Nonlinear Analysis: Modelling and Control, 2006, Vol. 11, No. 4, 393–411

Application of Clustering in the Non-Parametric Estimation of Distribution Density

T. Ruzgas[1], **R. Rudzkis**[1], **M. Kavaliauskas**[2]

[1]Institute of Mathematics and Informatics, Akademijos str. 4, LT-08663, Vilnius, Lithuania
tomas.ruzgas@ktu.lt; rudzkis@ktl.mii.lt

[2]Kaunas University of Technology, K. Donelaičio str. 20, LT-44239, Kaunas, Lithuania
snaiperiui@takas.lt

Received: 16.06.2006 **Revised:** 26.09.2006 **Published online:** 30.10.2006

Abstract. This paper discusses a multimodal density function estimation problem of a random vector. A comparative accuracy analysis of some popular non-parametric estimators is made by using the Monte-Carlo method. The paper demonstrates that the estimation quality increases significantly if the sample is clustered (i.e., the multimodal density function is approximated by a mixture of unimodal densities), and later on, the density estimation methods are applied separately to each cluster. In this paper, the sample is clustered using the Gaussian distribution mixture model and the EM algorithm. The highest efficiency in the analysed cases was reached by using the iterative procedure proposed by Friedman for estimating a density component corresponding to each cluster after the primary sample clustering mentioned. The Friedman procedure is based on both the projection pursuit of multivariate observations and transformation of the univariate projections into the standard Gaussian random values (using the density function estimates of these projections).

Keywords: non-parametric estimation, multivariate density function, sample clustering, projection pursuit, Monte-Carlo method.

1 Introduction

A large number of non-parametric methods designed for statistical estimation of the density function of random vectors are used in the modern data analysis. The kernel density estimators are the most common ones [1, 2]. Spline [3, 4] and semi-parametric [5, 6] algorithms are also popular. Application of many popular non-parametric estimation procedures in practice encounters a problem of optimal

parameter selection. The most important element of the kernel density estimators is the smoothing bandwidth. Spline knots selection for spline estimators is also a difficult task. Though there exists a lot of adaptive procedures for selection of mentioned parameters [2, 7–9], the efficiency is low in the case of a small sample size. It is advisable to apply the data projection technique [10–12] in this case, because the parameter selection problem becomes more difficult when the dimension of the observed random vectors increases.

Let X be a d-dimensional random vector with a density function $f(x)$. Let $T \subset \mathbb{R}^d$ be a unit sphere. For each $\tau \in T$, the scalar product $\tau' X$ will denote the projection of a random vector X onto a direction τ. Its density function will be denoted by $f_\tau(u), u \in \mathbb{R}^1$. Let $\mathbb{X} = (X(1), \ldots, X(n))$ be a sample of independent copies of X. The density function $f(x)$ could be estimated using the two-stage procedure:

1. The estimates $\widehat{f}_\tau(u)$ are calculated for each $\tau \in T_0$, where T_0 is a finite set of random points on T.

2. The density function $f(x)$ is estimated by $\{\widehat{f}_\tau(\cdot), \tau \in T_0\}$.

The multivariate density function estimate could be obtained using the inversion formula [12] if we have density function estimates for the large enough number of the univariate projections. One of such estimators is analysed in this paper (expressions (2) and (3); Section 2).

The idea proposed by J. H. Friedman [10] is more delicate. It facilitates overcoming much difficulty in applying the previously mentioned inversion formula, namely: selection of a smoothing parameter, a large number of projected density estimates, etc.

Friedman has developed the idea of Huber [13], who considered the Gaussian distribution to be least "interesting" (because it is so common), and proposed an iterative algorithm, based on both the sequential search of univariate projections, whose distribution function is most different from the Gaussian one, and transformation of those projections into the Gaussian random values. Let Z be a standardized random vector (i.e., random vector with zero mean and unit covariance matrix) with an unknown density function $f(z)$. The value Z is transformed after each step, $Z^{(k)} = Q_k(Z)$, $k = 1, 2, \ldots$. Let us define

$Z^{(0)} = Z$. $Z^{(k)}$ is obtained from $Z^{(k-1)}$ by the following procedure. Let $g_k(u)$, $u \in \mathbb{R}^1$ denote the density function of univariate projection $\tau' Z^{(k-1)}$, where the direction vector $\tau = \tau(k)$ is selected so that g_k differs most from the standard normal density φ. Let us denote the corresponding distribution functions by G_k and Φ. We define

$$Z^{(k)} = Z^{(k-1)} - (\tau' Z^{(k-1)})\tau + \Phi^{-1}\big(G_k(\tau' Z^{(k-1)})\big)\tau.$$

Thus, the random vector $Z^{(k-1)}$ is transformed in such a way that the projection of $Z^{(k)}$ onto the direction τ would have the distribution function Φ, and the projection to the direction orthogonal to τ would remain unchanged. Friedman has proved [10] that the random vector $Z^{(k)}$ converges in distribution to the standard Gaussian random vector as $k \to \infty$. Thus, for large enough M,

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

where $z^{(k)} = Q_k(z)$. Friedman's statistics is obtained by substituting statistical estimates for the unknown univariate density functions g_k into the right side of expression (1). Many-sided analysis results obtained by the authors and other scientists has showed sufficiently good properties of this density function estimator [14]. It is evident that, the more the analysed multivariate distribution is similar to the Gaussian distribution, the more accurate the estimator is. If this method is used to estimate multimodal density functions, larger errors are obtained. This conclusion can also be applied to other estimation methods under investigation.

One of the possible ways to increase the estimation accuracy is to reduce the problem of a multimodal density analysis to the estimation of unimodal densities by treating the density analysed as a mixture of unimodal densities. The authors suggest performing sample clustering at the first stage of analysis and estimating each component of distribution mixture separately at the second stage. The constructive procedure [15] based on approximation of the sample distribution by the Gaussian mixture can be used for sample clustering. The clustering can also be performed by EM algorithm with a random start. The idea of preliminary clustering is not new. Originally it has been used only for the popular kernel

density estimator. The authors are thankful to the referee for taking a note of papers [16] and [17]. The aim of this paper is to determine whether the usage of the preliminary sample clustering decreases estimation errors of multimodal densities. For such density functions, a comparative accuracy analysis of various non-parametric estimators is made by the Monte-Carlo method. This paper comprises the following sections: Section 2 reviews the density estimators; Section 3 describes the EM algorithm used for sample clustering; Section 4 contains the simulation results and conclusions. The accuracy of the estimators is presented (by means of figures and tables) in appendices.

2 The analysed algorithms

The comparative analysis of estimation accuracy was made using five different methods. The density function estimators were selected as representatives of popular different technique estimators which were studied experimentally by other researchers. The exception is the first procedure which is new. The Monte-Carlo method was used to analyse the following statistical estimators of the density function:

1. The inversion formula-based density estimator (IFDE), which is proposed by the authors of this paper.

2. The method based on projection pursuit and sequential normalization of projections proposed by Friedman (PPDE).

3. Silverman's adaptive kernel density estimator (AKDE). A separate bandwidth is used for each observation.

4. The semi-parametric kernel density estimator (SKDE) analysed by Hoti and Holmström, who decomposed a random vector into two subvectors. The density of one of these vectors is estimated by the kernel density estimator, while the density of the other is approximated by the normal density function.

5. The log-spline density estimator (LSDE) proposed by Kooperberg and Stone. The logarithm of the analysed density is approximated by the sum of cubic B-splines.

Before applying the above methods, the sample is standardized (except for the last method), i.e., the sample is transformed to have a zero mean and a unit covariance matrix. Let us describe these methods in more detail.

2.1 IFDE algorithm

Using the inversion formula and passing to spherical coordinates, we obtain

$$f(x) = c(d) \int\limits_{\{\tau \in T\}} ds \int\limits_0^\infty e^{-iu\tau' x} \psi(u\tau) u^{d-1}\, du, x \in \mathbb{R}^d, \tag{2}$$

where $\psi(x) \stackrel{def}{=} \mathbb{E}e^{it'X}$ is the characteristic function, $c(d) = d\, 2^{-d} \pi^{-\frac{d}{2}} / \Gamma(\frac{d}{2}+1)$, Γ is a Gamma function, and the outer integral is the surface integral over the unit sphere. Using expression (2), we obtain the estimator (originally proposed in [12])

$$\widehat{f}(x) = \frac{c(d)}{M} \sum_{\tau \in T_0} \int\limits_0^\infty e^{-iu\tau' x} \widehat{\psi}_\tau(u) u^{d-1} e^{-\lambda u^2}\, du; \tag{3}$$

here the set T_0 consists of M random points uniformly distributed on the sphere T, the factor $e^{-\lambda u^2}$ is used for additional smoothing and $\widehat{\psi}_\tau(\cdot)$ is the Fourier transform of the univariate projection $\tau' X$ density function estimate $\widehat{f}_\tau$. The estimate $\widehat{f}_\tau$ was obtained by AKDE procedure with the Gaussian kernel function. This enables us to calculate the integral on the right side of the expression (3) analytically. For each $\tau \in T_0$,

$$\widehat{f}_\tau(\nu) = \frac{1}{n} \sum_{j=1}^n \varphi\left(\frac{\nu - \tau' X(j)}{h_j} \right) / h_j, \quad h_j = h_j(\tau) \tag{4}$$

and

$$\widehat{\psi}_\tau(u) = \frac{1}{n} \sum_{j=1}^n \exp\left(iu\tau' X(j) - h_j^2 u^2 / 2 \right).$$

The smoothing parameter λ was selected using the cross-validation method [18], for $M = 10000$.

2.2 PPDE algorithm

This estimator is defined by equality (1). The projective estimator, on the basis of the Legendre orthogonal polynomial, was used for estimating densities g_k of the univariate projections. This estimator is identical to that used by Friedman. Let $\xi_1, \ldots, \xi_n$ be univariate random values with a density function $g(u)$. Applying the transformation $\eta_k = 2\Phi(\xi_k) - 1$, $\nu = 2\Phi(u) - 1$, we obtain random values $\eta_1, \ldots, \eta_n$ with density $g^*(\nu) = \frac{g(u)}{2\varphi(u)}$, which is supported on the interval $[-1, 1]$. Using the expansion in the Legendre polynomial basis $\{\psi_j\}_{j=0}^{\infty}$

$$g^*(\nu) = \sum_{j=0}^{\infty} b_j \psi_\tau(\nu)$$

and replacing the coefficients $b_j = (j+1/2)\mathbb{E}\psi_j(\eta_i)$ by their empirical analogues, we obtain the estimator

$$\widehat{g}(y) = \varphi(y) \sum_{y=0}^{s} \frac{2j+1}{n} \sum_{k=1}^{n} \psi_j(\eta_k)\psi_j(\cdot). \tag{5}$$

According to the recommendations [1], the order of expansion (5) was assumed to be $s \leq 6$. Projection directions, assuring the maximal absolute values of empirical skewness and kurtosis, were selected.

2.3 AKDE algorithm

The kernel density estimator with the variable bandwidth is defined by the following expression

$$\widehat{f}(z) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_j^d} K\left(\frac{z - Z(j)}{h_j}\right). \tag{6}$$

The algorithm is identical to the procedure defined in [1]. The standard Gaussian kernel function φ is used. The bandwidth is defined by

$$h_j = h\left(\widetilde{f}\big(Z(j)\big)/q\right)^{-\nu},$$

where $h = (\frac{4}{(2d+1)n})^{\frac{1}{d+4}}$, $\widetilde{f}(\cdot)$ is the kernel density estimator (6) obtained by substituting h for h_i, $q = \exp(\frac{1}{n}\log\sum_{j=1}^{n}\widetilde{f}(Z(j)))$ and ν is the sensitivity parameter. As proposed in [1], values of the parameter ν are chosen from the set $\{0.2, 0.4, 0.6, 0.8\}$ using the cross-validation method.

2.4 SKDE algorithm

The observed d-dimensional random vector X is decomposed into two sub-vectors $X = \binom{Y}{Z}$. Thus, the sample is decomposed $\mathbb{X} = \binom{\mathbb{Y}}{\mathbb{Z}}$. The density function $f_X(\cdot)$ is presented as the product of the density function of random vector Y and a conditional density function of random vector Z, i.e.,

$$f_X(x) = f_Y(y) f_{Z|Y}(z|y), \quad \text{where} \quad x = \begin{pmatrix} y \\ z \end{pmatrix} \in \mathbb{R}^d.$$

The density function $f_Y(y)$ is estimated using the kernel method, analogous to (6), with the constant kernel bandwidth h. Subvector Y and the kernel bandwidth h are selected by the cross-validation method [19] as suggested in [6]. The conditional density $f_{Z|Y}(\cdot|y)$ is approximated by the Gaussian distribution $\mathcal{N}(m(y), C(y))$. The conditional mean $m(y)$ and the conditional covariance matrix $C(y)$ of the random vector Y are defined by the equalities

$$\widehat{m}(y) = \sum_{j=1}^{n} W_{h,j}(y, \mathbb{Y}) Z(j)$$

and

$$\widehat{C}(y) = \sum_{j=1}^{n} W_{h,j}(y, \mathbb{Y}) \big(Z(j) - \widehat{m}(y)\big) \big(Z(j) - \widehat{m}(y)\big)',$$

where $W_{h,j}(y, \mathbb{Y}) = \frac{\varphi(\frac{y - Y(j)}{h})}{\sum_{i=1}^{n} \varphi(\frac{y - Y(j)}{h})}$.

2.5 LSDE algorithm

The log-spline density estimator approximates the logarithm of the multivariate density function by the sum of splines

$$\widehat{f}(x) = \exp\left(\sum_{j=1}^{n} \beta_j B_j(x) - C(\boldsymbol{\beta}) \right),$$

for the given set of basis functions $B_1, \ldots, B_s$ with the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s)$ and the normalizing factor $C(\boldsymbol{\beta})$. The procedure proposed by Kooperberg and Stone applies the cubic B-splines to estimate univariate densities.

The spline knots are selected using the Akaike information criterion [20], and the spline coefficients are calculated using the maximum likelihood method. The estimate of the multivariate density function is the product of univariate spline density estimates. To calculate this estimate, the software [21] is used.

3 Sample clustering using the EM algorithm

If the density function of the random vector X has q maxima, it can be approximated by a mixture of q unimodal densities

$$f(x) = \sum_{k=1}^{q} p_k \varphi_k(x). \tag{7}$$

Let the distribution of X depend on the random variable ν that assumes values $1, \ldots, q$ with probabilities $p_1, \ldots, p_q$, respectively. In the classification theory, ν is interpreted as the number of the class the object belongs to, and each observation $X(t)$, $t = 1, \ldots, n$ has a corresponding class number $\nu(t)$. The functions φ_k are treated as conditional densities given $\nu = k$. Using this approach, the soft clustering problem is equivalent to the estimation problem of posterior classification probabilities

$$\pi_k(x) = \mathbb{P}\{\nu = k | X = k\}$$

for each $x \in \{X(1), \ldots, X(n)\}$. A hard clustering problem is equivalent to the estimation problem of random variables $\nu(1), \ldots, \nu(n)$. In this paper, hard clustering is used for the density function estimation. The sample is decomposed into subsets using the following decision rule

$$\widehat{\nu}(t) = \arg \max_{k=1,\ldots,q} \widehat{\pi}_k\big(X(t)\big). \tag{8}$$

The estimates $\widehat{\pi}_k$ are obtained applying the approximation of unknown density components φ_k by the normal density function and using the EM algorithm. Let expression (7) be valid and φ_k be density functions of the normal distributions $\mathcal{N}(M(k), R(k))$, $k = 1, \ldots, q$. In this case, let us denote the right side of the expression (7) by $f(x, \theta)$, where $\theta = (p_k, M(k), R(k), k = 1, \ldots, q)$. Then the following expression holds

$$\pi_k(x) = \frac{p_k \varphi_k(x)}{f(x, \theta)}, k = \overline{1, q}. \tag{9}$$

Having the estimate of θ, the estimates of π_k are obtained from expression (9) by the "plug-in" method, i.e., by replacing unknown values on the right side of the expression with their statistical estimates. The EM algorithm is an iterative procedure for finding the maximum likelihood estimate θ^* of θ,

$$\theta^* = \arg\max_{\theta} L(\theta), \quad L(\theta) = \prod_{t=1}^{n} f\big(X(t), \theta\big) \tag{10}$$

and the corresponding estimates $\widehat{\pi}_k$. Assume that the estimates $\widehat{\pi}_k = \widehat{\pi}_k^{(r)}$ after r iterations of the procedure. Then a new value $\widehat{\theta} = \widehat{\theta}^{(r+1)}$ is defined by the equalities

$$\begin{aligned}
\widehat{p}_k &= \frac{1}{n}\sum_{t=1}^{n} \widehat{\pi}_k\big(X(t)\big), \\
\widehat{M}(k) &= \frac{1}{np_k}\sum_{t=1}^{n} \widehat{\pi}_k\big(X(t)\big)X(t), \\
\widehat{R}(k) &= \frac{1}{np_k}\sum_{t=1}^{n} \widehat{\pi}_k\big(X(t)\big)\big[X(t)-M(k)\big]\big[X(t)-M(k)\big]',
\end{aligned}$$

where $k = 1, \ldots, q$. By inserting $\widehat{\theta}^{(r+1)}$ into the right side of expression (9), we find $\widehat{\pi}^{(r+1)}(X(t))$, $k = \overline{1,q}$, $t = \overline{1,n}$. Using the above iterative procedure, we obtain a non-decreasing sequence $L(\widehat{\theta}^{(r)})$, whose convergence to the global maximum depends on the selection of the initial value $\widehat{\theta}^{(0)}$ (or $\widehat{\pi}^{(0)}$). The simplest solution of the initial value selection problem is the random start technique. The EM algorithm is repeatedly applied, using the random initial values $\widehat{\pi}^{(0)}$. Finally the estimate $\widehat{\theta}$ is selected if it gives maximum to $L(\widehat{\theta})$. The number of clusters is selected, using the cross-validation method [18]. Sufficiently good results are also obtained applying the automated procedure to select $\widehat{\pi}^{(0)}$.

4 Monte-Carlo simulation

The comparative analysis of the mentioned density estimation methods has been made exploring the data that was used by J.N. Hwang, S.R. Lay and A. Lippman in their paper. Mixtures of the multivariate ($d = \overline{2,5}$) Gaussian and Cauchy

distributions with independent components are used. So, the density functions of the data are defined as follows:

$$f(x) = \sum_{i=1}^{q} p_i f_N(x, M_i, \sigma_i) \quad \text{(Gaussian mixture)}$$

or

$$f(x) = \sum_{i=1}^{q} p_i f_C(x, M_i, u_i) \quad \text{(Cauchy mixture)}$$

with restrictions $\sum_{i=1}^{q} p_i = 1$, $p_i \geq 0$, $i = \overline{1,q}$. Here

$$f_N(x, M_i, \sigma_i) = \frac{1}{\prod\limits_{j=1}^{d} \sigma_{ij}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{j=1}^{d} \frac{(x_j - m_{ij})^2}{\sigma_{ij}^2}\right),$$

$$f_C(x, M_i, u_i) = \prod_{j=1}^{d} \frac{u_{ij}}{\pi\left(u_{ij}^2 + (x_j - m_{ij})^2\right)}.$$

Unimodal distribution

The very first data generated are of unimodal distribution with the following parameters: for the Gaussian distribution,

$$p = 1,\ m = (0.0, 0.0, 0.0, 0.0, 0.0)',\ \sigma^2 = (0.84, 1.02, 0.70, 1.20, 0.96)';$$

for the Cauchy distribution,

$$p = 1,\ m = (0.0, 0.0, 0.0, 0.0, 0.0)',\ u = (0.84, 1.02, 0.70, 1.20, 0.96)'.$$

In case $d = \overline{2,4}$, the parameters are defined by the first d elements of the given 5-dimensional parameter.

Slightly overlapping bimodal distribution

Data of the second type are of slightly overlapping bimodal distribution with the following parameters: for the Gaussian distribution,

$$p_1 = 0.65,\ m_1 = (0.0, 0.0, 0.0, 0.0, 0.0)',\ \sigma_1^2 = (0.42, 0.51, 0.35, 0.60, 0.48)',$$
$$p_2 = 0.35,\ m_2 = (2.0, 2.0, 2.0, 2.0, 2.0)',\ \sigma_2^2 = (0.33, 0.46, 0.53, 0.43, 0.45)';$$

for the Cauchy distribution,

$$p_1 = 0.65,\ m_1 = (0.0, 0.0, 0.0, 0.0, 0.0)',\ u_1 = (0.42, 0.51, 0.35, 0.60, 0.48)',$$
$$p_2 = 0.35,\ m_2 = (2.0, 2.0, 2.0, 2.0, 2.0)',\ u_2 = (0.33, 0.46, 0.53, 0.43, 0.45)'.$$

Highly overlapping bimodal distribution

Data of the third type are og highly overlapping bimodal distribution with the following parameters: for the Gaussian distribution,

$$p_1 = 0.65,\ m_1 = (0.0, 0.0, 0.0, 0.0, 0.0)',\ \sigma_1^2 = (0.84, 1.02, 0.70, 1.20, 0.96)',$$
$$p_2 = 0.35,\ m_2 = (2.0, 2.0, 2.0, 2.0, 2.0)',\ \sigma_2^2 = (0.66, 0.92, 1.06, 0.86, 0.90)';$$

for the Cauchy distribution,

$$p_1 = 0.65,\ m_1 = (0.0, 0.0, 0.0, 0.0, 0.0)',\ u_1 = (0.84, 1.02, 0.70, 1.20, 0.96)',$$
$$p_2 = 0.35,\ m_2 = (2.0, 2.0, 2.0, 2.0, 2.0)',\ u_2 = (0.66, 0.92, 1.06, 0.86, 0.90)'.$$

For each type of data, for both distributions (Gaussian and Cauchy) and for each dimension ($d = \overline{2,5}$), samples of sizes 200, 400, 800, 1600 and 3200 are generated. In each case, simulation is repeated 100 times.

The deviation of the approximation g of function f is measured by

$$\delta = \mathbb{E}\big(g(X) - f(X)\big)^2 / \mathbb{D}f(X).$$

This measure was proposed in [1], and we make use of it in order to obtain comparable results. We define

$$\delta = Err/Var$$

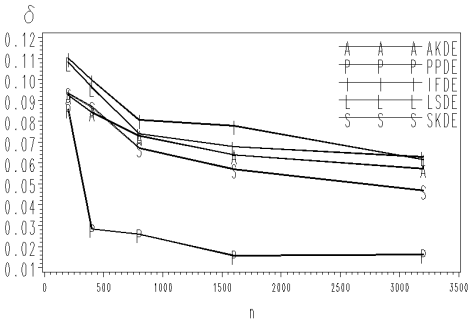
by substituting the density function for f, as well as the estimator $\widehat{f}$ for g, and by taking empirical analogues of unknown values. Here $Err = \frac{1}{n}\sum_{t=1}^{n}(\widehat{f}_t - f_t)^2$ stands for the mean square error, where $f_t = f(X(t))$ is a value of the true density at the observation point, and $Var = \frac{1}{n}\sum_{t=1}^{n}(f_t - \overline{f})^2$, where $\overline{f}$ signifies the average of $f_1, \ldots, f_n$.

Simulation results

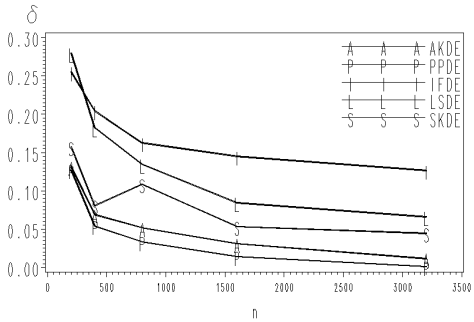
The results of errors of the analysed methods for the best selected values of the parameters are presented in appendices A, B, and C. The typical models are presented. For other data models, the accuracy analysis results are similar to the presented ones. For each method, the arithmetic mean of the error calculated using 100 simulations is presented in figures. Appendix A contains the density estimation results for AKDE, PPDE, IFDE, LSDE, and SKDE methods when the primary data clustering is used. The data clustering was performed, using automated clustering software (developed by Institute of Mathematics and Informatics, Vilnius) which is based on the EM algorithm. Appendix B contains the density estimation results for AKDE and PPDE methods, with the preliminary data clustering in use and without it. Appendix C contains the accuracy analysis results for the density estimators. The results, obtained by means of AKDE and PPDE, are similar to those obtained by J.N. Hwang, S.R. Lay and A. Lippman, i.e., in the case of small sample sizes and heavy tails (Cauchy samples), it is better to use the kernel density estimators, in the case of large data dimensions and large sample sizes (400 and more observations), or in the case of the Gaussian distribution, better results are obtained using the projection pursuit density estimator. In the case of the 5-dimensional Gaussian distribution, quite good results are obtained using the IFDE method, based on the inversion formula. The preliminary data clustering into homogeneous groups, using automatic EM algorithm, enabled us to reduce errors 2–3 times in the case of a small sample, and up to 5 times in some other cases. For large samples ($n = 1600, 3200$), the error reduction ratio equals 1.05–2. A conclusion can be drawn that unambiguously, *PPDE is the best estimator*. For unimodal Gaussian and Cauchy distributions, estimation errors decrease (especially in the case of small samples) up to 4.6 times, provided the preliminary data clustering is applied. SKDE estimations are good enough in the case of unimodal Gaussian densities. For 5-dimensional mixture densities, IFDE turned out to be a good estimator either. It has been found out that, in some cases, very accurate results could be obtained by the LSDE method, however, in the cases with outliers, LSDE yielded great errors that increased the overall averaged error of this method. The IFDE algorithm is very slow in comparison with other methods.

Appendix A

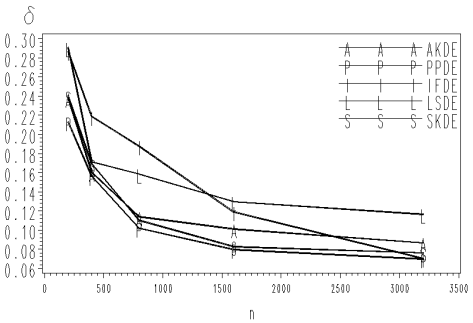
The density estimation results are presented. Preliminary data clustering is used. Each figure corresponds to a different sample distribution.



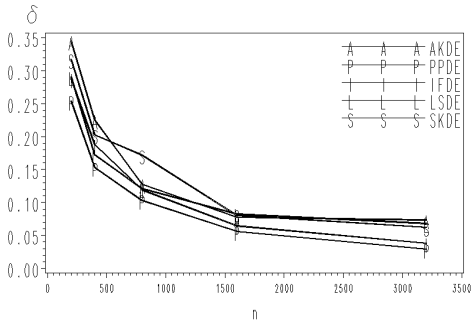
Single mode 2-d Gaussian



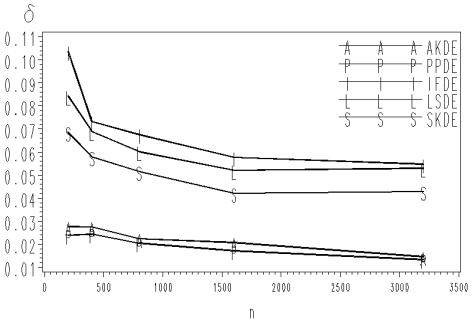
Single mode 5-d Gaussian



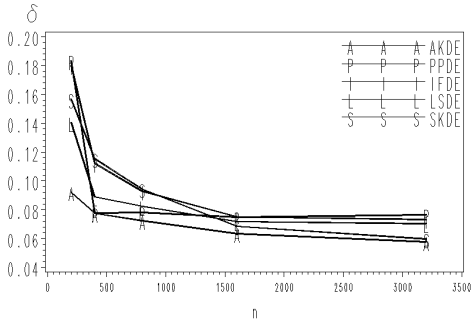
Bimodal slightly overlapping 5-d Gaussian



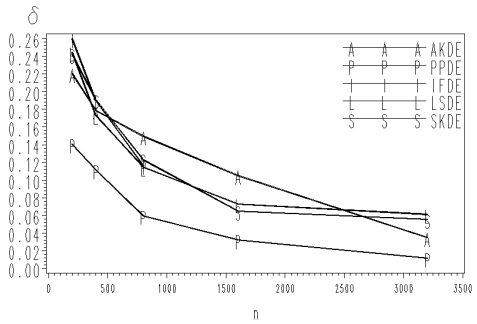
Bimodal highly overlapping 5-d Gaussian



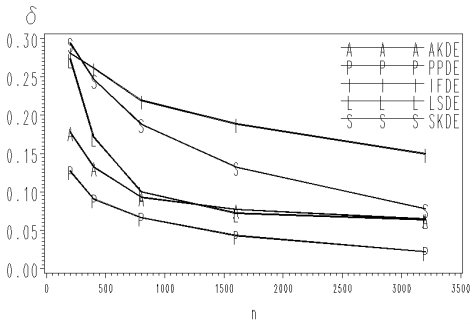
Single mode 2-d Cauchy



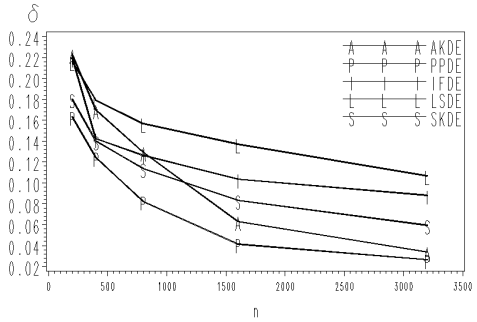
Single mode 3-d Cauchy



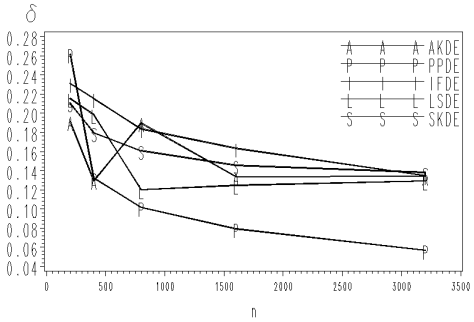
Single mode 4-d Cauchy



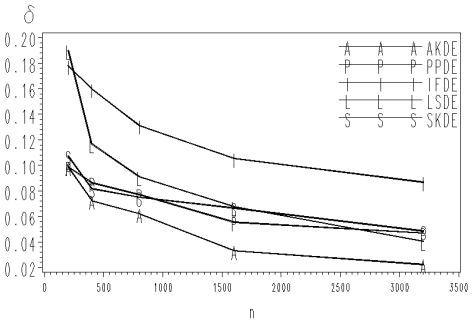
Single mode 5-d Cauchy



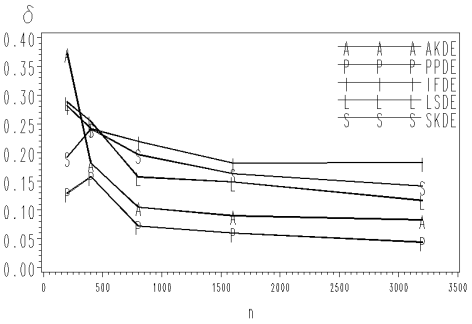
Bimodal slightly overlapping
2-d Cauchy



Bimodal slightly overlapping
5-d Cauchy



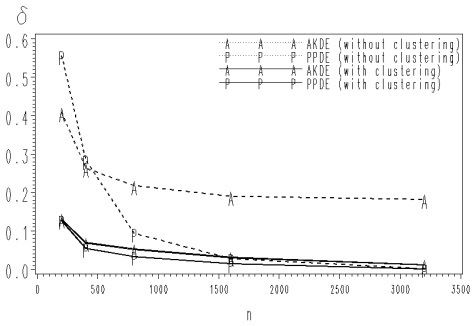
Bimodal highly overlapping
2-d Cauchy



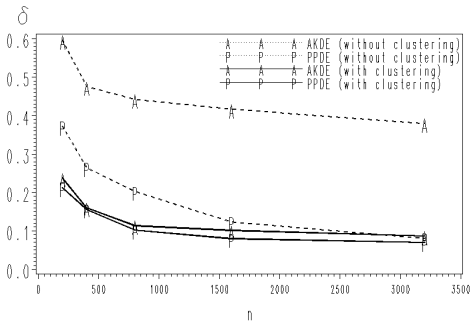
Bimodal highly overlapping
5-d Cauchy

Appendix B

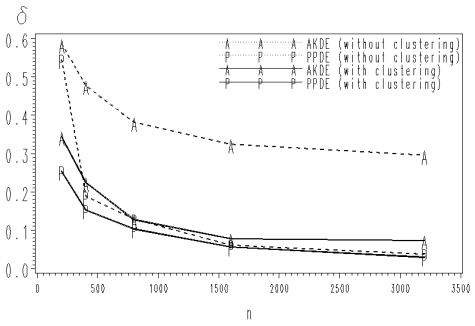
The efficiency analysis of the preliminary data clustering is presented. Each figure corresponds to a different sample distribution.



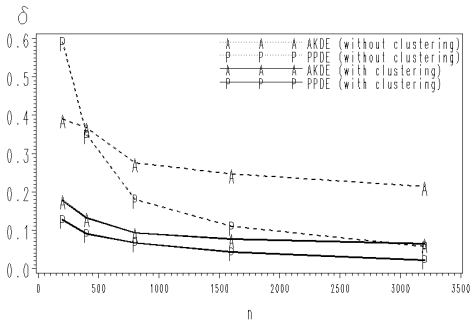
Single mode 5-d Gaussian



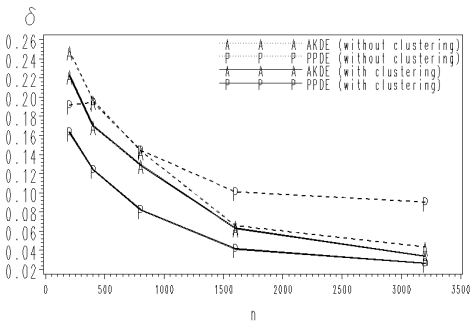
Bimodal slightly overlapping 5-d Gaussian



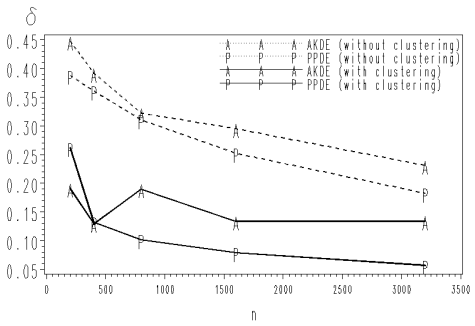
Bimodal highly overlapping 5-d Gaussian



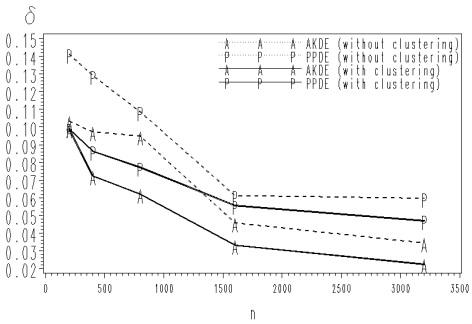
Single mode 5-d Cauchy



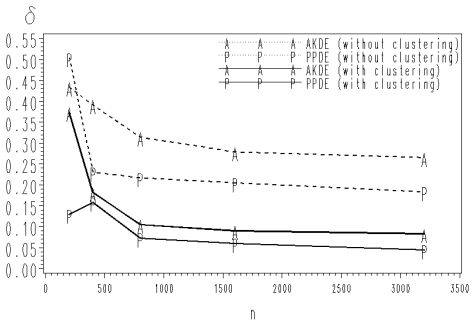
Bimodal slightly overlapping 2-d Cauchy



Bimodal slightly overlapping 5-d Cauchy



Bimodal highly overlapping
2-d Cauchy



Bimodal highly overlapping
5-d Cauchy

Appendix C

The tables illustrate the averaged errors (in bold) and their standard deviation.

Table 1. Single mode 4-dimensional distributions

Method	Gaussian distribution				Cauchy distribution			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.
AKDE	**0.2055**	**0.1178**	**0.1455**	**0.0845**	**0.1994**	**0.1787**	**0.1283**	**0.1052**
	0.0289	0.0121	0.0169	0.0169	0.0056	0.0324	0.0014	0.0051
PPDE	**0.1260**	**0.0668**	**0.0457**	**0.0243**	**0.1804**	**0.1115**	**0.0445**	**0.0323**
	0.0088	0.0141	0.0061	0.0061	0.0034	0.0109	0.0073	0.0031
IKDE	**0.1764**	**0.1661**	**0.1260**	**0.1159**	**0.2099**	**0.1900**	**0.0777**	**0.0719**
	0.0063	0.0178	0.0037	0.0059	0.0087	0.0278	0.0094	0.0118
LSDE		**0.1208**		**0.1099**		**0.1729**		**0.0729**
		0.0066		0.0126		0.0107		0.0045
SKDE		**0.0993**		**0.0541**		**0.1908**		**0.0647**
		0.0124		0.0015		0.0032		0.0071

Table 2. Bimodal slightly overlapping 4-dimensional mixtures

Method	Gaussian mixture				Cauchy mixture			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.
AKDE	**0.2963**	**0.2531**	**0.2495**	**0.1882**	**0.2173**	**0.1755**	**0.1706**	**0.1257**
	0.0550	0.0166	0.0706	0.0178	0.1644	0.0466	0.0861	0.0214
PPDE	**0.2219**	**0.0928**	**0.0590**	**0.0328**	**0.2106**	**0.2027**	**0.1834**	**0.1057**
	0.0242	0.0137	0.0229	0.0148	0.0804	0.0598	0.0266	0.0224
IKDE	**0.2530**	**0.2531**	**0.1841**	**0.1766**	**0.2270**	**0.2124**	**0.1851**	**0.1732**
	0.0621	0.0017	0.0316	0.0017	0.0685	0.0037	0.0847	0.0214
LSDE		**0.1281**		**0.0824**		**0.2130**		**0.1378**
		0.0148		0.0136		0.0283		0.0077
SKDE		**0.1393**		**0.0759**		**0.2011**		**0.1418**
		0.0107		0.0147		0.0313		0.0201

Table 3. Bimodal highly overlapping 4-dimensional mixtures

Method	Gaussian mixture				Cauchy mixture			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.
AKDE	**0.2526**	**0.1049**	**0.2039**	**0.0629**	**0.2478**	**0.1946**	**0.1416**	**0.1341**
	0.0471	0.0058	0.0729	0.0094	0.0889	0.0123	0.0434	0.0109
PPDE	**0.1684**	**0.0512**	**0.0591**	**0.0412**	**0.1879**	**0.1628**	**0.1403**	**0.0912**
	0.0278	0.0106	0.0050	0.0063	0.0078	0.0429	0.0165	0.0021
IKDE	**0.2563**	**0.2321**	**0.1808**	**0.1644**	**0.2496**	**0.2239**	**0.1455**	**0.1427**
	0.0122	0.0018	0.0050	0.0052	0.0258	0.0518	0.0373	0.0213
LSDE		**0.1772**		**0.1213**		**0.2184**		**0.1352**
		0.0061		0.0055		0.0299		0.0106
SKDE		**0.0801**		**0.0809**		**0.2193**		**0.1245**
		0.0078		0.0097		0.0047		0.0056

References

1. J. N. Hwang, S. R. Lay, A. Lippman, Nonparametric Multivariate Density Estimation: A Comparative Study, *IEEE Transactions on Signal Processing*, **42**(10), pp. 2795–2810, 1994.

2. D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: JohnWiley, 1992.

3. C. Kooperberg, Bivariate density estimation with an application to survival analysis, *Journal of Computational and Graphical Statistics*, **7**(3), pp. 322–341, 1998.

4. T. Takada, Nonparametric density estimation: A comparative study, *Economics Bulletin*, **3**(16), pp. 1–10, 2001.

5. A. R. Gallant, D. W. Nychka, Semi-nonparametric Maximum Likelihood Estimation, *Econometrica*, **55**(2), pp. 363–390, 1987.

6. F. Hoti, L. Holmström, Application of Semiparametric Density Estimation to Classification, in: *ICPR*, **3**, pp. 371–374, 2004.

7. C. Gu, C. Qiu, Smoothing spline density estimation: theory, *Annals of Statistics*, **21**(1), pp. 217–234, 1993.

8. W. Härdle, M. Müller, *Multivariate and semiparametric kernel regression*, New York: Wiley, 2000.

9. C. J. Stone, M. Hansen, C. Kooperberg, Y. K. Truong, Polynomial Splines and Their Tensor Products in Extended Linear Modeling, *Annals of Statistics*, **25**(4), pp. 1371–1470, 1997.

10. J. H. Friedman, W. Stuetzle, A. Schroeder, Projection pursuit density estimation, *Journal of the American Statistical Association*, **79**, pp. 599–608, 1984.

11. J. H. Friedman, Exploratory projection pursuit, *Journal of the American Statistical Association*, **82**(397), pp. 249–266, 1987.

12. M. Kavaliauskas, R. Rudzkis, Projection-based Estimation of Multivariate Distribution Density, *Lietuvos matematikos rinkinys*, **42**(spec. nr.), pp. 529–536, 2002.

13. P. J. Huber, Projection pursuit, *Annals of Statistics*, **13**(2), pp. 435–475, 1985.

14. J. Ćwik, J. Koronacki, Multivariate density estimation: A comparative study, *Neural Computing and Applications*, **6**(3), pp. 173–185, 1997.

15. R. Rudzkis, M. Radavicius, Statistical Estimations of a Mixture of Gaussian Distributions, *Acta Applicandae Mathematicae*, **38**(1), pp. 37–54, 1995.

16. T. Duong, *Bandwidth matrices for multivariate kernel density estimation*, PhD thesis, 2004.

17. B. Jeon, D. A. Landgrebe, Fast Parzen Density Estimation Using Clustering-Based Branch and Bound, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **16**(9), pp. 950–954, 1994.

18. M. J. van der Laan, S. Dudoit, S. Keles, Asymptotic Optimality of Likelihood-Based Cross-Validation, *Statistical Applications in Genetics and Molecular Biology*, **3**(1), 2004.

19. B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall, 1986.

20. H. Akaike, A new look at the statistical model identification, *IEEE Trans. AC*, **19**(6), pp. 716–723, 1974.

21. `http://bear.fhcrc.org/~clk/`

22. M. H. Hansen, C. Kooperberg, Spline Adaptation in Extended Linear Models, *Statistical Science*, **17**(1), pp. 2–20, 2002.

23. M. C. Jones, J. S. Marron, S. J. Sheather, A Brief Survey of Bandwidth Selection for Density Estimation, *Journal of the American Statistical Association*, **91**(433), pp. 401–407, 1996.