A Method for Detecting Discontinuous Probability Density Function from Data

Ioan V. Lemeni *

**University of Craiova, Craiova, Romania (Tel: 0251-435666; e-mail: ioan.lemeni@ cs.ucv.ro).*

Abstract: In real classification applications the patterns of different classes often overlap. In this situation the most appropriate classifier is the one whose outputs represent the class conditional probabilities. These probabilities are calculated in traditional statistics indirectly, in two steps: first the underlying prior probabilities are estimated and then the Bayes rule is applied. Popular methods for density estimation are Parzen Window and Gaussian Mixture. It is also possible to calculate directly the class conditional probabilities using the logistic regression, k-Nearest Neighbours algorithm or a Multilayer Perceptron Artificial Neural Network. Many methods, direct or indirect, perform poorly when the underlying prior probability densities are discontinuous along the support's border. This paper will present a method for detecting the discontinuity by analyzing samples drawn from the underlying density. Knowing the densities are discontinuous will help to choose an estimator insensitive to discontinuities.

Keywords: probability density function estimation, class conditional probability estimation

1. INTRODUCTION

In statistical pattern recognition, classification means assigning an object or a fact to a predefined class. The object or fact is represented by a subset of its attributes, say d. Supposing that the attributes are numerical or can be converted to numbers, each object's representation becomes a point, or a vector, in R^d. The classifier has to distinguish between classes trying to isolate homogenous regions. A region is considered homogenous if it contains vectors belonging to one class only.

Unfortunately, homogenous regions are rare; most of the times there is a degree of overlapping between them. This degree of overlapping is due to the fact that some essential attributes were not recorded. Financial data sets are well known for their high degree of overlapping. If the data set exhibits such a degree of overlapping, the best classifier is the one whose outputs represent a posteriori conditional probabilities or, in other words, the class conditional probability.

There are two approaches to calculate the class conditional probabilities. According to the first, these probabilities can be calculated using the well known Bayes rule from the statistics field:

$$P(\omega_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_k)P(\omega_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \omega_k)P(\omega_k)}{\sum_i [p(\mathbf{x} \mid \omega_i)P(\omega_i)]} \quad (1)$$

In (1), $P(\omega_k)$ is the probability of class ω_k, $p(\mathbf{x})$ is the probability density function (pdf) of the feature vector $\mathbf{x}$ and $p(\mathbf{x}|\omega_k)$ is the conditional probability density of $\mathbf{x}$ in class ω_k, known a priori. If we choose to follow the traditional statistical approach in order to calculate the class conditional probability, first we have to estimate the probability density of $\mathbf{x}$ in every class. Two well known methods used to estimate the probability density function from data are Parzen window and Gaussian Mixture Model (GMM).

The second approach relays on a special form of the error function. Ruck et al. (1990) showed that the outputs of a Multilayer Perceptron (MLP), when trained as a classifier using backpropagation, approximate the posterior conditional probabilities. This finding is not specific to MLP, but it holds true for any classifier that uses the sum-of-squares or the likelihood as error function. In this category also fall the logistic regression and k-Nearest Neighbors algorithm.

The performance of many classifiers deteriorates when the underlying prior probability densities $p(\mathbf{x}|\omega_k)$ are not fully supported in R^d and, additionally, these densities are not continuous along the support's frontier. For example, the standard uniform distribution is supported in the interval [0, 1] only and it is discontinuous in 0 and 1; the exponential distribution's support is [0, ∞), being discontinuous in 0.

The discontinuity does not affect the classifiers in the same degree: some classifiers are more affected than the others but there are classifiers that are not affected at all. If the closed form of the prior densities are known, it is easy to say if they are discontinuous or not and choose a classifier less sensitive. But in real life situations, the densities must be estimated from input data.

Consequently, the closed form is unknown, so we cannot say if they are discontinuous or not.

The remainder of this paper is organized as follows. Section 2 presents the effect of the discontinuity on density estimation and Section 3 presents a method that detects the discontinuity analyzing samples drawn from that discontinuous density.

2. THE EFFECT OF THE DISCONTINUITY

One widely used method for density estimation is the Parzen window. Let $\boldsymbol{x}_1$, $\boldsymbol{x}_2$,…., $\boldsymbol{x}_n$ be n independent and identical distributed samples drawn from some distribution with an unknown density f. Density f can be estimated as

$$\hat{f}_h = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i) \tag{2}$$

where $K(\cdot)$ is the kernel and $h>0$ is the smoothing parameter or bandwidth. This kind of estimation is called kernel density estimator or Parzen window. The kernel is a symmetric function that integrates to one. A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov and normal.

However, the estimator does not take into account the potential finite support of the feature vector **x**. When the support of some of its components is bounded, for example, in the case of nonnegative data, the standard kernel estimator gives weight outside the support. This causes a bias in the boundary region. In order to outline the Parzen window behaviour for discontinuous density, we will consider the standard exponential density. We drawn 1000 points and then we estimated the density using the Parzen window. The result is shown in Fig. 1.

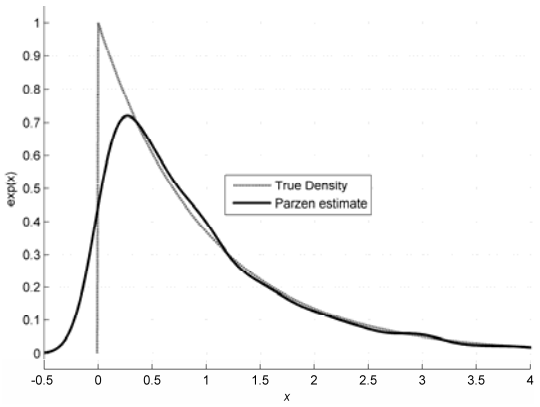

Fig. 1 Exponential density estimated from 1000 input vectors by a Parzen estimator.

The boundary bias problem of the standard kernel is well documented in the one-dimensional case, many solutions being proposed. In the one dimensional case the random variable is either positive, being discontinuous in zero, or is restricted to (0, 1) interval, being discontinuous in 0 and 1. An initial solution to the boundary problem is given by Schuster (1985), who proposes to generate data outside support by reflection. This solution is simple and easy to implement but it only works when the first derivative of the generator density is zero to the right or to the left of the discontinuity. As this requirement is rarely fulfilled, Cowling and Hall (1996) also generated new data outside the support but their location is determined now by interpolation, not by reflection. A different solution proposed by Marron and Ruppert (1994), consists in transforming the input data such as the discontinuity or discontinuities disappear. The function g(.) transforms the data such as g(0) and possibly g(1) are zero. Müller (1991) and many others suggest the use of adaptive boundary kernels at the edges and a fixed standard kernel in the interior region. More recently, Bouezmarni and Rombouts (2006) study the gamma kernels for univariate nonnegative data.

The boundary bias problem becomes more severe in the multivariate case because the boundary region increases with the dimension of the support and become more complex. In the one-dimensional case the frontier is a point and can be easily detected but in the multidimensional case it becomes a hypersuface. In this case any solution presented so far can be adapted only if the closed form of the boundary is known. In real life this is rarely the case, so Parzen estimator must be avoided if data are bounded and the boundary is not known.

Another popular method in probability density estimation is the Gaussian Mixture Model (GMM). The probability density function of the observed data is represented as a superposition of m multivariate Gaussian pdfs

$$p_{GM}(\boldsymbol{x}) = \sum_{j=1}^{m} \rho_j N(\boldsymbol{x}, \boldsymbol{\mu}_j, \mathbf{C}_j) \tag{3}$$

where $\boldsymbol{x}$ is a d-dimensional observation from the data set, ρ_j (j=1,…,m), are the component weights, summing up to unity, $\sum \rho_j = 1$ and $N(\boldsymbol{x}, \boldsymbol{\mu}_j, \mathbf{C}_j)$ is a multivariate normal density with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\mathbf{C}_j$. The negative log-likelihood of a data set made of n observations, given by

$$E = -\ln \mathrm{L} = -\sum_{i=1}^{n} \ln p(\boldsymbol{x}_i) = -\sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{m} \rho_j N(\boldsymbol{x}_i, \boldsymbol{\mu}_j, \boldsymbol{C}_j) \right\} \tag{4}$$

can be used as an error function. Its minimization through the expectation-maximization (EM) algorithm offers the optimal values for the GMM's parameters ρ_j, $\boldsymbol{\mu}_j$ and $\mathbf{C}_j$. A review of mixture models, maximum likelihood and EM algorithm has been given by Redner (1984).

GMM gives very good results if the densities to be estimated are continuous. It can be shown that any continuous pdf can be approximated arbitrarily closely by a Gaussian mixture density. But little work has been done in the field of discontinuous densities, being only two papers that address this problem.

Hedelin and Skoglund (2000), developing a method for high-rate vector quantization in speech coding, found that GMM performance is degraded when the density to be estimated has bounded support. In order to outline the GMM behaviour for discontinuous density, we'll consider the example used in Parzen window. Same 1000 points

were used to estimate the exponential density using a GM with 5 components. The result is shown in Fig. 2.

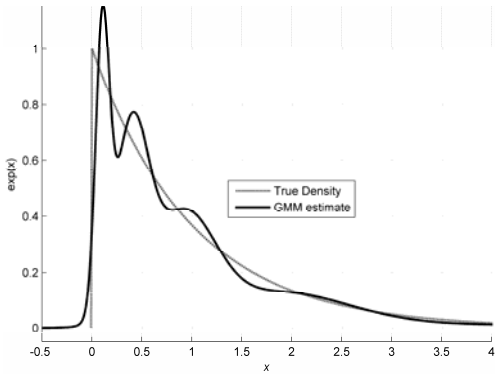

Fig. 2. Exponential density estimated from 1000 input vectors by a GMM with 5 components.

Hedelin and Skoglund proposed a version of the EM algorithm for such densities and named the new algorithm EMBS. Although the EMBS algorithm provides an interesting solution to the bounded support issue, it cannot be used when the analytical form of the support is not known.

Another author that emphasizes the poor performance of GMM for discontinuous densities is Likas. In Likas (2001) the author developed a method for pdf estimation based on a multilayer perceptron neural network and compared it against the traditional GMM. The densities estimated exhibit discontinuities that affect the performance of GMM.

A detailed analysis of GMM versus MLP performance for discontinuous densities can be found in Lemeni (2009). The paper outline the oscillating behaviour of GMM due to the discontinuity through one and two-dimensional examples and the good performance of MLP in the same examples. The author claims that MLP is superior to GMM for class conditional probability estimation when the underlying priors are discontinuous.

Discontinuous densities by definition, such as the exponential and the uniform densities, as well as truncated densities were used to asses the performance of two other methods: logistic regression and k-Nearest Neighbours (kNN). The results sowed that the logistic regression is not affected by the discontinuity but kNN is. Because some estimators are severely affected in the vicinity of the discontinuity, we need a method that is able to analyse samples drawn from an unknown density and tell if that density is discontinuous or not.

3. THE MASS CENTER METHOD FOR DISCONTINUITY DETECTION

In many real life situations, the densities to be estimated from input data originate from fully supported densities which have been truncated. Such a situation is common in the financial field.

Financial data are constrained, those constrains being either an effect of a law or specific to a particular company. For instance, quantities such as height or weight of a person are normally distributed with no restriction. On the other hand, other quantities such as the gross salary, must obey a legal constraint, namely the minimum living wage.

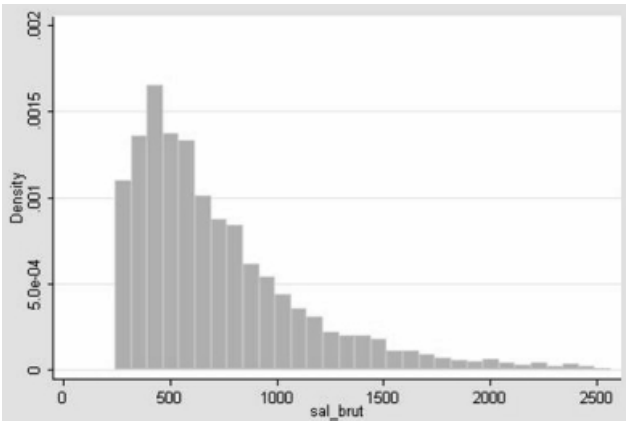

Fig. 3. Minimum living wage histogram in Romania, year 2005.

Fig. 3 presents the histogram of the minimum living wage in Romania, year 2005, taken from a study of the Group of Applied Economics (2008) on the impact of the flat income tax. The distribution's shape seems to be normal, except the missing left side of the graphic. The truncation is the effect of the minimum living wage law and makes the distribution discontinuous for this value. The minimum living wage was 330 RON in 2005.

Banks often use such quantities, e.g. gross salary and age, for classification and regression. If a bank accepted all the applications for a credit with no restriction, salary and age histograms would have the shape in Fig. 3, because there cannot be salaries less than the minimum living wage or clients younger than 18 years.

Generally, if the histogram of an attribute describing a real object, fact or event has the shape from Fig. 3, this is a clear indication of a discontinuous underlying density.

Constrains considered before are generated by the current legislation and the financial institutions must obey them. But there are a second type of constrains, this time self-imposed. For example, a bank trying to lend money only to non default customers, will accept an application if a certain criterion is met. Such a criterion is the credit score of the requester. Most of the time the credit score is calculated with Fisher's linear discriminant or one of its variants. The credit-score criterion will act as a knife, cutting the data space, so there will be records only for borrowers with a credit-score greater than a certain threshold. In the attribute space the credit-score criterion will act as a separating hypersurface: on one side there are no vectors, many vectors being on the other side in the vicinity of that hypersurface. Moreover, in real applications, the analytical form of the separating hypersurface may be not known. In many cases we do not even realize that the support has a border and the pdf to be estimated is discontinuous along this border. As it will be shown soon, the attributes' histograms are useless in situations like these.

As before, let's suppose that potential customers of a bank are described by two attributes, *salary* and *age*. After normalization they are Gaussian distributed with mean $\boldsymbol{\mu}$=(1, 1) and covariance matrix $\boldsymbol{\Sigma}$=**I**. Let's again suppose that the bank's credit-score is calculated as *nsalary*+*nage*≥2, where the leading *n* stands for "normalized". This criterion creates the distribution of the accepted customers. Its support consists of all the points in $\mathbf{R}^2$ for witch the score criterion is fulfilled. Thus the distribution of accepted customers is obtained from the distribution of potential customers by truncations along the line of equation *nsalary*+*nage*=2. This line is the support's frontier. In Fig. 4, 4000 random vectors were drawn from the distribution of the potential customers and only those fulfilling the score criterion were plotted.

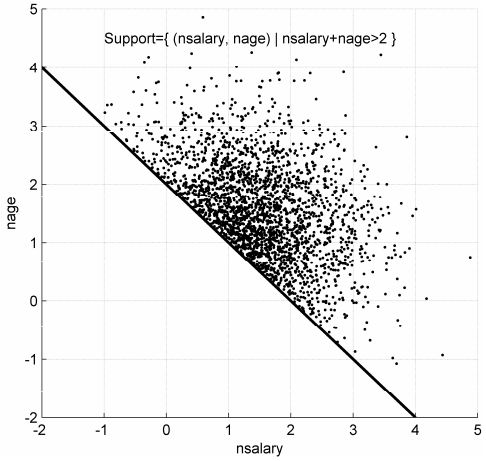

Fig. 4. Example of truncated data.

The data layout is typical for a financial institution that applied a score criterion. Hereinafter we want de determine how this type of truncation reflects in the attributes' histograms.

In Fig. 5 is presented the histogram of *nsalary* attribute after truncation.

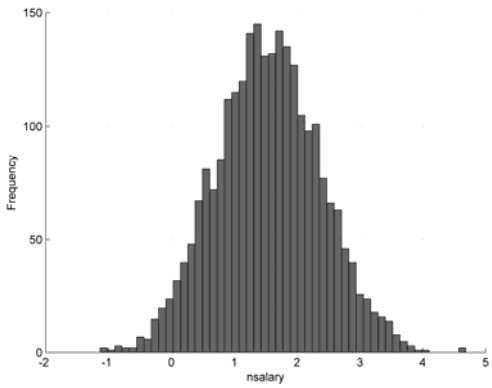

Fig. 5. Histogram of the *nsalary* attribute after truncation.

The histogram of *nage* has the same shape and has not been presented. The histogram rather suggests a normal distribution and doesn't offer any indication on the discontinuity at the support's border. For the two-dimensional example presented so far is easy to detect the discontinuity if we build a scatter plot as the one presented in Fig. 4. Unfortunately this approach is almost impossible for three dimensions and impossible for four dimensions or more.

For more than two dimensions the discontinuity can be detected if we analyze the spatial distribution of *k* neighbors of an input vector. The analysis must be conducted for all available input vectors. All those *k* neighbors lie in a *d*-sphere centered at the input vector with radius given by the distance between that input vector and the farthest neighbor. If the number *k* of neighbors is chosen so the generator density is constant in the vicinity of the input vector under analysis and every neighbor is seen as a small particle with one unit mass, then the mass center of all neighbors will overlap the geometric center of the *d*-sphere surrounding them. This situation is depicted in Fig. 6a.

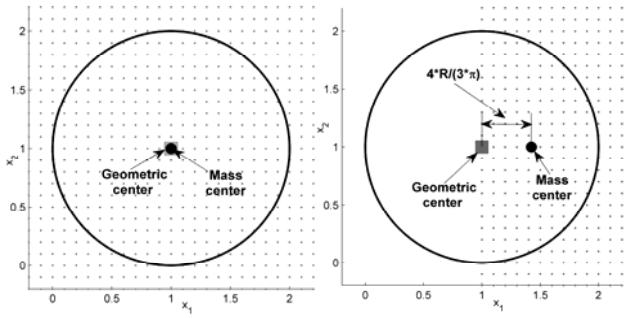

Fig. 6. Position of the neighbors' mass center relative to their geometric centre.

On the other hand, if the input vector under analysis lies exactly on the support's border and, additionally, the support's border can be approximated by a hyper plane, then the neighbors will form a homogenous hemi *d*-sphere with radius *R*. The position of the hemi *d*-sphere's mass center will no longer coincide with the geometric center. For *d*=2 the mass center lies at $4R/(3\pi)$ units apart from the geometric center and for *d*=3, at $3R/8$. Fig. 6b presents the position of the mass center for *d*=2.

In real conditions the neighbors' density is not perfect uniform, so a 5% deviation of mass center's position from the position corresponding to the uniform distribution is accepted.

Considering the example from Fig. 4, the input vectors with the mass center placed $4R/3\pi$ ±5% apart from the geometric center are represented as black squares in Fig. 7. The radius *R* is chosen such as the *d*-sphere (circle for *d*=2) of the input vector under analysis contains 40 neighbors. The optimal number of neighbors will be discussed later.

Analyzing the distribution of the black squares in figure, we notice two input vector categories: input vectors lying on the support's frontier, such as **v**1, and vectors lying in low density zones, such as **v**2. The mass center criterion indicates correctly the "on border" position for the vectors belonging to the first category and fails for the second. The category can be decided by the radius of the neighbors' circle. An analysis of the radius of the

neighbors' circle indicates a small radius for **v**1 and a big one for **v**2: **v**2's radius is almost five times bigger than **v**1's in figure. But a five time bigger radius means a 25 times lower density in **v**2's circle compared to the density in **v**1's circle. Because the area of the circle corresponding to **v**2 is too big, the underlying pdf is no longer constant in this zone. This is why the mass center criterion fails for such input vectors. In order to make this criterion work, the vectors lying in low density zones must not be analyzed.

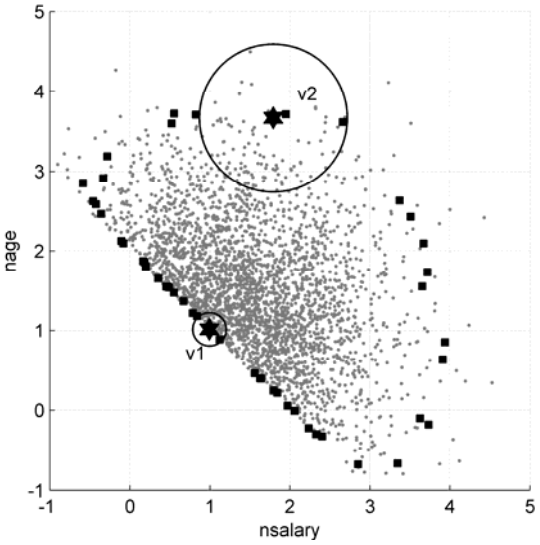

Fig. 7. Input vectors (black squares) lying on the support's frontier.

The vectors lying in low density zones can be filtered out based on the radius of the neighbors' *d*-sphere. Let's suppose that the *d*-sphere contains *k* neighbors. First we calculate the radius corresponding to the largest density. As density and radius are inversely proportional, the highest density zone contains the vectors with the smallest associated radii. In order to eliminate an exceptional small radius, due to an exceptional distribution, we sort the radii of the *d*-spheres which encompass *k* neighbors in ascending order and then we chose the tenth radius as the minimum radius. This value has been chosen after many simulations for different pdfs and vector numbers has been carried out.

Let ρ_{max} denote the largest density and R_{min} the corresponding radius. Then $\rho_{\max} = \frac{k}{V_{d-sfera}} = \frac{k}{pR_{\min}^d}$, where *d* represents the number of dimensions and *p* is a constant depending on *d*. For *d*=2 the *d*-sphere is a circle and its "volume" is πR^2, so $p = \pi$. In 3 dimension the *d*-sphere is a ball and its volume is $4/3\pi R^3$, so $p = 4/3\pi$. If we analyze the position of the mass center only for vectors lying in zones with density at least ρ_{max}/n, a threshold radius can be expressed as follows:

$$\rho_{th} = \frac{k}{pR_{th}^d} = \frac{\rho_{\max}}{n} = \frac{k}{npR_{\min}^d} \Rightarrow pR_{th}^d = npR_{\min}^d \Rightarrow R_{th} = \sqrt[d]{n}R_{\min}.$$

For *n*=3 the threshold radius is $R_{th} = \sqrt{3}R_{\min}$. This value has been chosen after many simulations for different pdfs and vector numbers. The filtering based on R_{th} was applied to the vectors from Fig. 4, only the vectors lying in zones with density greater than ρ_{th} being analyzed. From the selected vectors, those with the mass center placed $4R/3\pi \pm 5\%$ apart from the geometric center are represented as black squares in Fig. 8.

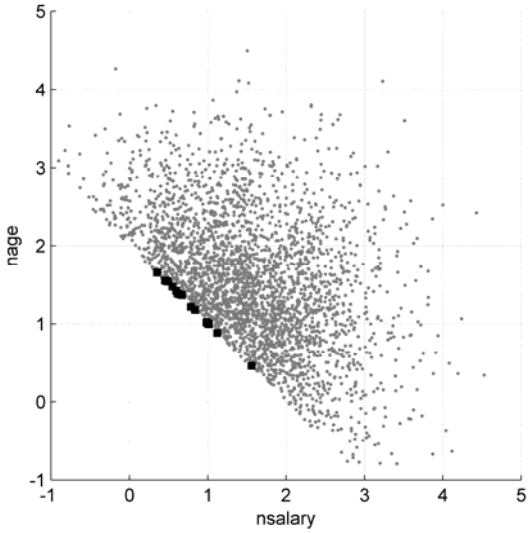

Fig. 8. Input vectors (black squares) from high density zones lying on the support's frontier.

Comparing Fig. 7 and Fig. 8 we notice that the vectors from low density zones, previously wrongly indicated as lying on border, were eliminated.

The last parameter we have to discuss is the number of neighbors *k*. If this number is too small, the estimation of the mass center's position will be erroneous. For example, considering only one neighbor, we can obtain any mass center deviation between 0 and 1. A small number of neighbors cannot correctly describe the underlying density and will result in a large number of erroneous "on border" indications. On the other side a large number of neighbors means a large corresponding volume where the vectors' density is no longer constant. This situation will result again in a large number of erroneous "on border" indications. Concluding, the erroneous indication will be obtained for a small number as well as for a large number of neighbors. We can guess that there is an optimal number of neighbors for witch the "on border" indication is correct. For a larger or a smaller number of neighbors the mass center method will indicate more vectors lying on border. This is why the graph of the number of vectors lying on border as a function of the number of neighbors should have a V shape. In order to test this hypothesis, we plot this graph for the example in Fig. 4. The number of neighbors varies between 10 and 200 with a step of 10. The resulting graph is presented in Fig. 9a. Figure also plots same function for the original (not truncated) density $N(\boldsymbol{\mu} = (1,1), \boldsymbol{\Sigma} = \mathbf{I})$. Regardless the number of neighbors *k*, the non zero number of "on border" vectors and the V-shape graph for the truncated data set confirm the above hypothesis. Furthermore, we notice that the number of "on border" vectors for the original, non truncated, data set is zero for a large number of *k* values.

Now, the graph of the number of vectors lying on border as a function of the number of neighbors has an L shape.

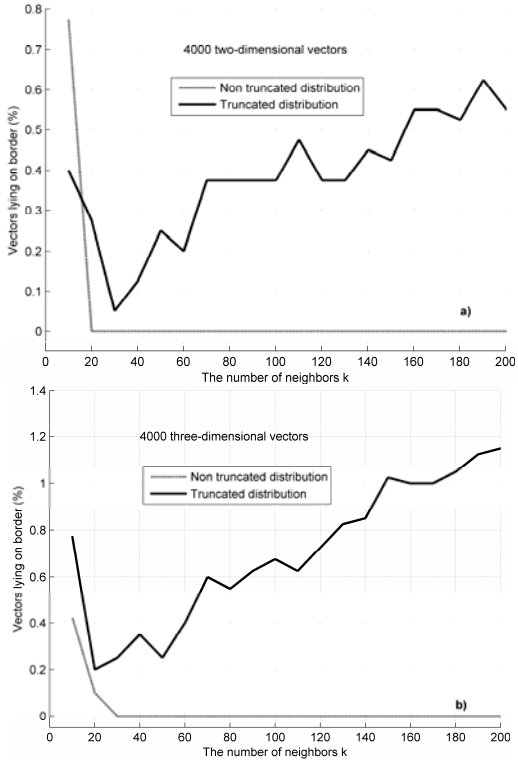

Fig. 9 Vectors lying on border as a function of the number of neighbors.

Similar results have been obtained for three dimensional data. The generator density used in this example was $N(\boldsymbol{\mu} = (1,1,1), \boldsymbol{\Sigma} = \mathbf{I})$ and the truncation criterion was $x_1 + x_2 + x_3 > 3$. The graph of the number of vectors lying on border as a function of the number of neighbors for both the truncated and non truncated densities are presented in Fig. 9b. We notice a remarkable similarity with the graph corresponding to the two dimensional case.

The next test has been carried out on real data originating from the credit data base of CEC Bank.

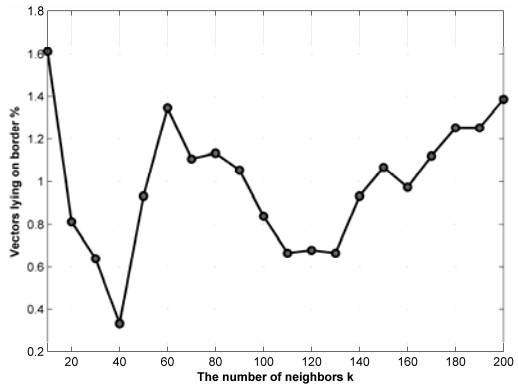

Fig. 10 Vectors lying on border as a function of the number of neighbors for CEC Bank credit records.

The records from data base were preprocessed and then used to create a data set consisting of 7502 four-dimensional vectors. This data set was tested for discontinuity by the mass center method. The graph of the number of vectors lying on border as a function of the number of neighbors is shown in Fig. 10. The graph has a V shape and the number of "on border" vectors is never zero. These two characteristics are a clear indication that the CEC data set is discontinuous.

4. CONCLUSIONS

Many data sets originating especially from financial field are truncated due to some legal restrictions or a self impose criterion. The discontinuity along the support's border affects the performance of some estimators such as Parzen window, GMM or kNN while other estimators such as MLP or logistic regression are less affected.

This paper will present the mass center method for detecting the discontinuity. Knowing the densities are discontinuous will help to choose an estimator insensitive to discontinuities. The performance of the mass center method was tested in many runs on artificial data drawn from different densities, with different dimensions, truncation criteria and size. The method was always able to indicate the presence of the discontinuity.

REFERENCES

Bouezmarni, T., and Rombouts, J. (2006). "Nonparametric Density Estimation for Positive Time Series," CORE discusion paper 2006/85.

Group of Applied Economics (2008). Available on line at: http://www.gea.org.ro/documente/ro/studii/cota_unica/cota_unica_ppt.ppt

Hedelin, P. and Skoglund, J. (2000). "Vector quantization based on Gaussian mixture models," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 4, pp. 385-401.

Lemeni, I. (2009) "Multilayer Perceptron versus Gaussian Mixture for Class Probability Estimation with Discontinuous Underlying Prior Densities" in *Proc. IEEE Computing in the Global Information Technology,* ICCGI '09, pp. 240-245

Likas, A. (2001) "Probability Density Estimation Using Artificial Neural Networks," *Computer Physics Communications*, vol. 135, no. 2, pp. 167-175.

Marron, J., and Ruppert, D. (1994). "Transformations to Reduce Boundary Bias in Kernel Density Estimation," *Journal of the Royal Statistical Society*, Series B, 56, 653–671.

Müller, H. (1991). "Smooth Optimum Kernel Estimators near Endpoints," *Biometrika*, 78, 521–530.

Redner, R. A. and Walker, H. F. (1984). "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239

Ruck, D. W. , Rogers, S. K. and Other (1990). "The multilayer perceptron as an approximation to a Bayesoptimal discriminant function," *IEEE Trans. Neural Networks,* vol. 1, pp. 296–298.

Schuster, E. (1985): "Incorporating Support Constraints into Nonparametric Estimators of Densities," *Communications in Statistics - Theory and Methods*, 14, 1123–1136.