Accuracy of Nonparametric Density Estimation for Univariate Gaussian Mixture Models: a Comparative Study

Jurgita Arnastauskaitė[a,b] and Tomas Ruzgas[a]

[a]*Department of Applied Mathematics, Kaunas University of Technology*
  K. Donelaičio g. 73, 44249 Kaunas, Lithuania
[b]*Department of Computer Sciences, Kaunas University of Technology*
  K. Donelaičio g. 73, 44249 Kaunas, Lithuania
E-mail(*corresp.*): `jurgita.arnastauskaite@ktu.lt`
E-mail: `tomas.ruzgas@ktu.lt`

Abstract. Flexible and reliable probability density estimation is fundamental in unsupervised learning and classification. Finite Gaussian mixture models are commonly used for this purpose. However, the parametric form of the distribution is not always known. In this case, non-parametric density estimation methods are used. Usually, these methods become computationally demanding as the number of components increases. In this paper, a comparative study of accuracy of some nonparametric density estimators is made by means of simulation. The following approaches have been considered: an adaptive bandwidth kernel estimator, a projection pursuit estimator, a logspline estimator, and a *k*-nearest neighbor estimator. It was concluded that data clustering as a pre-processing step improves the estimation of mixture densities. However, in case data does not have clearly defined clusters, the pre-preprocessing step does not give that much of advantage. The application of density estimators is illustrated using municipal solid waste data collected in Kaunas (Lithuania). The data distribution is similar (i.e., with kurtotic unimodal density) to the benchmark distribution introduced by Marron and Wand. Based on the homogeneity tests it can be concluded that distributions of the municipal solid waste fractions in Kutaisi (Georgia), Saint-Petersburg (Russia), and Boryspil (Ukraine) are statistically indifferent compared to the distribution of waste fractions in Kaunas. The distribution of waste data collected in Kaunas (Lithuania) follows the general observations introduced by Marron and Wand (i.e., has one mode and certain kurtosis).

Keywords: univariate probability density, nonparametric density estimation, homogeneity test, sample clustering, Monte Carlo method, municipal solid waste.

AMS Subject Classification: 62G07.

1 Introduction

The problem under consideration is closely related to distribution analysis. Which is an important branch of data analysis and is being used to solve various other problems (discriminant analysis, image recognition, etc.). The methodology for estimating distributions is receiving increasing attention due to emerging applications: processing of genetic information, analysis of astronomical objects, research of computer equipment and its peripheral data, etc. Even though there are already many methods for estimating data distributions, new advancements are being proposed by various authors.

In case data distribution is multimodal and the sample size is small, in practice, it is not easy to choose a robust density estimation method. Kernel estimators are the most popular method for density estimation (see, for example [19, 26, 33]). Other estimators are frequently used by researchers (see in [6, 11, 18, 22]). As big data become commonplace in almost every field of data science lately, various approaches to estimate density are being investigated, including machine learning algorithms. Academia and industry focus on the development of innovative density estimation procedures (see in [25, 30]). In some cases the accuracy of the assessment can be significantly increased (according to [32]), if the observations are clustered at first (i.e., treating the multimodal density as a mixture of unimodal densities), and the popular nonparametric estimators are applied to each cluster separately.

The present work extended the pilot research presented in [32]. The following new research areas are investigated: 1) new density estimation methods based on inverse formula are formulated, 2) in order to compare the robustness of estimators the wide set of distributions proposed by Marron and Wand [26] is used to study densities, 3) in order to obtain results with a reasonably high level of confidence a relatively high (100000) number of independent samples have been generated.

The primary purpose of this paper is to assess the performance of several density estimators, by using benchmark densities [26] that represent different types of problems that can arise for unimodal and multimodal densities. The relationship between the estimation accuracy and complex structures in case of univariate Gaussian mixture models (GMM) are analysed by Monte Carlo method. Applications of GMM distributions are popular and are used in various scientific fields for examination of important problems. For example, Younghong [14] in medical studies for limb mobility classification, (using mioelectrical signals) compared GMM with three other frequently used classification methods of linear discriminant analysis, linear multilayer perceptrons network and multilayer neural network. GMM demonstrated the exceptional classification accuracy in [14]. Henthorn [16] examined the ability to hydroxylate (the degree of urinal metabolism) debrisoquin and O-dimethilath dekstrometorfan with the help of GMM. Zimmerman [37] used the GMM for EKG analysis using ST segments and T waves for determining the ischemia. Kovacs [23] applied it on simulations of molecular dynamics of membrane pore formation. Stachniss [35] applied GMM in geological investigations for gas concentration prediction based on carried mining. Gruszczyński [13] used ap-

proximation by GMM in ecological research on the soil contamination. Chernova and Veloso [4] used the learning results of interactive policy formation that learners demonstrate later, the results are described by GMM in the field of applied sociology.

In this work, after the clustering pre-processing step, the mixture components were identified using: 1) a kernel estimator with the adaptively selected smoothing width, 2) Friedmans proposed algorithm based on the projection pursuit, 3) Kooperberg and Stone log-spline estimator based on the approximation of logarithm of distribution density by the sum of cubic B-splines, 4) Fix and Hodges k-nearest neighbor density estimator, and 5) the proposed inversion formula modification for density estimator, where the structure is given in (3.6). A pilot comparative study of several non-parametric estimators accuracy [32] showed that the Friedman procedure is more robust in the majority of examined Gaussian multivariate mixtures cases where the components can be separated. The kernel estimator is more accurate even with a small sample size. Initial clustering of the sample allows to get less biased estimates of density.

The scientific novelty of this research is the comparative study accuracy of density estimation methods. The performance of various non-parametric estimators has been compared using different data sets. The representative density estimators of various techniques, analysed by other researchers, has been selected. In addition, this study reveals the practical value of the homogeneity tests.

This paper is organized as follows: Section 2 describes the EM algorithm used for sample clustering before density estimation. Section 3 summarizes the five selected density estimators. Section 4 contains the simulation results and states the findings in the context of the municipal solid waste data. The detailed results of the Monte Carlo study can be found in appendixes. Concluding remarks are presented in Section 5.

2 Sample clustering with the EM algorithm

Although previous numerical simulations [32] have shown that initial data clustering is beneficial for evaluating multimodal distribution density, the question on the optimal clustering method is still open. A comparative study in [32] showed that probabilistic clustering methods are more robust compared to geometric clustering methods (e.g., k-means, hierarchical, etc.) applied in conjunction with the transformation of mixture components to spherical [34]. This work is limited by the study of the sample clustering based on considered density approximation by Gaussian mixture distribution.

Let $X(1), ..., X(n)$ be the observed independent and identically distributed random variables with unknown distribution density $f(x)$. If density is multimodal, it can be analysed as a mixture of several unimodal densities:

$$f(x) = \sum_{k=1}^{q} p_k f_k(x). \tag{2.1}$$

Suppose the observed random variable X depends on the random variable v,

taking the values $1, ..., q$ (interpreted as a class number of the observed object). Weights $p_k = \mathbf{P}v = k$ in classification theory are called *a priori* probabilities of observed object belonging to class k and weights $\pi_k(x) = \mathbf{P}v = kjX = x$ are called *a posteriori* probabilities. The function f_k is treated as X conditional distribution density at the condition $v = k$. The assessment of the values $\pi_k X(t)$ for all $k = 1, ..., q$, $t = 1, ..., n$, is called sample soft clustering. In this paper the sample is strictly clustered: it is split into subsets using estimates $\widehat{v}(1), ..., \widehat{v}(n)$, where $v(t)$ denotes the class number of observation $X(t)$.

GMM is a well studied model in model-based clustering theory and practice; therefore it was selected for this research. The application of GMM is based on the assumption that the functions $f_k(x)$ are Gaussian densities with mean parameters μ_k and variances σ_k^2. In this case the right side of equality (2.1) is marked as $f(\theta, x)$, where $\theta = ((p_k, \mu_k, \sigma_k^2), \quad k = 1, \ldots, q)$. Since the equalities are valid

$$\pi_k(x) = \frac{p_k f_k(x)}{f(\theta, x)}, k = \overline{1, q}, \tag{2.2}$$

the *a posteriori* probability estimates can be simply obtained by a *plug-in* method which replaces unknown parameter vector θ on the right-hand side of (2.2) with its maximum likelihood estimate $\theta^* = \arg\max_{\theta} L(\theta)$, $L(\theta) = \prod_{t=1}^{n} f(\theta, X(t))$. The EM algorithm is applied in this work to find the estimate of the parameter vectors.

Given the estimates $\hat{\pi}_k = \hat{\pi}_k^{(r)}$ after r cycles, the new estimate $\hat{\theta} = \hat{\theta}^{(r+1)}$ is obtained by:

$$\hat{p}_k = \frac{1}{n} \sum_{t=1}^{n} \hat{\pi}_k(X(t)), \quad \hat{\mu}_k = \frac{1}{n\hat{p}_k} \sum_{t=1}^{n} \hat{\pi}_k(X(t)) \cdot X(t),$$

$$\hat{\sigma}_k^2 = \frac{1}{n\hat{p}_k} \sum_{t=1}^{n} \hat{\pi}_k(X(t)) \left[X(t) - \hat{\mu}_k\right] \cdot \left[X(t) - \hat{\mu}_k\right]'$$

for all $k = 1, \ldots, q$. Inserting $\hat{\theta}^{(r+1)}$ into (2.2) the right side $\hat{\pi}_k^{(r+1)}(X(t))$, $k = 1, ..., q$, $t = 1, ..., n$ is found. Non-decreasing sequence of likelihood $L(\widehat{\theta}^{(r)})$ is obtained as a result of this iterative procedure. However, the convergence to the global maximum depends on the choice of the initial estimate $\hat{\theta}^{(0)}$ (or $\hat{\pi}^{(0)}$). The simplest strategy is to apply the EM algorithm on a number of randomly generated initial estimates $\hat{\pi}^{(0)}$. The resulting estimate with the highest likelihood function $L(\hat{\theta})$ is taken. The sequential mixture components extraction method described in [7, 27, 31] could be used here as well.

A variety of model adequacy tests can be used for selection of the clusters' number q [28]. Let $f^*(x) = f(\theta^*, x)$, where θ^* is the estimate obtained by a maximum likelihood estimator. Lets us denote the distribution function as $F^* = \int_{-\infty}^{x} f^*(t)dt$, and the empirical distribution function as $\tilde{F}(x) = n^{-1} \sum_{t=1}^{n} 1_{X(t)<x}$. Let $\hat{f}$ be a non-parametric density estimate of f, that we will present in Section 3. Defining (for more details we refer the

reader to [31])

$$\psi = \int \frac{\hat{f}(x)}{f^*(x)} d\tilde{F}(x) - 1 = n^{-1} \sum_{t=1}^{n} \hat{f}(X(t)) \Big/ f^*(X(t)) - 1$$

we do not reject the adequacy hypothesis of model (2.1), if the following condition $\psi < \varepsilon$ is fulfilled. Given significance level α, it is desirable to choose the level ε as close as possible to the quantile u_α : $P\{\psi > u_\alpha\} = \alpha$. The *bootstrap* method (see [3, 15, 24, 36]) can be used for this purpose. Let $G(u)$ denote the distribution function of the statistic ψ, then, with a fixed n, it is completely defined by the distribution function F of the random variable X, i.e., $P\{\psi > u\} = G(u, F)$, for all u. The variable ε is defined by equality $G(\varepsilon, F^*) = 1 - \alpha$ for the given α. Function $G(\cdot, F^*)$ can be obtained by *bootstrap* method.

3 The density estimation algorithms analysed

The comparative study of accuracy of density estimation algorithms has been performed based on four different types of statistics, which are used by other researchers, as well. In this paper the following density estimators for probability distribution density function using Monte Carlo method were examined:

1. Density estimator (PPDE), proposed by Friedman (see [11, 12]), which is based on the target design and projection consistent gausianization;

2. The kernel density estimator (AKDE) adapted for sing different kernel width for various observations by Silverman [33];

3. The logspline distribution density estimator (LSDE), approximating density logarithm of the specimen by the cubic B-spline sum by proposed Kooperberg and Stone [22];

4. The k-nearest neighbor distribution density estimator (NNDE), which is based on the distance to the nearest k observation position evaluation to the tested random dimension and has been examined by Devroye and Krzyżak [9];

5. Modified distribution density estimator (MIDE) with the original version proposed in [20] and analysed in [32] based on application of the inversion formula.

The methods are used on a standardized sample. The examined algorithms will be described in detail.

PPDE algorithm. J.H. Friedman proposed an iterative algorithm of multidimensional density estimation. It is based on consistent search of one dimensional projections [17], where distributions are mostly different from normal. These projections are transformed into Gaussian values. Let X be the standardized random variable (i.e., with zero mean and the unit variance) with the unknown distribution density $f(x)$. The size of X is transformed $Z^{(k)} = Q_k(X)$,

$k = 1, 2, \ldots$ after each iteration. The transformation is performed so that $Z^{(k)}$ is projected to the direction τ, where the density g_k differs from the standard normal density φ at most, a distribution would be Φ and the projection τ to orthogonal $d - 1$ measurement $\mathbf{R}^d$ subspace remains unchanged. Friedman proved [12] that $Z^{(k)}$ over $k \to \infty$ converges to the standard multi-dimensional Gaussian random variable according to the distribution. Therefore, for a sufficiently large M

$$f(z) \cong \varphi(z^{(M)}) \prod_{k=1}^{M} \frac{g_k(\tau'(k) z^{(k-1)})}{\varphi(\tau'(k) z^{(k)})}, \tag{3.1}$$

where $z^{(k)} = Q_k(x)$. The Friedman statistic is obtained by substituting statistical estimates for the unknown distribution densities g_k on the right-hand side of the expression (3.1).

The direction is $\tau = 1$ and $M = 1$, using one dimentional random variables. Friedman's proposed projection estimator on the base of Legendre orthogonal polynomials was applied to statistical estimation of density g_k. Let $\xi_1, \ldots, \xi_n$ be random variables with distribution density $g(u)$. After transformation $\eta_t = 2\Phi(\xi_t) - 1$, $y = 2\Phi(u) - 1$ the random variables $\eta_1, \ldots, \eta_n$ having distribution with density $g^*(y) = \frac{g(u)}{2\varphi(u)}$ (concentrated in the interval [-1,1]) are obtained. Using the expansions in Legendre polynomial $\{\psi_j\}_{j=0}^{\infty}$ base

$$g^*(y) = \sum_{j=0}^{\infty} b_j \psi_j(y)$$

and replacing the coefficients $b_j = (j + \frac{1}{2}) \mathbf{E} \psi_j(\eta_t)$ to their empirical analogues, the following estimator is obtained

$$\hat{g}(y) = \varphi(y) \sum_{j=1}^{s} \frac{2j + 1}{n} \sum_{t=1}^{n} \psi_j(\eta_t) \psi_j(y). \tag{3.2}$$

According to the recommendation [18], the expansion (3.2) order should be $s \leq 6$.

AKDE algorithm. Kernel density distribution statistical estimator with the local bandwidth has the following form

$$\hat{f}(x) = \frac{1}{n} \sum_{t=1}^{n} \frac{1}{h_t} K\left(\frac{x - X(t)}{h_t}\right). \tag{3.3}$$

The same procedure (as in the article [18]) is used in this paper, while the standard Gaussian distribution density φ is used as a kernel function, and the bandwidth [19] is defined by

$$h_j = h\left(\tilde{f}(X(j))/q\right)^{-v},$$

where $h = (4/(3n))^{1/5}$, $\tilde{f}(\cdot)$ is kernel estimator (3.3) [19], obtained after the change h_j to h, $q = \exp\{\frac{1}{n} \log \sum_{j=1}^{n} \tilde{f}(X(j))\}$, v is the sensitivity parameter.

The authors in [18] proposed choosing latter value from a set of $0.2, 0.4, 0.6, 0.8$, the specific v value is determined by the means of cross-validation.

LSDE algorithm. Logspline distribution density estimators approximate density logarithm by the sum of splines

$$\hat{f}(x) = \exp\left\{\sum_{j=1}^{s} \beta_j B_j(x) - C(\beta)\right\}$$

for a given set of basis functions $B_1, \ldots, B_s$ with a certain coefficient vector $\beta = (\beta_1, \ldots, \beta_s)$ and the normalization constant $C(\beta)$.

The procedure applies cubic B-splines proposed by Kooperberg and Stone. Spline knots are chosen by [1], the coefficients are obtained by maximum likelihood method. Computer program which computes the sample density estimate is given in [21] an has been used in our study.

NNDE algorithm. The main idea behind this algorithm is revealed in multivariate case by formula [10]:

$$\hat{f}(x) = k / \left(n V_d \left\| x - X_{(k,x)} \right\|^d \right),$$

where $V_d = \pi^{d/2} / \Gamma(\frac{d}{2} + 1)$, $\Gamma(z) = \int_0^{-\infty} x^{z-1} e^{-x} dx$, $X_{(k,x)}$ is the k-th nearest x neighbor in the whole sample of observations $\{X(t), t = 1, ..., n\}$. k is the nearest neighbor density estimator depending on the coordinate system, as the distances vary after their changes $\|x - X(t)\|$, $1 \leq t \leq n$. In order to avoid this dependence, the following changes are introduced [9]:

$$\hat{f}(x) = \frac{k(d-1)!}{n \log^{d-1}(n/k(d-1)!) \, 2^d \prod_{j=1}^{d} \left| x_j - X_{(k)j} \right|},$$

where $x = (x_1, ..., x_d)$, and $X(1), \ldots, X(n)$ is a permutation $X_1, \ldots, X_d$ according to increasing values of product $\prod_{j=1}^{d} |x_j - X_j(t)|$. This permutation is invariant under linear transformations of the coordinate axes (excluding rotations). The standard k-nearest neighbor estimator is obtained when $d = 1$. It is recommended to choose $k = \log(n)$ [9].

MIDE algorithm. The density function $f(x)$ can be obtained from the characteristic function using the inversion formula

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-it'x} \psi(t) dt, \tag{3.4}$$

where $\psi(t) = E e^{it'X}$ denotes the characteristic function of a random vector X. Let us denote the characteristic functions projected data by ψ_τ. Since there is a one-to-one correspondence between densities and characteristic functions, and

$$\psi(t) = E e^{it'X} = E e^{i|t|\tau'X} = \psi_\tau(|t|),$$

where $\tau = t/|t|$, we obtain an equation for

$$f(x) = \frac{1}{(2\pi)^d} \int_{\tau:|\tau|=1} ds \int_0^\infty e^{-iu\tau'} \psi_\tau(u) u^{d-1} du \tag{3.5}$$

using the inversion formula (3.4) and performing the transformation to a spherical coordinate system. The first integral in expression (3.5) denotes the surface integral over unit sphere. By using the estimates of ψ_τ and replacing the surface integral by a sum, we obtain the formula for calculation of an estimate $\hat{f}(x)$

$$\hat{f}(x) = \frac{c(d)}{M} \sum_{\tau \in T_0} \int_0^\infty e^{-iu\tau'x} \hat{\psi}_\tau(u) u^{d-1} e^{-\lambda u^2}, \tag{3.6}$$

where $c(d) = d2^{-d}\pi^{-d/2}/\Gamma(\frac{d}{2}+1)$, T_0 consists of M random points uniformly distributed on a sphere T, multiplier $e^{-\lambda u^2}$ performs the additional smoothing function, and the estimator $\hat{\psi}_\tau(\bullet)$ is one dimensional $\tau'X$ distribution density estimator $\hat{f}_\tau$ Fourier transformation. Estimate $\hat{f}_\tau$ is obtained in [32] by the help of AKDE procedure, a normal distribution density φ using as kernel function enabling to calculate analytically the integral on the right-hand side (3.6).

The disadvantage of the inversion formula method (3.6) is that the Gaussian distribution mixture model described by this estimator only gives good estimation of the density of observations similar to it. Inversion formula density estimator often becomes complicated due to large components with a small number of *a priori* probabilities approximating the analysed density by Gaussian distribution mixture. Their value can be reduced by introducing the noise cluster. Let the characteristic function estimator be constructed as Gaussian mixture distributions and uniform distribution of characteristic functions with *a priori* probabilities in modified density estimator equation (3.6):

$$\hat{\psi}_\tau(u) = \sum_{k=1}^{\hat{q}_\tau} \hat{p}_k(\tau) e^{iu\hat{m}_k(\tau) - u^2 \hat{\sigma}_k^2(\tau)/2} + \hat{p}_0(\tau) \frac{2}{(b-a)u} \sin \frac{(b-a)u}{2} \cdot e^{\frac{iu(a+b)}{2}},$$

where the second term describes the cluster of an uniform noise distribution, $\hat{p}_0$ is the noise cluster weight, $a = a(\tau)$, $b = b(\tau)$. We can write

$$a = (\tau'x)_{\min} - \frac{(\tau'x)_{\max} - (\tau'x)_{\min}}{2(n-1)}, \quad b = (\tau'x)_{\max} + \frac{(\tau'x)_{\max} - (\tau'x)_{\min}}{2(n-1)}$$

based on the given equivalent distribution parameter estimates and the designed data. The direction is $\tau = 1$, and $M = 1$, and the calculations become simpler using one dimensional ($d = 1$) random variables. The proposed noise cluster weight is 0.1.

We will consider the application of these methods in estimating the conditional density in case the sample has been clustered. As noted at the beginning of Section 2, we assume that the observed random value X depends on a latent random variable v that takes on the values $1, \ldots, q$, and that the conditional density $f_i(x)$ under the condition $\{v = i\}$ is unimodal for $i = \overline{1, q}$. Since the formula (2.1) holds, we use $\hat{f}(x) = \sum_{i=1}^q \hat{p}_i \hat{f}_i(x)$ for the estimation of density $f(x)$ when the sample is clustered. With the posterior probability $\pi_i(x) = \mathbf{P}\{v = i | X = x\}$ estimates $\hat{\pi}_i(x)$ and on the basis of equality

$p_i = \mathbf{E}\pi_i(x)$ are obtained:

$$\hat{p}_i = \frac{1}{n}\sum_{t=1}^{n}\hat{\pi}_i(X(t)), i = 1,\ldots, q.$$

Strict sample clustering would be the evaluation of random variables $v(1),\ldots,$ $v(n)$ i.e., the sample is split into subsets according to $\hat{v}(t) = \arg\max_{i=1,\ldots,q}\hat{\pi}_i(X(t))$.

4   The analysis of estimation accuracy

This study was performed by comparing previously described density estimation methods. As benchmark distributions the following Gaussian mixture models (introduced by Marron and Wand [26]) were used (also see Figure 1):
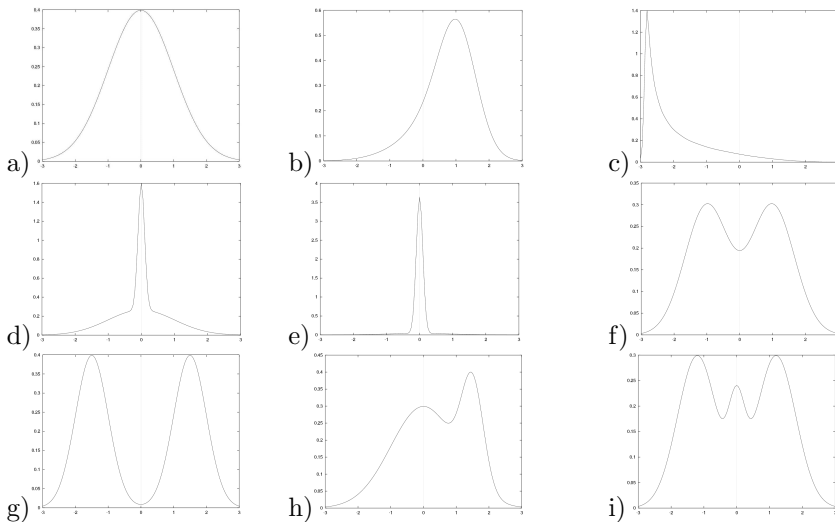

Figure 1.   Benchmark densities for nonparametric density estimation.

a)   Gaussian $N(0, 1)$;

b)   skewed unimodal $\frac{1}{5}N(0, 1) + \frac{1}{5}N(\frac{1}{2}, (\frac{2}{3})^2) + \frac{3}{5}N(\frac{13}{12}, (\frac{5}{9})^2)$;

c)   strongly skewed $\sum_{l=0}^{7}\frac{1}{8}N(3\{(\frac{2}{3})^l - 1\}, (\frac{2}{3})^{2l})$;

d)   kurtotic unimodal $\frac{2}{3}N(0, 1) + \frac{1}{3}N(0, (\frac{1}{10})^2)$;

e)   outlier $\frac{1}{10}N(0, 1) + \frac{9}{10}N(0, (\frac{1}{10})^2)$;

f)   bimodal $\frac{1}{2}N(-1, (\frac{2}{3})^2) + \frac{1}{2}N(1, (\frac{2}{3})^2)$;

g)   separated bimodal $\frac{1}{2}N(-\frac{3}{2}, (\frac{1}{2})^2) + \frac{1}{2}N(\frac{3}{2}, (\frac{1}{2})^2)$;

h)   skewed bimodal $\frac{3}{4}N(0, 1) + \frac{1}{4}N(\frac{3}{2}, (\frac{1}{3})^2)$;

i) trimodal $\frac{9}{20}N(-\frac{6}{5}, (\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5}, (\frac{3}{5})^2) + \frac{1}{10}N(0, (\frac{1}{4})^2)$.

These distributions have been carefully chosen because they thoroughly represent many different types of challenges to non-parametric estimators. The first five ones represent different types of problems that can arise for unimodal densities. The densities of the rest distributions are multimodal. Densities of distributions from f) to i) are mildly multimodal and one might hope to be able to estimate them fairly well with a data set of a moderate size. Small and medium-size samples (50, 100, 200, 400, 800, 1 600, 3 200) were used for modeling. 100 000 replications were generated in each case.

The following error measures have been computed to compare accuracy of estimators:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \left| f(X(t)) - \hat{f}(X(t)) \right| \cong \int \left| f(x) - \hat{f}(x) \right| f(x) dx$$

and

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{f(X(t)) - \hat{f}(X(t))}{f(X(t))} \right| \cong \int \left| f(x) - \hat{f}(x) \right| dx.$$

The results of analysis. The obtained error measures results of the estimators with the best parameters are given in appendices A and B. This is the example of typical models, in other cases the tendency of errors are similar to the provided cases. Arithmetic mean obtained from 100 000 generated samples are given for each estimator. Appendix A graphically shows the results of density estimation without original data clustering and after clustering process. The symbols A, P, L, N, and M mark AKDE, PPDE, LSDE, NNDE, and MIDE algorithms respectively (the last two ones, because of significantly larger error values are not given in all graphs); the distribution density estimation errors without original data clustering are indicated by a solid line, while the errors after the initial data clustering with a dashed line. The tables in Appendix B provide the accuracy results of density estimators.

The conducted computer experiment results qualitatively confirmed the conclusions of [32] on the appropriateness of clustering when the mixture components are separated easily, and as 'separated bimodal' in the case of densities. Clustering is useful for larger samples (from 800 observations), in case of asymmetric *skewed unimodal* densities. When clustering fails to isolate the components of a mixture, when density has strongly expressed single apex or distinguished observations, such as *kurtotic* and *outlier* densities, then primary clustering does not benefit.

In this section, we discuss the results of applying density estimators in more detail:

1. PPDE algorithm gave the best estimators when components of mixtures were separated well enough during clustering, i.e., are the most close to Gaussian densities, without the *Gaussian* density they are *separated bimodal* densities. Contrary to the paper [32] where multidimensional densities were used. The design-based estimates used in this study had the

disadvantage that only one projection could be used for a one-dimensional case.

2. MAE and MAPE are up to 1.5 and 1.7 times lower values compared to other best estimator results. The smallest errors are obtained using the AKDE algorithm in other cases, when peak densities of the mixture are highly expressed. Additionally this method has also been good for a *kurtotic unimodal* densities in small sample cases (up to 200 observations).

3. The advantages of LSDE algorithm were revealed while evaluating complex multimodal density structures. MAE and MAPE for *strongly skewed* density were respectively up to 2.95 and 2.7 times lower values compared to the best results of other estimates. This estimate is also superior to the distribution density which has heavy tails, such as *kurtotic* and *outlier* densities.

4. Although NNDE algorithm is very fast and requires significantly less of CPU time for computations compared to other algorithms (for the largest samples at least 2 times as compared to LSDE; up to 30 times compared to MIDE), it gives an approximate estimation for the small samples. It was competitive with other investigated methods only for individual samples starting from 1600 observations.

5. The MIDE algorithm is competitive with other examined algorithms when the samples have a *Gaussian* distribution or are in multimodal form (e.g., *trimodal* density) initial clustering isolates components of such a distribution.

We analysed a wide range of different types of univariate distribution densities. Based on results, we can state that there is no universal estimator suitable for most cases. The selection of a particular density estimator depends on shape of the density function, which can be characterized by descriptive statistics: skewness, kurtosis, percent of outliers' in the sample, number of modes, etc.

Empirical examples. We compared density estimates of 519 observations of weekly municipal solid waste data (kg/capita) change in a medium-scaled Eastern European city (Kaunas, Lithuania) from 2000 to 2009, which are analysed in [29]. We chose the three density estimators of the highest accuracy as suggested by our comparative study results.

The top row of panels of Figure 2 illustrate the density estimates without data clustering (solid line) and after initial data clustering (dashed line). The peak height is significantly lower for the PPDE estimates, which is inferred as underestimate, considering the tendency of the PPDE algorithm to fit transformations which reveal that distribution structures are too low in univariate case. The same density estimates are illustrated in the bottom row of panels in the Figure 2, with a log-scaled vertical axis to emphasize tail fitting performance. The bumps generated by the outliers are not visible in the tails. Similar pattern is seen in the shape of kurtotic unimodal density that was exhibited
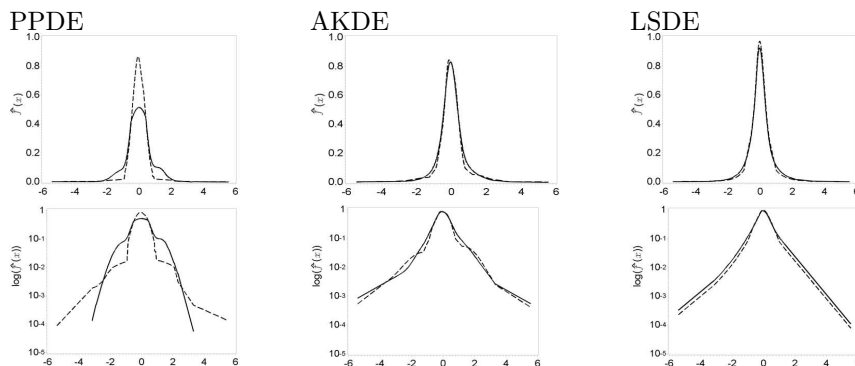
PPDE AKDE LSDE

Figure 2. Density of the municipal solid waste data (kg capita^{-1} week^{-1}) changes.

in Figure 1. LSDE is the best estimator for this density type when the sample size is more than 400 observations.

Statistical data analysis [29] showed that the amount of mixed municipal waste was changing significantly and that statistically important periodic variations (dependability on seasons) exist. By studying the time series data and the distribution density form of the waste amount changes generated in households, we can see that the highest changes are observed before the main holidays of the year and after them. It can be explained by the increased and decreased consumption by the capita. The increase of waste is also observed in the middle of spring and the beginning of autumn, and a sharp drop in the early summer and winter. The relation between the amount of accumulated municipal waste and GDP can be influenced not only by specificity of production of municipal waste but also by the fact that fluctuation of national income does not necessarily change the main household expenses [5]. For example, decrease of GDP can have no influence on consumption producing waste yet it can reduce saving. Thus, accumulation of municipal waste can be explained by the part of GDP spent on consumption by private economies. The amount of waste generated by the population stabilizes quickly after a change in their household habits; it causes the form of density peak.

We expand short-term forecasting of the generation of municipal solid waste (kg capita^{-1} month^{-1}) fractions study [8]. In the above study bootstrapping technique was applied with assumption that distribution densities of Kutaisi (Georgia), Saint-Petersburg (Russia), Boryspil (Ukraine) municipal solid waste fractions have very similar shapes like those of Kaunas (Lithuania). We used the data of municipal solid waste composition research obtained in the framework of the research project "Seasonality of Municipal Waste Generation and Composition and Corresponding Fluctuations of Various Environmental Indicators for Waste Management and Treatment Facilities" conducted during the period of 2010-2013 and including field investigations of the years 2009-2011. Within this section the set of values of the municipal solid waste fractions has been analysed retrospectively in an empirical context in order to estimate the

homogeneity tests.

For the homogeneity we are using some tests based on density estimators described in previous section. For given two samples of independent observations with unknown probability density functions $f(x)$ and $g(x)$ it is required to test homogeneity hypothesis mentioned in paper [2] by Bakshaev (original version proposed by Rudzkis):

$$H_0 : f(x) = g(x).$$

The test statistic for the hypothesis of homogeneity of two samples X and Y have the form:

$$T_{n,m} = \frac{2}{mn} \sum_{i,j} K(X_i, Y_j) - \frac{1}{n^2} \sum_{i,j} K(X_i, X'_j) - \frac{1}{m^2} \sum_{i,j} K(Y_i, Y'_j),$$

where X' and Y' are the randomly splited parts of sample X and Y, respectively, $K(u, v)$ is kernel function of the difference between u and v. K is a non-negative function that integrates to 1. Usually, but not always, K will be a symmetric probability density function. The critical region of the test statistic $T_{n,m}$ established by the next statement.

Under the null hypothesis statistics $T_{n,m}$ will asymptotically ($n, m \to \infty$ and $n/m \to \rho \neq 0$) have the Gaussian distribution with mean zero and variance σ^2 (see the proof of the theorem in [2]).

In this study we used two tests of homogeneity: Kolmogorov-Smirnov test (denote as K-S) and Rudzkis-Bakshaev's test (denote as R-B), for three different density estimators. One of the steps, leading to the main result, was to check the homogeneity between the densities of Kaunas municipal solid waste fractions and densities of other cities municipal solid waste fractions. All results (p-values) of the homogeneity between the density of municipal solid waste fractions are shown in Tables 1–3.

This study has shown that the densities of Kutaisi, Saint-Petersburg, and Boryspil municipal solid waste fractions are similar (are not statistically significant) to the densities of Kaunas municipal solid waste fractions.

5 Conclusions and future work

This paper examines several statistical estimation procedures of a mixture of Gaussian densities. In practice, it is not easy to select an robust estimation procedure if the data density is multimodal and the sample volume is small. The authors have introduced an inversion formula method with an uniform noise component for density estimation. The proposed method, together with other popular estimators, was validated by means of computer simulation.

The quality of multimodal distribution density statistical estimation increases significantly if the sample is pre-clustered (i.e., the multimodal density function is approximated by a mixture of unimodal densities), and the density estimation methods are applied to each cluster afterwards.

The application of statistical methods revealed the relationships and dependencies of municipal solid waste data, and the results could be used in supporting decision making.

As a future work it might be worthwhile to overview the current recommendation systems and propose one for the selection of the most appropriate density estimators for the given sample.

Table 1. The *p*-values results of homogeneity test for Kaunas and Kutaisi municipal solid waste fractions

Fraction	Homogeneity test	Density estimator		
		PPDE	AKDE	LSDE
Paper and cardboard	K-S	0.888	0.897	0.897
	R-B	0.917	0.94	0.939
Plastics	K-S	0.954	0.965	0.965
	R-B	0.987	0.993	0.993
Ferrous metals	K-S	0.901	0.937	0.937
	R-B	0.941	0.972	0.971
Other metals	K-S	0.795	0.811	0.809
	R-B	0.828	0.857	0.856
Glass	K-S	0.922	0.941	0.941
	R-B	0.983	0.987	0.987
Tetrapacks	K-S	0.818	0.821	0.819
	R-B	0.831	0.849	0.848
Food waste	K-S	0.816	0.820	0.818
	R-B	0.830	0.848	0.848
Yard waste	K-S	0.925	0.937	0.937
	R-B	0.950	0.971	0.971
Other unburnable	K-S	0.934	0.952	0.952
	R-B	0.961	0.978	0.978
Total MSW	K-S	0.964	0.966	0.966
	R-B	0.992	0.993	0.993

Table 2. The *p*-values results of homogeneity test for Kaunas and S. Peterburg municipal solid waste fractions

Fraction	Homogeneity test	Density estimator		
		PPDE	AKDE	LSDE
Paper and cardboard	K-S	0.885	0.895	0.895
	R-B	0.916	0.939	0.938
Plastics	K-S	0.910	0.923	0.923
	R-B	0.936	0.951	0.951
Ferrous metals	K-S	0.936	0.951	0.95
	R-B	0.970	0.978	0.977
Other metals	K-S	0.838	0.859	0.858
	R-B	0.869	0.882	0.881
Glass	K-S	0.908	0.926	0.926
	R-B	0.933	0.956	0.956
Tetrapacks	K-S	0.724	0.762	0.761
	R-B	0.743	0.783	0.782
Food waste	K-S	0.897	0.901	0.902
	R-B	0.919	0.942	0.941
Yard waste	K-S	0.922	0.934	0.935
	R-B	0.947	0.962	0.962
Other unburnable	K-S	0.803	0.810	0.808
	R-B	0.856	0.857	0.857
Total MSW	K-S	0.922	0.937	0.936
	R-B	0.984	0.971	0.970

Table 3. The p-values results of homogeneity test for Kaunas and Boryspil municipal solid waste fractions

Fraction	Homogeneity test	Density estimator		
		PPDE	AKDE	LSDE
Paper and cardboard	K-S	0.836	0.857	0.854
	R-B	0.865	0.879	0.878
Plastics	K-S	0.89	0.897	0.897
	R-B	0.918	0.931	0.93
Ferrous metals	K-S	0.815	0.819	0.817
	R-B	0.83	0.848	0.847
Other metals	K-S	0.827	0.834	0.833
	R-B	0.846	0.851	0.85
Glass	K-S	0.998	0.999	0.999
	R-B	0.999	0.999	0.999
Tetrapacks	K-S	0.955	0.967	0.967
	R-B	0.979	0.998	0.998
Food waste	K-S	0.953	0.964	0.964
	R-B	0.976	0.988	0.988
Yard waste	K-S	0.955	0.966	0.966
	R-B	0.978	0.993	0.994
Other unburnable	K-S	0.998	0.999	0.999
	R-B	0.999	0.999	0.999
Total MSW	K-S	0.971	0.976	0.976
	R-B	0.982	0.986	0.985

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6):716–723, 1974. https://doi.org/10.1109/TAC.1974.1100705.

[2] A. Bakshaev. Goodness of fit and homogeneity tests on the basis of N-distances. *Journal of Statistical Planning and Inference*, **139**(11):3750–3758, 2009. ISSN 0378-3758. Special Issue: The 8th Tartu Conference on Multivariate Statistics & The 6th Conference on Multivariate Distributions with Fixed Marginals. https://doi.org/10.1016/j.jspi.2009.05.014.

[3] M.D. Burke and E. Gombay. The bootstrapped maximum likelihood estimator with an application. *Statistics & Probability Letters*, **12**(5):421–427, 1991. https://doi.org/10.1016/0167-7152(91)90031-L.

[4] S. Chernova and M. Veloso. Confidence-based policy learning from demonstration using Gaussian mixture models. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS'07, pp. 1–8, New York, NY, USA, 2007. Association for Computing Machinery. https://doi.org/10.1145/1329125.1329407.

[5] K.M. Christiansen and C. Fischer. Baseline projections of selected waste streams. *Development of Methodology, European Environment Agency, Copenhagen*, 1999. Technical Report No. 28.

[6] J. Ćwik and J. Koronacki. Multivariate density estimation: A comparative study. *Neural Computing & Applications*, **6**(3):173–185, 1997. https://doi.org/10.1007/BF01413829.

[7] A.P. Dempster, N.M. Laird and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1):1–22, 1977. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[8] G. Denafas, T. Ruzgas, D. Martuzevičius, S. Shmarin, M. Hoffmann, V. Mykhaylenko, S. Ogorodnik, M. Romanov, E. Neguliaeva, A. Chusov et al. Seasonal variation of municipal solid waste generation and composition in four East European cities. *Resources, Conservation and Recycling*, **89**:22–30, 2014. https://doi.org/10.1016/j.resconrec.2014.06.001.

[9] L. Devroye and A. Krzyżak. New multivariate product density estimators. *Journal of Multivariate Analysis*, **82**(1):88–110, 2002. https://doi.org/10.1006/jmva.2001.2021.

[10] E. Fix and J.L. Hodges. *Discriminatory analysis: nonparametric discrimination, consistency properties.* Interscience Publishers, New York, 1951.

[11] J.H Friedman. Exploratory projection pursuit. *Journal of the American statistical association*, **82**(397):249–266, 1987. https://doi.org/10.1080/01621459.1987.10478427.

[12] J.H. Friedman, W. Stuetzle and A. Schroeder. Projection pursuit density estimation. *Journal of the American Statistical Association*, **79**(387):599–608, 1984. https://doi.org/10.1080/01621459.1984.10478086.

[13] S. Gruszczyński. The assessment of variability of the concentration of chromium in soils with the application of neural networks. *Polish Journal of Environmental Studies*, **14**(6):743–751, 2005.

[14] Yonghong H., K.B. Englehart, B. Hudgins and A.D.C. Chan. A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses. *IEEE Transactions on Biomedical Engineering*, **52**(11):1801–1811, 2005. https://doi.org/10.1109/TBME.2005.856295.

[15] P. Hall. *The bootstrap and Edgeworth expansion.* Springer, New York, NY, 1992. https://doi.org/10.1007/978-1-4612-4384-7.

[16] T.K. Henthorn, J. Benitez, M.J. Avram, C. Martinez, A. Llerena, J. Cobaleda, T.C. Krejcie and R.D. Gibbons. Assessment of the debrisoquin and dextromethorphan phenotyping tests by Gaussian mixture distributions analysis. *Clinical Pharmacology & Therapeutics*, **45**(3):328–333, 1989. https://doi.org/10.1038/clpt.1989.36.

[17] P.J. Huber. Projection pursuit. *The Annals of Statistics*, **13**(2):435–475, 1985. https://doi.org/10.1214/aos/1176349519.

[18] J.-N. Hwang, Sh.-R. Lay and A. Lippman. Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing*, **42**(10):2795–2810, 1994. https://doi.org/10.1109/78.324744.

[19] M.C. Jones, J.S. Marron and S.J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, **91**(433):401–407, 1996. https://doi.org/10.1080/01621459.1996.10476701.

[20] M. Kavaliauskas, R. Rudzkis and T. Ruzgas. The projection-based multi-variate distribution density estimation. *Acta Comment. Univ. Tartu. Math., Tartu University Press*, **8**:135–141, 2004.

[21] C Kooperberg. Software available at, 2013. Available from Internet: `http://bear.fhcrc.org`.

[22] Ch. Kooperberg and C.J. Stone. A study of logspline density estimation. *Computational Statistics & Data Analysis*, **12**(3):327–347, 1991. https://doi.org/10.1016/0167-9473(91)90115-I.

[23] J.A. Kovacs, C. Helmick and W. Wriggers. A balanced approach to adaptive probability density estimation. *Frontiers in Molecular Biosciences*, **4**:25, 2017. https://doi.org/10.3389/fmolb.2017.00025.

[24] Ch. Léger, D. N. Politis, and O.P. Romano. Bootstrap technology and applications. *Technometrics*, **34**(4):378–398, 1992. https://doi.org/10.1080/00401706.1992.10484950.

[25] D. Li, K. Yang and W.H. Wong. Density estimation via discrepancy based adaptive sequential partition. In D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon and R. Garnett(Eds.), *Advances in neural information processing systems 29*, pp. 1091–1099. Curran Associates, Inc., 2016. Available from Internet: `http://papers.nips.cc/paper/6217-density-estimation-via-discrepancy-based-adaptive-sequential-partition.pdf`.

[26] J.S. Marron and M.P. Wand. Exact mean integrated squared error. *Ann. Statist.*, **20**(2):712–736, 06 1992. https://doi.org/10.1214/aos/1176348653.

[27] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, Inc., 2008. https://doi.org/10.1002/9780470191613.

[28] G.J. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, Inc., 2000. https://doi.org/10.1002/0471721182.

[29] I. Rimaitytė, T. Ruzgas, G. Denafas, V. Račys and D. Martuzevicius. Application and evaluation of forecasting methods for municipal solid waste generation in an eastern-European city. *Waste Management & Research*, **30**(1):89–98, 2012. https://doi.org/10.1177/0734242X10396754.

[30] J. Rothfuss, F. Ferreira, S. Walther and M. Ulrich. Conditional density estimation with neural networks: Best practices and benchmarks. *arXiv preprint arXiv:1903.00954*, 2019.

[31] R. Rudzkis and M. Radavičius. Statistical estimation of a mixture of Gaussian distributions. *Acta Applicandae Mathematica*, **38**(1):37–54, 1995. https://doi.org/10.1007/BF00992613.

[32] T. Ruzgas, R. Rudzkis and M. Kavaliauskas. Application of clustering in the nonparametric estimation of distribution density. *Nonlinear Analysis: Modeling and Control*, **11**(4):393–411, 2006. https://doi.org/10.15388/NA.2006.11.4.14741.

[33] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[34] R. Smidtaite. *Application of nonlinear statistics for distribution density estimation of random vectors*. Kaunas University of Technology, 2008. (MS thesis).

[35] C. Stachniss, C. Plagemann and A.J. Lilienthal. Learning gas distribution models using sparse Gaussian process mixtures. *Autonomous Robots*, **26**(2):187–202, 2009. https://doi.org/10.1007/s10514-009-9111-5.

[36] M.A. Wong. A bootstrap testing procedure for investigating the number of subpopulations. *Journal of Statistical Computation and Simulation*, **22**(2):99–112, 1985. https://doi.org/10.1080/00949658508810837.

[37] M.W. Zimmerman, R.J. Povinelli, M.T. Johnson and K.M. Ropella. A reconstructed phase space approach for distinguishing ischemic from non-ischemic ST changes using Holter ECG data. In *Computers in Cardiology, 2003*, pp. 243–246, 2003. https://doi.org/10.1109/CIC.2003.1291136.

Appendix A

The preliminary data clustering performance analysis is provided there. Each figure corresponds a different density error dependence on the sample size.
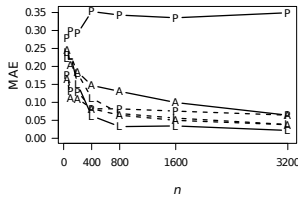

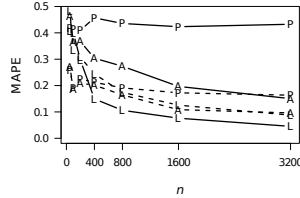**Figure 3.** Estimation accuracy based on MAE for skewed unimodal density

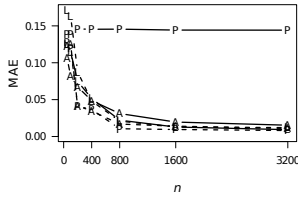**Figure 4.** Estimation accuracy based on MAPE for skewed unimodal density

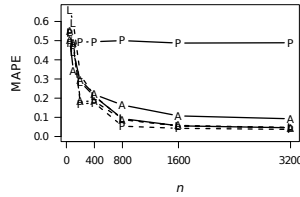**Figure 5.** Estimation accuracy based on MAE for kurtotic unimodal density

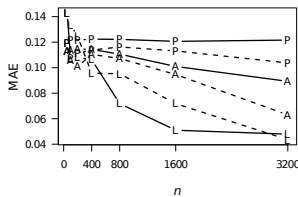**Figure 6.** Estimation accuracy based on MAPE for kurtotic unimodal density

Figure 7. (Estimation accuracy based on MAE for outlier density
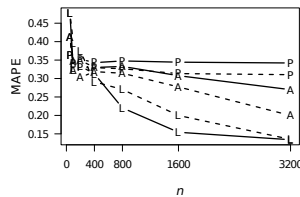

Figure 8. Estimation accuracy based on MAPE for outlier density
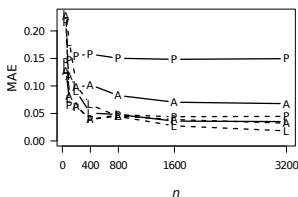

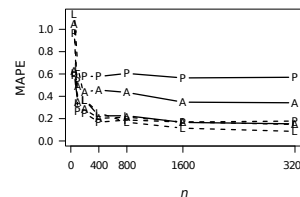**Figure 9.** Estimation accuracy based on MAE for separated bimodal density

Figure 10. Estimation accuracy based on MAPE for separated bimodal density
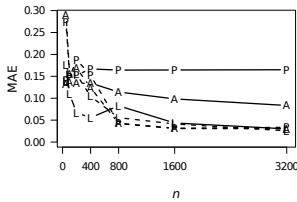
Figure 11. Estimation accuracy based on MAE for skewed bimodal density
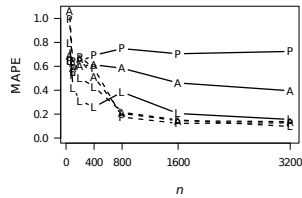

Figure 12. Estimation accuracy based on MAPE for skewed bimodal density

Appendix B

The density error estimators are given in the tables. The error without the initial clustering is given in the first row of each method, and the second row shows the error after the initial data clustering.

Table 4. Performance of density estimators for strongly skewed density

	MAE			MAPE		
	n=200	n=800	n=3200	n=200	n=800	n=3200
AKDE	0.1822	0.1458	0.1035	0.2721	0.2430	0.1824
	0.1697	0.0997	0.0584	0.2287	0.1846	0.1354
PPDE	0.2788	0.3297	0.3319	0.3700	0.4775	0.4877
	0.1983	0.1483	0.1135	0.3611	0.2645	0.2432
LSDE	0.1185	0.0338	0.0419	0.2287	0.0689	0.0783
	0.2430	0.1077	0.0433	0.4667	0.1981	0.1185
NNDE	0.4115	0.3535	0.2398	0.9014	0.6515	0.4109
	0.5020	0.4078	0.3214	1.0365	0.7319	0.5622
MIDE	0.2635	0.2701	0.2818	0.3231	0.3428	0.3561
	0.2831	0.2892	0.3013	0.3623	0.3887	0.4029

Table 5. Performance of density estimators for bimodal density

	MAE			MAPE		
	n=200	n=800	n=3200	n=200	n=800	n=3200
AKDE	0.0195	0.0090	0.0063	0.1116	0.0633	0.0352
	0.0181	0.0084	0.0029	0.1018	0.0594	0.0313
PPDE	0.0469	0.0428	0.0401	0.2093	0.1808	0.1662
	0.0456	0.0273	0.0237	0.2009	0.1279	0.1160
LSDE	0.0571	0.0203	0.0085	0.2674	0.0922	0.0405
	0.0751	0.0209	0.0065	0.3178	0.0978	0.0332
NNDE	0.1484	0.1069	0.0950	0.6445	0.4607	0.3974
	0.1719	0.1355	0.1068	0.7385	0.5763	0.4454
MIDE	0.0372	0.0340	0.0308	0.1483	0.1024	0.0713
	0.0389	0.0234	0.0127	0.1661	0.1001	0.0631

Table 6. Performance of density estimators for trimodal density

	MAE			MAPE		
	n=200	n=800	n=3200	n=200	n=800	n=3200
AKDE	0.0256	0.0175	0.0116	0.1388	0.1021	0.0626
	0.0247	0.0165	0.0102	0.1351	0.0926	0.0551
PPDE	0.0563	0.0560	0.0530	0.2455	0.2311	0.2192
	0.0391	0.0260	0.0363	0.1922	0.1317	0.1774
LSDE	0.0263	0.0126	0.0144	0.1199	0.0952	0.0717
	0.0486	0.0365	0.0138	0.2124	0.0985	0.0775
NNDE	0.1469	0.1106	0.0954	0.6306	0.4800	0.4112
	0.1797	0.1309	0.1232	0.7507	0.5743	0.5361
MIDE	0.0371	0.0368	0.0306	0.1672	0.1461	0.1188
	0.0284	0.0225	0.0159	0.1557	0.1302	0.1094