o^k,\ \sum_{l\in K} X_l = m\,, \tag{6}$$

from the class Ω_i (where its class index $D^o = i$ is fixed) can be described by the conditional Tukey-Huber p.d. (e.g., Huber 1981):

$$\mathbf{P}\{\mathbf{X} = \mathbf{x} \mid D^o = i\} = p(\mathbf{x};\theta_i^o,m)$$
$$= (1-\varepsilon_i)q(\mathbf{x};\theta_i^o,m) + \varepsilon_i q(\mathbf{x};\theta_i^+,m)\,;$$
$$\mathbf{x} \in Q^k(m),\ 0 \le \varepsilon_i \le \varepsilon_{+i} < 1,\ i \in S\,,$$

where ε_{+i} is the so-called contamination level for Ω_i; $q(\mathbf{x};\theta_i^+,m)$, $\theta_i^+ \neq \theta_i^o$, is the contaminating multinomial p.d. If $\varepsilon_{+i} = 0$, then there are no outliers in the class Ω_i. The unconditional p.d. of an observation $\mathbf{X}$:

$$\mathbf{P}\{\mathbf{X} = \mathbf{x}\} = p^{\pi}(\mathbf{x}) = \sum_{i\in S} \pi_i p(\mathbf{x};\theta_i^o,m) \tag{7}$$
$$= \sum_{i\in S} \pi_i^* q(\mathbf{x};\theta_i^o,m) + \pi^* h^+(\mathbf{x})$$

is in this situation a mixture of $L+1$ p.d.'s, which determine $L+1$ classes $\{\Omega_0,\Omega_1,\ldots,\Omega_L\}$. As under the hypothetical model (where the contamination level $\varepsilon_+ = \max_{i\in S}\varepsilon_{+i}$ is equal to zero) the classes $\{\Omega_i\}_{i\in S}$ have the hypothetical conditional p.d.'s $\{q(\mathbf{x};\theta_i^o,m)\}_{i\in S}$, but their prior probabilities are different from (3):

$$\pi_i^* = \mathbf{P}\{d^o = i\} = \pi_i(1-\varepsilon_i), \quad i \in S\,. \tag{8}$$

The additional class Ω_0 containing outliers has the prior probability:

$$\pi^* = 1 - \sum_{i\in S} \pi_i^* = \sum_{i\in S} \pi_i\,\varepsilon_i\,, \tag{9}$$

and is described by the mixture of contaminating p.d.'s:

$$h^+(\mathbf{x}) = \sum_{i\in S} \frac{\pi_i\,\varepsilon_i}{\pi^*}\, q(\mathbf{x};\theta_i^+,m)\,. \tag{10}$$

Assuming outliers while solving this cluster analysis problem we need to construct the robust (vis-à-vis outliers) decision rule (DR):

$$\hat{\mathbf{D}} = \hat{\mathbf{D}}\,(\tilde{\mathbf{X}}^n) = (\hat{D}_1,\ldots,\hat{D}_n)' \in S_o^n\,,$$

$$S_o = \{0\} \cup S = \{0,1,\ldots,L\}\,,$$

which classifies the observations from the contaminated sample $\tilde{\mathbf{X}}^n = \{\mathbf{X}_t\}_{t=1}^n$ into the $L+1$ classes $\{\Omega_i\}_{i\in S_o}$. The observations attributed to the class Ω_0 are considered as outliers.

3. Optimal Discrimination of Multinomial Observations for the Hypothetical Model

First, consider the hypothetical model where outliers are absent ($\varepsilon_+ = 0$) and all class characteristics $\{\pi_i,\theta_i^o\}_{i\in S}$ are known *a priori*. In this case we need to classify the observation $\mathbf{X}$ defined by (6) and described by models (3)-(5).

Note, that the Bayesian DR (see Kharin 1996, Section 1.5)

$$d_o(\mathbf{X}) = \arg\max_{i\in S} \{\pi_i\, q(\mathbf{X};\theta_i^o,m)\} \tag{11}$$

$$= \arg\max_{i\in S}\{\ln\pi_i + \sum_{l\in K} X_l \ln\theta_{il}^o\}$$

has the minimal risk (classification error probability):

$$r_o = \mathbf{P}\,\{d_o(\mathbf{X}) \neq D^o\} \tag{12}$$

$$= 1 - \sum_{\mathbf{x}\in Q^L(m)} \max_{i\in S} \{\pi_i\, q(\mathbf{x};\theta_i^o,m)\}\,.$$

Let us investigate the asymptotics $m \to +\infty$. We introduce the following notation:

$$K^*(\theta_i^o, \theta_j^o) = \sum_{l \in K} \theta_{il}^o \ln \frac{\theta_{il}^o}{\theta_{jl}^o} \geq 0$$

is the directed Kullback (1959) divergence between classes Ω_i and Ω_j $(i,j \in S)$; $\Phi(\cdot)$ is the standard Gaussian distribution function with the probability density function:

$$\phi(z) = \frac{d}{dz}\Phi(z) = (2\pi)^{-1/2} \exp(-z^2/2), \quad z \in \mathbf{R},$$

and $N_k(\mu,\Sigma)$ is k-variate Gaussian probability distribution law with expectation vector $\mu \in \mathbf{R}^k$ and covariance $(k \times k)$-matrix Σ, which can be singular; if $\det(\Sigma) = 0$, then $N_k(\mu,\Sigma)$ is the k-variate singular Gaussian distribution law. Also let us introduce the vector $\mathbf{Y}$ of the empirical frequencies of the events (1) corresponding to the observation $\mathbf{X}$ from (^):

$$\mathbf{Y} = \frac{1}{m}\mathbf{X} = (Y_1, \ldots, Y_k)'; \; Y_l = \frac{X_l}{m} \geq 0, \; \sum_{l \in K} Y_l = 1. \tag{13}$$

Theorem 1. *Let the prior class probabilities (3) and the class-specific conditional probabilities (5) (the elements of the k-vectors $\{\theta_i^o\}_{i \in S}$) be separated from zero:*

$$\pi_i > 0; \quad \theta_{il}^o > 0, \; l \in K, \quad i \in S. \tag{14}$$

Then for $m \to +\infty$ the DR

$$d_*(\mathbf{Y}) = \arg\max_{i \in S} \{\sum_{l \in K} Y_l \ln \theta_{il}^o\} \tag{15}$$

is asymptotically optimal in the sense that the risk $r_* = \mathbf{P}\{d_*(\mathbf{Y}) \neq D^o\}$ satisfies the asymptotics:

$$r_*/r_o \to 1, \tag{16}$$

where r_o is the minimal risk value (12) attained by the Bayesian DR (11).

Also the following asymptotic statement holds:

$$r_*/\tilde{r} \to 1, \tag{17}$$

where

$$\tilde{r} = \tilde{r}(m) = 1 \tag{18}$$

$$- \sum_{i \in S} \pi_i \mathbf{P}\left\{\bigcap_{\substack{j \in S \\ j \neq i}} \left\{\sum_{l \in K} z_l^{(i)} \ln \frac{\theta_{il}^o}{\theta_{jl}^o} + \sqrt{m}\, K^*(\theta_i^o, \theta_j^o) \geq 0\right\}\right\};$$

$\mathbf{Z}^{(i)} = (z_1^{(i)}, \ldots, z_k^{(i)})' \in \mathbf{R}^k$ is a random k-vector with singular Gaussian distribution $N_k(\mathbf{0}_k, \mathbf{W}(\theta_i^o))$, where $\mathbf{0}_k$ is the null vector of $\mathbf{R}^k$ and $\mathbf{W}(\theta_i^o)$ is the following singular covariance $(k \times k)$-matrix $(i \in S)$:

$$\mathbf{W}(\theta_i^o) = (w_{lj}(\theta_i^o))_{l,j \in K},$$

$$w_{lj}(\theta_i^o) = \begin{cases} -\theta_{il}^o \theta_{ij}^o, & \text{if } l \neq j; \\ \theta_{il}^o (1 - \theta_{il}^o), & \text{if } l = j. \end{cases}$$

Proof. The result in (16) is proved by applying the risk asymptotic expansion method (cf., Kharin 1996) to the Bayesian risk (12) at $m \to +\infty$ under condition (14).

As in Kharin (1996, Section 1.5), using the normal (Gaussian) approximation for the conditional p.d. (given the fixed class index $D^o = i$, $i \in S$) of the random vector (13):

$$\sqrt{m}\,(\mathbf{Y} - \theta_i^o) \sim> N_k(\mathbf{0}_k, \mathbf{W}(\theta_i^o)), \quad m \to +\infty,$$

we obtain (17), (18). ∎

Corollary 1. *If under the conditions of Theorem 1 a value $v^* > 0$ exists such that for $i \neq j \in S$*

$$V(\theta_i^o, \theta_j^o) = \sqrt{\sum_{l \in K} \theta_{il}^o \left| \ln \frac{\theta_{il}^o}{\theta_{jl}^o} \right|^2 - \left| \sum_{l \in K} \theta_{il}^o \ln \frac{\theta_{il}^o}{\theta_{jl}^o} \right|^2} > v^*, \qquad (19)$$

then for $L \geq 2$ classes the following inequalities hold $(m \to +\infty)$:

$$r_o \leq r_* \leq \sum_{i \in S} \pi_i \Phi \left[-\sqrt{m} \min_{\substack{j \in S \\ j \neq i}} \frac{K^*(\theta_i^o, \theta_j^o)}{V(\theta_i^o, \theta_j^o)} \right],$$

and for the case of two classes $(L = 2)$:

$$\tilde{r} = \pi_1 \Phi \left[-\sqrt{m}\, \frac{K^*(\theta_1^o, \theta_2^o)}{V(\theta_1^o, \theta_2^o)} \right] + \pi_2 \Phi \left[-\sqrt{m}\, \frac{K^*(\theta_2^o, \theta_1^o)}{V(\theta_2^o, \theta_1^o)} \right]. \qquad (20)$$

From the results of Theorem 1 it follows that for a large number of experiments $(m \to +\infty)$ in the series from which the observation $\mathbf{X}$ is originated the asymptotic DR (15) can be used instead of the exact Bayesian DR (11). At $m \to +\infty$ the value $\tilde{r} = \tilde{r}(m)$ from (18), (20) can be used as the approximation of the risk. Note, unlike the Bayesian DR $d_o(\cdot)$ the DR $d_*(\cdot)$ doesn't depend on the prior class probabilities $\{\pi_i\}_{i \in S}$.

4. Robust Clustering Procedure Based on the Truncation Principle

4.1 The Truncation Principle and the Robust Clustering Procedure

Now, let us investigate the situation with outliers ($\varepsilon_+ > 0$), and use the truncation principle already applied in Kharin and Zhuk (1993) to construct the robust DR based on the truncated minimum contrast estimators of unknown parameters for the case, where the classes $\{\Omega_i\}_{i\in S}$ are described by continuous p.d.'s.

To classify the sample $\tilde{\mathbf{X}}^n = \{\mathbf{X}_t\}_{t=1}^n$ with outliers, we use the truncated DR $d_*^C(\mathbf{Y};\theta^o) \in S_o$, $\theta^o = ((\theta_1^o)', \ldots, (\theta_L^o)')'$, which is obtained from the asymptotic DR (15) by the truncation principle:

$$d_*^C(\mathbf{Y};\theta^o) = \begin{cases} \arg\max\limits_{i\in S}\{\sum\limits_{l\in K} Y_l \ln \theta_{il}^o\}, & \text{if } \min\limits_{i\in S} K(\mathbf{Y},\theta_i^o) \le C; \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Here

$$K(\mathbf{Y},\theta_i^o) = \begin{cases} \sum\limits_{l\in K}(Y_l - \theta_{il}^o)\ln\dfrac{Y_l}{\theta_{il}^o}, & \text{if } \min\limits_{l\in K} Y_l > 0; \\ +\infty, & \text{otherwise;} \end{cases}$$

is the symmetric Kullback distance between the observed frequencies $\mathbf{Y} = (Y_1, \ldots, Y_k)'$ and the class-specific probabilities $\theta_i^o = (\theta_{i1}^o, \ldots, \theta_{ik}^o)'$; $C \in (0, +\infty]$ is a truncation parameter which will be further evaluated. If $C = +\infty$, then $d_*^C(\cdot;\theta^o) \equiv d_*(\cdot)$ is the DR (15).

The clustering procedure based on the truncated DR (21) consists of the following steps.

1. The observed sample $\tilde{\mathbf{X}}^n = \{\mathbf{X}_t\}_{t=1}^n$ formed by the n observations (2) is transformed to the sample $\tilde{\mathbf{Y}}^n = \{\mathbf{Y}_t\}_{t=1}^n$, where

$$\mathbf{Y}_t = \frac{1}{m_t}\mathbf{X}_t = (Y_{t1}, \ldots, Y_{tk})',$$

$$Y_{tl} = \frac{X_{tl}}{m_t} = \frac{X_{tl}}{\sum_{j\in K} X_{tj}},$$

is the k-vector of the empirical frequencies of the events (1) in the

t-th series of experiments. Then $L \geq 2$ points from $\tilde{\mathbf{Y}}^n = \{\mathbf{Y}_t\}_{t=1}^n$ are arbitrarily chosen as the initial estimates $\{\theta_i^{(0)}\}_{i \in S}$ of the unknown parameters $\{\theta_i^o\}_{i \in S}$.

2. On the p-th step ($p = 1,2,...$) the estimate $\hat{\mathbf{D}}^{(p)} = (\hat{D}_1^{(p)}, \ldots, \hat{D}_n^{(p)})' \in S_o^n$ for the unknown classification vector $\mathbf{D}^o \in S_o^n$ is evaluated:

$$\hat{D}_t^{(p)} = d_*^C(\mathbf{Y}_t; \theta^{(p-1)}), \quad t = \overline{1,n},$$

where the applied DR $d_*^C(\cdot; \theta^{(p-1)})$ is obtained from the asymptotic truncated DR (21) by substituting $\theta^{(p-1)} = ((\theta_1^{(p-1)})', \ldots, (\theta_L^{(p-1)})')'$ instead of the unknown composite vector $\theta^o = ((\theta_1^o)', \ldots, (\theta_L^o)')'$ of the parameters $\{\theta_i^o\}_{i \in S}$. Then the estimates for $\{\theta_i^o\}_{i \in S}$ are adjusted:

$$\theta_i^{(p)} = \left[\sum_{t=1}^n \delta_{\hat{D}_t^{(p)},i}\right]^{-1} \sum_{t=1}^n \delta_{\hat{D}_t^{(p)},i}\, \mathbf{Y}_t, \quad i \in S.$$

Here $\delta_{j,i} = \{1, \text{if } j = i; 0, \text{if } j \neq i\}$ is the Kronecker symbol.

3. If $\hat{\mathbf{D}}^{(p)} = \hat{\mathbf{D}}^{(p-1)}$ ($p \geq 2$), then this iterative process is terminated, and $\hat{\mathbf{D}} := \hat{\mathbf{D}}^{(p)} \in S_o^n$ is the final estimate for $\mathbf{D}^o \in S_o^n$, which determines $L + 1$ clusters $\{\Omega_0, \Omega_1, \ldots, \Omega_L\}$:

$$\hat{\Omega}_i = \{\mathbf{X}_t : \hat{D}_t^{(p)} = i\}, \quad i \in S_o.$$

The observations from the cluster $\hat{\Omega}_0$ are considered as outliers. Note, the estimate $\hat{\theta} := \theta^{(p)}$ for the unknown $\theta^o = ((\theta_1^o)', \ldots, (\theta_L^o)')'$ is also constructed by this procedure.

4.2 Evaluation of the Truncation Parameter

The main problem in the clustering procedure proposed above is to determine the truncation parameter $C \in (0, +\infty]$. To solve this problem, let us investigate the behavior of the risk:

$$r_{\varepsilon^*}^C = \mathbf{P}\{d_*^C(\mathbf{Y}; \theta^o) \neq D^o\}, \quad \mathbf{Y} = \frac{1}{m}\mathbf{X}, \tag{22}$$

for the truncated DR (21) for classifying an observation $\mathbf{X}$ described by the model (7)-(10) ($D^o \in S_o$) considering C as a parameter.

First, let us prove the following helpful lemma.

Lemma 1. *If a random k-vector* $\mathbf{X}$ *has the multinomial p.d.:*

$$\mathbf{P}\{\mathbf{X} = \mathbf{x}\} = q(\mathbf{x};\theta,m),\ \mathbf{x} \in Q^k(m)\,;$$

$$\theta = (\theta_1, \ldots, \theta_k)',$$

then for $m \to +\infty$*:*

$$\mathbf{P}\{m \cdot K(\mathbf{Y},\theta) \le z\} \to F_{\chi^2_{k-1}}(z),\ \ z \ge 0\,, \tag{23}$$

where $F_{\chi^2_{k-1}}(\cdot)$ *is the probability distribution function of the* X^2*-distribution with* $k-1$ *degrees of freedom.*

For any $\theta^* = (\theta_1^*, \ldots, \theta_k^*)'$, $\Sigma_{l\in K}\,\theta_l^* = 1$, $0 < \theta_l^* < 1$, *if the Euclidean norm of the vector* $\theta^* - \theta$ *is separated from zero:* $|\theta^* - \theta| > 0$, *then at* $m \to +\infty$*:*

$$\mathbf{P}\left\{\frac{\sqrt{m}\,(K(\mathbf{Y},\theta^*) - K(\theta,\theta^*))}{\tilde{V}(\theta,\theta^*)} \le z\right\} \to \Phi(z),\ \ z \in R\,, \tag{24}$$

where

$$\tilde{V}(\theta,\theta^*) = \sqrt{\sum_{l\in K} \frac{(\theta_l - \theta_l^*)^2}{\theta_l} + 2K(\theta,\theta^*) + (V(\theta,\theta^*))^2}\,,$$

and $V(\theta,\theta^*)$ *is defined by (19).*

Proof. According to Lemma 1 a random k-vector $\mathbf{X}$ has the multinomial p.d. $q(\mathbf{x};\theta,m)$, and for $\mathbf{Y} = \frac{1}{m}\mathbf{X}$ we have:

$$\sqrt{m}\,(\mathbf{Y} - \theta) \dashrightarrow N_k(\mathbf{0}_k, \mathbf{W}(\theta)),\ \ m \to +\infty\,; \tag{25}$$

$$\mathbf{W}(\theta) = (w_{lj}(\theta))_{l,j\in K},\ \ w_{lj}(\theta) = \begin{cases} -\theta_l\theta_j, & \text{if } l \ne j; \\ \theta_l(1-\theta_l), & \text{if } l = j. \end{cases}$$

To prove (24), let us use the well known Cramér theorem (see, for example, Anderson 1958, Theorem 4.2.5):

$$\sqrt{m}\,(K(\mathbf{Y},\theta^*) - K(\theta,\theta^*)) \dashrightarrow \tag{26}$$

$$N_1(0, (\nabla_\theta K(\theta,\theta^*))'\mathbf{W}(\theta)\,\nabla_\theta K(\theta,\theta^*))\,,$$

where for $\nabla_\theta K(\theta,\theta^*) = (\frac{d}{d\theta_l} K(\theta,\theta^*))_{l\in K}$:

$$\frac{d}{d\theta_l} K(\theta,\theta^*) = \frac{d}{d\theta_l}\left[(\theta_l - \theta_l^*)\ln\frac{\theta_l}{\theta_l^*}\right] = 1 - \frac{\theta_l^*}{\theta_l} - \ln\frac{\theta_l^*}{\theta_l},\ \ l \in K\,, \tag{27}$$

and $|\nabla_\theta K(\theta,\theta^*)| \ne 0$ because of $|\theta^* - \theta| > 0$. Note that

$$(\nabla_\theta K(\theta,\theta^*))' \mathbf{W}(\theta)\, \nabla_\theta K(\theta,\theta^*) = (\tilde{V}(\theta,\theta^*))^2 ,$$

and the relation (24) holds.

In attempting to prove (23) under assumption that $\theta^* = \theta$, the asymptotic relation (26) can't be used: $\nabla_\theta K(\theta,\theta^*) \mid_{\theta^*=\theta} = \mathbf{0}_k$. Use the fact that for $m \to +\infty$ a random variable $\sqrt{m}(K(\mathbf{Y},\theta) - K(\theta,\theta)) = \sqrt{m}\, K(\mathbf{Y},\theta)$ has the p.d. which coincides with the p.d. of the random variable

$$\frac{1}{2}\sqrt{m}(\mathbf{Y}-\theta)' \nabla^2_\theta K(\theta,\theta^*) \mid_{\theta^*=\theta} (\mathbf{Y}-\theta)\sqrt{m} . \tag{28}$$

Evaluate $\nabla^2_\theta K(\theta,\theta^*) \mid_{\theta^*=\theta}$. From (27) we obtain:

$$\nabla^2_\theta K(\theta,\theta^*) = \left[\frac{d}{d\theta_j d\theta_l} K(\theta,\theta^*) \right]_{l,j\in K} ;$$

$$\frac{d}{d\theta_j d\theta_l} K(\theta,\theta^*) = \begin{cases} 0, & \text{if } l \neq j; \\ \dfrac{\theta_l^*}{\theta_l^2} + \dfrac{1}{\theta_l}, & \text{if } l = j, \end{cases}$$

and

$$\nabla^2_\theta K(\theta,\theta^*) \mid_{\theta^*=\theta} = 2 \cdot diag \left\{ \frac{1}{\theta_1}, \ldots, \frac{1}{\theta_k} \right\} .$$

From (28) and the last equation it follows that for $m \to +\infty$ the p.d. of random variable $\sqrt{m}\, K(\mathbf{Y},\theta)$ coincides with the p.d. of random variable

$$\sum_{l\in K} \frac{(\sqrt{m}(Y_l - \theta_l))^2}{\theta_l} = \sum_{l\in K} \frac{(X_l - m\theta_l)^2}{m\theta_l} ,$$

which has for $m \to +\infty$ the X^2-distribution with $k-1$ degrees of freedom. ▪

Theorem 2. *If under the conditions of Theorem 1:*

$$| \theta_i^o - \theta_j^+ | > 0, \quad i,j \in S ,$$

then the risk (22) of the DR (21) has the following representation:

$$r^C_{\varepsilon_*} = r^{+\infty}_{\varepsilon_*} + \alpha^C_{\varepsilon_*} - H^C_{\varepsilon_*} , \tag{29}$$

where $r^{+\infty}_{\varepsilon_*} = \mathbf{P}\{d_*(\mathbf{Y}) \neq D^o\}$ *is the risk of the asymptotic DR (15) (without truncation:* $C = +\infty$*),*

$$\alpha_{\varepsilon_*}^C = \sum_{i \in S} \pi_i (1 - \varepsilon_i) \tag{30}$$

$$\times \mathbf{P}\{\{d_*(\mathbf{Y}) = i\} \cap \{\min_{j \in S} K(\mathbf{Y}, \theta_j^o) \geq C\} \mid D^o = i\} ;$$

$$\alpha_{\varepsilon_*}^C \leq (1 - \varepsilon_+)\tilde{\alpha}^C, \ \tilde{\alpha}^C = 1 - F_{\chi^2_{l-1}}(m \cdot C) ,$$

and

$$H_{\varepsilon_*}^C = \sum_{j \in S} \pi_j \varepsilon_j \cdot \mathbf{P}\{\min_{i \in S} K(\mathbf{Y}, \theta_i^o) \geq C \mid D^o = 0\} \leq \tilde{H}_{\varepsilon_*}^C ; \tag{31}$$

$$\tilde{H}_{\varepsilon_*}^C = \min_{i \in S} \sum_{j \in S} \pi_j \varepsilon_j \Phi \left(\frac{\sqrt{m}(K(\theta_j^+, \theta_i^o) - C)}{\tilde{V}(\theta_j^+, \theta_i^o)} \right) .$$

Proof. The relation (29) is the representation of the risk (22) of the DR (21) which is directly obtained from (22):

$$r_{\varepsilon_*}^C = 1 - \sum_{i \in S} \pi_i (1 - \varepsilon_i) \mathbf{P}\{d_*^C(\mathbf{Y}; \theta^o) = i \mid D^o = i\}$$

$$- \sum_{j \in S} \pi_j \varepsilon_j \cdot \mathbf{P}\{d_*^C(\mathbf{Y}; \theta^o) = 0 \mid D^o = 0\}$$

$$= r_{\varepsilon_*}^{+\infty} + \alpha_{\varepsilon_*}^C - H_{\varepsilon_*}^C .$$

The inequalities (30) and (31) follow from the results (23) and (24) of Lemma 1 respectively. For $\alpha_{\varepsilon_*}^C$ we have:

$$\alpha_{\varepsilon_*}^C \leq \sum_{i \in S} \pi_i (1 - \varepsilon_i) \mathbf{P}\{\min_{j \in S} K(\mathbf{Y}, \theta_j^o) \geq C \mid D^o = i\}$$

$$\leq \sum_{i \in S} \pi_i (1 - \varepsilon_i) \mathbf{P}\{K(\mathbf{Y}, \theta_i^o) \geq C \mid D^o = i\} \leq (1 - \varepsilon_+)\tilde{\alpha}^C .$$

The inequality (31) is proved analogously. ▪

Let us analyze expressions (29)-(31). The value $H_{\varepsilon_*}^C \geq 0$ characterizes the positive effect of truncation: the larger $H_{\varepsilon_*}^C$ the more effective the truncated DR (21) (the smaller its risk $r_{\varepsilon_*}^C$). The value $\alpha_{\varepsilon_*}^C \geq 0$ describes the negative effect: because of truncation we forfeit some information, and the risk increases by the value $\alpha_{\varepsilon_*}^C$.

Corollary 2. *If the truncation parameter C satisfies the asymptotics:*

$$m \to +\infty, \ \ \varepsilon_+ \to 0, \ \ C \to +\infty, \ \text{and} \ \ (1 - F_{\chi^2_{l-1}}(m \cdot C))/\varepsilon_+ \to 0 , \tag{32}$$

then

$$r_{\varepsilon_+}^{C} = r_{\varepsilon_+}^{+\infty} - H_{\varepsilon_+}^{C} + o(\varepsilon_+)\,, \tag{33}$$

where $H_{\varepsilon_+}^{C} = O(\varepsilon_+) \geq 0$ *is the value from (31), and* $O(\cdot)$, $o(\cdot)$ *are the Landau symbols:* $O(\varepsilon_+)/\varepsilon_+ \to \nu$, $|\nu| > 0$, $o(\varepsilon_+)/\varepsilon_+ \to 0$ *for* $\varepsilon_+ \to 0$.

Proof. Under the asymptotics (32) for the value $\tilde{\alpha}^{C}$ from (30) we have: $\tilde{\alpha}^{C} = o(\varepsilon_+)$, and in the formula (29): $\alpha_{\varepsilon_+}^{C} = o(\varepsilon_+)$ because of the inequality (30) holds. From (31) it follows that $H_{\varepsilon_+}^{C} = O(\varepsilon_+)$. ▪

Note that the asymptotics in (32) allow one to determine the truncation parameter. But in practice the contamination level ε_+ is often unknown. In this situation we propose to determine the truncation parameter by the $X84$ rule of Hampel (Hampel et al. 1986): all the points which are far from their mean more than 5.2 times the standard deviation must be rejected. In our case using the results of Lemma 1 and the properties of the X^2-distribution, we obtain:

$$C = C(m,k) = \frac{5.2\sqrt{2(k-1)}}{m}, \quad k \geq 2\,. \tag{34}$$

Let us analyze how this choice (34) conforms to the asymptotics in (32). Under (34) for the value $\tilde{\alpha}^{C} = 1 - F_{\chi^2_{k-1}}(m \cdot C)$, we have: $\tilde{\alpha}^{C} = 1 - F_{\chi^2_{k-1}}(5.2\sqrt{2(k-1)})$. The values of $\tilde{\alpha}^{C} = \tilde{\alpha}^{C}(k)$ are presented in Table 1.

From Table 1 it is seen that, for example, under $\varepsilon_+ \geq 0.1$ and $k \leq 11$, the value $\tilde{\alpha}^{C}$ is less than ε_+ in order, and the choice (34) is acceptable.

Note that the truncation parameter depends on the number of experiments m in the series to which the observation $\mathbf{X}$ corresponds. In the clustering procedure constructed above we must evaluate the truncation parameter by the relation (34) for each observation: $C = C(m_t,k)$, $t = 1,n$.

5. Computer Results

We now investigate the proposed clustering procedure experimentally. The performance is characterized by the experimental frequency of error decisions in classifying hypothetical observations:

$$\gamma = \frac{1}{n - \tilde{n}} \sum_{t=1}^{n} \begin{cases} 1, & \text{if } \hat{D}_t \neq D_t^o \text{ and } D_t^o \neq 0; \\ 0, & \text{otherwise,} \end{cases} \tag{35}$$

Table 1: Values of $\tilde{\alpha}^C = \tilde{\alpha}^C(k)$ for the choice of the truncation parameter C by the $X84$ rule

k	2	3	4	5	6	7	11	16
$\tilde{\alpha}^C$	0.0067	0.0055	0.0052	0.0053	0.0057	0.0062	0.0098	0.0187

Table 2: Experimental results

Sample number	Number of outliers	Frequency of errors	
		Robust algorithm	Classical algorithm
1	6	0.088	0.235
2	4	0.056	0.167
3	4	0.111	0.361
4	7	0.061	0.273
5	5	0.114	0.200
6	3	0.054	0.135
7	9	0.161	0.355
8	6	0.118	0.206
9	8	0.094	0.219
10	5	0.086	0.171
Total	57	0.093	0.230

where

$$\tilde{n} = \sum_{t=1}^{n} \delta_{D_t^o,0}$$

is the number of outliers in the sample $\tilde{\mathbf{X}}^n = \{\mathbf{X}_t\}_{t=1}^n$ of size n.

As the example consider the case of two ($L = 2$) equiprobable ($\pi_1 = \pi_2 = 0.5$) hypothetical classes Ω_1, Ω_2 in the presence of outliers ($\varepsilon_1 = \varepsilon_2 = 0.15$):

$$\theta_1^o = (0.4, 0.2, 0.2, 0.2)' ;$$

$$\theta_2^o = (0.2, 0.2, 0.2, 0.4)' ;$$

$$\theta_1^+ = \theta_2^+ = (0.25, 0.25, 0.25, 0.25)' .$$

Using a Monte Carlo method, ten independent samples are generated. Each sample has the size $n = 40$ and contains observations produced by a series of $m = 50$ experiments. For the each sample the classical algorithm (without truncation: $C = +\infty$) and the robust truncated procedure (C is evaluated by the formula (34): $m = 50$, $k = 4$) are applied, and the experimental frequency of error decisions γ is evaluated by formula (35). The results are presented in Table 2.

To summarize the numerical results the total experimental frequency of error decisions is evaluated (all $n = 400$ observations are considered as one sample). Table 2 demonstrates that the robust algorithm essentially improves the clustering performance (approximately twofold). Note that for this numerical example, according to formula (20): $\tilde{r} = \tilde{r}(50) = 0.030$, where $\tilde{r}$ is the asymptotic risk value attained by the asymptotically optimal DR (15) for the hypothetical model (when outliers are absent and all class characteristics are known *a priori*). Table 2 also shows that for the robust algorithm the total experimental frequency of error decisions is close to this asymptotic value.

References

ABUSEV, R. A. (1998), "Statistical Group Classification of Multinomial Populations," in *Computer Data Analysis and Modeling*, Vol. 3 (in Russian), Eds., S.A. Aivazyan, and Yu.S. Kharin, Minsk: Belarusian State University, 18-23.

ANDERBERG, M. R. (1973), *Cluster Analysis for Applications*, New York: Academic Press.

ANDERSON, T. W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

BOCK, H.-H. (1989), "Probabilistic Aspects in Cluster Analysis," in *Conceptual and Numerical Analysis of Data*, Ed., O. Opitz, Berlin: Springer-Verlag, 12-44.

BOCK, H.-H. (1996), "Probability Models and Hypotheses Testing in Partitioning Cluster Analysis," in *Clustering and Classification*, Eds., P. Arabie, L.J. Hubert, and G. De Soete, River Edge, New Jersey: World Scientific, 377-463.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J., and STAHEL, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

HARTIGAN, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

HUBER, P. J. (1981), *Robust Statistics*, New York: Wiley.

KENDALL, M. G., and STUART, A. (1967), *Advanced Theory of Statistics*, Vol. 2: *Inference and Relationship*, New York: Hafner.

KHARIN, YU. S. (1996), *Robustness in Statistical Pattern Recognition*, Dordrecht: Kluwer.

KHARIN, YU. S., and ZHUK E. E. (1993), "Asymptotic Robustness in Cluster Analysis for the Case of Tukey-Huber Distortions," in *Information and Classification: Concepts, Methods and Applications*, Eds., O. Opitz, B. Lausen, and R. Klar, New York: Springer-Verlag, 31-39.

KULLBACK, S. (1959), *Information Theory and Statistics*, New York: Wiley.

LUI, K.-J. (1996), "Hypothesis Testing Procedures for a Series of Independent Fourfold Tables Under Inverse Sampling," *Biometrical Journal, 38*, 347-357.

McLACHLAN, G. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.