mathematics

MDPI

Article

Nonparametric Multivariate Density Estimation: Case Study of Cauchy Mixture Model

Tomas Ruzgas *, Mantas Lukauskas * and Gedmantas Čepkauskas

Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, 44249 Kaunas, Lithuania; gedmantas.cepkauskas@ktu.edu

* Correspondence: mantas.lukauskas@ktu.lt (M.L.); tomas.ruzgas@ktu.lt (T.R.)

Abstract: Estimation of probability density functions (pdf) is considered an essential part of statistical modelling. Heteroskedasticity and outliers are the problems that make data analysis harder. The Cauchy mixture model helps us to cover both of them. This paper studies five different significant types of non-parametric multivariate density estimation techniques algorithmically and empirically. At the same time, we do not make assumptions about the origin of data from any known parametric families of distribution. The method of the inversion formula is made when the cluster of noise is involved in the general mixture model. The effectiveness of the method is demonstrated through a simulation study. The relationship between the accuracy of evaluation and complicated multidimensional Cauchy mixture models (CMM) is analyzed using the Monte Carlo method. For larger dimensions (d ~ 5) and small samples (n ~ 50), the adaptive kernel method is recommended. If the sample is n ~ 100, it is recommended to use a modified inversion formula (MIDE). It is better for larger samples with overlapping distributions to use a semi-parametric kernel estimation and more isolated distribution-modified inversion methods. For the mean absolute percentage error, it is recommended to use a semi-parametric kernel estimation when the sample has overlapping distributions. In the smaller dimensions (d = 2) and a sample is with overlapping distributions, it is recommended to use the semi-parametric kernel method (SKDE) and for isolated distributions, it is recommended to use modified inversion formula (MIDE). The inversion formula algorithm shows that with noise cluster, the results of the inversion formula improved significantly.

Keywords: cauchy mixture model; nonparametric density estimation; density estimation algorithms; adapted kernel density estimate; logspline estimation;

MSC: 62G05; 62G07; 62G30

Citation: Ruzgas, T.; Lukauskas, M.; Čepkauskas, G. Nonparametric Multivariate Density Estimation: Case Study of Cauchy Mixture Model. *Mathematics* **2021**, *9*, 2717. https://doi.org/10.3390/math9212717

Academic Editor: Antonio Di Crescenzo

Received: 1 September 2021
Accepted: 21 October 2021
Published: 26 October 2021

1. Introduction

Estimation of probability density functions (pdf) is considered an essential part of statistical modelling. It expresses random variables as functions of other variables, making it possible to detect hidden relationships between data [1]. In a significant number of machine learning algorithms, it is essential to determine a previously unknown function of the distribution density of the data. The function of the distribution density is applied in the Bayesian classifier [2,3], in density-based clustering algorithms [4–6], or information-based feature selection algorithms [7,8]. Effective density estimates must be carefully created in advance to obtain unknown functions of probability density. Nowadays, there is still much focus on developing innovative density estimation procedures [9,10]. Density estimation is an open research topic in the fast-growing area of deep learning. Scientists have begun proposing robust density estimators based on neural networks such as Parzen neural networks [11], soft-constrained neural networks [12], and others [13].

Let us say that the random vector $X \in R^d$ satisfies the distribution mixture model if its distribution density $f(x)$ satisfies the equation $f(x) = \sum_{k=1}^{q} p_k f_k(x) = f(x, \theta)$. The parameter q is the number of mixture clusters, and p_k is the a priori probability. These conditions must also be met: $p_k > 0$ and $\sum_{k=1}^{q} p_k = 1$. The function $f_k(x)$ is a function of the distribution density, and θ is a multidimensional parameter of the model. Suppose X is a d-dimensional random vector with a distribution density $f(x)$, and there is a sample of independent copies of X, where $X = (X(1), \ldots, X(n))$. It can be argued that the sample satisfies the mixture model if $X(t)$ satisfies $f(x) = \sum_{k=1}^{q} p_k f_k(x) = f(x, \theta)$.

One of the statistical tasks is to estimate the density of the observed random variable. Suppose the available sample's distribution type is known (Normal, Poisson, and others). In that case, the distribution density of the data can be estimated simply using mean and covariance matrix estimates, fitting them to a defined distribution [14–16]. Thus, the standard parametric method is applied when the assumptions about the density form are met. When estimating density in a parametric way, the value of the multidimensional distribution parameter θ needs to be found, which is not straightforward because the number of parameters increases rapidly as the dimension d increases. For example, in the case of a mixture of Gaussian distributions, $\dim \theta = \frac{1}{2qd(d+1)} + qd + q - 1$, and even with a small dimension $d = q = 5$, the model will consist of $\dim(\theta) = 104$ parameters. When searching for parameter estimates, it may be necessary to solve the optimization problem in the 104-dimensional space. In practice, the number of clusters q may also be unknown, and it needs to be estimated. The parametric method is not proper when the random size distribution is unknown. In this case, non-parametric methods are used to determine certain forms of density estimates [17–19].

The histogram is one of the simplest and oldest estimates of density. To the best of our knowledge, data in the form of histograms (without graphical representation) were first presented in 1661 to determine mortality probabilities in different age groups [20]. To approximate the density $f(x)$ in the area Ω, the number of observations $X(t)$ falling into Ω is calculated and divided by n and the volume of the area Ω. The area of space to which all observations fall is first found. That means the fluctuation intervals of all X projections on the axes $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$ are found. The fluctuation intervals of the observations are divided into l partial intervals and in the hypercubes $\Omega_j (j = 1, \ldots, r)$ bounded by them, the density estimate is calculated as

$$\hat{f}(x) = \frac{n(\Omega_j)}{n \cdot h_1 \cdot h_2 \ldots \cdot h_d}. \quad (1)$$

Here $n(\Omega_j)$ is the number of observations entering the hypercube Ω_j and $h_j, j = 1, \ldots, d$ are the edges of the hypercube [21,22]. It is recommended to select the number of hypercubes [17,23,24], and to choose $r \cong 1 + 3.32 \log(n)$, and $l = \sqrt[d]{r}$ has to be an integer number, so r is chosen that $\left\lceil \sqrt[d]{1 + 3.32 \log n} \right\rceil^d$.

A histogram is one of the simplest means of presenting data that is easy to understand and convenient. This estimate is a function that acquires non-negative values, and its integral is equal to one. However, it is not continuous. That poses problems when knowing the density estimate derivatives is essential, mainly when density estimation is used in intermediate steps of other methods, such as clustering using a gradient algorithm or plotting high-measurement data-level lines. Remarkably, the histogram stood as the only non-parametric density estimator until the 1950's when substantial and simultaneous progress was made in density estimation and spectral density estimation. In 1951, in a little-known paper, Fix and Hodges [25] introduced the basic algorithm of non-parametric density estimation; an unpublished technical report was formally published as a review by Silverman and Jones in 1989 [26]. They addressed the problem of statistical discrimination when the parametric form of the sampling density was not known. During the following decade, several general algorithms and alternative theoretical modes of analysis were introduced by Rosenblatt in 1956 [27], Parzen in 1962 [28], and Cencov in 1962 [29]. Then followed the second wave of essential and primarily theoretical papers by Watson

and Leadbetter in 1963 [30], Loftsgaarden and Quesenberry in 1965 [31], Schwartz in 1967 [32], Epanechnikov in 1969 [33], Tarter and Kronmal in 1970 [34], and Kimeldorf and Wahba in 1971 [35]. Next, Cacoullos introduced the natural multivariate generalization in 1966 [36]. Finally, in the 1970s, the first papers focusing on the practical application of these methods were published by Scott et al. in 1979 [24] and Silverman in 1978 [37]. These and later multivariate applications awaited the computing revolution.

Modern data analysis uses several non-parametric methods for statistically estimating the distribution density of multivariate random variables. Kernel estimates are particularly prevalent [38,39]. Quite popular and spline [40,41] and semi-parametric [42–46] algorithms. However, detailed comparisons of the effectiveness of existing popular estimates for multimodal density are lacking. With the most popular non-parametric estimation procedures, optimal selection of their parameters is encountered in practice. The most crucial element in the design of kernel estimates is the width of the smoothing. It is not easy to select the nodes of the spline estimates. Although several adaptive procedures for the selection of these parameters have been developed [39,47–52], however, they are not efficient enough when the sample volume is not large, especially then the observational dimension is large. In the latter case, it is appropriate to apply data design [53–56] because of the more extensive the dimension of the observed random vectors, the more complex the task of parameter selection.

The main idea of this paper is to estimate the performance of different density estimators by using density mixtures to show another type of problem, which may result from data heteroscedasticity and outliers. The relationship between the accuracy of evaluation and complicated multidimensional Cauchy mixture models (CMM) is analyzed using the Monte Carlo method. For example, Kalantan and Einbeck [57] used engineering data and, for computer vision, used CMM, comparing it with the Gaussian mixture model. Azzari and Foi [58] used harmony between Gauss and heavy-tailed Cauchy to find noise-model parameters that make outlier estimation robust when imaged dominated by texture. Finally, Teimouri [59] analyzed patients with Cushing's syndrome and their diagnostic tests. The focus was on the tetra hydrocortisone urine release rate (mg/24 h) and evaluating parameters in the EM algorithm and Cauchy mixture model.

Scientific novelty. Evaluation accuracy comparative analysis is made by using different probability density estimation procedures. Density function estimates are chosen as popular different technique estimates, which other researchers have already analyzed. This research is essential because it focuses on Cauchy distributions.

2. The Density Estimation Algorithms

This section aims to present the density estimation algorithms used in the study theoretically. All algorithms are presented with algorithms theoretical substantiation. When making the histogram, each $X(t)$ can be imagined as a separate column with a height of $1/n$. Then it makes sense to change the centre of the column to $X(t)$ itself and get the following function:

$$\hat{f}(x) = \frac{1}{n \cdot h_1 \cdot h_2 \cdots h_d} \sum_{t=1}^{n} I_{C_h(X(t))}(x). \quad (2)$$

Here C_h is a hypercube with centre $X(t)$, and the lengths of the edges are $h_1, \ldots, h_d$. In summary, instead of the indicator function, a smooth "prominence"—the kernel function—can be used at each observed point. The multidimensional fixed-width bandwidth estimate with the kernel function K and the fixed (global) kernel width parameter h, which can be used to estimate the density $\hat{f}(x)$ of the multidimensional data $X \in R^d$, is then defined as follows:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{t=1}^{n} K\left(\frac{x - X(t)}{h}\right). \quad (3)$$

These are some of the most common non-parametric estimates of distribution density [38,39,60,61]. The kernel function is selected to meet the following condition:

$$\int_{R^d} K(x)dx = 1. \quad (4)$$

The standard normal distribution density function φ is often used as the kernel [62,63]:

$$\phi(x) = (2\pi)^{-d/2}\, exp(-\frac{1}{2}x'x). \quad (5)$$

Often the observations are not evenly distributed in all directions. Therefore, it is desirable to scale the data by eliminating the most significant dispersion differences in the different coordinate directions. One suitable method for this [64] is data standardization. That means the sample's effect on a linear transformation. The mean of the transformed data is zero, the covariance matrix is unitary, and (3) apply the Equation to already standardized data. For example, suppose Z is a standardized random vector,

$$Z = S^{-\frac{1}{2}}(X - \bar{X}), \quad (6)$$

here $\bar{X}$ is the empirical mean of the sample, and $S \in R^{d\times d}$ is the empirical covariance matrix. Based on the fixed kernel width density estimate (3), a more complex standardized data density estimate has been constructed:

$$\widehat{f_z}(z) = \frac{1}{nh^d}\sum_{t=1}^{n} K\left(\frac{z-Z(t)}{h}\right). \quad (7)$$

$$\acute{f}(x) = \frac{(detS)^{-1/2}}{nh^d}\sum_{t=1}^{n} K\left(S^{-1/2}\frac{x-X(t)}{h}\right). \quad (8)$$

The optimal kernel width h^* for a fixed core width is determined by minimizing the average integral root mean square error (MISE) [65]. For example, when the distribution of observations is normal with a unit covariance matrix in Gaussian kernel, the expression h^* proposed by [65] is $h^* = An^{-\frac{1}{d+4}}$, here $A = [4/(2d+1)]^{\frac{1}{d+4}}$. More sophisticated kernel width selection methods (such as the least-squares cross-checking method) are obtained by more complex and lengthy calculations [66–70].

In practical research, the kernel width is often selected experimentally. If the value of h is small, the density function estimate has more modes that correspond to the layout of the observed data. A higher value of h means more significant smoothing of the estimate.

Although fixed-core width density estimates are widely used to estimate non-parametric densities, they often have some practical drawbacks [65]. For example, fixed-core width density estimates do not ensure the distribution ends' integrity without over-smoothing the underlying bulk density.

2.1. Adapted Kernel Density Estimate (AKDE)

A good improvement on the fixed kernel width density estimate is the adapted kernel density estimate [65]. The adapted kernel density estimate is constructed similarly to the fixed kernel width density estimate. The kernel describes the density at each observed point. In this case, the kernel width is already considered when moving from one observation to another. In areas of different smoothness, it is appropriate to take different kernel widths. This method consists of two steps: estimation of the adapted kernel width and density estimation by the kernel method, using the information obtained in the first step. The algorithm can be summarized as follows:

Step 1: The elements of sample $X = (X(1), \ldots, X(n))$ are standardized to $Z = (Z(1), \ldots, Z(n))$ such that $\hat{E}[Z] = 0$ and $\hat{E}[ZZ'] = I$.

Step 2: Estimates $\tilde{f}_Z(z)$ of the fixed kernel density estimate (3) satisfying the condition $\tilde{f}_Z(Z(t)) > 0$, $\forall$t.

Step 3: The local width parameter is determined $\lambda_t = \left(\frac{\tilde{f}_Z(Z(t))}{g}\right)^{-\gamma}$, where g is $\tilde{f}_Z(z)$ the geometric mean, $\log g = \frac{1}{n}\sum_{i=1}^{n}\log \widehat{f_Z}(Z(t))$ and γ is the sensitivity parameter: $0 \le \gamma \le 1$.

Step 4: An adapted kernel estimate is made with variable-width kernels: $\hat{f}_Z(z) = \frac{1}{n}\sum_{t=1}^{n} h^{-d}\lambda_t^{-d} K\left(\frac{z-Z(t)}{h\lambda_t}\right)$.

Where h is the same global smoothness parameter as in Equation (3), the higher γ, the more sensitive the density selection. Quite often, the parameter value is selected as follows $\gamma = \frac{1}{2}$ [65,71].

2.2. Semi-Parametric Kernel Density Estimate (SKDE)

When data are scarce, parametric estimates are often applied even when the unknown density is not parameterized. Therefore, it is essential to mention the combination of parametric and non-parametric estimates. For example, one of the semi-parametric estimates of kernel density was examined by F. Hoti and L. Holmström [46]. This estimate divides the random vector into two subvectors and estimates the distribution density of one of them by the kernel method. Afterward, another relative density is approximated by the Normal distribution density [46]. For example, suppose d and s are positive integers $d \geq 2$, $1 \leq s \leq d-1$. Using this method, the d-dimensional vector $X \in R^d$ is decomposed into two s and (d–s) dimensioned subvectors $X = \binom{Y}{Z}$, and the sample is decomposed accordingly: $X = \binom{Y}{Z}$, where $Y \in R^s$, $Z \in R^{d-s}$. The evaluated density function is expressed as the product of the distribution density of the random vector Y and the conditional distribution density of the random vector Z: $f_X(x) = f_{(Y,Z)}(y,z) = f_Y(y) f_{Z|Y=y}(z|y)$, $x = \binom{y}{z} \in R^d$. Here f_X and f_Y are the densities of X and Y. $f_{Z|Y=y}$ is the density of Z when $Y = y$.

Suppose that the relative density $Y = y$ is multidimensional normal Gaussian, but the density f_Y does not belong to any family of parametric functions. The density f_X is then obtained by estimating f_Y in a non-parametric manner and applying a multidimensional Normal density to each $f_{Z|Y=y}$. The density function $f_Y(y)$, as with (8), is evaluated by the kernel method [65]. Since the sample elements are not standardized, the smoothness parameter is not the same in all directions. Therefore, using the kernel method, it is replaced by the s-dimensional matrix H:

$$\hat{f}(y) = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{det(H)} K(H^{-1}(y - Y(t))). \quad (9)$$

Usually, the shape of H is chosen diagonally—$H = diag(h_1, \ldots, h_s)$, and the smoothness parameters are calculated as follow

$$h_j = \left(\frac{4}{s+2}\right)^{1/(s+4)} n^{-1/(s+4)} \sigma_j. \quad (10)$$

It should be noted that this form, when $s = 1$, was proposed by B. W. Silverman [65]. Replacing the standard deviation σ_j of the component Y_j with its estimate $\hat{\sigma}_j = \sqrt{\frac{\sum (X_j - \overline{X_j})^2}{n_j}}$ and by the rule of D. W. Scott [39] first multiplier is always between 0.924 and 1.059 $\hat{h}_j$ can be calculated as follows

$$\hat{h}_j = n^{-1/(s+4)} \hat{\sigma}_j. \quad (11)$$

This Scott's rule is easy to summarize for the smoothness matrix H:

$$\hat{H} = n^{-1/(s+4)} \hat{\Sigma}^{1/2}. \quad (12)$$

Here $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_s^2)$ is the diagonal matrix of Y empirical variances.

The conditional density $f_{Z|Y}(\cdot\,|y)$ is approximated by the Gaussian distribution $N(m(y), C(y))$, where $m(y)$, $C(y)$ denote the conditional mean of the vector Y and the conditional covariance matrix: $m(y) = E(Z|Y = y)$, $y \in R^s$, $C(y) = E[(Z - m(y))'(Z - m(y))|Y = y]$, $y \in R^s$. For the estimation of $m(y)$ and $C(y)$, it is proposed to apply the kernel smoothing:

$$\hat{m}(y) = \frac{\sum_{t=1}^{n} K_{H_2}(y - Y(t)) Z(t)}{\sum_{j=1}^{n} K_{H_2}(y - Y(j))} = \sum_{t=1}^{n} W_{H_2}(y - Y(t)) Z(t), \; y \in R^s. \quad (13)$$

Here are the weights $W_{H_2}(y - Y(t)) = \frac{K_{H_2}(y-Y(t))}{\sum_{j=1}^{n} K_{H_2}(y-Y(j))}$.

The sum of which is equal to one. The formula (13) can be understood as a regression estimate of the conditional mean function m of Nadaraya and Watson [72,73]. The conditional covariance matrix can be evaluated similarly $\hat{C}(y) = \sum_{t=1}^{n} W_{H_3}(y - Y(t))(Z(t) - \hat{m}(y))'(Z(t) - \hat{m}(y))$, $y \in R^s$. The parametric estimate of the relative density $f(Z \mid Y) = y$ looks akin to this $\hat{f}_{Z|Y=y}(z) = [(2\pi)^{d-s}\, \widehat{det\, C}\,(y)]^{-1/2}\, exp\left\{-\frac{1}{2}(z - \hat{m}(y))\hat{C}(y)^{-1}(z - \hat{m}(y))'\right\}$, $z \in R^{d-s}$. The estimate of the distribution density f_X of X then is: $\hat{f}_X(x) = \hat{f}_{(Y,Z)}(y, z) = \hat{f}_Y(y)\hat{f}_{Z|Y=y}(z)$, $x = (y,z) \in R^d$.

The procedure described above is called the semi-parametric kernel density estimate. In practice, even if the conditional assumption of the normality of several random vector components is satisfied. The decomposition dimensions also influence the accuracy of the density estimation results, and the choice of the coordinates influences the accuracy of the density estimation results. One way to select them is to use the least-squares method or the maximum likelihood cross-entropy method recommended by original method authors [46]. The authors propose the parameters H_2 and H_3 [46] to select $2H$.

2.3. Logspline Estimation (LSDE)

This subsection describes the logspline estimation (LSDE) calculation. One-dimensional polynomial splines are called partial polynomials of a certain degree. Breakpoints that contain a transition from one polynomial to another are called nodes. Suppose that the vector $t = (t_1, \ldots, t_K) \in R^K$ defines a set of such K points. Splines describe smooth connections, showing how different areas are separated by nodes [74]. These constraints are precisely defined by expressing partial polynomials in the number of continuous derivatives s. These include partially linear curves. If there are no restrictions, breakpoints are allowed in the nodes of these functions. Assuming that the functions are globally continuous, it is required that the individual linear parts meet at each node. If greater smoothness is needed (for continuous first-order derivatives), then the flexibility of the nodes is lost. Moreover, the curves are considered simple linear functions. The term "linear spline" is applied to a continuous partial linear function in the literature on approximation theory. Accordingly, the term "cubic spline" is assigned to continuous cubic functions with second-order continuous derivatives and nodes that allow jumps of third-order derivatives. If the polynomial degree is b and the vector of the nodes is t, then the set of polynomial splines with s continuous derivatives forms a linear space. For example, a set of linear splines with nodes in the sequence t is defined by function

$$1, x, (x - t_1)+, \ldots, (x - t_K) +. \quad (14)$$

Here $(\cdot)_+ = \max(\cdot, 0)$. We will rely on this set as the base of space. The base of the spline space of degree b and s smoothness consists of monomial whose form $(x - t_k)_+^{s+j}$, here $1 \le j \le b - s$. Using this formula, in the case of classical cubic splines, where $b = 3$ and $s = 2$, the base consists of elements

$$1, x, x^2, x^3, (x - t_1)^3+, \ldots, (x - t_K)^3 +. \quad (15)$$

From the model point of view, this base is convenient because the individual functions at the nodes are merged. In expressions (14) and (15), each function is precisely associated with one of the nodes, and removing this function essentially corresponds to removing the node itself. It is known that the numerical properties of functions (14) and (15) are poor. For example, the solution matrix deteriorates as rapidly as the number of nodes decreases in linear regression problems. A practical alternative is the so-called B-spline base [75,76]. These functions are designed to be supported in several contiguous intervals defined by nodes ($b + 1$ contiguous intervals are used for the smoothest splines). Suppose we can find the basis for splines of space $B_1(x; t)$, …, $B_J(x; t)$ b with smoothness s and a sequence of nodes t so that any function in space can be written as $g(x; \beta, t) =$

$\beta_1 B_1(x;t)+\ldots+\beta_J B_J(x;t)$. Where the corresponding coefficient vector is $\beta = (\beta_1, \ldots, \beta_J)'$. As seen from (14) and (15), then spline spaces of maximum smoothness are used $J = K + b + 1$.

According to the title of the subsection, the object of this analysis is the logarithmic density. Suppose X is a random vector that takes values from the interval (L, U). In the individual case, L and U can be $\pm\infty$. The parameters L and U are set to $2t_1 - t_2$ and $2t_K - t_{K-1}$, respectively. If $\beta_1 \geq 0$ or $\beta_{K-1} \geq 0$, then the adjustment is made $2L_{old} - t_1$ and $U_{new} = 2U_{old} - t_K$ is performed. The method of Kooperberg and Stone [52,77–79], known as logspline, is implemented with cubic spline. The cubic spline is described in (15). These functions are also continuously differentiated, and the partial polynomials are defined accordingly in the sequence of nodes $t = (t_1, \ldots, t_K)$. In each interval $[t_1, t_2], \ldots, [t_{K-1}, t_K]$ cubic splines are also cubic polynomials, but at the edges $(L, t_1]$ and $[t_K, U)$ are linear functions. The minimum number of nodes is $K \geq 3$ (otherwise, a linear function or constant can be obtained). The basis form is $1, B_1(x; t), \ldots, B_J(x; t)$, where $J = K - 1$.

It is said that the vector $\beta = (\beta_1, \ldots, \beta_J)' \in R^J$ exists then $C(\beta, t) = log\left(\int_L^U exp(\beta_1 B_1(x;t)+\ldots+\beta_J B_J(x;t))dx\right) < \infty$. Suppose B denotes a set of such possible vectors. After selecting $\beta \in B$, the family of positive density functions in the interval (L, U) is defined the form of which is

$$g(x;\beta,t) = exp(\beta_1 B_1(x;t)+\ldots+\beta_J B_J(x;t) - C(\beta,t)), L < x < U. \quad (16)$$

Now, having a random sample n of magnitude $X(1), \ldots, X(n)$ from the interval (L, U) with an unknown density function f, the logical probability function corresponding to the model of logsplines (16) is

$$l(\beta,t) = \sum_i \log(g(X_i;\beta,t)) \ = \sum_i \sum_j \beta_j B_j(X_i;t) - nC(\beta,t), \beta \in B. \quad (17)$$

where estimation of maximum likelihood $\hat{\beta} = argmax_{\beta \in B}\, l\,(\beta, t)$ and an estimate of density $\hat{f} = g(x;\hat{\beta},t)$, $L < x < U$.

Let us say that during the stepwise determination procedure, the sequence of models is denoted by v, the vth model has J_v base functions. The Generalized Akaike Information Criterion (AIC) selects the best model [80]. Suppose that $\hat{l}_v$ defines the estimate of the logic-likelihood function (17) for the vth model. The Equation defines the Akaike information criterion $\mathrm{AIC}_{a,v}(t) = -2\hat{l}_v(t) + aJ_v$ for which the model has a loss parameter a. From many models, the one whose value of v minimizes $\mathrm{AIC}_{a,v}$. Stone [52] recommends the use of $a = \log n$.

2.4. PPDE Algorithm. Estimation of the Projection Density of the Target

The projection pursuit density estimator (PPDE) proposed by Friedman is based on the target projection and consistent projection Gaussianization. The essence of J. H. Friedman and coauthors [54,55,81] in estimating the target projection density is to search for "interesting", small-measurement data projections. The distribution structures, where the projections have distributions that are very different (in the sense of some projection index) from Gaussian. Huber [82] made a heuristic proposal to consider the Gaussian distribution as the least interesting. This proposal is based on the facts that:

- The multidimensional Gaussian distribution is entirely defined by its linear structure (mean and covariance matrices). Therefore, it is desired to feel a data structure independent of the correlation and linear transformations such as the scale parameter.
- All projections of a multidimensional Gaussian distribution are also Gaussian distributions. Thus, if the projection differs insignificantly from the Gaussian distribution, it indicates that distribution is also close to the Gaussian.
- For multidimensional data with a structure in multiple projection directions, many projections will have a distribution close to normal. This statement follows from the central limit theorem.
- In the case of constant variance, the Gaussian distribution is considered to be the least informative.

Friedman developed Huber's idea and proposed an algorithm called exploratory target projection to estimate multidimensional non-parametric density. This procedure consists of five steps:

(1) Data standardization simplifies layout, scalability, and correlation structures;
(2) Projection index: the degrees of 'interest' in various directions are determined.
(3) Optimization strategy: search for the direction in which the projection index is the largest.
(4) Data transformation: the one-dimensional density is calculated in the chosen direction, and the data are multiplied.
(5) Density formation: multidimensional density is formed from the calculated one-dimensional densities. Multidimensional density is a function of one-dimensional densities.

The following projection index construction has been proposed. It is known that all projections of a multidimensional Gaussian distribution are one-dimensional Gaussian distributions. If the distribution in one direction is not Gaussian, then the multidimensional distribution is also not Gaussian. Therefore, the projection index $I(\tau)$ shows how far the one-dimensional density $f_\tau(y)$ is in the direction $\tau(Y = \tau' Z)$ from the Gaussian distribution when Z is a standardized quantity [83]:

$$\tilde{I}(\tau) = \int_{-\infty}^{\infty}(f_\tau(y) - \phi(y))^2 dy, \text{ where } \phi(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}. \quad (18)$$

The projection direction τ, which maximizes the projection of a distribution $\tilde{I}(\tau)$, is chosen to highlight the multimodal or other nonlinear structure of that distribution. We transform the data y by equality $R = 2\Phi(Y) - 1 = 2\Phi(\tau' Z) - 1, R \in [-1, 1]$, where $\Phi(u)$ is a function of the standard normal distribution $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{\frac{-t^2}{2}} dt$. The distribution density of the transformed quantity R, function $f_R(r)$ can be rewritten as

$$f_R(r) = \frac{f_\tau(y)}{\left|\frac{\partial r}{\partial y}\right|} = \frac{f_\tau(y)}{2\phi(y)}. \quad (19)$$

Equation (18) can be rewritten by changing the variable y to r: $\tilde{I}(\tau) = \int_{-1}^{1} 2\phi(y)(f_R(r) - 1/2)^2 dr = \int_{-1}^{1} 2\phi(\Phi^{-1}(\frac{R+1}{2}))(f_R(r) - 1/2)^2 dr$. Friedman [55] proposed a slightly different form of the projection index $I(\tau)$, taking the integrated square error as a measure of R inequality:

$$I(\tau) = \int_{-1}^{1}(f_R(r) - 1/2)^2 dr = \int_{-1}^{1} f_R^2(r) dr - 1/2. \quad (20)$$

Note that if the distribution of Y is Gaussian, then $f_R(r) = \frac{1}{2}$, ∀r, and the projection index $I(\tau)$ is zero. The more the Y distribution differs from the normal, the higher the value of the index $I(\tau)$. Since $R \in [-1, 1]$, $f_R(r)$ can be decomposed by orthogonal Lagrangian polynomials, $\{\psi_j\}_{j=0}^{\infty}$, i.e., $f_R(r) = \sum_{j=0}^{\infty} b_j \psi_j(r)$:

$$I(\tau) = \int_{-1}^{1} f_R^2(r) dr - 1/2 = \int_{-1}^{1}\left[\sum_{j=0}^{\infty} b_j \psi_j(r)\right] f_R(r) dr - 1/2. \quad (21)$$

An iterative expression defines orthogonal Lagrangian polynomials. $\psi_0(r) = 1$ and $\psi_1(r) = r.\psi_j(r) = \frac{(2j-1)r\psi_{j-1}(r) - (j-1)\psi_{j-2}(r)}{j}$, then $j \geq 2$. It follows from the orthogonality property that the coefficients b_j can be calculated as follows $b_j = \frac{2j+1}{2}\int_{-1}^{1} \psi_j(r) f_R(r) dr = \frac{2j+1}{2} E_R[\psi_j(r)] = \frac{2j+1}{2}\frac{1}{n}\sum_{t=1}^{n} \psi_j(2\Phi(Y(t)) - 1)$, where $\int_{-1}^{1} \psi_j(r) f_R(r) dr = E_R[\psi_j(r)]$ is the mean of the sample approximates the expression. Thus, equality can be written as

$$I(\tau) = \int_{-1}^{1} f_R^2(r) dr - 1/2 = \sum_{j=1}^{s} \frac{2j+1}{2} E_R^2[\psi_j(r)]. \quad (22)$$

It should be noted that the infinite amount has been changed to finite. Such a change has advantages: the sum is calculated faster, giving robustness to the projection index. By summing only a finite number of members, the slowly fading "tails" of the projection

distributions have a more negligible effect on the value of the projection index. Therefore, it is suggested to choose $4 \le s \le 7$.

There are many methods for finding "interesting" projections. The method used in this research for finding the 'most interesting' projection direction is a mixed optimization strategy [55,64,84]. After defining the analytical expression of the projection index, its gradient in the projection direction τ is obtained as follows

$$\frac{\partial I}{\partial \tau} = \frac{2}{\sqrt{2\pi}} \sum_{j=1}^{s} (2j+1) E[\psi_j(r)] E[\psi_j'(r) e^{-y^2/2} (z - \tau y)]. \quad (23)$$

Here, the Lagrangian polynomial derivative is calculated by an iterative formula: $\psi_1'(r) = 1$, then $\psi_j'(r) = r\psi_{j-1}'(r) + j\psi_{j-1}(r)$, then $j \ge 1$. Initially, an approximate step optimizer is found by searching in the directions of the main components and their combinations so that the initial convergence to the maximum can be achieved quickly. Then, the approximate step optimizer (steepest ascent) quickly selects the projections required to ascend to the (local) maximum of the projection index. The projection index is used to search for 'interesting' data projections. However, it is usually not enough to find a single projection to reasonably estimate the multidimensional density. In general, "interesting" directions do not have to be orthogonal and may require more projection directions than the data dimension. Therefore, when estimating density by targeted projection, the so-called deletion of the data structure is applied. A nonlinear scale transformation is performed, found in the projection direction, so the distribution of the transformed data becomes normal. This operation ensures that the same direction as before was not found when searching for another projection direction.

The deletion of the data structure is based on the fact that if the projection of one-dimensional data projection $\tau' Z$ has a distribution density $f_\tau(y)$ and a corresponding distribution function F_τ, then the random variable is equal to

$$\tilde{Y} = \Phi^{-1}(F_\tau(Y)), \quad (24)$$

where Φ^{-1} is the inverse of the standard normal distribution. Friedman [55] proposed to calculate the empirical estimate of the distribution function as follows $\hat{F}_\tau(y) = rank(Y)/n - \frac{1}{2n}$, where $rank(y)$ is the rank of Y in the whole sample of size n. Unfortunately, this estimate is not accurate and often results in a very uneven density function. By denoting $Z^{(0)} = Z$, we will discuss how $Z^{(k-1)}$ is obtained from $Z^{(k)}$. Based on Equation (14), $Z^{(k)}$ can be defined as

$$Z^{(k)} = Z^{(k-1)} + [\Phi^{-1}\left(F_\tau\left(\tau' Z^{(k-1)}\right)\right) - \tau' Z^{(k-1)}]\tau. \quad (25)$$

The same procedure is performed to find the 'most interesting' projection with $Z^{(k)}$ searching for a new direction. This sequence is repeated until the multidimensional distribution becomes close to the Gaussian distribution in all directions. It has been observed [55] that gaussianization in one direction disrupts normalcy in the directions previously studied so that their projection index $I(\tau)$ is no longer zero. However, studies show [54] that the changes in results are minimal. Multidimensional density is calculated from one-dimensional density estimates.

The relationship between the multidimensional densities $Z^{(k)}$ and $Z^{(k-1)}$ (where $Z^{(k)}$ is the structure of the distant data $Z^{(k-1)}$ along the k-th projection $\tau(k)$) is $f_{\tau(k)}(z^{(k)}) = \frac{f_{\tau(k-1)}(z^{(k-1)})}{|J_k(z^{(k-1)})|}$ and $f_{\tau(k-1)}(z^{(k-1)}) = f_{\tau(k)}(z^{(k)})|J_k(z^{(k-1)})|$, here is the Jacobian $J_k(z^{(k-1)}) = \frac{\partial z^{(k)}}{\partial z^{(k-1)}} = \frac{\partial (Uz^{(k)})}{\partial (Uz^{(k-1)})} = \frac{\partial y^{(k)}}{\partial y^{(k-1)}} = \frac{f_{\tau(k)}(y^{(k-1)})}{\phi(y^{(k)})} = \frac{f_{\tau(k)}(\tau'(k) z^{(k-1)})}{\phi(\tau'(k) z^{(k)})} \ge 0$.

Starting from the initial multidimensional data $Z^{(0)}$ gaussianization procedure is performed for each "interesting" projection found by $I(\tau)$. After a certain number, the projections' multidimensional data $Z^{(M)}$ differ slightly from the normal distribution. Density $Z^{(0)}$ can be calculated as follows

$$f(z^{(0)}) = f_{\tau(1)}(z^{(1)})J_1(z^{(0)}) = f_{\tau(2)}(z^{(2)})J_2(z^{(1)})J_1(z^{(0)}) = f_{\tau(M)}(z^{(M)})\prod_{k=1}^{M} J_k(z^{(k-1)})$$
$$\approx \phi(z^{(M)})\prod_{k=1}^{M} J_k(z^{(k-1)}) = \phi(z^{(M)})\prod_{k=1}^{M}\frac{f_{\tau(k)}(\tau'(k)z^{(k-1)})}{\phi(\tau'(k)z^{(k)})}. \quad (26)$$

The one-dimensional density of the projected data $f_{\tau(k)}(\tau'(k)z^{(k-1)})$ is calculated according to Equation (18) or more precisely

$$f_{\tau(k)}\left(\tau'(k)z^{(k-1)}\right) = \phi\left(\tau'(k)z^{(k-1)}\right)\sum_{j=0}^{s}\frac{2j+1}{n}\sum_{t=1}^{n}\psi_j\left(r_t^{(k-1)}\right)\psi_j\left(r^{(k-1)}\right). \quad (27)$$

Then, replacing the unknown one-dimensional distribution densities on the right-hand side (26) with their statistical estimates, we obtain

$$\hat{f}(z) = \phi(z^{(M)})\prod_{k=1}^{M}\frac{\hat{f}_{\tau(k)}(\tau'(k)z^{(k-1)})}{\phi(\tau'(k)z^{(k)})}. \quad (28)$$

The target projection density estimate is calculated relatively quickly because of the shape of the multivariate projection index and the iterative relationship between polynomials.

2.5. Inversion Formula

When examining approximations of parametric methods, it should be emphasized that as the data dimension increases, the number of model parameters increases rapidly, making it more difficult to find accurate parameter estimates. One-dimensional data projections $X_\tau = \tau' X$ density f_τ is much easier to find than multidimensional data density f because there exists a mutually unambiguous correspondence, $f \leftrightarrow \{f_\tau, \tau \in R^d\}$. It is quite natural to try to find the multidimensional density f using the density estimates $\hat{f}_\tau$ of one-dimensional observational projections. It should be noted that in the case of the mixture, when the distributions are Gaussian, the projections of observations are also distributed according to the (one-dimensional) Gaussian mixture model

$$f_\tau(x) = \sum_{k=1}^{q} p_{k,\tau}\varphi_{k,\tau}(x) = f_\tau(x, \theta_\tau). \quad (29)$$

Here $\varphi_{k,\tau}(x) = \varphi\left(x;\, m_{k,\tau}, \sigma_{k,\tau}^2\right)$– one-dimensional Gaussian density. The parameter θ of the multidimensional mixture. The distribution parameters of the data projections $\theta_\tau = \left(p_{k,\tau}, m_{k,\tau}, \sigma_{k,\tau}^2\right),\ k = 1, \ldots, q$ are related by equations: $p_{j,\tau} = p_j, m_{j,\tau} = \tau' M_j$ and $\sigma_{j,\tau}^2 = \tau' R_j \tau$. Using the inversion formula

$$f(x) = \frac{1}{(2\pi)^d}\int_{R^d} e^{-it'x}\psi(t)dt, \quad (30)$$

where $\psi(t) = Ee^{it'x}$ denotes the characteristic function of the random variable X. Marking $u = |t|,\ \tau = t/|t|$ and changing the variables to a spherical coordinate system, density is written

$$f(x) = \frac{1}{(2\pi)^d}\int_{\tau:\,|\tau|=1} ds \int_0^\infty e^{-iu\tau'x}\psi(u\tau)u^{d-1}du. \quad (31)$$

Here, the first integral is understood as the surface integral of the unit sphere. After noting the characteristic function of the projection of the observed random variable as $\psi_\tau(u) = Ee^{iu\tau' X}$. Then equality $\psi(u\tau) = \psi_\tau(u)$ holds. By selecting the set T of projection directions evenly spaced on the sphere and replacing the characteristic function with its estimate ($\hat{f}(x)$) a formula

$$\hat{f}(x) = \frac{A(d)}{\#T}\sum_{\tau\in T}\int_0^\infty e^{-iu\tau'x}\hat{\psi}_\tau(u)u^{d-1}e^{-hu^2}du, \quad (32)$$

is obtained to calculate the estimate [85,86]. Here and # continue to denote the number of elements in the set T. Using the d-meter ball volume

$$V_d(R) = \frac{\pi^{\frac{d}{2}}R^d}{\Gamma\left(\frac{d}{2}+1\right)} = \begin{cases} \dfrac{\pi^{\frac{d}{2}}R^d}{\left(\frac{d}{2}\right)!}, & \text{then } d \bmod 2 \equiv 0 \\ \dfrac{2^{\frac{d+1}{2}}\pi^{\frac{d-1}{2}}R^d}{d!!}, & \text{then } d \bmod 2 \equiv 1 \end{cases}, \tag{33}$$

the constant $A(d)$ depending on the data dimension can be calculated using

$$A(d) = \frac{\left(V_d(1)\right)'_R}{(2\pi)^d} = \frac{d2^{-d}\pi^{-\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}. \tag{34}$$

Computer simulation studies have shown that the density estimates obtained using the inversion formula are not smooth. Therefore, in formula (32), an additional multiplier e^{-hu^2} is used below the integral sign. This multiplier further smoothes the estimate $\hat{f}(x)$ (32) with the Gaussian kernel function. This form of the multiplier allows the value of the integral to be calculated analytically. The number of clusters and Gaussian mixture parameters was selected using the constructive procedure and software developed at the Lithuanian Institute of Mathematics and Informatics, applying the w^2 type criterion [87]. Formula (32) can be used for various estimates of the characteristic function of the projected data. We will discuss the two methods used in this work.

One of them is based on the density approximation of the Gaussian distribution mixture model. In the present case, after replacing the parameters of the Gaussian mixture with their statistical estimates ($\hat{p}_{k,\tau} = p_k, \hat{m}_{k,\tau} = \tau' M_k, \sigma^2_{k,\tau} = \tau' R_k \tau$) (Page 10), the following parametric estimate

$$\hat{\psi}_\tau(u) = \sum_{k=1}^{\hat{q}_\tau} \hat{p}_{k,\tau} e^{iu\hat{m}_{k,\tau} - u^2\hat{\sigma}^2_{k,\tau}/2} \tag{35}$$

of the characteristic function is used, and adding (32) to (35) gives

$$\begin{aligned} \hat{f}(x) &= \frac{A(d)}{\#T}\sum_{\tau\in T}\sum_{k=1}^{\hat{q}_\tau}\hat{p}_{k,\tau}\int_0^\infty e^{iu(\hat{m}_{k,\tau}-\tau' x)-u^2\left(h+\hat{\sigma}^2_{k,\tau}/2\right)}u^{d-1}du \\ &= \frac{A(d)}{\#T}\sum_{\tau\in T}\sum_{k=1}^{\hat{q}_\tau}\hat{p}_{k,\tau}I_{d-1}\left(\frac{\hat{m}_{k,\tau}-\tau' x}{\sqrt{\hat{\sigma}^2_{k,\tau}+2h}}\right)\left(\sqrt{\hat{\sigma}^2_{k,\tau}+2h}\right)^{-d} \end{aligned} \tag{36}$$

and where $I_j(y)$ can be written as

$$I_j(y) = \mathrm{Re}\left[\int_0^\infty e^{iyz-z^2/2}z^j dz\right]. \tag{37}$$

It should be noted that only the real part of the expression can be considered here. The sum of the imaginary parts must be equal to zero. Because the density estimate $\hat{f}(x)$ can acquire only real values. The chosen form of the smoothing multiplier e^{-hu^2} allows relating the smoothing parameter h to the variances of the projection clusters—in the calculations, the variances are increased by $2h$. How to calculate expression (37) is given in Appendix B.

2.6. Modified Density Estimate of the Inversion Formula

One of the disadvantages of the inversion formula method defined in (32) is that the Gaussian distribution mixture model described by this estimate (where $f_k = \varphi_k$) evaluates well only the density of observations close to it. However, when approximating the

density under study with a mixture of Gaussian distributions, the estimation of the density of the inversion formula often becomes complicated due to a large number of components with low a priori probabilities. Their number can be reduced by introducing a noise cluster—the modified algorithm based on a multidimensional Gaussian distribution mixture model. Let us use the inversion formula (30). The parametric estimate of the characteristic function of uniform distribution density can be calculated as follows

$$\hat{\psi}(u) = \frac{2}{(b-a)u}\sin\frac{(b-a)u}{2}\cdot e^{\frac{iu(a+b)}{2}}. \tag{38}$$

In uniform distribution density function (38), b is the maximum value, and a is the minimum value. In the density estimate calculation formula (32), construct the estimation of the characteristic function as a union of the characteristic functions of a mixture of Gaussian distributions and a uniform distribution with corresponding a priori probabilities as follows

$$\hat{\psi}_\tau(u) = \sum_{k=1}^{\hat{q}_\tau} \hat{p}_{k,\tau} e^{iu\hat{m}_{k,\tau} - u^2\hat{\sigma}^2_{k,\tau}/2} + \hat{p}_{0,\tau}\frac{2}{\big(b(\tau)-a(\tau)\big)u}\sin\frac{\big(b(\tau)-a(\tau)\big)u}{2}\cdot e^{\frac{iu\left(a(\tau)+b(\tau)\right)}{2}}. \tag{39}$$

Here the second term describes a uniform distributed noise cluster and $\hat{p}_0$ is the weight of the noise cluster. Based on the parameters of the uniform distribution and the projected data, we can write

$$a(\tau) = (\tau' x)_{\min} - \frac{(\tau' x)_{\max} - (\tau' x)_{\min}}{2(n-1)} \text{ and} \tag{40}$$

$$b(\tau) = (\tau' x)_{\max} + \frac{(\tau' x)_{\max} - (\tau' x)_{\min}}{2(n-1)}. \tag{41}$$

Using notations such as (36), we can write

$$\begin{aligned}\hat{f}(x) = \frac{A(d)}{\#T}\sum_{\tau\in T}\Bigg[&\sum_{k=1}^{\hat{q}_\tau}\hat{p}_{k,\tau}I_{d-1}\left(\frac{\hat{m}_{k,\tau}-\tau' x}{\sqrt{\hat{\sigma}^2_{k,\tau}+2h}}\right)\left(\hat{\sigma}^2_{k,\tau}+2h\right)^{-\frac{d}{2}}\\ &+\frac{2\hat{p}_{0,\tau}}{b(\tau)-a(\tau)}J_{d-2}\left(\frac{a(\tau)+b(\tau)-2\tau' x}{2\sqrt{2h}},\frac{b(\tau)-a(\tau)}{2\sqrt{2h}}\right)\cdot(2h)^{-\frac{d-1}{2}}\Bigg].\end{aligned} \tag{42}$$

where the expression $I_j(y)$ is the same as (37) and its value is $I_j(y) = C_j(y)$ and

$$J_j(y,z) = \mathrm{Re}\left[\int_0^\infty e^{iyu-u^2/2}\cdot\sin zu\cdot u^j du\right]. \tag{43}$$

By integrating, we get

$$\begin{aligned}&\int_0^\infty e^{iyu-u^2/2}\cdot\sin zu\cdot u^j du = \int_0^\infty(\cos yu + i\sin yu)\cdot\sin zu\cdot e^{-u^2/2}\cdot u^j du\\ =&\int_0^\infty\left(\frac{\sin(y+z)u+\sin(z-y)u}{2}+i\frac{\cos(y-z)u-\cos(y+z)u}{2}\right)\cdot e^{-u^2/2}\cdot u^j du\\ &=\frac{1}{2}S_j(y+z)+\frac{1}{2}S_j(z-y)+i\frac{1}{2}C_j(y-z)-i\frac{1}{2}C_j(y+z).\end{aligned} \tag{44}$$

the above formula uses the variables $S_j(y)$ and $C_j(y)$, the calculation of which is given in formulas (52) and (53) in Appendix B.

3. Materials and Methods

Density estimation algorithms were presented in the previous section. The Monte Carlo method was used in this study. Such a comparison of algorithms allows us to measure the real observation density values and evaluate algorithms' efficiency. For the research, we used multidimensional (d = 2, 5, 10, 15) distributions of the Cauchy mixture

$$\sum_{j=1}^{q} p_j C(x, m_j, u_j). \quad (45)$$

Additionally, $C(x, m_j, u_j)$ is defined as follows

$$C(x, m_j, u_j) = \prod_{k=1}^{d} \frac{u_{jk}}{\pi[u_{jk}^2 + (x_k - m_{jk})^2]}. \quad (46)$$

Calculations were performed using sample sizes of n = 50, 100, 200, 400, 800 while changing the number of distributions, their weights, and centers (see. Table 1). In each case, 100,000 samples were generated.

Table 1. Parameters table.

Number of Components	Proportions of Components	Location Parameters	Separation Size of Locations
q = 2	$p_1 = (1 - p_2)$, p_2 = 0.1, 0.3, 0.5	m_1 = (0, 0), m_2 = (0.5i, 0.5i)	i = 1, 2,…, 6
q = 3	$p_1 = p_2 = (1 - p_3)/2$, p_3 = 0.1, 1/3, 0.8	m_1 = (0, 0), m_2 = (0.5i, 0.5i), m_3 = (0.5i, 0)	i = 1, 2,…, 6
q = 4	$p_1 = p_2 = p_3 = (1 - p_4)/3$, p_4 = 0.1, 0.25, 0.7	m_1 = (0, 0), m_2 = (0.5i, 0.5i), m_3 = (0.5i, 0), m_4 = (0, 0.5i)	i = 1, 2,…, 6
q = 2	$p_1 = (1 - p_2)$, p_2 = 0.1, 0.2, 0.3, 0.4, 0.5	m_1 = (0, 0, 0, 0, 0), m_2 = (0.5i, 0.5i, 0.5i, 0.5i, 0.5i)	i = 1, 2,…, 6
q = 3	$p_1 = p_2 = (1 - p_3)/2$, p_3 = 0.1, 0.2, 1/3, 0.4, 0.6, 0.8	m_1 = (0, 0, 0, 0, 0), m_2 = (0.5i, 0.5i, 0.5i, 0.5i, 0.5i), m_3 = (0.5i, 0.5i, 0, 0, 0)	i = 1, 2,…, 6
q = 4	$p_1 = p_2 = p_3 = (1 - p_4)/3$, p_4 = 0.1, 0.16, 0.25, 0.4, 0.7	m_1 = (0, 0, 0, 0, 0), m_2 = (0.5i, 0.5i, 0.5i, 0.5i, 0.5i), m_3 = (0.5i, 0.5i, 0, 0, 0), m_4 = (0, 0, 0.5i, 0.5i, 0.5i)	i = 1, 2,…, 6

In cases of d = 10, 15, the same weights were used as in d = 5. Additionally, centres are located on the apexes of the hypercube.

Algorithms used in the research: AKDE—adaptive kernel, PPDE—projection pursuit, LSDE—logspline, SKDE—semi-parametric kernel, IFDE—inversion formula, MIDE—inversion formula with noise cluster. In IFDE and MIDE methods are used mixture parameters, calculated with a program made in an institute of Mathematics and Informatics (Vilnius) [87].

Selection of parameters in the density estimation procedure. In this study, the Monte Carlo method aimed to perform the accuracy of the non-parametric estimates of distribution density previously described in the methodological sections (AKDE, PPDE, LSDE, SKDE, IFDE, MIDE) comparative analysis. The authors [34] propose to collect the value of the sensitivity parameter (γ, see. AKDE method step 3) used in the AKDE method from the set {0.2; 0.4; 0.6; 0.8}. The specific value of the parameter is determined by a probabilistic cross-check [88,89]. In the SKDE, all possible values of the sub-vector Y dimension s ($1 \leq s \leq d-1$, where d is dimensions, see page 5) and their corresponding coordinates were reselected. The most factual errors were used to compare the results with other studied methods. The LSDE method minimizes the Akaike information criterion by selecting the number of baseline spline points [78]. The computer program for calculating this estimate is provided in the R package and was used in the study. Akaike information criterion $AIC = -2l(t) + aJ(t)$, J—degree of spline, $a = \log(n)$, l—probability function used to select the spline coefficients. The MIDE method has a smoothing parameter, h. The chosen form of the smoothing multiplier e^{-hu^2} allows relating the smoothing parameter h to the

variances of the projection clusters. Modelling studies have shown that this method is sensitive to parameter selection. If h is set too low, the estimate becomes very slick and has large errors. Excessive smoothing of the density estimate does not greatly affect its quality. In the studies, it was observed that the estimation becomes uneven due to the similarity of the values of the observations projected in some directions, thus distinguishing low-weight components with small dispersions. The smoothing parameter (h) as well as the specific value of the noise cluster weight (probability) from the set {0.05; 0.1; 0.15; 0.2; 0.3; 0.4} are selected by cross-checking the least squares [65]. The vector of the estimate parameters is searched for in such a way that it minimizes the integrated square error

$$\Theta = \arg\min_{\Theta} \int_{-\infty}^{\infty} (\widehat{f_\Theta}(x) - f(x))^2 dx = arg \min_{\Theta}\{\left\|\widehat{f_\Theta}(x)\right\|_2^2 - \frac{2}{n}\sum_{t=1}^{n} \widehat{f_\Theta}(X(t))\}, \quad (47)$$

where Θ is the evaluated parameter and $F(x)$ is the observed random variable distribution function. Changing an unknown distribution function to an empirical distribution function yields an expression for the parameter estimate

$$\widehat{\Theta} = \arg\min_{\Theta}(\left\|\widehat{f_\Theta}(x)\right\|_2^2 - \frac{2}{n}\sum_{t=1}^{n} \widehat{f_\Theta}(X(t)|t), \quad (48)$$

where $\widehat{f_\Theta}(x|t)$ is the value of the estimate at point x, which is calculated by subtracting the value of $X(t)$ from the observations. In addition, empirical research suggests that it is better to look for a maximum local minimum point rather than a global minimum [90]. Using PPDE method and following the recommendation of the paper [38], the order of the spread was $4 \le s \le 6$ (see Page 9), and the projection directions were chosen to maximize the value of the estimate of the design index (2) recommended by J. H. Friedman

$$I(\alpha) = \int_{-1}^{1} f_r^2(r)dr - \frac{1}{2} = \sum_{j=1}^{J} \frac{2j+1}{2} E_r^2[\psi_j(r)]. \quad (49)$$

4. Results and Discussion

This section presents the main results obtained during the simulations. We calculate the mean absolute error and (50) mean absolute percentage error (51) to evaluate the accuracy.

$$\delta_1 = \frac{1}{n}\sum_{t=1}^{n}\left|f(x(t)) - \hat{f}(x(t))\right| \cong \int\left|f(x) - \hat{f}(x)\right|f(x)dx. \quad (50)$$

$$\delta_2 = \frac{2}{n}\sum_{t=1}^{n}\left|\frac{f(x(t)) - \hat{f}(x(t))}{f(x(t)) + \hat{f}(x(t))}\right| \cong \int\left|f(x) - \hat{f}(x)\right|dx. \quad (51)$$

The result tables (Tables 2 and A1–A10) provide 100,000 samples densities mean absolute percentage error. The values in parentheses provide information about the standard deviation of errors. The best results in these tables are bolded and underlined. According to Table 2, it is concluded that when $n = 100$, $d = 5$, the best results are obtained by SKDE and MIDE methods. Based on Table A2, it can be observed that when $q = 2$, $n = 200$, the best results are obtained using SKDE and MIDE methods. According to Table A3, it is concluded that when $q = 3$, $n = 200$, in the case of highly overlapping distributions ($i = 1$, 2), the best results are obtained by the SKDE method, and in the case of more isolated distributions ($i \ge 3$)—by the MIDE method. Based on Table A4, it can be observed that when $q = 3$, $n \ge 400$, the best results are obtained by SKDE, while the second-best method is MIDE. According to Table A5, it is concluded that when $q = 4$, $n = 400$, in the case of highly overlapping distributions ($i \le 3$), the best results are obtained by the SKDE method and in the case of more isolated distributions ($i \ge 4$)—by the MIDE method. Table A6 shows results of $q = 4$, $n \ge 400$, it can be noticed that, in the case of highly overlapping or average isolated distributions ($i \le 5$), the best results are obtained by the SKDE method and in the case of more isolated distributions ($i = 6$)—by the MIDE method. Tables A7 and

A8 show results of $q = 2$ and $n = 50$, It can be noticed that in all cases highly overlapping or isolated distributions, the best results are obtained by AKDE method and in the case of more isolated distributions ($i = 6$) with $p_1 = 0.6$; $p_2 = 0.4$—by the MIDE method. Tables A9 and A10 show results $q = 3$ and $n = 50$; the best results are obtained by the AKDE method in all cases (highly overlapping or isolated distributions).The LSDE method with huge outliers ($|x - m_j| > 100uj$) is grouped with a more significant number of values closer to the centre of the distribution. With the help of the calculated spline coefficients, the density in the outliers is estimated at a value close to 10^{100}. That is incorrect, and in such cases, the use of this method is not recommended.

Table 2. An example of mean absolute percentage error.

Evaluation Methods		Density $d = 5$; $p_1 = p_2 = p_3 = 1/3$; $n = 100$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	0.8268	0.8257	0.8198	0.8128	0.8066	0.8075
	SD	(0.0760)	(0.0814)	(0.0848)	(0.0827)	(0.0788)	(0.0731)
PPDE	Mean	0.9243	0.9319	0.9303	0.9300	0.9284	0.9250
	SD	(0.0500)	(0.0364)	(0.0375)	(0.0387)	(0.0410)	(0.0433)
LSDE	Mean	0.8043	0.8162	0.8583	0.8611	0.8613	0.8711
	SD	(0.0534)	(0.0540)	(0.0490)	(0.0349)	(0.0434)	(0.0577)
SKDE	Mean	**0.7158**	**0.7144**	**0.7088**	**0.7071**	0.7179	0.7227
	SD	(0.0260)	(0.0905)	(0.0905)	(0.0830)	(0.0631)	(0.0499)
IFDE	Mean	0.94593	0.8886	0.7857	0.8463	0.8761	0.8312
	SD	(0.0362)	(0.1318)	(0.0706)	(0.0380)	(0.1110)	(0.0538)
MIDE	Mean	0.7389	0.7332	0.7235	0.7149	**0.7121**	**0.7219**
	SD	(0.0280)	(0.0221)	(0.0338)	(0.0195)	(0.0208)	(0.0203)

The results for the smaller dimensions ($d = 2$) are presented in Table A1. It can be seen that the best results are obtained using the SKDE method, both in large- and small-scale overlapping cases ($i < 4$). On the other hand, in the case of isolated distributions ($i \geq 5$), good results were obtained by the MIDE method.

In the case of mean absolute percentage error, recommended using the semiparametric kernel when the sample has overlapping distributions. In the case of two dimensions (d~2) and a sample is with overlapping distributions, it is recommended to use the semiparametric kernel method and for isolated distributions, to use the adaptive kernel method.

5. Conclusions

This paper reviewed the most popular and most often used nonparametric density estimation algorithms. The density estimation inversion formula was also presented in this article. It was observed that when a noise cluster is included, the results of the inversion formula improved statistically significantly. Based on the mean absolute error, in the case of higher dimension (d~5) and small samples (n~50), it is recommended to use the adaptive kernel method. If the sample is n~100, then the modified inversion formula method showed the best results. For larger samples with overlapping distributions it is recommended to use a semi-parametric kernel and for more isolated distribution—modified inversion methods. Based on the mean absolute percentage error, it is recommended to use the semiparametric kernel when the sample is with overlapping distributions. In the case of two dimensions (d~2) and a sample is with overlapping distributions, it is recommended to use the semiparametric kernel method. For isolated distributions, it is recommended to use the adaptive kernel method.

Author Contributions: Conceptualization, T.R. and M.L.; methodology, T.R.; software, T.R. and M.L.; formal analysis, T.R. and M.L.; investigation, T.R. and M.L.; writing—original draft preparation, T.R., M.L. and G.Č.; writing—review and editing, M.L. and G.Č.; supervision, T.R. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are thankful to the area editor and the reviewers for giving valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. 10^2 times upscaled mean absolute error with $d = 2$; $p_1 = 0.5$; $p_2 = 0.5$; $n = 100$.

Evaluation Methods		Density $d = 2$; $p_1 = 0.5$; $p_2 = 0.5$; $n = 100$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	1.74	1.41	1.16	1.08	1.02	0.99
	SD	(0.94)	(0.80)	(0.70)	(0.59)	(0.54)	(0.50)
PPDE	Mean	2.21	1.85	1.52	1.32	1.25	1.21
	SD	(0.49)	(0.41)	(0.40)	(0.44)	(0.37)	(0.31)
LSDE	Mean	0.87	0.71	0.78	0.63	0.69	0.69
	SD	(0.43)	(0.20)	(0.08)	(0.08)	(0.04)	(0.09)
SKDE	Mean	**0.63**	**0.61**	**0.52**	**0.52**	**0.51**	**0.51**
	SD	(0.12)	(0.17)	(0.07)	(0.05)	(0.06)	(0.04)
IFDE	Mean	1.69	1.31	0.97	0.75	0.61	0.53
	SD	(0.06)	(0.10)	(0.08)	(0.01)	(0.04)	(0.06)
MIDE	Mean	0.69	0.66	0.57	0.55	**0.51**	**0.51**
	SD	(0.06)	(0.10)	(0.08)	(0.01)	(0.04)	(0.06)

Table A2. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.7$; $p_2 = 0.3$; $n = 200$.

Evaluation Methods		Density d=5; $p_1 = 0.7$; $p_2 = 0.3$; $n = 200$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	0.979	0.801	0.703	0.664	0.655	0.654
	SD	(0.171)	(0.146)	(0.145)	(0.155)	(0.158)	(0.159)
PPDE	Mean	1.001	0.822	0.722	0.681	0.671	0.669
	SD	(0.174)	(0.152)	(0.151)	(0.160)	(0.163)	(0.163)
LSDE	Mean	5.039	4.185	2.632	0.944	0.665	0.660
	SD	(1.265)	(6.747)	(1.081)	(0.138)	(0.140)	(0.112)
SKDE	Mean	**0.857**	**0.759**	0.705	**0.658**	**0.649**	**0.638**
	SD	(0.087)	(0.069)	(0.076)	(0.085)	(0.083)	(0.083)
IFDE	Mean	0.912	0.801	0.721	0.681	0.667	0.666
	SD	(0.133)	(0.149)	(0.151)	(0.160)	(0.162)	(0.163)
MIDE	Mean	0.956	0.788	**0.694**	0.661	0.657	0.640
	SD	(0.162)	(0.154)	(0.144)	(0.152)	(0.158)	(0.163)

Table A3. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.45$; $p_2 = 0.45$; $p_3 = 0.1$; $n = 200$.

Evaluation Methods		Density $d = 5; p_1 = 0.45; p_2 = 0.45; p_3 = 0.1; n = 200$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	0.970	0.725	0.576	0.533	0.504	0.500
	SD	(0.127)	(0.089)	(0.078)	(0.073)	(0.069)	(0.066)
PPDE	Mean	0.992	0.746	0.594	0.528	0.506	0.500
	SD	(0.137)	(0.092)	(0.077)	(0.072)	(0.069)	(0.066)
LSDE	Mean	1.057	0.775	0.652	0.590	0.508	0.503
	SD	(0.164)	(0.203)	(0.650)	(0.491)	(0.067)	(0.081)
SKDE	Mean	**0.6245**	**0.6274**	0.6312	0.629	0.630	0.628
	SD	(0.072)	(0.025)	(0.027)	(0.049)	(0.049)	(0.050)
IFDE	Mean	0.990	0.743	0.589	0.525	0.497	0.499
	SD	(0.136)	(0.091)	(0.076)	(0.071)	(0.071)	(0.067)
MIDE	Mean	0.993	0.746	**0.574**	**0.525**	**0.496**	**0.490**
	SD	(0.137)	(0.092)	(0.077)	(0.072)	(0.069)	(0.066)

Table A4. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.4$; $p_2 = 0.4$; $p_3 = 0.2$; $n = 400$.

Evaluation Methods		Density $d = 5; p_1 = 0.4; p_2 = 0.4; p_3 = 0.2; n = 400$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	0.916	0.689	0.527	0.445	0.410	0.396
	SD	(0.099)	(0.068)	(0.048)	(0.044)	(0.048)	(0.052)
PPDE	Mean	0.937	0.709	0.545	0.461	0.423	0.407
	SD	(0.109)	(0.074)	(0.049)	(0.044)	(0.048)	(0.052)
LSDE	Mean	0.815	0.549	0.511	0.443	0.404	0.401
	SD	(0.007)	(0.063)	(0.151)	(0.094)	(0.040)	(0.030)
SKDE	Mean	**0.655**	**0.499**	**0.413**	**0.388**	**0.385**	**0.384**
	SD	(0.064)	(0.049)	(0.034)	(0.031)	(0.028)	(0.027)
IFDE	Mean	0.937	0.709	0.544	0.460	0.423	0.404
	SD	(0.109)	(0.074)	(0.049)	(0.044)	(0.048)	(0.052)
MIDE	Mean	0.757	0.509	0.415	0.391	0.391	0.388
	SD	(0.109)	(0.074)	(0.049)	(0.044)	(0.048)	(0.052)

Table A5. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.25$; $p_2 = 0.25$; $p_3 = 0.25$; $p_4 = 0.25$; $n = 400$.

Evaluation Methods		Density $d = 5; p_1 = 0.25; p_2 = 0.25; p_3 = 0.25; p_4 = 0.25; n = 400$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	0.912	0.645	0.447	0.351	0.309	0.290
	SD	(0.128)	(0.068)	(0.029)	(0.019)	(0.026)	(0.030)
PPDE	Mean	0.934	0.665	0.464	0.365	0.321	0.299
	SD	(0.145)	(0.089)	(0.048)	(0.031)	(0.033)	(0.035)
LSDE	Mean	0.934	0.676	0.464	0.365	0.321	0.293
	SD	(0.145)	(0.064)	(0.048)	(0.031)	(0.033)	(0.039)
SKDE	Mean	**0.658**	**0.472**	**0.372**	0.345	0.316	0.290
	SD	(0.071)	(0.031)	(0.020)	(0.017)	(0.019)	(0.018)
IFDE	Mean	0.933	0.665	0.464	0.365	0.321	0.299
	SD	(0.145)	(0.089)	(0.048)	(0.031)	(0.033)	(0.035)
MIDE	Mean	0.889	0.622	0.433	**0.341**	**0.307**	**0.281**
	SD	(0.118)	(0.074)	(0.037)	(0.026)	(0.019)	(0.027)

Table A6. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.1$; $p_2 = 0.1$; $p_3 = 0.1$; $p_4 = 0.7$; $n = 400$.

Evaluation Methods		Density $d = 5; p_1 = 0.1; p_2 = 0.1; p_3 = 0.1; p_4 = 0.7; n = 400$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	0.957	0.800	0.678	0.617	0.587	0.571
	SD	(0.131)	(0.127)	(0.103)	(0.090)	(0.087)	(0.087)
PPDE	Mean	0.979	0.821	0.697	0.634	0.603	0.586
	SD	(0.141)	(0.137)	(0.112)	(0.099)	(0.094)	(0.093)
LSDE	Mean	0.979	0.821	0.697	0.634	0.596	0.586
	SD	(0.141)	(0.137)	(0.112)	(0.099)	(0.098)	(0.093)
SKDE	Mean	**0.687**	**0.580**	**0.514**	**0.496**	**0.491**	0.489
	SD	(0.076)	(0.070)	(0.058)	(0.058)	(0.058)	(0.056)
IFDE	Mean	0.979	0.820	0.697	0.634	0.602	0.585
	SD	(0.141)	(0.137)	(0.112)	(0.098)	(0.094)	(0.093)
MIDE	Mean	0.924	0.770	0.652	0.597	0.533	**0.488**
	SD	(0.135)	(0.131)	(0.108)	(0.093)	(0.092)	(0.091)

Table A7. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.5$; $p_2 = 0.5$; $n = 50$.

Evaluation Methods		Density $d = 5; p_1 = 0.5; p_2 = 0.5; n = 50$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	**1.093**	**0.828**	**0.758**	**0.741**	**0.739**	**0.740**
	SD	(0.095)	(0.099)	(0.114)	(0.123)	(0.130)	(0.134)
PPDE	Mean	1.147	0.872	0.794	0.770	0.764	0.762
	SD	(0.157)	(0.150)	(0.157)	(0.156)	(0.159)	(0.160)
LSDE	Mean	2.100	1.997	2.002	2.010	2.013	2.014
	SD	(0.078)	(0.028)	(0.017)	(0.019)	(0.024)	(0.025)
SKDE	Mean	1.149	0.875	0.797	0.773	0.765	0.763
	SD	(0.160)	(0.154)	(0.160)	(0.160)	(0.161)	(0.162)
IFDE	Mean	1.145	0.864	0.780	0.765	0.763	0.757
	SD	(0.163)	(0.137)	(0.140)	(0.150)	(0.163)	(0.156)
MIDE	Mean	1.094	0.860	0.759	0.767	0.742	0.752
	SD	(0.142)	(0.167)	(0.160)	(0.156)	(0.163)	(0.160)

Table A8. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.6$; $p_2 = 0.4$; $n = 50$.

Evaluation Methods		Density $d = 5; p_1 = 0.6; p_2 = 0.4; n = 50$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	**1.138**	**0.872**	**0.770**	**0.748**	**0.745**	0.746
	SD	(0.105)	(0.085)	(0.126)	(0.143)	(0.152)	(0.155)
PPDE	Mean	1.192	0.918	0.808	0.778	0.771	0.769
	SD	(0.150)	(0.136)	(0.172)	(0.178)	(0.181)	(0.182)
LSDE	Mean	2.114	1.995	1.977	1.983	1.986	1.987
	SD	(0.101)	(0.083)	(0.080)	(0.079)	(0.082)	(0.084)
SKDE	Mean	1.195	0.919	0.810	0.780	0.772	0.770
	SD	(0.154)	(0.138)	(0.174)	(0.182)	(0.183)	(0.183)
IFDE	Mean	1.183	0.906	0.802	0.778	0.765	0.769
	SD	(0.142)	(0.125)	(0.163)	(0.185)	(0.176)	(0.185)
MIDE	Mean	1.152	0.882	0.782	0.754	0.747	**0.742**
	SD	(0.136)	(0.124)	(0.155)	(0.175)	(0.176)	(0.180)

Table A9. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.33$; $p_2 = 0.33$; $p_3 = 0.33$; $n = 50$.

Evaluation Methods		Density $d = 5; p_1 = 0.33; p_2 = 0.33; p_3 = 0.33; n = 50$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	**1.166**	**0.828**	**0.634**	0.547	**0.512**	**0.500**
	SD	(0.120)	(0.086)	(0.057)	(0.064)	(0.074)	(0.080)
PPDE	Mean	1.224	0.879	0.677	0.581	0.540	0.523
	SD	(0.184)	(0.128)	(0.107)	(0.108)	(0.108)	(0.109)
LSDE	Mean	2.075	1.934	1.921	1.939	1.938	1.937
	SD	(0.127)	(0.094)	(0.058)	(0.054)	(0.048)	(0.044)
SKDE	Mean	1.226	0.881	0.678	0.583	0.542	0.524
	SD	(0.186)	(0.130)	(0.109)	(0.110)	(0.110)	(0.110)
IFDE	Mean	1.215	0.839	0.649	0.554	0.522	0.513
	SD	(0.175)	(0.099)	(0.110)	(0.102)	(0.110)	(0.111)
MIDE	Mean	1.182	0.834	0.638	**0.545**	0.518	0.501
	SD	(0.167)	(0.124)	(0.097)	(0.101)	(0.106)	(0.106)

Table A10. 10^4 times upscaled mean absolute error with $d = 5$; $p_1 = 0.45$; $p_2 = 0.45$; $p_3 = 0.1$; $n = 50$.

Evaluation Methods		Density $d = 5; p_1 = 0.45; p_2 = 0.45; p_3 = 0.1; n = 50$					
		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$
AKDE	Mean	**1.126**	**0.838**	**0.690**	**0.633**	**0.618**	**0.615**
	SD	(0.112)	(0.126)	(0.063)	(0.053)	(0.061)	(0.067)
PPDE	Mean	1.182	0.882	0.727	0.660	0.640	0.634
	SD	(0.156)	(0.132)	(0.091)	(0.085)	(0.087)	(0.088)
LSDE	Mean	2.101	2.002	1.985	2.010	2.014	2.015
	SD	(0.105)	(0.063)	(0.012)	(0.029)	(0.028)	(0.025)
SKDE	Mean	1.183	0.885	0.729	0.663	0.642	0.635
	SD	(0.157)	(0.134)	(0.094)	(0.089)	(0.090)	(0.090)
IFDE	Mean	1.170	0.859	0.702	0.649	0.624	0.619
	SD	(0.142)	(0.129)	(0.074)	(0.088)	(0.083)	(0.086)
MIDE	Mean	1.142	0.850	0.696	0.639	0.620	0.618
	SD	(0.141)	(0.125)	(0.080)	(0.084)	(0.083)	(0.087)

Appendix B

Calculate expression (36). Marked

$$C_j(y) = \int_0^\infty \cos yz \cdot e^{-z^2/2} \cdot z^j dz \text{ and} \tag{A1}$$

$$S_j(y) = \int_0^\infty \sin yz \cdot e^{-z^2/2} \cdot z^j dz\,. \tag{A2}$$

The Equation holds

$$\int_0^\infty e^{-iyz-z^2/2} z^j dz = C_j(y) + iS_j(y). \tag{A3}$$

Integration in parts results in

$$C_j(y) = e^{-\frac{z^2}{2}} z^{j-1} \cos yz \Big|_0^\infty + \int_0^\infty e^{-\frac{z^2}{2}} \left((j-1) z^{j-2} \cos yz - yz^{j-1} \sin yz \right) dz = \\ = 1_{\{j=1\}} + (j-1) C_{j-2}(y) - y S_{j-1}(y), j \geq 1. \tag{A4}$$

Analogously expressing $S_j(y)$ and taking into account the constraints of the j index, recursive equations are obtained

$$C_j(y) = (j-1) C_{j-2}(y) - y S_{j-1}(y), j \geq 2 \text{ and} \tag{A5}$$

$$C_1(y) = 1 - y S_0(y) \text{ also} \tag{A6}$$

$$S_j(y) = (j-1) S_{j-2}(y) - y C_{j-1}(y), j \geq 2 \text{ and} \tag{A7}$$

$$S_1(y) = y C_0(y), \quad then\ j = 1. \tag{A8}$$

To calculate the functions $C_0(y)$ and $S_0(y)$ it is used that

$$\left(S_0(y) \right)'_y = \int_0^\infty z \cos yz \cdot e^{-z^2/2} dz = C_1(y). \tag{A9}$$

From (A7) and (A10), it is obtained that S_0 satisfies the differential equation $S_0'(y) = 1 - y S_0(y)$. This Equation is solved by spreading S_0 by Taylor series

$$S_0'(y) = \sum_{l=0}^\infty c_{l+1} (l+1) y^{l+1} = 1 - \sum_{l=2}^\infty c_{l-1} y^l. \tag{A10}$$

Comparing the coefficients to similar members, their values are found $c_0 = 0,\ c_1 = 1,\ c_l = -c_{l-2}/l,\ l \geq 2$. Thus,

$$S_0(y) = \sum_{l=0}^\infty \frac{(-1)^l y^{2l+1}}{(2l+1)!!} = y - \frac{y^3}{3!!} + \frac{y^5}{5!!} - \frac{y^7}{7!!} + \cdots. \tag{A11}$$

C_0 is found from expression (50)

$$C_0(y) = \int_0^\infty \cos yz \cdot e^{-z^2/2} dz = \frac{1}{2} \int_{-\infty}^\infty \cos yz \cdot e^{-z^2/2} dz \\ = \frac{1}{2} \int_{-\infty}^\infty (\cos yz - i \sin yz) \cdot e^{-z^2/2} dz = \sqrt{\frac{\pi}{2}} e^{-y^2/2}. \tag{A12}$$

Seeking integral (32) value $I_j(y) = C_j(y)$.

References

1. Duda, R.O.; Hart, P.E. *Pattern Classification and Scene Analysis*; Wiley: New York, NY, USA, 1973; Volume 3.
2. John, G.; Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–20 August 1995.
3. Wang, X.-Z.; He, Y.-L.; Wang, D.D. Non-Naive Bayesian Classifiers for Classification Problems with Continuous Attributes. *IEEE Trans. Cybern.* **2013**, *44*, 21–39, doi:10.1109/tcyb.2013.2245891.
4. Azzalini, A.; Menardi, G. Clustering via nonparametric density estimation: The R package pdf Cluster. *arXiv* **2013**, arXiv:1301.6559.
5. Cuevas, A.; Febrero-Bande, M.; Fraiman, R. Cluster analysis: A further approach based on density estimation. *Comput. Stat. Data Anal.* **2001**, *36*, 441–459, doi:10.1016/s0167-9473(00)00052-9.
6. Campello, R.J.; Kröger, P.; Sander, J.; Zimek, A. Density-based clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1343.
7. Kwak, N.; Choi, C.-H. Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1667–1671, doi:10.1109/tpami.2002.1114861.

8. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238, doi:10.1109/tpami.2005.159.
9. Li, D.; Yang, K.; Wong, W.H. Density Estimation via Discrepancy Based Adaptive Sequential Partition. In Proceedings of the 30th Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Advances in Neural Information Processing Systems 29; pp. 1091–1099.
10. Rothfuss, J.; Ferreira, F.; Walther, S.; Ulrich, M. Conditional density estimation with neural networks: Best practices and benchmarks. *arXiv* **2019**, arXiv:1903.00954.
11. Trentin, E.; Lusnig, L.; Cavalli, F. Parzen neural networks: Fundamentals, properties, and an application to forensic anthropology. *Neural Netw.* **2018**, *97*, 137–151.
12. Trentin, E. Soft-Constrained Neural Networks for Nonparametric Density Estimation. *Neural Process. Lett.* **2017**, *48*, 915–932, doi:10.1007/s11063-017-9740-1.
13. Huynh, H.T.; Nguyen, L. Nonparametric maximum likelihood estimation using neural networks. *Pattern Recognit. Lett.* **2020**, *138*, 580–586, doi:10.1016/j.patrec.2020.09.006.
14. Archambeau, C.; Verleysen, M. Fully nonparametric probability density function estimation with finite gaussian mixture models. In Proceedings of the 5th ICPAR Conference, Calcutta, India, 10–13 December 2003; pp. 81–84.
15. Priebe, C.E. Adaptive mixtures. *J. Am. Stat. Assoc.* **1994**, *89*, 796–806.
16. Scott, D.W. Remarks on fitting and interpreting mixture models. *Comput. Sci. Stat.* **1999**, 104–109.
17. Delicado, P.; Del Río, M. A generalization of histogram type estimators. *J. Nonparametr. Stat.* **2003**, *15*, 113–135, doi:10.1080/10485250306036.
18. Peel, D.; MacLahlan, G. *Finite Mixture Models*; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2000. http://dx.doi.org/10.1002/0471721182.
19. Minnotte, M.C. Achieving higher-order convergence rates for density estimation with binned data. *J. Am. Stat. Assoc.* **1998**, *93*, 663–672.
20. Tapia, R.A.; Thompson, J.R. *Nonparametric Probability Density Estimation*; Johns Hopkins University Press: Baltimore, MD, USA, 1978; p. 176.
21. Jones, M.C.; Samiuddin, M.; Al-Harbey, A.H.; Maatouk, T.A.H. Miscellanea. The edge frequency polygon. *Biometrika* **1998**, *85*, 235–239, doi:10.1093/biomet/85.1.235.
22. Simonoff, J.S. The anchor position of histograms and frequency polygons: Quantitative and qualitative smoothing. *Commun. Stat. Simul. Comput.* **1995**, *24*, 691–710, doi:10.1080/03610919508813267.
23. Scott, D.W. Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. *Ann. Stat.* **1985**, *13*, 1024–1040, doi:10.1214/aos/1176349654.
24. Scott, D.W. On optimal and data-based histograms. *Biometrika* **1979**, *66*, 605–610.
25. Fix, E.; Hodges, J. An important contribution to nonparametric discriminant analysis and density estimation. *Int. Stat. Rev.* **1951**, *3*, 233–238.
26. Silverman, B.W.; Jones, M.C.; Fix, E. An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *Int. Stat. Rev.* **1989**, *57*, 233, doi:10.2307/1403796.
27. Rosenblatt, M. A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **1956**, *42*, 43–47, doi:10.1073/pnas.42.1.43.
28. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076, doi:10.1214/aoms/1177704472.
29. Cencov, N.N. Estimation of an unknown distribution density from observations. *Soviet Math.* **1962**, *3*, 1559–1566.
30. Watson, G.S.; Leadbetter, M. On the estimation of the probability density, I. *Ann. Math. Stat.* **1963**, *34*, 480–491.
31. Loftsgaarden, D.O.; Quesenberry, C.P. A Nonparametric Estimate of a Multivariate Density Function. *Ann. Math. Stat.* **1965**, *36*, 1049–1051, doi:10.1214/aoms/1177700079.
32. Schwartz, S.C. Estimation of Probability Density by an Orthogonal Series. *Ann. Math. Stat.* **1967**, *38*, 1261–1265, doi:10.1214/aoms/1177698795.
33. Epanechnikov, V.A. Non-Parametric Estimation of a Multivariate Probability Density. *Theory Probab. Its Appl.* **1969**, *14*, 153–158, doi:10.1137/1114019.
34. Tarter, M.; Kronmal, R. On Multivariate Density Estimates Based on Orthogonal Expansions. *Ann. Math. Stat.* **1970**, *41*, 718–722, doi:10.1214/aoms/1177697122.
35. Kimeldorf, G.; Wahba, G. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **1971**, *33*, 82–95, doi:10.1016/0022-247x(71)90184-3.
36. Cacoullos, T.; Sobel, M. An inverse sampling procedure for selecting the most probable event in a multinomial distribution. In *Multivariate Analysis;* Academic Press: New York, NY, USA, 1966; pp. 423–455.
37. Silverman, B. Choosing the window width when estimating a density. *Biometrika* **1978**, *65*, 1–11.
38. Hwang, J.-N.; Lay, S.-R.; Lippman, A. Nonparametric multivariate density estimation: A comparative study. *IEEE Trans. Signal. Process.* **1994**, *42*, 2795–2810, doi:10.1109/78.324744.
39. Scott, D. Multivariate Density Estimation. *Ann. Stat.* **1992**, *20*, 1236–1265.
40. Kooperberg, C. Bivariate density estimation with an application to survival analysis. *J. Comput. Graph. Stat.* **1998**, *7*, 322–341.
41. Takada, T. Nonparametric density estimation: A comparative study. *Econ. Bull.* **2001**, *3*, 1–10.

42. Delgado, M.A.; Robinson, P.M. Nonparametric and semiparametric methods for economic research. *J. Econ. Surv.* **1992**, *6*, 201–249, doi:10.1111/j.1467-6419.1992.tb00151.x.
43. Gill, R.D.; Wellner, J.A.; Præstgaard, J. Non-and semi-parametric maximum likelihood estimators and the von mises method (part 1) [with discussion and reply]. *Scand. J. Stat.* **1989**, *16*, 97–128.
44. Gill, R.D.; Van Der Vaart, A.W. Non-and semi-parametric maximum likelihood estimators and the von Mises method: II. *Scand. J. Stat.* **1993**, *20*, 271–288.
45. Hyndman, R.J.; Yao, Q. Nonparametric Estimation and Symmetry Tests for Conditional Density Functions. *J. Nonparametr. Stat.* **2002**, *14*, 259–278, doi:10.1080/10485250212374.
46. Holmström, L.; Hoti, F. Application of semiparametric density estimation to classification. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004; Volume 3, pp. 371–374.
47. Castellana, J.; Leadbetter, M. On smoothed probability density estimation for stationary processes. *Stoch. Process. Their Appl.* **1986**, *21*, 179–193, doi:10.1016/0304-4149(86)90095-5.
48. Chiu, S.-T. Bandwidth Selection for Kernel Density Estimation. *Ann. Stat.* **1991**, *19*, 1883–1905, doi:10.1214/aos/1176348376.
49. Gu, C.; Qiu, C. Smoothing Spline Density Estimation: Theory. *Ann. Stat.* **1993**, *21*, 217–234, doi:10.1214/aos/1176349023.
50. Härdle, W.; Müller, M. *Multivariate and Semiparametric Kernel Regression*; SFB 373 Discussion Paper; Springer: Heidelberg/Berlin, Germany, 1997.
51. Jones, M.C.; Marron, J.S.; Sheather, S.J. A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc.* **1996**, *91*, 401–407.
52. Stone, C.J.; Hansen, M.H.; Kooperberg, C.; Truong, Y.K. Polynomial splines and their tensor products in ex-tended linear modeling: 1994 Wald memorial lecture. *Ann. Stat.* **1997**, *25*, 1371–1470.
53. Aladjem, M. Projection pursuit mixture density estimation. *IEEE Trans. Signal. Process.* **2005**, *53*, 4376–4383, doi:10.1109/TSP.2005.857007.
54. Friedman, J.H.; Stuetzle, W.; Schroeder, A. Projection pursuit density estimation. *J. Am. Stat. Assoc.* **1984**, *79*, 599–608.
55. Friedman, J.H. Exploratory projection pursuit. *J. Am. Stat. Assoc.* **1987**, *82*, 249–266.
56. Rudzkis, R.; Radavičius, M. Testing Hypotheses on Discriminant Space in the Mixture Model of Gaussian Distributions. *Acta Appl. Math.* **2003**, *79*, 105–114, doi:10.1023/a:1025891613510.
57. Kalantan, Z.I.; Einbeck, J. Quantile-Based Estimation of the Finite Cauchy Mixture Model. *Symmetry* **2019**, *11*, 1186, doi:10.3390/sym11091186.
58. Azzari, L.; Foi, A. Gaussian-Cauchy mixture modeling for robust signal-dependent noise estimation. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 5357–5361.
59. Teimouri, M. Statistical Inference for Mixture of Cauchy Distributions. *arXiv* **2018**, arXiv:1809.05722.
60. Jones, M.C. Discretized and interpolated kernel density estimates. *J. Am. Stat. Assoc.* **1989**, *84*, 733–741.
61. Lambert, C.G.; Harrington, S.E.; Harvey, C.R.; Glodjo, A. Efficient on-line nonparametric kernel density estimation. *Algorithmica* **1999**, *25*, 37–57.
62. Gasser, T.; Müller, H.-G.; Mammitzsch, V. Kernels for Nonparametric Curve Estimation. *J. R. Stat. Soc. Ser. B* **1985**, *47*, 238–252, doi:10.1111/j.2517-6161.1985.tb01350.x.
63. Marron, J.; Nolan, D. Canonical kernels for density estimation. *Stat. Probab. Lett.* **1988**, *7*, 195–199, doi:10.1016/0167-7152(88)90050-8.
64. Fukunaga, K.; Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* **1975**, *21*, 32–40, doi:10.1109/tit.1975.1055330.
65. Silverman, B.W. Density Estimation for Statistics and Data Analysis. In *Monographs on Statistics and Applied Probability*; Chapman & Hall: London, UK, 1994.
66. Breiman, L.; Meisel, W.; Purcell, E. Variable kernel estimates of multivariate densities. *Technometrics* **1977**, *19*, 135–144.
67. Hall, P.; Sheather, S.J.; Jones, M.; Marron, J.S. On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **1991**, *78*, 263–269.
68. Hart, J.D.; Vieu, P. Data-Driven Bandwidth Choice for Density Estimation Based on Dependent Data. *Ann. Stat.* **1990**, *18*, 873–890, doi:10.1214/aos/1176347630.
69. Marron, J.S. An Asymptotically Efficient Solution to the Bandwidth Problem of Kernel Density Estimation. *Ann. Stat.* **1985**, *13*, 1011–1023, doi:10.1214/aos/1176349653.
70. Wand, M.P.; Jones, M.C. Multivariate plug-in bandwidth selection. *Comput. Stat.* **1994**, *9*, 97–116.
71. Abramson, I.S. On Bandwidth Variation in Kernel Estimates-A Square Root Law. *Ann. Stat.* **1982**, *10*, 1217–1223, doi:10.1214/aos/1176345986.
72. Nadaraya, E.A. On estimating regression. *Theory Probab. Its Appl.* **1964**, *9*, 141–142. https://doi.org/10.1137/1109020.
73. Watson, G.S. Smooth regression analysis. *Sankhyā Indian J. Stat. Ser. A* **1964**, *26*, 359–372.
74. Smith, P.W.; Schumaker, L. Spline Functions: Basic Theory. *Math. Comput.* **1982**, *38*, 652, doi:10.2307/2007301.
75. De Boor, C.; De Boor, C. *A Practical Guide to Splines*; Springer: New York, NY, USA, 1978; Volume 27.
76. Koo, J.-Y. Bivariate B-splines for tensor logspline density estimation. *Comput. Stat. Data Anal.* **1996**, *21*, 31–42, doi:10.1016/0167-9473(95)00003-8.
77. Hansen, M.H.; Kooperberg, C. Spline Adaptation in Extended Linear Models (with comments and a rejoinder by the authors). *Stat. Sci.* **2002**, *17*, 2–51, doi:10.1214/ss/1023798997.

78. Kooperberg, C.; Stone, C.J. A study of logspline density estimation. *Comput. Stat. Data Anal.* **1991**, *12*, 327–347, doi:10.1016/0167-9473(91)90115-i.
79. Kooperberg, C.; Stone, C.J. Logspline density estimation for censored data. *J. Comput. Graph. Stat.* **1992**, *1*, 301–328.
80. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control.* **1974**, *19*, 716–723, doi:10.1109/tac.1974.1100705.
81. Friedman, J.; Tukey, J. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Trans. Comput.* **1974**, *100*, 881–890, doi:10.1109/t-c.1974.224051.
82. Huber, P.J. Projection pursuit. *Ann. Stat.* **1985**, *13*, 435–475.
83. Hall, P. On Polynomial-Based Projection Indices for Exploratory Projection Pursuit. *Ann. Stat.* **1989**, *17*, 589–605, doi:10.1214/aos/1176347127.
84. Goffe, W.L.; Ferrier, G.; Rogers, J. Global optimization of statistical functions with simulated annealing. *J. Econ.* **1994**, *60*, 65–99, doi:10.1016/0304-4076(94)90038-8.
85. Ruzgas, T. The Nonparametric Estimation of Multivariate Distribution Density Applying Clustering Procedures. Ph.D. Thesis, Matematikos ir Informatikos Institutas, Vilnius, Lithuania, 2007.
86. Kavaliauskas, M.; Rudzkis, R.; Ruzgas, T. The projection-based multi-variate distribution density estimation. *Acta Comment. Univ. Tartu. Math.* **2004**, *8*, 135–141.
87. Rudzkis, R.; Radavičius, M. Statistical estimation of a mixture of Gaussian distributions. *Acta Appl. Math.* **1995**, *38*, 37–54, doi:10.1007/bf00992613.
88. Van deLaan, M. *Efficient and Inefficient Estimation in Semiparametric Models*; CWI Tracts: Amsterdam, The Netherlands, 1995.
89. Van Der Laan, M.J.; Dudoit, S.; Keles, S. Asymptotic Optimality of Likelihood-Based Cross-Validation. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 1–23, doi:10.2202/1544-6115.1036.
90. Hall, P. Large Sample Optimality of Least Squares Cross-Validation in Density Estimation. *Ann. Stat.* **1983**, *11*, 1156–1174, doi:10.1214/aos/1176346329.