Opuscula Mathematica • Vol. 29 • No. 1 • 2009

http://dx.doi.org/10.7494/OpMath.2009.29.1.41

Jolanta Jarnicka

MULTIVARIATE KERNEL DENSITY ESTIMATION WITH A PARAMETRIC SUPPORT

Abstract. We consider kernel density estimation in the multivariate case, focusing on the use of some elements of parametric estimation. We present a two-step method, based on a modification of the EM algorithm and the generalized kernel density estimator, and compare this method with a couple of well known multivariate kernel density estimation methods.

Keywords: density estimation, kernel, bandwidth, kernel density estimator, EM algorithm.

Mathematics Subject Classification: 62G07, 65C99.

1. INTRODUCTION

Most of mathematical tools and techniques have been introduced and improved in view of possible applications. Thanks to technical development, numerous processes may be better understood, and numerous phenomena may be explained. In some situations the probability distribution of a random variable or its estimator is a helpful tool.

In this paper we consider distributions of absolutely continuous d-dimensional random variables ($d \geq 1$) and methods of estimation of their density functions. We focus on the case of independent random variables.

Let $(\Omega, \mathfrak{F}, P)$ be a probability space, $\boldsymbol{X}_i : \Omega \longrightarrow \mathbb{R}^d$, $i = 1, 2, \ldots, n$, $d \geq 1$ – a sequence of independent and identically distributed random variables with an unknown density function f. Assume also that $\{\boldsymbol{x}_i\}_{i=1}^n$ is a sequence of observations on $\boldsymbol{X}_i$.

To estimate the unknown density function f, one can use parametric or nonparametric methods. The parametric ones, e.g., the maximum likelihood method, require assumptions on the form of the unknown density. Then, the only problem is to estimate the parameters. But sometimes, having no additional information about the distribution, we should use nonparametric methods, like the histogram or the kernel estimator.

The kernel density estimator, introduced in [12] (in the univariate case), is characterized by two components: the bandwidth $h(n)$ and the kernel K. We consider its multivariate version, $d \geq 1$ (see e.g., [14]).

Definition 1.1. *Let $\{h(n)\}_{n=1}^{\infty} \subset (0, +\infty)$ with $h(n) \longrightarrow 0$. Let $K : \mathbb{R}^d \longrightarrow \mathbb{R}$ be a measurable, nonnegative function. The kernel density estimator is given by the formula*

$$\widehat{f}_h(\mathbf{x}) = \frac{1}{nh^d(n)} \sum_{i=1}^{n} K\Big(\frac{\mathbf{x} - \boldsymbol{x_i}}{h(n)}\Big), \quad \mathbf{x} \in \mathbb{R}^d. \tag{1.1}$$

We call h the bandwidth (window width or smoothing parameter). We call K the kernel.

The kernel K determines the regularity, the symmetry about zero, and the shape of the estimator, while the bandwidth h – the amount of smoothing. In particular, $\widehat{f}_h$ is a density, provided that $K \geq 0$ and $\int_{\mathbb{R}^d} K(\mathbf{t})d\mathbf{t} = 1$.

Considerable research has been carried out on the question of how one should select K in order to optimize the properties of $\widehat{f}_h$ (see e.g., [5, 8, 10, 16], and [4]). In the case of $d \geq 1$, the most often choice is a density function, symmetric about zero, and such that

$$\int_{\mathbb{R}^d} t_i K(\mathbf{t})d\mathbf{t} = 0 \ \text{ and } \ \int_{\mathbb{R}^d} t_i t_j K(\mathbf{t})d\mathbf{t} = \begin{cases} \boldsymbol{\mu_2}, & i = j, \\ 0, & i \neq j, \end{cases} \quad \boldsymbol{\mu_2} < \infty, \quad i, j = 1, \ldots, d, \tag{1.2}$$

like, e.g., the Gaussian kernel

$$K(\mathbf{x}) = \frac{1}{(2\pi)^{-\frac{d}{2}}} \exp\Big(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\Big), \quad \mathbf{x} \in \mathbb{R}^d. \tag{1.3}$$

In some situations also nonsymmetric kernels are used (see e.g., [8]), giving better results.

Still the main problem of the kernel density estimation is the choice of the bandwidth $h(n)$. The usual criterion for the optimal bandwidth is the asymptotic version of the *mean integrated squared error* (see e.g. [4, 5, 9], and [14]), defined by

$$\mathrm{MISE}(\widehat{f}_h) = E \int_{\mathbb{R}^d} \big(\widehat{f}_h(\mathbf{x}) - f(\mathbf{x})\big)^2 d\mathbf{x},$$

which can also be written as a sum of the integrated squared bias and the integrated variance of $\widehat{f}_h$:

$$\mathrm{MISE}(\widehat{f}_h) = \int_{\mathbb{R}^d} (E(\widehat{f}_h(\mathbf{x})) - f(\mathbf{x}))^2 d\mathbf{x} + \int_{\mathbb{R}^d} V(\widehat{f}_h(\mathbf{x}))d\mathbf{x}.$$

We want to keep both the bias and the variance as small as possible, so the optimal bandwidth $h(n)$ should be chosen by the minimization of MISE.

Under additional assumption on the density f,

$$R(\Delta^2 f) := \int_{\mathbb{R}^d} (\Delta^2 f(\mathbf{x}))^2 \, d\mathbf{x} < \infty, \quad \Delta^2 f(\mathbf{x}) = \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}(\mathbf{x}), \tag{1.4}$$

we obtain the formula for the optimal bandwidth in the following form:

$$h_{\text{opt}} = \Big(\frac{d \cdot k_2}{n R(\Delta^2 f) \mu_2^2} \Big)^{\frac{1}{d+4}}, \tag{1.5}$$

where $k_2 = \int_{\mathbb{R}^d} K^2(\mathbf{t}) d\mathbf{t}$, and $\mu_2 = \int_{\mathbb{R}^d} t_1^2 K(\mathbf{t}) d\mathbf{t}$ (see e.g. [14]). The problem is that, formula (1.5) depends on the unknown density f. Hence, in practice we get the optimal bandwidth if we use an estimate for $R(\Delta^2 f)$.

Several ways of solving this problem have appeared in the literature. Most of the known methods are based on the idea of constructing a pilot function (quite often using some elements of parametric estimation) and then use it for actual estimation (see e.g., [1, 2, 7, 8, 10, 13, 14], and [6]). But there are also methods using their own data-based procedures for the bandwidth choice (see e.g., [14, 15]). The problem is that, none of these methods guarantees good results for a wide class of estimated density functions.

In this paper, we compare a couple of multivariate kernel density estimation methods mentioned above with the *two-step method* (introduced in [7, 8]), that generalizes the method considered in [6].

In Section 2 we recall a couple of multivariate kernel density estimation methods. Section 3 is dedicated to the *two-step method*. The idea of the method is presented in the case of independent d-dimensional random variables and a kernel symmetric about zero (for results in some cases of weak dependence and in the case of nonsymmetric kernel see [8]). The last Section contains a comparative study of the methods considered (with the use of results of a computer simulation).

2. MULTIVARIATE KERNEL DENSITY ESTIMATION METHODS

Let $\boldsymbol{x}_i$, $i = 1, \ldots, n$, be a sequence of independent d-dimensional data points from absolutely continuous distribution with an unknown density f. Assume that the kernel $K : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a density satisfying (1.2) and consider the kernel density estimator (1.1).

The choice of the optimal bandwidth $h(n)$ is crucial also in the case of multivariate kernel density estimation (see Figs 1–3).

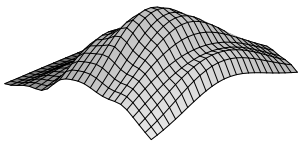

Fig. 1. Bandwidth choice $h = 0.958$, $n = 100$

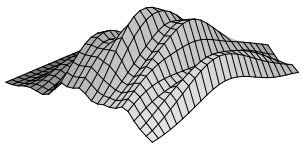

Fig. 2. Bandwidth choice $h = 0.656$, $n = 100$

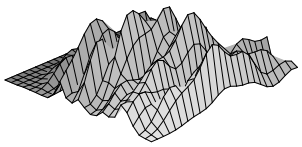

Fig. 3. Bandwidth choice $h = 0.362$, $n = 100$

The estimated density function f is presented below (Fig. 4).

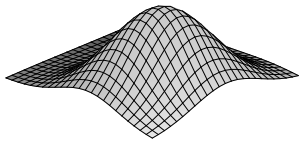

Fig. 4. Density function f of

$$N_2\Big(\begin{bmatrix}2\\-10\end{bmatrix}, \begin{bmatrix}1.5 & -1.2\\-1.2 & 10.1\end{bmatrix}\Big)$$

As mentioned above, most of the known kernel density estimation methods are based on the use of an estimator for the unknown factor (1.4) in formula (1.5). In a few methods some modification of the kernel density estimator (1.1) was necessary (see e.g., [6, 14]). Many of these methods are known only in the univariate case (see e.g., [10,13]), some were generalized to the multivariate case. Let us recall a couple of the latter ones. We choose those that, contain some elements of parametric estimation and are therefore often called *semiparametric*.

1st method. The choice of the bandwidth under the assumption that f is a density of the multivariate normal distribution (see [14]). Then

$$R(\Delta^2 f) = \int_{\mathbb{R}^d} (\Delta^2 f(\mathbf{x}))^2 d\mathbf{x} = (2\sqrt{\pi})^{-d}\Big(\frac{d}{2} + \frac{d^2}{4}\Big),$$

and hence the bandwidth

$$h_{\mathbf{1}} = \left(\frac{d \cdot k_2}{n(2\sqrt{\pi})^{-d}\Big(\frac{d}{2} + \frac{d^2}{4}\Big)\mu_2^2}\right)^{\frac{1}{d+4}},$$

where $k_2 = \int_{\mathbb{R}^d} K^2(\mathbf{t})d\mathbf{t}$ and $\mu_2 = \int_{\mathbb{R}^d} t_i^2 K(\mathbf{t})d\mathbf{t}$, $i = 1, \ldots, d$.

2nd method. The adaptive kernel method (see [1, 14]). In the initial step one has to find a pilot estimate $\widehat{f}_{\mathbf{h}_0}$, usually in the form of (1.1) (with h_0 calculated as in the first method). Then the adaptive estimator has the form

$$\widehat{f}_{\boldsymbol{h}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i^d(n)} K\Big(\frac{\mathbf{x} - \boldsymbol{x}_i}{h_i(n)}\Big), \quad \mathbf{x} \in \mathbb{R}^d,$$

where

$$h_i(n) = h_0\lambda_i, \quad \lambda_i = \Big(\frac{\widehat{f}_{\mathbf{h}_0}(\mathbf{x})}{G}\Big)^{-\alpha}, \quad \alpha = \frac{1}{d}, \quad G = \sqrt[n]{\widehat{f}_{\mathbf{h}_0}(\boldsymbol{x}_1)\cdots\widehat{f}_{\mathbf{h}_0}(\boldsymbol{x}_n)}.$$

3rd method. The nonparametric density estimation with a parametric start. The method was introduced in [6] in the univariate case, but can be generalized to the multivariate case (see [15]), by taking

$$\widehat{f}(\mathbf{x}) = \frac{1}{n} f_0(\mathbf{x}|\widehat{\boldsymbol{\theta}}) \sum_{i=1}^{n} \frac{K\Big(\frac{\mathbf{x}-\boldsymbol{x}_i}{h(n)}\Big)}{f_0(\boldsymbol{x}_i|\widehat{\boldsymbol{\theta}})},$$

where f_0 is for instance, a multivariate density of a normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, estimated by the use of the maximum likelihood method.

For some modifications and other known methods, see e.g., [14, 15].

Note that, the simplest element of the parametric estimation appearing in these methods is the normality assumption. This assumption is used in the first method and in the second one – in the initial step. The third method is based on the maximum likelihood estimation. Hence it most deserves to be called *semiparametric*.

Let us look at the effectiveness of the above methods. Consider the following example.

Example 2.1. *Suppose we are given $n = 250$ data points from a bivariate absolutely continuous distribution with an unknown density f.*

As an example, let us consider the distribution

$$\frac{1}{3}N_2\Big(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\Big) + \frac{2}{3}N_2\Big(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\Big),$$

where:

$$\boldsymbol{\mu}_1 = \begin{bmatrix}3\\0\end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix}\frac{5}{2}\\-\frac{5}{2}\end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix}\frac{9}{4} & \frac{3}{100}\\ \frac{3}{100} & \frac{1}{100}\end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix}\frac{25}{100} & \frac{3}{40}\\ \frac{3}{40} & \frac{9}{4}\end{bmatrix}.$$

The results of the estimation using all three methods are presented below (Figs 5–7).

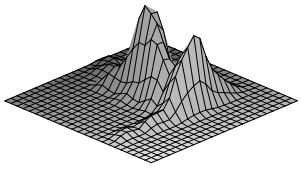

Fig. 5. Kernel estimator first method, $n = 250$

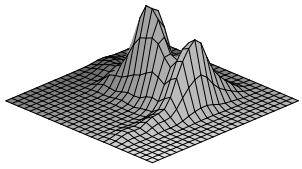

Fig. 6. Kernel estimator second method, $n = 250$

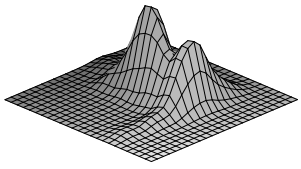

Fig. 7. Kernel estimator third method, $n = 250$

The estimators obtained have shapes close to the shape of the estimated density (Fig. 8)*, but some irregularities are noticeable. We can obtain better results using the two-step method* (Fig. 9).

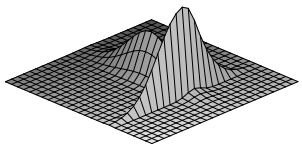

Fig. 8. Estimated density f

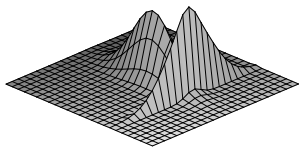

Fig. 9. The two-step method, $n = 250$

3. THE TWO-STEP METHOD

Let $(\Omega, \mathfrak{F}, P)$ be a probability space, $\boldsymbol{X}_i : \Omega \longrightarrow \mathbb{R}^d$, $i = 1, 2, \ldots, n$, $d \geq 1$ – a sequence of independent and identically distributed random variables. Assume that f is an unknown density function of $\boldsymbol{X}_i$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ is a sequence of observations on $\boldsymbol{X}_i$.

Consider a family of measurable functions $\{\phi_j(\mathbf{x})\}_{j=1}^m$ such that

$$0 \leq \phi_j(\mathbf{x}) \leq 1, \quad \sum_{j=1}^m \phi_j(\mathbf{x}) = 1, \quad \mathbf{x} \in \mathbb{R}^d,$$

and a sequence of corresponding bandwidths $\{h_j(n)\}_{j=1}^m$ such that

$$h_j(n) > 0,\ h_j(n) \longrightarrow 0, \text{ and } nh_j^d(n) \longrightarrow \infty, \text{ as } n \longrightarrow \infty,\ j = 1, \ldots, m.$$

Assume also that, the kernel $K : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a density function satisfying (1.2).

Definition 3.1. *The generalized kernel density estimator is defined as follows*

$$\widehat{f}_\phi(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \frac{\phi_j(\boldsymbol{x}_i)}{h_j^d(n)} K\Big(\frac{\mathbf{x} - \boldsymbol{x}_i}{h_j(n)}\Big), \quad \mathbf{x} \in \mathbb{R}^d. \tag{3.1}$$

The idea of the two-step method

The method consists of two steps – a parametric one and a nonparametric one. First, we have to divide randomly the sample $\{\boldsymbol{x}_i\}_{i=1}^n$ into $\{\boldsymbol{x}_{1,i}\}_{i=1}^{n_1}$ and $\{\boldsymbol{x}_{2,i}\}_{i=1}^{n_2}$, where $n_1 + n_2 = n$. This technical operation allows us to treat components of estimator (3.1) as independent random variables (as in the case of (1.1)).

1st step: parametric estimation

For $\{\boldsymbol{x}_{1,i}\}_{i=1}^{n_1}$ one estimates the parameters:

$$m \in \mathbb{N}, \quad \boldsymbol{\mu}_j \in \mathbb{R}^d, \quad \boldsymbol{\Sigma}_j \in M(d, d; \mathbb{R}), \quad \alpha_j \in (0, 1), \quad j = 1, \ldots, m, \quad \sum_{j=1}^m \alpha_j = 1, \tag{3.2}$$

of the pilot function f_0 using an *adaptive algorithm* (based on the EM algorithm), where

$$f_0(\mathbf{x}) = \sum_{j=1}^{m} \alpha_j f_j(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{3.3}$$

$$f_j(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_j|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)}, \quad j = 1, \ldots, m. \tag{3.4}$$

2nd step: nonparametric estimation
For $\{\boldsymbol{x}_{2,i}\}_{i=1}^{n_2}$ one considers the generalized kernel density estimator (3.1), with a sequence of bandwidths $h_j(n_2)$, $j = 1, \ldots, m$, and functions ϕ_j of the form:

$$\phi_j(\mathbf{x}) = \frac{\widehat{\alpha}_j f_j(\mathbf{x}|\widehat{\boldsymbol{\mu}}_j, \widehat{\boldsymbol{\Sigma}}_j)}{f_0(\mathbf{x})} = \frac{\widehat{\alpha}_j f_j(\mathbf{x}|\widehat{\boldsymbol{\mu}}_j, \widehat{\boldsymbol{\Sigma}}_j)}{\sum_{k=1}^{m} \widehat{\alpha}_k f_k(\mathbf{x}|\widehat{\boldsymbol{\mu}}_j, \widehat{\boldsymbol{\Sigma}}_j)}, \quad j = 1, \ldots, m, \tag{3.5}$$

where f_j and f_0 are calculated in the first step.

We choose the bandwidths $h_j(n_2)$, $j = 1, \ldots, m$, minimizing the asymptotic version of the mean integrated squared error, given by

Theorem 3.2 ([7, 8]).

$$\begin{aligned} \mathrm{AMISE}(\widehat{f}_\phi) = \frac{1}{4} \sum_{j,k=1}^{m} h_j^2(n_2) h_k^2(n) \mu_2^2 \int_{\mathbb{R}^d} \Delta^2(\phi_j f)(\mathbf{x}) \Delta^2(\phi_k f)(\mathbf{x}) d\mathbf{x} \\ + \frac{1}{n_2} \sum_{j,k=1}^{m} \frac{1}{h_j^d(n_2) h_k^d(n_2)} \Big(\int_{\mathbb{R}^d} K\Big(\frac{\mathbf{t}}{h_j(n_2)}\Big) K\Big(\frac{\mathbf{t}}{h_k(n_2)}\Big) d\mathbf{t} \Big) \times \\ \times \int_{\mathbb{R}^d} \phi_j(\mathbf{y}) \phi_k(\mathbf{y}) f(\mathbf{y}) d\mathbf{y}, \end{aligned} \tag{3.6}$$

where $\Delta^2(\phi_j f)(\mathbf{x}) = \sum_{i=1}^{d} \frac{\partial^2(\phi_j f)}{\partial x_i^2}(\mathbf{x})$, $j = 1, \ldots, m$.

The **adaptive algorithm** in the first step, based on the EM algorithm (see [3]) and the mutual information (see [17]) can be stated as follows:
We set initial value of m.
(A1) For a given m, we use the EM algorithm:
We set initial values α_j^0, $\boldsymbol{\mu}_j^0$, and $\boldsymbol{\Sigma}_j^0$.
For $j = 1, 2, \ldots, m$, we iterate ($t = 1, 2, \ldots$):

E-step: we calculate

$$q_{ij}^t = \frac{\widehat{\alpha}_j^{t-1} f_j(\boldsymbol{x}_{1,i}|\widehat{\boldsymbol{\mu}}_j^{t-1}, \widehat{\boldsymbol{\Sigma}}_j^{t-1})}{\sum_{j=1}^{m} \widehat{\alpha}_j^{t-1} f_j(\boldsymbol{x}_{1,i}|\widehat{\boldsymbol{\mu}}_j^{t-1}, \widehat{\boldsymbol{\Sigma}}_j^{t-1})}, \quad i = 1, \ldots, n_1, \quad j = 1, \ldots, m,$$

M-step: we calculate

$$\begin{aligned}\widehat{\alpha}_j^t =&\frac{1}{n_1}\sum_{i=1}^{n_1} q_{ij}^t,\\ \widehat{\boldsymbol{\mu}}_j^t =&\frac{\sum_{i=1}^{n_1} q_{ij}^t \boldsymbol{x}_{1,i}}{\sum_{i=1}^{n_1} q_{ij}^t},\\ \widehat{\boldsymbol{\Sigma}}_j^t =&\frac{\sum_{i=1}^{n_1} (\boldsymbol{x}_{1,i}-\widehat{\boldsymbol{\mu}}_j^t)(\boldsymbol{x}_{1,i}-\widehat{\boldsymbol{\mu}}_j^t)^T q_{ij}^t}{\sum_{i=1}^{n_1} q_{ij}^t}.\end{aligned}$$

(A2) We check whether m is the optimal number of components in (3.3), using the system mutual information

$$I(\Theta)=\sum_{i=1}^{m} p_i I(i,\Theta^{-i}),$$

where the component mutual information

$$I(i,\Theta^{-i})=\sum_{j\neq i} p_{ij}\log\frac{p_{ij}}{p_i p_j} \quad\text{and}\quad p_j:=\alpha_j,\ \ p_{ij}:=f_j(\boldsymbol{\mu}_i|\boldsymbol{\mu}_j,\Sigma_j),\ i,j=1,\dots,m.$$

If $I(\Theta)\leq 0$ and $\forall i\ I(i,\Theta^{-i})\leq 0$, then m is the optimal number of components in mixture (3.3) – we have the pilot f_0.
If $I(\Theta)>0$, then there exists an i such that $I(i,\Theta^{-i})>0$. So, we take a β, such that $I(\beta,\Theta^{-\beta})=\max_i\{I(i,\Theta^{-i})>0\}$. Since $I(\beta,\Theta^{-\beta})>0$, there exists a j ($j\neq\beta$), such that $p_{\beta j}\log\frac{p_{\beta j}}{p_\beta p_j}>0$. We take a γ such that $p_{\beta\gamma}\log\frac{p_{\beta\gamma}}{p_\beta p_\gamma}=\max\limits_j\{p_{\beta j}\log\frac{p_{\beta j}}{p_\beta p_j}>0\}$ and remove γth or βth component. We repeat (A1)–(A2) for $m-1$ components.

Example 3.3. *Consider the bivariate observations* $\{\boldsymbol{x}_i\}$, $i=1,\dots,250$, *from the example presented in Section 2. We use the two-step method to estimate the density* f.

First, we randomly divide the sample into $\{\boldsymbol{x}_{1,i}\}$, $i=1,\dots,n_1$ *and* $\{\boldsymbol{x}_{2,i}\}$, $i=1,\dots,n_2$, *taking* $n_1=n_2=125$.

For $\{\boldsymbol{x}_{1,i}\}$, $i=1,\dots,n_1$, *we estimate parameters* (3.2) *of the pilot* f_0, *given by* (3.3) – (3.4), *using the adaptive algorithm described in Section 3. We start with* $m=4$, *using 75 iterations of the EM algorithm. We obtain*

$$\begin{aligned}&\widehat{\alpha}_1=0.708, && \widehat{\boldsymbol{\mu}}_1=\begin{bmatrix}2.495\\-2.464\end{bmatrix}, && \widehat{\boldsymbol{\Sigma}}_1=\begin{bmatrix}0.286 & 0.140\\0.140 & 1.803\end{bmatrix},\\ &\widehat{\alpha}_2=0.019, && \widehat{\boldsymbol{\mu}}_2=\begin{bmatrix}-0.033\\-0.149\end{bmatrix}, && \widehat{\boldsymbol{\Sigma}}_2=\begin{bmatrix}0.206 & 0.004\\0.004 & 0.0006\end{bmatrix},\\ &\widehat{\alpha}_3=0.190, && \widehat{\boldsymbol{\mu}}_3=\begin{bmatrix}1.927\\-0.029\end{bmatrix}, && \widehat{\boldsymbol{\Sigma}}_3=\begin{bmatrix}0.749 & -0.03\\-0.03 & 0.004\end{bmatrix},\\ &\widehat{\alpha}_4=0.083, && \widehat{\boldsymbol{\mu}}_4=\begin{bmatrix}3.886\\0.113\end{bmatrix}, && \widehat{\boldsymbol{\Sigma}}_4=\begin{bmatrix}2.198 & 0.141\\0.141 & 0.020\end{bmatrix}.\end{aligned}$$

We check the accuracy of this model, calculating

$$I(1,\Theta^{-1}) = 0.12\cdot 10^{-1032}, \qquad I(2,\Theta^{-2}) = -0.5\cdot 10^{-3},$$
$$I(3,\Theta^{-3}) = -0.015, \qquad I(4,\Theta^{-4}) = -0.004,$$

and $I(\Theta) = 0.000\,32$. *Since* $I(\Theta) > 0$ *and* $I(1,\Theta^{-1}) > 0$, *we take* $m = 3$ *and repeat the procedure, getting*

$$\widehat{\alpha}_1 = 0.693, \qquad \widehat{\alpha}_2 = 0.047, \qquad \widehat{\alpha}_3 = 0.260,$$
$$\widehat{\boldsymbol{\mu}}_1 = \begin{bmatrix} 2.525 \\ -2.488 \end{bmatrix}, \quad \widehat{\boldsymbol{\mu}}_2 = \begin{bmatrix} 1.901 \\ -0.245 \end{bmatrix}, \quad \widehat{\boldsymbol{\mu}}_3 = \begin{bmatrix} 2.358 \\ -0.019 \end{bmatrix},$$
$$\widehat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 0.278 & 0.205 \\ 0.205 & 1.827 \end{bmatrix}, \quad \widehat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 0.032 & 0.058 \\ 0.058 & 0.202 \end{bmatrix}, \quad \widehat{\boldsymbol{\Sigma}}_3 = \begin{bmatrix} 2.547 & 0.020 \\ 0.020 & 0.008 \end{bmatrix}.$$

In this case

$$I(1,\Theta^{-1}) = 0.012\cdot 10^{-3701} > 0, \;\; I(2,\Theta^{-2}) = -0.005, \;\; I(3,\Theta^{-3}) = -1.418,$$

and $I(\Theta) = 0.000\,25 > 0$, *so we reduce the number of components to* $m = 2$ *and estimate parameters* $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, $\widehat{\boldsymbol{\mu}}_1$, $\widehat{\boldsymbol{\mu}}_2$, $\widehat{\boldsymbol{\Sigma}}_1$, *and* $\widehat{\boldsymbol{\Sigma}}_2$. *As a result we obtain*

$$\widehat{\alpha}_1 = 0.728, \qquad \widehat{\alpha}_2 = 0.272, \tag{3.7}$$
$$\widehat{\boldsymbol{\mu}}_1 = \begin{bmatrix} 2.495 \\ -2.389 \end{bmatrix}, \qquad \widehat{\boldsymbol{\mu}}_2 = \begin{bmatrix} 2.339 \\ -0.018 \end{bmatrix}, \tag{3.8}$$
$$\widehat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 0.284 & 0.138 \\ 0.138 & 1.954 \end{bmatrix}, \qquad \widehat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 2.429 & 0.019 \\ 0.019 & 0.008 \end{bmatrix} \tag{3.9}$$

and

$$I(1,\Theta^{-1}) \approx 0, \quad I(2,\Theta^{-2}) = -0.028, \quad \textit{and} \quad I(\Theta) = -0.008 < 0.$$

Therefore, $m = 2$ *proved to be the optimal number of components in mixture* (3.3).

The pilot functions f_0 *in the cases of* $m = 4$, $m = 3$, *and* $m = 2$ *are presented below* (Figs 10–12).

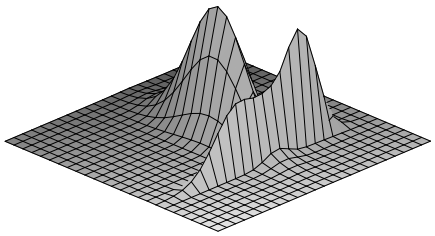

Fig. 10. The two-step method pilot f_0, $m = 4$, $n_1 = 125$

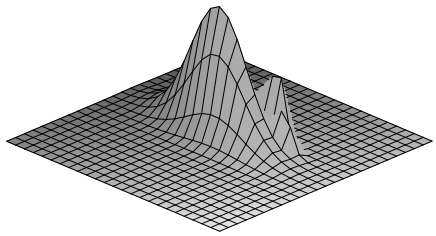

Fig. 11. The two-step method pilot f_0, $m = 3$, $n_1 = 125$

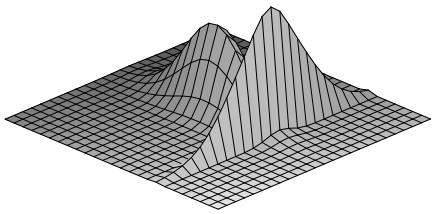

Fig. 12. The two-step method pilot f_0, $m = 2$, $n_1 = 125$

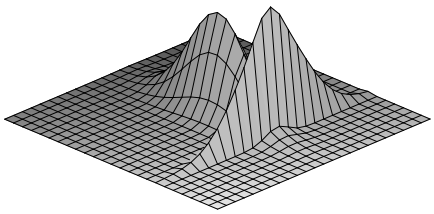

Fig. 13. The two-step method final result, $n_2 = 125$

Having obtained f_0 in the form

$$f_0(\mathbf{x}) = \widehat{\alpha}_1 f_1(\mathbf{x}|\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}}_1) + \widehat{\alpha}_2 f_2(\mathbf{x}|\widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}_2),$$

with f_1, f_2 given by (3.4)*, and parameters* (3.7)–(3.9)*, we use the generalized kernel density estimator*

$$\widehat{f}_\phi(\mathbf{x}) = \sum_{i=1}^{n_2} \Big(\frac{\phi_1(\boldsymbol{x}_{2,i})}{n_2 h_1^2(n_2)} K\Big(\frac{\mathbf{x} - \boldsymbol{x}_{2,i}}{h_1(n_2)}\Big) + \frac{\phi_2(\boldsymbol{x}_{2,i})}{n_2 h_2^2(n_2)} K\Big(\frac{\mathbf{x} - \boldsymbol{x}_{2,i}}{h_2(n_2)}\Big)\Big), \quad \mathbf{x} \in \mathbb{R}^2,$$

for $\{\boldsymbol{x}_{2,i}\}$, $i = 1, \dots, n_2$ ($n_2 = 125$). We take the Gaussian kernel (1.3) *and functions ϕ_1, ϕ_2 in form* (3.5)*.*

Two bandwidths $h_1 = 0.825$ and $h_2 = 3.920$ are calculated by numerical minimization of (3.6)*. The final result of the estimation is presented in Figure 13.*

4. COMPARATIVE STUDY – SIMULATION RESULTS, $d = 2$

Consider bivariate normal distributions, with the densities of the form:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)},$$

with parameters

$$\mu_1, \mu_2 \in \Big(-\frac{3}{2}, \frac{3}{2}\Big), \quad \sigma_1, \sigma_2 \in (0, 2), \quad \text{and} \quad \rho \in (0, 1).$$

Consider also mixtures of two and three distributions of the above form, with parameters

$$\mu_{1j}, \mu_{2j} \in \Big(-\frac{3}{2}, \frac{3}{2}\Big), \ \ \sigma_{1j}, \sigma_{2j} \in (0, 2), \ \ \rho_j \in (0, 1), \quad j = 1, \dots, m, \tag{4.1}$$

$$\alpha_j \in (0, 1), \ \ \sum_{j=1}^{m} \alpha_j = 1. \tag{4.2}$$

We use the above distributions for simulation and the *integrated squared error*

$$\mathrm{ISE}(\widehat{f}) = \int_{\mathbb{R}^2} (\widehat{f}(x_1, x_2) - f(x_1, x_2))^2 dx_1 dx_2,$$

as a criterion of comparing the methods. A method is better than another one if it gives a smaller value of ISE (for the same sample).

The following simulation is conducted:

We randomly choose:

— the number of components in the mixture (3.3) (m from 1 to 3),

— the parameters (4.1) and (4.2) (from uniform distribution on the given interval),

— the samples of $n = 100$, $n = 200$, and $n = 500$ elements from the chosen distribution.

For each sample, we apply three density estimation methods from Section 2 and the two-step method as well. Each time we take the Gaussian kernel (1.3). In each case we compute ISE.

The whole procedure is repeated 250 times.

The results of the simulation – the values of ISE – are presented as tables and as box plots. Each box plot contains a box with the central line showing the median, the lower line showing the first quartile Q_1, the upper line showing the third quartile Q_3, two lines extending from the central box bounding 90% of the data, and outliers. The tables contain mean values of ISE for $n = 100$, $n = 200$, and $n = 500$ (Tab. 1) and sample quartiles (Q_1, me, and Q_3) calculated for the obtained values of ISE (Tabs 2–4).

Table 1
Mean values of ISE, ($n = 100$, $n = 200$, $n = 500$)

n	Two-step method	1st method	2nd method	3rd method
100	0.0219	0.0510	0.0342	0.0331
200	0.0091	0.0180	0.0161	0.0146
500	0.0026	0.0042	0.0046	0.0037

Comparison of mean values of ISE shows that the two-step method gives the best results, having the smallest mean value. We get the greatest mean in the case of the first method, taking the last place.

Let us look at further analysis, (Fig. 14) starting with $n = 100$.

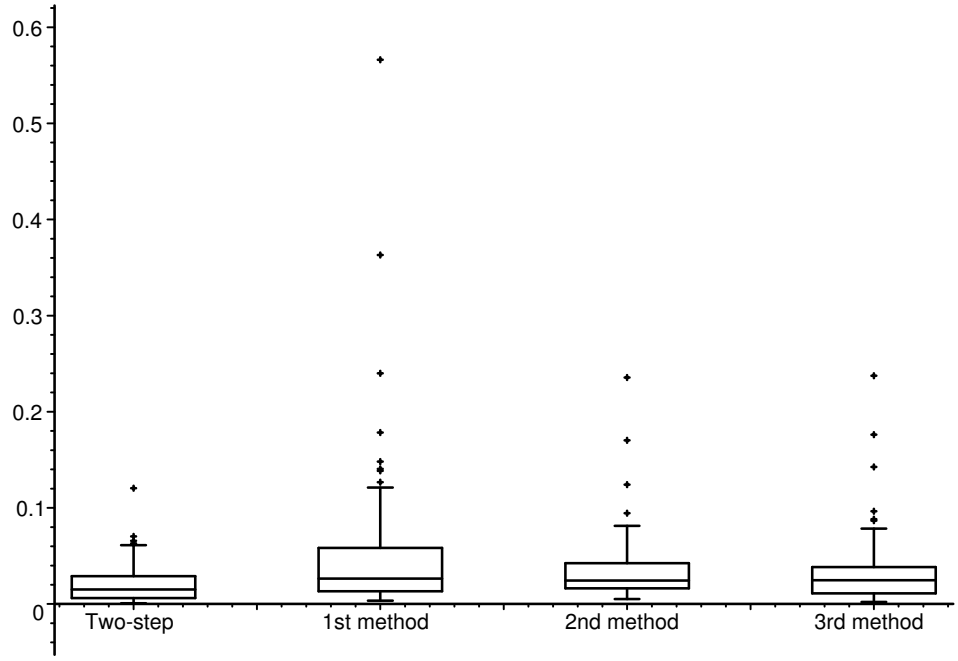

Fig. 14. Values of ISE for $n = 100$

For a convenience, we present the values of Q_1, me, and Q_3 in Table 2.

Table 2
Values of ISE, $n = 100$, $d = 2$

Method	Two-step method	1st method	2nd method	3rd method
Q_1	0.0062	0.0133	0.0163	0.0110
me	0.0150	0.0253	0.0243	0.0247
Q_3	0.0289	0.0564	0.0424	0.0383

Analyzing the box plots obtained and the values of ISE in Table 2, we can observe that, the two-step method gives the best results. In a half of the cases we got values of ISE from the interval $(0.0062, 0.0289)$, while in case of the best of the remaining methods – the third one – this interval has the form $(0.0110, 0.0383)$. The first method seems to be the worst.

And now for the results in the case $n = 200$ (Fig. 15).

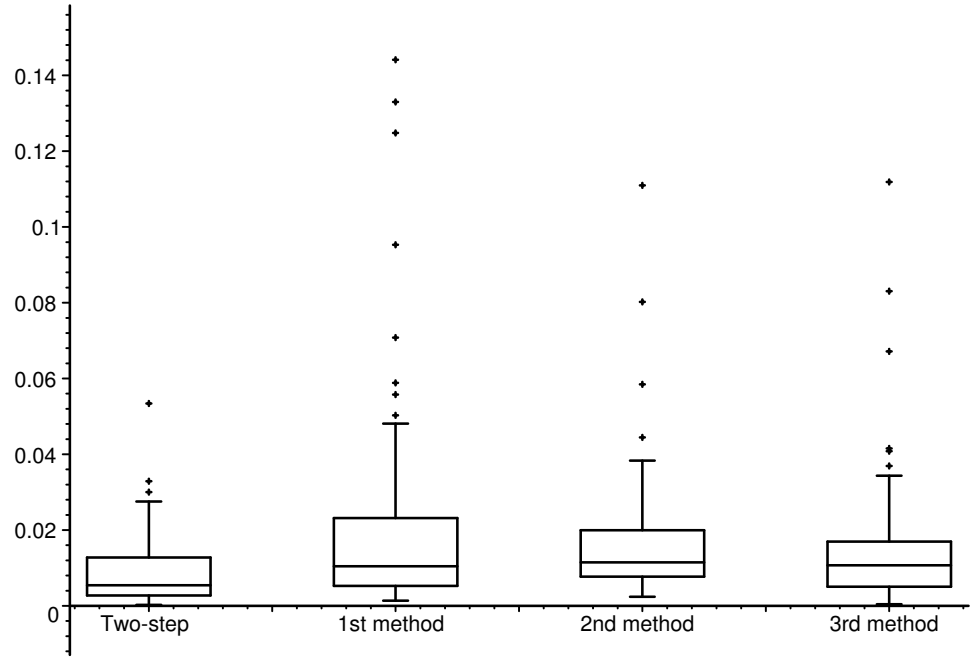

Fig. 15. Values of ISE for $n = 200$

Observe that, the scale of the box plots has changed in comparison to Figure 14. Additionally, the differences between the results obtained by the use of respective methods have decreased. It is related to the larger sample size (this will be visible even better in Figure 16, in the case of $n = 500$).

Table 3
Values of ISE, $n = 200$, $d = 2$

Method	Two-step method	1st method	2nd method	3rd method
Q_1	0.0027	0.0052	0.0077	0.0051
me	0.0053	0.0097	0.0115	0.0110
Q_3	0.0121	0.0220	0.0187	0.0177

It can be seen that the two-step method takes the first place, as previously. We should admit that the third method – the one with a parametric start – is also good, giving results just a little bit worse than the two-step method. The second method – the adaptive one – takes the next place. The worst is once again the first, simplest method. Although the median is less than the one in the case of the second and third method, the interquartile range is essentially greater.

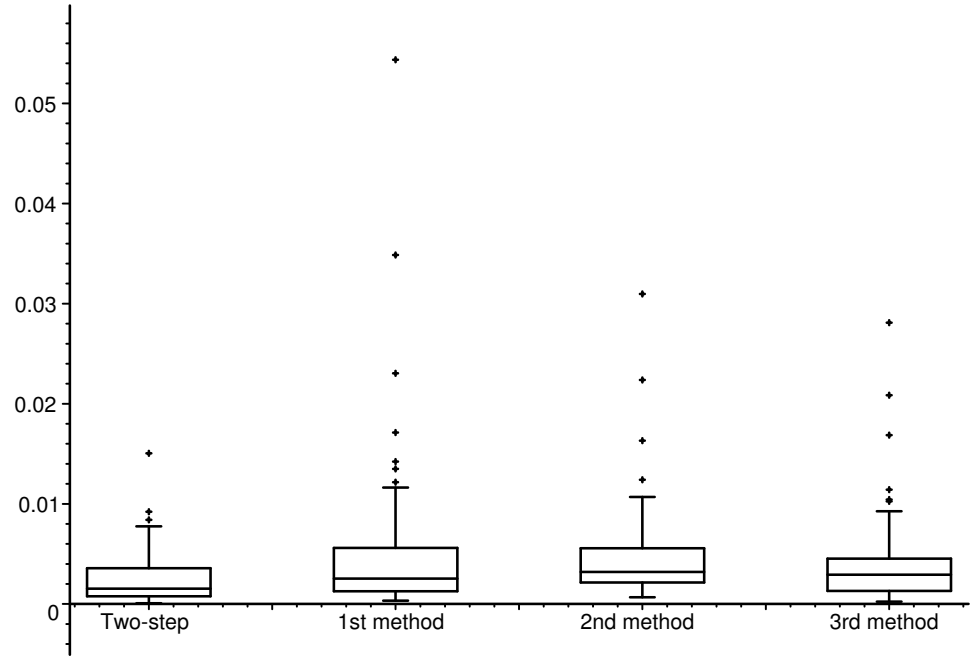

Fig. 16. Values of ISE for $n = 500$

Table 4
Values of ISE, $n = 500$, $d = 2$

Method	Two-step method	1st method	2nd method	3rd method
Q_1	0.0008	0.0013	0.0021	0.0013
me	0.0015	0.0025	0.0032	0.0029
Q_3	0.0034	0.0053	0.0055	0.0045

The results obtained in this case are comparable, thanks to the large sample. Still, the two-step method gives the best results, although differences between the methods are slight.

Consequently, the two-step method has proved to be the best one when compared with other methods mentioned. It can be recommended for applications.

REFERENCES

[1] I.S. Abramson, *Arbitrariness of the pilot estimator in adaptive kernel methods,* J. Multivariate Anal. **12** (1982), 562–567.

[2] L. Breiman, W. Meisel, E. Purcell, *Variable kernel estimates of multivariate densities,* Technometrics, 1977, 135–144.

[3] A.P. Dempster, N.M. Laird, D.B. Rubin, *Maximum-likelihood from incomplete data via EM algorithm,* J. Roy. Statist. Soc. Ser. B **39** (1977), 1–38.

[4] L. Devroye, L. Gjörfi, *Nonparametric density estimation: the L^1 view,* Willey, New York, 1985.

[5] P. Hall, J.S. Marron, *Choice of kernel order in density estimation,* Ann. Statis. **16** (1987), 161–173.

[6] N.L. Hjort, I.K. Glad, *Nonparametric density estimation with a parametric start*, Ann. Statist. **23** (1995), 882–904.

[7] J. Jarnicka, *On some two-step density estimation method*, Univ. Iagel. Acta Math. **43** (2005), 67–91.

[8] J. Jarnicka, *Density estimation – the two-step method*, Thesis, the Jagiellonian University, Cracow, 2006 [in Polish].

[9] J.S. Marron, M.P. Wand, *Exact mean integrated squared error,* Ann. Statist. **20** (1992), 712–736.

[10] B.U. Park, J.S. Marron, *Comparison of data-driven bandwidth selectors*, J. Amer. Statist. Assoc. **85** (1990), 66–72.

[11] E. Parzen, *On estimation of a probability density function and mode,* Ann. Math. Statist. **33** (1962), 1065–1076.

[12] M. Rosenblatt, *Remarks on some nonparametrics estimates of a density function,* Ann. Math. Statist. **27** (1956), 827–837.

[13] S.J. Sheater, M.C. Jones, *A reliable data-based bandwidth selection method for kernel density estimation,* J. Roy. Statist. Soc B **53** (1991), 683–690.

[14] B.W. Silverman, *Density estimation for statistics and data analysis,* Chapman and Hall, London, New York, 1986.

[15] J.S. Simonoff, *Smoothing methods in statistics,* Springer-Verlag, 1996.

[16] G.R. Terrel, D.W. Scott, *Variable kernel density estimation,* Ann. Statist. **20** (1992), 1236–1265.

[17] Z.R. Yang, M. Zwolinski, *Mutual information theory for adaptive mixture models,* Trans. Pattern Anal. Machine Intellegence 23, 40 (2001), 396–403.

Jolanta Jarnicka
jarnicka@wms.mat.agh.edu.pl
jarnicka@ibspan.waw.pl

AGH University of Science and Technology
Faculty of Applied Mathematics
Mickiewicza 30, 30-059 Cracow, Poland

Polish Academy of Sciences
Systems Research Institute
Newelska 6, 01-447 Warsaw, Poland

Received: March 12, 2008.
Accepted: September 25, 2008.