Control and Cybernetics

vol. **37** (2008) No. 3

A new method for identifying outlying subsets of data*

by

Marta Zalewska[1], Antoni Grzanka[2], Wojciech Niemiro[3] and Bolesław Samoliński[4]

[1]Medical University of Warsaw
Department of Prevention of Environmental Hazards and Allergology
ul. Żwirki i Wigury 61, 02-091 Warszawa, Poland

[2]Warsaw University of Technology, Institute of Electronic Systems
Pl. Politechniki 1, 00-661 Warszawa, Poland

[3]Nicolaus Copernicus University,
Faculty of Mathematics and Computer Sciences
ul. Gagarina 11, 87-100 Toruń, Poland

Abstract:
In various branches of science, e.g. medicine, economics, sociology, it is necessary to identify or detect outlying subsets of data. Suppose that the set of data is partitioned into many relatively small subsets and we have some reason to suspect that one or several of these subsets may be atypical or aberrant. We propose applying a new measure of separability, based on the ideas borrowed from the discriminant analysis. In our paper we define two versions of this measure, both using a jacknife, leave-one-out, estimator of classification error. If a suspected subset is significantly well separated from the main bulk of data, then we regard it as outlying. The usefulness of our algorithm is illustrated on a set of medical data collected in a large survey "Epidemiology of Allergic Diseases in Poland" (ECAP). We also tested our method on artificial data sets and on the classical IRIS data set. For a comparison, we report the results of a homogeneity test of Bartoszyński, Pearl and Lawrence, applied to the same data sets.

Keywords: multidimensional homogeneity test, misclassification error, discriminant analysis, medical data.

1. Introduction

Let us consider data of the form of an array $X = [x_{i,j}]_{i=1,...,n;j=1,...,d}$ with n rows and d columns. The data describe n objects. Every row $x_i^T = [x_{i,1}, ..., x_{i,d}]$ consists of values of d features (or attributes) for a single object.

*Submitted: June 2007; Accepted: October 2008.

Suppose that the set of objects is partitioned into many relatively small subsets and we have some reason to suspect that one or several of these subsets may be atypical or aberrant. Our motivating example is a set of questionnaires partitioned into subsets corresponding to pollsters. Similar situations occur very frequently if data concerning, e.g., patients, are partitioned into subsets corresponding to different hospitals (with some of the hospitals possibly atypical), or students, partitioned into subsets corresponding to schools, etc. The problem is particularly important if we have very large sets of data. Although there is extensive literature on identifying individual outliers among data points (Barnett and Toby, 1994; Hampel et al., 1986; Renze, no date), detecting atypical subsets has not received enough attention yet. This problem is closely related to discriminant analysis (Morrison, 1967; Koronacki, 2005; Lachenbruch, 1975; Ripley, 1996), discordancy tests, homogeneity tests, goodness-of-fit tests (Mardia, Kent and Bibby, 1979; Venables and Ripley, 2002) and block procedures for multiple outliers (Barnett and Toby, 1994).

For simplicity let us focus on just one subset, marked out. We are to decide if this subset is abnormal, unrepresentative, e.g. includes some errors or differs from the rest of data with respect to the mean or covariance structure. In order to verify or falsify our supposition we perform a test of discordancy. We will construct a suitable new measure J, which quantifies separability between our suspected subset and the rest of data. Small value of J indicates good separation and thus supports our supposition. The measure J is normalized so that it takes values in the interval [0,1], with 0 corresponding to perfect separability. Therefore if the value of J is significantly small, this is an evidence of atypicality of the subset under consideration. In fact, we will define two versions of measure J, denoted *Jd* and *Jw*. Precise definitions are given in Section 3. In view of our applications, both these measures are related to quadratic discrimination and estimation of classification error (Koronacki, 2005; Lachenbruch, 1967, 1975; Lachenbruch and Mickey, 1968). In principle our idea of quantifying separability can be applied more generally, with other methods of discrimination used instead of quadratic discrimination.

Formally, the problem, which we consider in this paper, can be regarded as a special case of testing homogeneity between two samples. However, we should point out some differences. We have in mind situations where a relatively small subset may stand out from the homogeneous main bulk of data. Moreover, in most applications we should perform simultaneous tests of multiple hypotheses, corresponding to several suspected subsets. Let us also emphasize that we assume an *a priori* given and known partition of data into subsets; we are only to detect *which of them* are outlying. In this respect our procedure differs from detection of multiple outliers (Barnett and Toby, 1994).

2. The general scheme of the algorithm

Our algorithm consists of the following two steps:

Step 1. We perform the principal component analysis (Mardia, Kent and Bibby, 1979; Morrison, 1967) in order to reduce the dimensionality of data. We retain only a limited number of pricipal components. It is necessary if the number of objects is not too large. Let us note that the quadratic discrimination requires estimation of covariance matrices from two samples. To ensure reasonable precision of estimation, the ratio of the sample size to the dimension cannot be too small.

Step 2. We fix a threshold C. For the considered subset of objects, we compute the measure J, which indicates how well this subset is separated from the rest of data. If $J < C$, then we decide that the subset is atypical. Otherwise we do not have enough evidence to suspect its atypicality. Let us remark that our approach fits in the classical framework of statistical tests of significance.

3. Description of the algorithm and simulations

3.1. Definition of the measures *Jd* and *Jw*

Recall that we have an $n \times d$ matrix $X = [x_{i,j}]_{i=1,\ldots,n;j=1,\ldots,d}$ with a specified subset of n_1 rows. We try to separate this subset from the remaining $n_2 = n - n_1$ rows, using the quadratic discriminant function (QDF). Let us first recall the basic formulas and the background of classical discriminant analysis. Suppose that we have two populations (classes) described by multivariate normal distributions $N(\mu_k, V_k)$ for $k = 1, 2$. We consider functions given by

$$D_k(x) = \ln(\pi_k p_k(x)) = -\frac{1}{2}(x-\mu_k)^T V_k^{-1}(x-\mu_k) - \frac{1}{2}\ln|V_k| + \ln\pi_k + const$$

for $k = 1, 2$ where $p_k(x)$ is the density of the probability distribution in the kth class and π_k is the prior probability of the kth class. The QDF is defined as $D(x) = D_2(x) - D_1(x)$. The posterior probability of the two classes is given by

$$p(1|x) = \frac{\pi_1 p_1(x)}{\pi_1 p_1(x) + \pi_2 p_2(x)} = \frac{1}{1 + e^{D(x)}},$$
$$p(2|x) = \frac{\pi_2 p_2(x)}{\pi_1 p_1(x) + \pi_2 p_2(x)} = \frac{1}{1 + e^{-D(x)}}.$$

The Bayes classification rule assigns (the object described by) vector x to class 1 or 2 according to $p(1|x) > p(2|x)$ or $p(1|x) \leqslant p(2|x)$, respectively. This decision rule is also called MAP (maximum a posteriori) estimate of the class:

$$\mathrm{MAP}(x) = \begin{cases} 1 & if \quad D(x) < 0; \\ 2 & if \quad D(x) \geqslant 0. \end{cases}$$

The MAP decision rule is known to be optimal, i.e. it minimizes the probability of misclassification. Since the parameters of the classes are usually unknown, in

practice QDF with estimated parameters is used. It is obtained in the following way. We regard data as a set of row vectors $X = \{x_i\}$, $i = 1, ..., n$ partitioned into two classes, C_1, C_2. Here x_i denotes the d-dimensional vector of attributes of the ith object. We will write $i \in C_k$ if ith object belongs to kth class. Symbol $\hat{D}(x|X)$ will denote empirical QDF, given by a formula analogous to that for $D(x)$ with the population parameters μ_k and Σ_k replaced by their estimates:

$$\hat{\mu}_k = \bar{x}_k = \frac{1}{n_k} \sum_{i \in C_k} x_i, \quad \hat{V}_k = \frac{1}{n_k - 1} \sum_{i \in C_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T.$$

Of course, it would be possible to replace also the prior probabilities by their estimates, but for our purposes we decided to use fixed priors $\pi_1 = \pi_2 = \frac{1}{2}$.

Finally, we are in a position to precisely define the measures *Jd* and *Jw*. They are based on the leave-one-out estimators of the class assignments $\mathrm{MAP}(x_i)$ and the posteriors $p(k|x_i)$ for all data points x_i, $i = 1, ..., n$. Let $\hat{D}_{CV}(x_i|X - x_i)$ stand for the QDF estimated from the data with removed vector x_i, evaluated at x_i. Explicitly, we define for $k = 1, 2$ and $i = 1, ..., n$,

$$\begin{aligned}\hat{D}_{k,CV}(x_i|X - x_i) &= \ln(\pi_k \hat{p}_{k,CV[-i]}(x_i)) \\ &= -\frac{1}{2}(x_i - \hat{\mu}_{k,CV[-i]})^T \hat{V}^{-1}_{k,CV[-i]}(x_i - \hat{\mu}_{k,CV[-i]}) - \frac{1}{2} \ln \left| \hat{V}_{k,CV[-i]} \right| \\ &\quad + \ln \pi_{k,CV[-i]} + const,\end{aligned}$$

and

$$\hat{D}_{CV}(x_i|X - x_i) = \hat{D}_{2,CV}(x_i|X - x_i) - \hat{D}_{1,CV}(x_i|X - x_i),$$

where subscript CV or $CV[-i]$ indicates the leave-one-out cross validation estimates, i.e.

$$\hat{\mu}_{k,CV[-i]} = \frac{1}{n_k - 1} \sum_{r \in C_k, r \neq i} x_r \quad \text{if } i \in C_k \quad \text{and}$$

$$\hat{\mu}_{k,CV[-i]} = \hat{\mu}_k = \bar{x}_k \text{ otherwise,}$$

$$\hat{V}_{k,CV[-i]} = \frac{1}{n_k - 2} \sum_{r \in C_k, r \neq i} (x_r - \bar{x}_k)(x_r - \bar{x}_k)^T \text{ if } i \in C_k \text{ and}$$

$$\hat{V}_{k,CV[-i]} = \hat{V}_k \text{ otherwise,}$$

Quantities $\hat{p}_{CV}(k|x_i; X - x_i)$ and $\widehat{\mathrm{MAP}}_{CV}(k|x_i; X - x_i)$ are defined in an obvious way in terms of $\hat{D}_{CV}(x_i|X - x_i)$:

$$\hat{p}_{CV}(1|x_i; X - x_i) = \frac{1}{1 + \exp[\hat{D}_{CV}(x_i|X - x_i)]},$$

$$\hat{p}_{CV}(2|x_i; X - x_i) = \frac{1}{1 + \exp[-\hat{D}_{CV}(x_i|X - x_i)]}.$$

$$\widehat{MAP}_{CV}(x_i; X - x_i) = \begin{cases} 1 & if \quad \hat{D}_{CV}(x_i|X - x_i) < 0; \\ 2 & if \quad \hat{D}_{CV}(x_i|X - x_i) \geqslant 0. \end{cases}$$

Finally, writing $l_i(k) := \hat{p}_{CV}(k|x_i; X - x_i)$ and $m_i := \widehat{\mathrm{MAP}}_{CV}(k|x_i; X - x_i)$ for brevity, we define:

$$Jd = \frac{1}{2}\left[\frac{\#\{i : i \in C_1, m_i = 2\}}{n_1} + \frac{\#\{i : i \in C_2, m_i = 1\}}{n_2}\right],$$

$$Jw = \frac{1}{2}\left[\frac{1}{n_1}\sum_{i \in C_1} l_i(2) + \frac{1}{n_2}\sum_{i \in C_2} l_i(1)\right].$$

Note that *Jd* is the usual leave-one-out estimator of the probability of misclassification (Koronacki, 2005; Lachenbruch, 1967, 1975; Lachenbruch and Mickey, 1968). The measure *Jw* can be regarded as a weighted or fuzzy version of *Jd*. If we replaced $l_i(1)$ by 1 or 0 according to $l_i(1) > l_i(2)$ or $l_i(1) \leqslant l_i(2)$ – and $l_i(2)$ analogously – then we would obtain exactly the formula for *Jd*.

Let us sum up the above considerations. We estimate the probability of incorrect classification by the cross validation leave-one-out method. In this way we construct the measure *Jd*. An alternative measure *Jw* is defined analogously, but we use estimated *posterior probabilities* of the two classes instead of the class indicators. It is interesting to note that in our simulation experiments described in the next section, the measure *Jw* turned out to be better (more sensitive) than *Jd*.

We should emphasize that computation of *Jd* and *Jw* makes sense *even if the probability distributions in both classes are not normal.* In fact, *Jd* is an unbiased estimator for the probability of misclassification of QDF based on the learning sample of size n-1 (Lachenbruch, 1967; Lachenbruch and Mickey, 1968). Moreover, in the definition of our separability measure we can use virtually any algorithm of classification instead of QDF. In this way the whole family of separability measures can be introduced, based on the same general idea. In this paper we have chosen to work with QDF, because we think it is most suitable for application to our survey data.

3.2. Choosing the value of C

We select the threshold C according to the classical theory of testing statistical hypotheses (Koronacki, 2005; Venables and Ripley, 2002; Watała, 2002). The null hypothesis is that the given subset is *not* different from the rest of data (i.e. objects belonging to the subset under consideration do not differ systematically from the remaining objects). The test rejects the null hypothesis if the test statistic falls below the critical value ($Jd < C$ or $Jw < C$). We should choose C so that the test has the given level of significance α. Of course, analytical computation of C is impossible. In the era of easily available powerful computers and flexible statistical software, this difficulty can be overcome by simulation methods. In our work we use R software environment for statistical computing (Becker, Chambers and Wilks, 1988; Venables and Ripley, 2002).

We repeatedly select marked out subsets *at random*, each subset consisting of n_1 rows, from the whole set of data. For each random selection, we perform computations described in Subsection 3.1, i.e. we compute the measure of separability *Jd* or *Jw*. The histogram of these values is an empirical approximation to the probability distribution of the random variable (*Jd* or *Jw*, respectively) under the null hypothesis. Clearly, the quantile of order $1-\alpha$ of this distribution is the sought critical threshold C.

The empirical probability distribution under the null hypothesis is shown in the upper parts of Fig. 1 (histogram J represents the distribution of *Jd*) and Fig. 2 (histogram of *Jw*). The computations are performed on an artificial set of data, generated from a multivariate normal distribution, for n=1000, n_1=20, d=10. The quantile of *Jd* of order $1-\alpha$=0.99 is equal to C=0.3928571 and for $1-\alpha$=0.95 we have C=0.4250000.

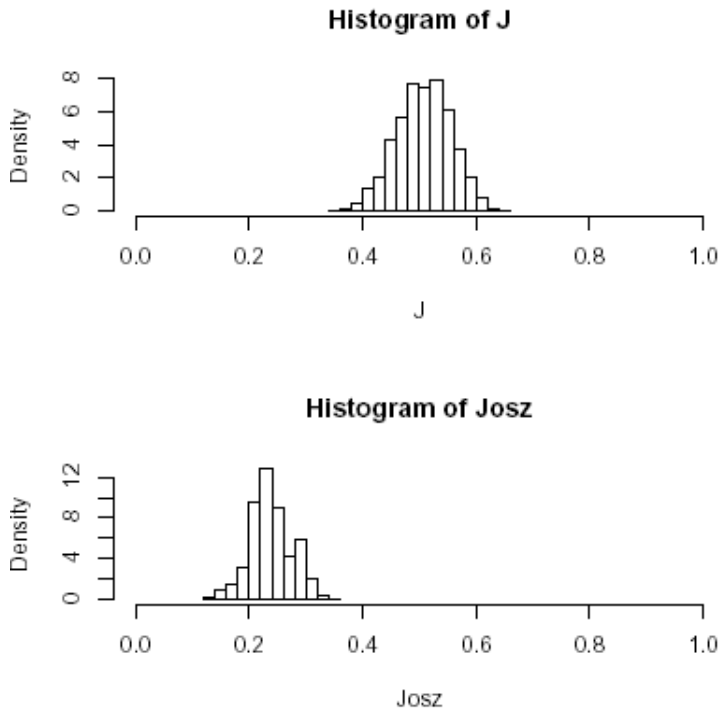

Figure 1. Measure *Jd*: empirical probability distribution under the null hypothesis (histogram J) and under an alternative hypothesis (histogram $Josz$), for n=1000, n_1=20, d=10.

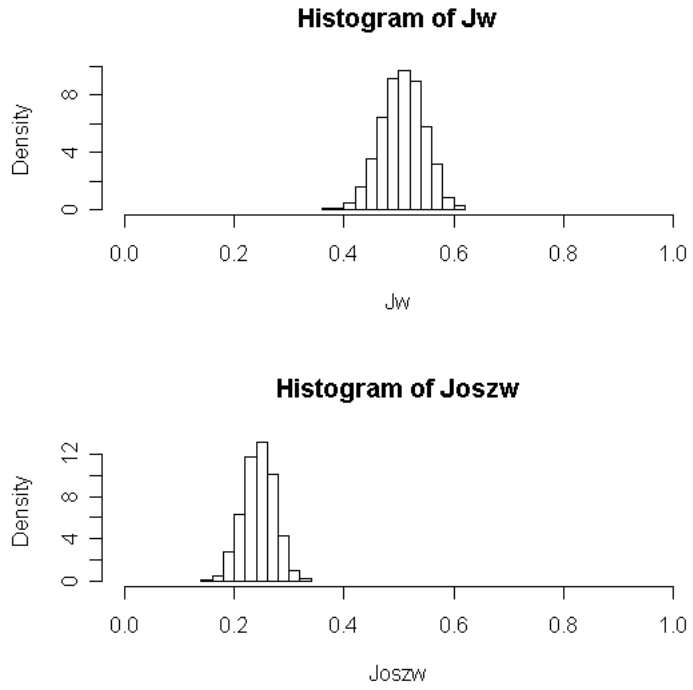

Figure 2. Measure *Jw*: the empirical probability distribution under the null hypothesis (histogram *Jw*) and under an alternative (histogram *Joszw*), for n=1000, n_1=20, d=10.

4. Computation of the power of the test for alternative hypotheses

Let us now examine the distribution of our separability measures when there is some systematic difference between objects in the marked out subset and the rest of data. Namely, we distort all the objects in the marked out subset according to the formula $x'_{i,j} = x_{i,j}/2 + 1/2$ (if row i belongs to the subset, $x'_{i,j} = x_{i,j}$ otherwise). The computations are quite analogous to the previously considered ones. The results for *Jd* are shown in the lower part of Fig. 1 (histogram *Josz*). In this way we compute the power of the test. For the special form of alternative described above, the power is very close to 100% (for the tests at standard levels of significance α=0.05 and even $\alpha = 0.01$).

Analogous computations are conducted also for the second version of our measure, *Jw*. The results are shown in Fig. 2 (histogram *Joszw*). By comparing Fig. 1 with Fig. 2 we can see that the properties of *Jd* and *Jw* are similar. Both

of our measures can be used to quantify separability of data subsets in much the same way. However, *Jw* is more sensitive and thus tests based on *Jw* are more powerful than those based on *Jd*. Therefore, in our further analyses we concentrate on *Jw*.

5. Analysis of real-life data

5.1. Description of the ECAP data set and preliminary analysis

In our work we use data collected in a preliminary part of Polish Allergic Survey, ECAP 2007. Array $X = [x_{i,j}]_{i=1,\ldots,n;j=1,\ldots,d}$ of dimensions n=2240 (respondents) and d= 17 (features or attributes) is partitioned into 21 subsets of different size. These subsets correspond to different pollsters. The cardinality of the subsets is given in Fig. 3. The problem is to identify which subsets are atypical.

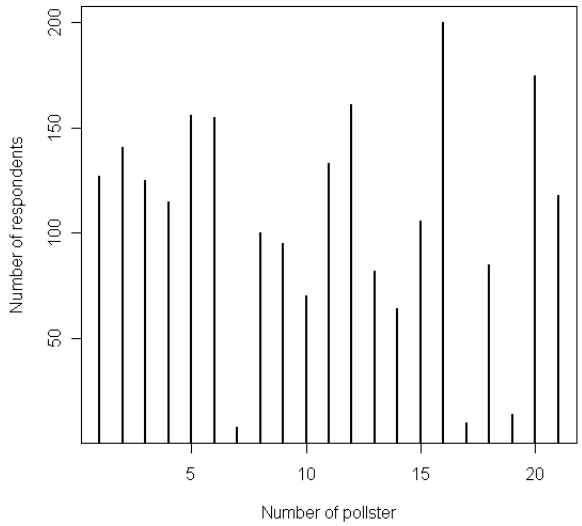

Figure 3. Numbers of respondents belonging to the 21 subsets. Vertical bars give the number of respondents questioned by each of the 21 pollsters.

Before applying our main algorithm, we conducted the principal component analysis. The goal was to reduce dimensionality. The standard deviations corresponding to the principal components are:

30.96199029	26.56101213	8.36990581	6.26343886	1.58521482
0.92421190	0.48679767	0.45873532	0.27564555	0.24208383
0.16604877	0.13678143	0.11012109	0.08888029	0.07883676

These values are shown in Fig. 4.

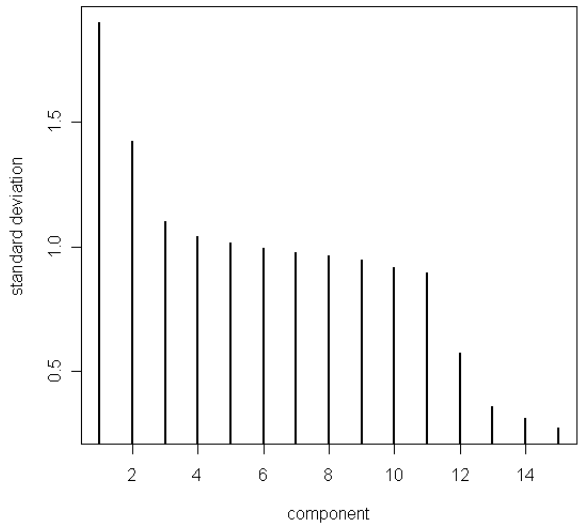

Figure 4. Standard deviations for subsequent principal components

On the basis of the above results we decided to use only the first four principal components in further analysis. Therefore, the dimension d of our data was reduced from 17 to 4 and discriminant analysis was conducted in 4-dimensional space. The first two principal components of our set of data are presented in Fig. 5 and the other two (components no. 3 and 4) - in Fig. 6. The points belonging to subset no. 13 are shown as bigger circles in both figures. The reasons why subset no. 13 is singled out is explained later in this section.

In Fig. 5 points are clearly placed on several parallel straight lines. This phenomenon simply reflects the discrete structure of data. Many of the features are either binary or take only a small number of possible values.

5.2. Results of application of our method

We exclude subsets no. 7, 17 and 19 from further considerations, because they contain too few respondents (less than 10). For each of the remaining 18 subsets we conduct a statistical test of the null hypothesis described in Subsection 3.2 (that a given subset is not significantly different from the rest of data). Let us note that the threshold C considered in Subsection 3.2 depends not only on the given significance level α but also on n and n_1 (the size of the data set and the subset) and on the overall structure of data. Therefore, we had to repeat computations described in 3.2 on our data set separately for each value of n_1 (i.e. 18 times). In all these computations we used the array X containing real data set described in Section 4.1. We obtained 18 distinct (but similar in shape) probability distributions of Jw under the null hypothesis. Nine of these distributions are sketched in Fig. 7 and one of them is shown in detail in Fig. 8.

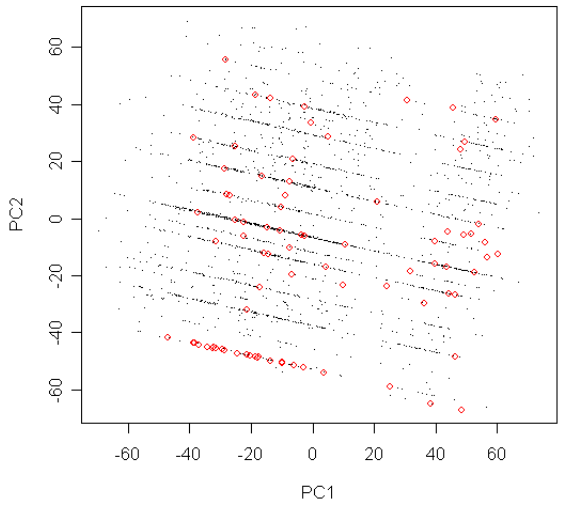

Figure 5. Data set in the space of the first and second principal components. Points belonging to subset no. 13 are marked with bigger circles

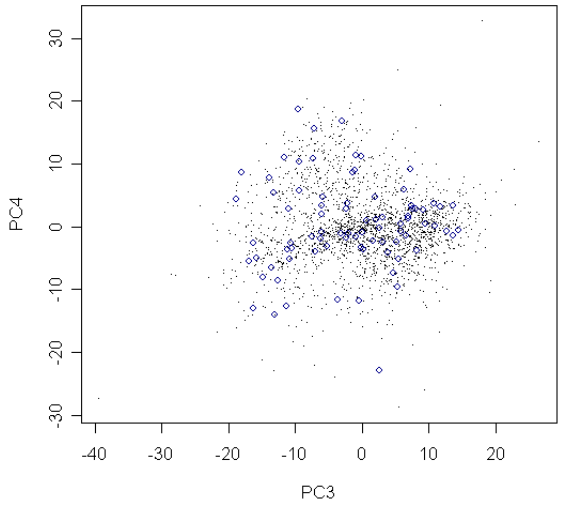

Figure 6. Data set in the space of the third and fourth principal components. Points belonging to subset no. 13 are marked with bigger circles

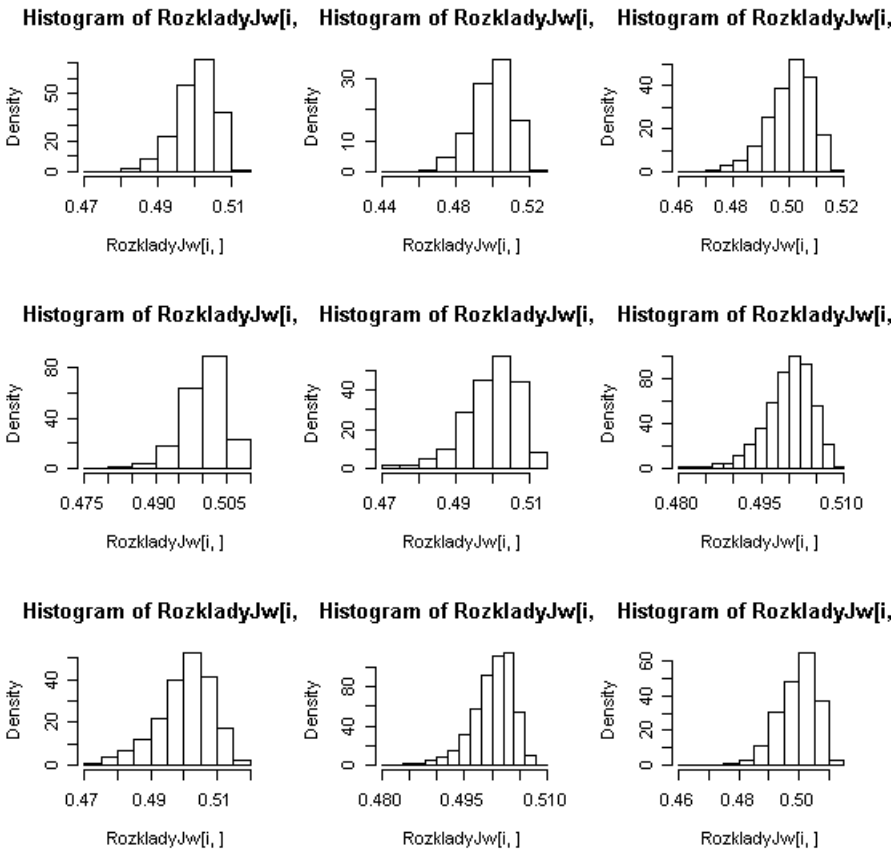

Figure 7. Examples of probability distributions of index Jw under the null hypothesis for the real data set X for nine chosen values of n_1.

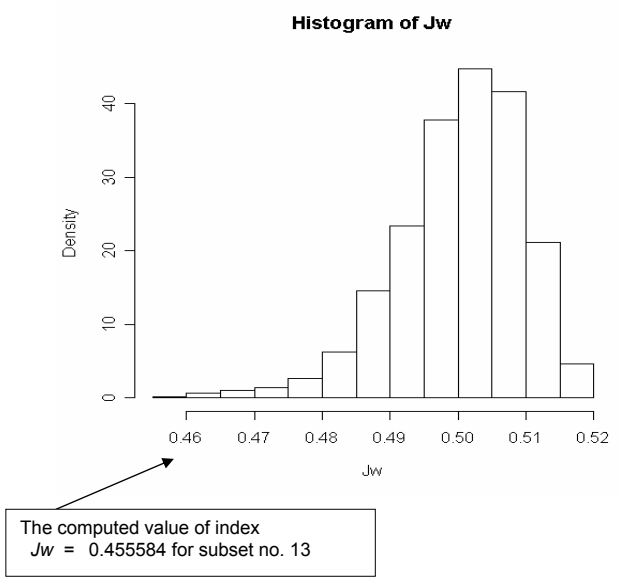

Figure 8. Empirical probability distribution of index Jw under the null hypothesis for the real data set X with n=2240, n_1=82, d=4. The actual value of Jw for the subset no. 13 is indicated by the arrow (this subset contains 82 respondents).

The obtained values of the separability index Jd and Jw for all 18 subsets under consideration are shown in Table 1.

Let us now explain the meaning of the obtained results, focusing on subset no. 13 (marked by two stars and boldface in Table 1). We have chosen this subset because of the smallest p-value. The value of Jd is equal 0.3995287 and the value of Jw is equal 0.4555840. Among 1000 subsets of 82 elements selected at random from all 2240 respondents there is no subset with the value Jd below 0.3995287. Analogously, only one of 1000 subsets has the value Jw below 0.4555840. Thus the p-values of the two tests (Monte Carlo approximations) are p=0.000 and p=0.001, respectively. Fig. 4 shows the probability distribution of Jw. In this figure we also placed the actual value of the test statistic Jw for subset no. 13. The results mean that subset no. 13 is significantly atypical (marked with **). In Table 1 we can also see a few other subsets (marked with *) which seem to be atypical, but not as much as no. 13.

Summing up, the analysis leads to the conclusion that at least subset no. 13 and to a lesser extent also a few other subsets are significantly different from the main bulk of data. On the basis of available information it is difficult to identify the source of these differences. Maybe some of the pollsters were assigned to atypical districts or regions. There is also a possibility that some of the pollsters did not question the respondents in an honest way. Further investigation confirmed that some of the pollsters were not adequately trained.

Table 1. Results of the test for 18 marked out subsets of data. Subsequent columns of the table indicate: ordinal number of the subset, number of respondents in this subset, indices *Jd* and *Jw*, p-values of the tests based on *Jd* and *Jw*, respectively.

NP	n_1	*Jd*	*Jw*	p-value (*Jd*)	p-value (*Jw*)
1	127	0.4892	0.4968	0.147	0.273
2	141	0.4837	0.4953	0.332	0.083
3	125	0.5067	0.4978	0.843	0.659
4	115	0.4782	0.5038	0.728	0.713
5	156	0.4379	0.4889	0.006 (*)	0.014 (*)
6	155	0.4446	0.4838	0.005 (*)	0.008 (*)
8	100	0.4706	0.4882	0.365	0.049 (*)
9	95	0.4194	0.4792	0.020 (*)	0.028 (*)
10	70	0.4465	0.4611	0.018 (*)	0.001 (*)
11	133	0.4399	0.4862	0.009 (*)	0.060
12	161	0.4517	0.4891	0.001 (*)	0.042 (*)
13	**82**	**0.3995**	**0.4556**	**0.001 (*)**	**0.000 (**)**
14	64	0.4099	0.4781	0.129	0.485
15	106	0.4081	0.4767	0.003 (*)	0.009 (*)
16	200	0.4914	0.4917	0.373	0.177
18	85	0.4604	0.4947	0.038	0.119
20	175	0.4425	0.4918	0.219	0.032
21	118	0.4202	0.4691	0.005 (*)	0.001 (*)

5.3. Application of the test of Bartoszyński et al. to ECAP data

We compared our method with another multivariate test of homogeneity, and selected for this purpose the test due to Bartoszyński, Pearl and Lawrence (1997), BPL further on, mainly because of its conceptual simplicity and beauty. In principle, it was designed as a goodness–of–fit test, but a minor modification allows us to use it as a test of homogeneity. Below we describe the basic idea of the version of the BPL, test which we applied. As before, we assume that the data consist of n points in d-dimensional space, with n_1 points belonging to the marked out subset. We consider all triangles with two vertices in this subset (the side joining these points we call the *base*) and the third vertex belonging to the set of the remaining $n_2 = n - n_1$ points. Altogether we have $N = n_2 n_1(n_1 - 1)/2$ such triangles. We count the triangles of three types: those in which the base is the shortest of the three sides, of intermediate length, and the longest. Under the null hypothesis there is approximately $N/3$ triangles of every type. The chi-square statistic, based on the counts of triangles has an asymptotic exponential distribution. Unfortunately, in general, the scale parameter depends on the underlying distribution of data. We estimate this parameter empirically, using Monte Carlo bootstrap-like experiments in much the same way as we did

for our method, see the previous section. The results of the analysis are given in Table 2.

Table 2. Results of the BPL test applied to the ECAP data. Three columns indicate the number of the pollster, the value of test statistic and the p-value

NP	chi2	p-value
1	33879.44	0.2419
2	597.90	0.9779
3	13423.02	0.5644
4	1110.76	0.9495
5	17461.74	0.5557
6	177375.27	0.0025 (*)
8	854.09	0.9547
9	1069.36	0.9405
10	31445.05	0.0790
11	34975.81	0.2480
12	56285.84	0.1601
13	18582.02	0.2851
14	16096.25	0.2369
15	71278.53	0.0265 (*)
16	139123.65	0.0272 (*)
18	19344.23	0.2850
20	19982.50	0.5512
21	52728.29	0.0916

We can see that the BPL test detects three atypical subsets at the level of significance 0.05, namely subsets numbered 6, 15 and 16. The results seem to be less decisive than those of our test. However, subsets number 6 and 15 are selected by both methods.

Let us mention that the computational complexity of the BPL test is high. Counting triangles is very time consuming. Our test, based on QDF turned out to be much faster.

5.4. Additional analysis of Iris data

To enable evaluation of our method, we applied our test as well as the BPL test also to the classical IRIS data set available in R. We chose this well-known set of data despite the fact that it is suitable rather for using discriminant analysis than for detecting outlying subsets. However, we can use these data to examine our methodology of testing homogeneity and computing p-values using Monte Carlo bootstrap-like experiments. We show the results in Table 3 and Fig. 9 in the analogous way as in Table 1 and Fig. 8. The histograms present the empirical distributions of our measures J and Jd for a subset (of 50 objects) randomly chosen from the IRIS set (of 150 objects). The arrows indicate the

Table 3. Results of our test for IRIS data. Three columns of the table indicate the name the of subset and the indices Jw and J. The p-values are practically zero.

Species	Jw	J
Iris setosa	0.000028	0.000
Iris versicolor	0.100752	0.055
Iris virginica	0.055071	0.035

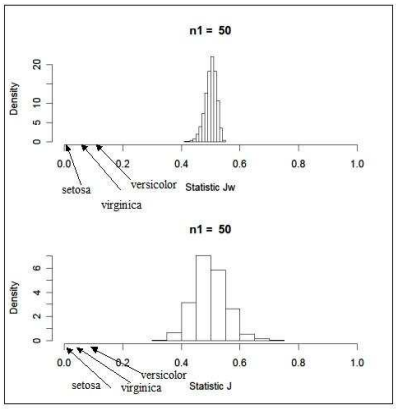

Figure 9. Empirical probability distributions of indices Jw and J under the null hypothesis for IRIS data set with n=150, n_1=50, d=4. The actual values of Jw and J for subsets corresponding to the three species are indicated by arrows.

values for the three species. All p-values are practically zero, as one should expect.

For comparison, we applied also the BPL test to the Iris data. Values of the chi-square statistic are shown in Table 4. All three species are perfectly separated. For this data set the results of our method and of the BPL test are very similar. This is hardly surprising, since the IRIS set is a well-known example of easily separated data.

Table 4. Results of the BPL test for IRIS data. Two columns of the table indicate the name of subset and *chi2* statistic. The p-values are practically zero.

Species	*chi2*
Iris setosa	244964.0
Iris versicolor	144664.4
Iris virginica	128913.4

6. Concluding remarks

1. The method presented in this paper can be used to identify atypical subsets of data in various medical and other applications. It is particularly useful when we deal with large data sets. Our test is much faster than that of Bartoszyński, Pearl and Lawrence and seems to be more sensitive.

2. Apart from simulation studies, which confirmed the usefulness of the proposed algorithm, our method was successfully applied to real medical data collected in a big epidemiological programme. In some cases the p-values are very small, what indicates high significance of the results. Further investigation revealed the fact consistent with the results of our analysis: some of the pollsters had not been enough competent.

3. The general idea of the separability measure, defined in this paper, can be easily adapted to discrimination methods other than QDF. In this way we can obtain a flexible tool for evaluating atypicality of subsets of data.

References

Barnett, V. and Toby, L. (1994) *Outliers in statistical data*, 3rd ed. Wiley.

Bartoszyński, R., Pearl, D.K. and Lawrence, J. (1997) A Multidimensional Goodness-of-Fit Test Based on Interpoint Distances. *Journal of the American Statistical Association* **92**, 577-586.

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: The Approach based on Influence Functions*. John Wiley, New York.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press, London.

Morrison D.F. (1967) *Multivariate Statistical* Methods. Mc Graw Hill, New York.

Koronacki J. (2005) *Statystyczne systemy uczące się*. WNT, Warszawa.

Lachenbruch, P.A. (1967) An Almost Unbiased Method of Obtaining Confidence Intervals for the Probability of Misclassification in Discriminant Analysis. *Biometrics* **23**, 639-645.

Lachenbruch, P.A. (1975) *Discriminant Analysis*. Hafner, New York.

Lachenbruch, P.A. and Mickey, M.R. (1968) Estimation of Error Rates in Discriminant Analysis. *Technometrics* **10**, 1-11.

Renze, John (no date) "Outlier." From MathWorld - A Wolfram Web Resource, created by Eric W. Weisstein.
http://mathworld.wolfram.com/Outlier.html

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.

VENABLES, W.N. and RIPLEY, B.D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.

WATAŁA, C. (2002) Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych. α-medica press, Bielsko-Biała.