A semi-parametric density estimation with application in clustering

Mahdi Salehi
Department of Mathematics and Statistics, University of Neyshabur, Iran
Department of Statistics, University of Pretoria, South Africa
and
Andriette Bekker
Department of Statistics, University of Pretoria, South Africa
and
Mohammad Arashi
Department of Statistics, Ferdowsi University of Mashhad, Iran

March 14, 2024

Abstract

The idea behind density-based clustering is to associate groups to the connected components of the level sets of the density of the data to be estimated by a non-parametric method. This approach claims some advantages over both distance- and model-based clustering. Some researchers developed this technique by proposing a graph theory-based method for identifying local modes of the underlying density being estimated by the well-known kernel density estimation (KDE) with normal and t kernels. The present work proposes a semi-parametric KDE with a more flexible family of kernels including skew-normal (SN) and skew-t (ST). We show that the proposed estimator not only reduces boundary bias but it is also closer to the actual density compared to that of the usual estimator employing the Gaussian kernel. Finding optimal bandwidth for one-dimensional and multidimensional cases under the mentioned asymmetric kernels is another main result of this paper where we shrink the bandwidth more than the one obtained under the normal assumption. Finally, through a comprehensive numerical study, we will illustrate the application of the proposed semi-parametric KDE on the density-based clustering using some simulated and real data sets.

Keywords: Asymmetric kernels, Boundary bias, Density-based clustering, Density-based Silhouette, Kernel density estimation, Optimum bandwidth.

1 Introduction

The model-based clustering, which introduces a finite mixture of parametric distributions, addresses some major challenges of unsupervised classification, e.g. finding the number of clusters and working with high dimensional matrix-valued data. In this context, the multivariate normal mixture model has been the most frequently used parametric distribution (Bouveyron et al, 2019; McNicholas, 2016; Fraley and Raftery, 2002). But, single Gaussian component is not capable to cover skewed and heavy-tailed data. Moreover, the number of clusters in the model-based clustering may be overestimated when the components themselves are wrongly assumed to be symmetric (Loperfido, 2019; Malsiner-Walli et al, 2017). Thus, much attention has been paid to develop mixture models using a more flexible parametric family of distributions for the components. For instance, Lin et al (2007) used the skew-t distribution as the components of the finite mixture model. Interested readers are referred to Lee and McLachlan (2014) for more details on finite mixtures of multivariate skew-t distributions. To tackle the overestimation of the number of clusters, Malsiner-Walli et al (2017) also identified the mixture of mixtures model within a Bayesian framework through a hierarchical prior construction and proposed a method to select a suitable number of clusters. Recently, Millard (2019) developed semi-parametric model-based clustering. He proposes a clustering approach using the finite mixture of the exponential distributions in the context of the generalized linear modeling where the link function is unknown and has to be estimated.

On the other hand, the mechanism behind the density-based clustering is associating groups to the connected regions of the density. Thus, it is clear that these approaches involve somewhat different notions of cluster: the density-based clustering falls within the nonparametric context, while the model-based clustering has a parametric setting.

The two concepts are different, even if they have the same outcome. The density-based clustering claims some advantages with respect to both classical clustering approaches (which mostly work based on a distance) and model-based clustering. Firstly, the flexibility of nonparametric density estimators allows for detection of groups with arbitrary shapes. In contrast, the shape of model-based clusters depends on the components of the mixture model, and distance-based methods are known to favour specific clustering structures. Secondly, another problem that arises in most classical clustering algorithms is that the number of clusters is selected by the user (as in k-means) or left undetermined (as in hierarchical clustering). On the contrary, the number of clusters in the density-based approach is determined by the method itself. More specifically, clusters correspond to the regions around the modes of the data distribution, as a result, their number is conceptually well-defined and, then, estimable in practice (Menardi and Azzalini, 2014; Chacon, 2015).

Classical kernel methods are based on symmetric densities, say around zero, as in the case of the normal kernel. However, a deficiency arises by using such kernels when we try to estimate densities with bounded supports (Marron and Ruppert, 1994). The main benefit of using asymmetric kernels, rather than the classical symmetric ones, are that the formers have varying shape and are flexible in smoothing over the domain of the population. This property implies that they reduce the boundary bias (as it will be shown by Figure 2). Moreover, if the kernel is also bounded, then, it even avoids assigning weight outside the density support. For more details in this regard, one can refer to Chen (1999, 2000), Fernandez and Monteiro (2005), Punzo (2010), Saulo et al (2013), Mazza and Punzo (2011, 2013a,b, 2014, 2015) and Tomarchio and Punzo (2019). There is some research on the development of the kernel density estimator (KDE) with asymmetric kernels in the literature. Among all, Abadir and Lawford (2004) provided a reason for the use of asymmetric

kernels arguing that density estimators in moderately-sized samples tend to inherit salient properties of their kernels, thus, if a density is suspected to be highly skewed, it makes sense to utilize asymmetric kernels rather than a symmetric one. Saulo et al (2013) used the skew-symmetric Birnbaum-Saunders density as an asymmetric kernel for estimating asymmetric densities. See Chen (2000), Bouezmarni and Scaillet (2005), Kuruwita et al (2010) for additional references in this subject area. The R package `kdensity` covers most of the asymmetric kernels proposed in the literature so far (Moss and Tveten, 2019).

As mentioned before, Azzalini and Torelli (2007) proposed a density-based clustering procedure via KDE. They exploited spatial tessellation to generalize the concept of adjacency among points in several dimensions and determine the connected level sets. Then, clusters are associated to the maximally connected components with estimated density above a threshold. As the threshold varies, these clusters may be represented according to a hierarchical structure in the form of a tree. Their algorithm was feasible when data dimensionality is low to moderate (up to 6 for instance). In order to improve this method, Menardi and Azzalini (2014) found a viable solution to the problem of detecting connected regions in higher dimensional spaces. Both methods are aggregated in a unique algorithm and are implemented by the R package `pdfCluster` (Azzalini and Menardi, 2014). Our main goal in this paper is to improve the density estimation part of this algorithm by utilizing some skew-symmetric kernels rather than normal and t_7 used in `pdfCluster`. As its terminology, the skew-symmetric (skew-modulated) distributions can generate both symmetric and asymmetric densities: see for example Salehi and Azzalini (2018). Among so many distributions belonging to this family, the skew-normal and skew-t, as its most well-known members, are flexible enough to be employed as a kernel in the KDE. We will demonstrate that our approach utilizing such skew-symmetric kernels offers more shrinkage

for the bandwidth which is practically shown to be desirable in `pdfCluster` (Azzalini and Torelli, 2007).

The rest of this paper is organized as follows: In Section 2, a semi-parametric KDE algorithm is proposed. In Section 3, the optimal bandwidths for both one-dimensional and multi-dimensional cases under the assumption of skew kernels are derived. Furthermore, the boundary bias is investigated via a simulation study. Some real examples on the usage of the proposed semi-parametric KDE algorithm in `pdfCluster` are presented in Section 4. This section also presents a relatively extensive simulation study under different scenarios to check the behaviour of the `pdfCluster` with various symmetric and asymmetric kernels. Finally, some concluding remarks and discussions are given by Section 5.

2 A semi-parametric KDE

Suppose that $\mathbf{X}_i := (X_{i1}, \ldots, X_{id})^\top$, $i = 1, \ldots, n$, is a random sample of size n from a d-dimensional density function f and $\mathbf{x}_i := (x_{i1}, \ldots, x_{id})^\top$ is an observation. Azzalini and Menardi (2014) performed the density estimation step of their clustering algorithm by the kernel method employing a product KDE of the form

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_{i,j}} \kappa \left(\frac{x_j - X_{ij}}{h_{i,j}} \right), \tag{1}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^\top$, $\kappa(\cdot)$ is a symmetric kernel (that can either be chosen to be a normal or a t_7 density in `pdfCluster`) and $h_{i,j}$ stands for the adaptive bandwidth corresponding to the X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, d$.

Here, we allow the $\kappa(\cdot)$ to vary for each variable with aiming to have a more efficient KDE. In more details, we intend to employ a product kernel of the following form

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j} \kappa_j \left(\frac{x_j - X_{ij}}{h_j} \right), \tag{2}$$

where $\kappa_j(\cdot)$ can be either a skew-normal density or a skew-t density. More precisely, in the former case, we have

$$\kappa_j(z) = 2\phi(z)\Phi(\lambda_j z), \qquad z \in R, \tag{3}$$

where λ_j, $j = 1, \ldots, d$, is a slant parameter, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and the cumulative distribution function of the standard normal distribution, respectively. Our second proposal for the $\kappa_j(\cdot)$ is the standard skew-t density given by

$$\kappa_j(z) = 2t(z;\nu_j)T\left(\lambda_j z\sqrt{\frac{\nu_j+1}{\nu_j+z^2}};\nu_j+1\right), \qquad z \in R, \tag{4}$$

with the slant parameter λ_j and the degrees of freedom ν_j, $j = 1, \ldots, d$. Also, $t(\cdot;\nu_j)$ and $T(\cdot;\nu_j)$ are the density and cumulative distribution function of the classical Student's t distribution with ν_j degrees of freedom, respectively. We turn the reader's attention to different degrees of freedom with respect to each kernel. Hence, here we have some unknown parameters to be estimated. To this end, the maximum penalized likelihood estimate (MPLE) can be a reasonable approximation for λ_j in the first case assuming that $z_{1j} \ldots, z_{nj}$ are observations from the SN distribution given by (3), where

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad i = 1, \ldots, n,\ j = 1, \ldots, d, \tag{5}$$

where $\bar{x}$ and s_j are the sample mean and sample standard deviation of the observations $x_{1j} \ldots, x_{nj}$. The routine `sn.mple()` in the R package `sn` can be utilized for obtaining the MPLE of the λ_j. For more details on the MPLE, one can refer to Azzalini and Arellano-Valle (2013).

For the skew-t case, we have $2d$ unknown parameters to be approximated. Here we can use either a one-stage or a two-stage approximation procedure depending on the dimension of the data (d). If the dimension is high, one can use only a quick preliminary estimate for the pair (λ_j, ν_j) given by (4). Such initial estimates are proposed by Azzalini and Salehi

(2020) and can be implemented by the function `st.prelimFit()` in the R package `sn`. Otherwise, for small/moderate dimensions one can use a double-step optimization method: using the initial estimates given by the mentioned approach for the subsequent numerical maximization of the penalized log-likelihood function which can be performed employing the routine `st.mple()`. This approach is summarized by Algorithm 1.

The algorithm just described is no longer a non-parametric density estimation but a semi-parametric one. However, Hjort and Glad (1995) were probably the first researchers to use a semi-parametric approach by mixing the nonparametric version of the density estimation with a parametric start where their nuisance parameters were to be estimated. Here, another problem of interest is to find optimum values for the bandwidth h_j in (2) through minimizing the approximated mean integrated square error (MISE). To address this, we start with the one-dimensional case first.

3 Finding an optimal bandwidth

In this section, by considering $\kappa_j(\cdot)$ to be the density (3), we obtain the optimal bandwidth involved in (2) through minimizing the MISE of $\hat{f}$ given by Wand and Jones (1995)

$$\begin{aligned} \text{MISE}_h \hat{f}(\mathbf{x}) &= \text{E} \int \left\{ \hat{f}(\mathbf{x}) - f(\mathbf{x}) \right\}^2 d\mathbf{x} \\ &= \int \text{Var} \hat{f}(\mathbf{x}) d\mathbf{x} + \int \text{Bias}^2 \hat{f}(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{6}$$

But, a problem with the MISE in (6) is that it does depend on the bandwidth in a complicated way. This fact makes it difficult to obtain the optimal bandwidth based on the MISE. Hence, we employ the AMISE which is an approximation of the MISE.

For convenience, we start with the one-dimensional case and then give the results for the multi-dimensional case.

Algorithm 1: The semi-parametric KDE

1 Take $\mathbf{z}_j := (z_{1j}, \ldots, z_{nj})^\top$ whose components are given by (5).

2 Choose the desired kernel between (3) or (4) and then estimate their unknown parameters in Steps 3 or 4;

3 For the SN kernel, plug-in the λ_j with its MPLE assuming $\mathbf{z}_j$ is a random sample from (3).

4 For the ST kernel, do the following procedure assuming $\mathbf{z}_j$ as a random sample from (4):

(i) *For high dimensions:* approximate (λ_j, ν_j) with quick preliminary estimates using `st.prelimFit()`.

(ii) *For low dimensions:* plug-in (λ_j, ν_j) with their MPLEs via incorporating the routines `st.prelimFit()` and `st.mple()`.

5 Compute $\hat{f}(\mathbf{x})$ in (2) with $\kappa_j(\cdot)$ whose parameters are approximated either in step 3 or 4 and with an optimal bandwidth which will be given in Section 3.

3.1 One-dimensional case

Suppose that $f(\cdot)$ is a univariate density function to be estimated using (2) considering $d = 1$. Assume further that $\kappa(\cdot)$ is a one-dimensional asymmetric kernel density satisfying the following regularity assumptions

$$\begin{aligned} (A_1) \quad & \kappa(z) \geq 0, \quad z \in R, \\ (A_2) \quad & \int \kappa(z)dz = 1, \\ (A_3) \quad & \int z\kappa(z)dz = \mu_1(\kappa) \neq 0. \end{aligned} \tag{7}$$

Note, unlike symmetric kernels, $\mu_1(\kappa)$ is not equal to zero. This is an inherit property emanated from asymmetric kernels. Also, we define $\mu_2(\kappa) := \int z^2\kappa(z)dz$. From Silverman (1986) we have

$$\begin{aligned} \text{Bias}\hat{f}(x) &= \int \frac{1}{h}\kappa\left(\frac{x-y}{h}\right) f(y)dy - f(x) \\ &= \int \kappa(t)\left[f(x-ht) - f(x)\right]dt. \end{aligned} \tag{8}$$

Using the Taylor series expansion we simply get

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) + \cdots. \tag{9}$$

Then, we have

$$\begin{aligned} \text{Bias}\hat{f}(x) &= -h\mu_1(\kappa)f'(x) + \frac{1}{2}h^2\mu_2(\kappa)f''(x) + O(h^3) \\ &\approx -h\mu_1(\kappa)f'(x) + \frac{1}{2}h^2\mu_2(\kappa)f''(x) \end{aligned} \tag{10}$$

where $\mu_1(\kappa)$ and $\mu_2(\kappa)$ are given by (7). Accordingly, from (10) we get

$$\begin{aligned} \int \text{Bias}^2\hat{f}(x)dx &\approx h^2\mu_1^2(\kappa)\int [f'(x)]^2dx + \frac{1}{4}h^4\mu_2^2(\kappa)\int [f''(x)]^2dx \\ &\quad -h^3\mu_1(\kappa)\mu_2(\kappa)\int f'(x)f''(x)dx. \end{aligned} \tag{11}$$

The variance of the estimator evaluated at x is simply obtained as

$$\begin{aligned}
\mathrm{Var}\hat{f}(x) &= \frac{1}{n}\int \frac{1}{h^2}\kappa^2\left(\frac{x-y}{h}\right) f(y)dy - \frac{1}{n}[f(x) + \mathrm{Bias}\hat{f}(x)]^2 \\
&\approx \frac{1}{nh}\int f(x-ht)\kappa^2(t)dt - \frac{1}{n}[f(x) - h\mu_1(\kappa)f'(x) \\
&\quad + \frac{1}{2}h^2\mu_2(\kappa)f''(x)]^2 \\
&= \frac{1}{nh}\int [f(x) - htf'(x) + \cdots]\kappa^2(t)dt - \frac{1}{n}[f(x) + O(h)]^2 \\
&= \frac{1}{n}\left\{\frac{1}{h}f(x)\int \kappa^2(t)dt - f'(x)\int t\kappa^2(t)dt\right\} + O\left(n^{-1}\right),
\end{aligned} \tag{12}$$

and accordingly

$$\int \mathrm{Var}\hat{f}(x)dx = \frac{1}{nh}\int \kappa^2(t)dt + O\left(n^{-1}\right), \tag{13}$$

Thus, by substituting (11) and (13) in (6), we obtain

$$\begin{aligned}
\mathrm{AMISE}_h\hat{f}(x) &= \frac{1}{nh}\int \kappa^2(t)dt + h^2\mu_1^2(\kappa)\int [f'(x)]^2dx \\
&\quad + \frac{1}{4}h^4\mu_2^2(\kappa)\int [f''(x)]^2dx \\
&\quad - h^3\mu_1(\kappa)\mu_2(\kappa)\int f'(x)f''(x)dx.
\end{aligned} \tag{14}$$

The optimal bandwidth is obtained by minimizing the $\mathrm{AMISE}_h\hat{f}(x)$ given by (14). Let us consider two approaches for the purpose of minimization: (I) Drop the terms in the order of h^k, $k \geq 3$, (II) Work with all the terms involved in (14).

The second approach does not give any close form for the optimal bandwidth. However, in either case, the values of h_{opt} are related to the density function being estimated and its derivatives. The optimal bandwidth following Approach (I) is derived as

$$h_{\mathrm{opt}} = \left\{\frac{\int \kappa^2(t)dt}{2\mu_1^2(\kappa)\int [f'(x)]^2dx}\right\}^{\frac{1}{3}}\left(\frac{1}{n}\right)^{\frac{1}{3}}, \tag{15}$$

and then the value of AMISE at h_{opt} is obtained as follows

$$\text{AMISE}_{h_{\text{opt}}}\hat{f}(x) = \left(2^{\frac{1}{3}} + 2^{-\frac{1}{3}}\right)\left\{\mu_1(\kappa)\int \kappa^2(t)dt\right\}^{\frac{2}{3}}\left\{\int [f'(x)]^2 dx\right\}^{\frac{1}{3}} n^{-\frac{2}{3}}. \tag{16}$$

Silverman (1986) suggested using a standard family of distributions to obtain an approximation for the terms which are related to the target density itself. A choice for this purpose is $f(x) = 1/\sigma\phi(x/\sigma)$, where $\phi(\cdot)$ denotes the density of the standard normal distribution. Then, we can simply obtain

$$\int [f'(x)]^2 dx = \frac{1}{4}\pi^{-\frac{1}{2}}\sigma^{-3}. \tag{17}$$

Also we have (Silverman, 1986, pp. 45)

$$\int [f''(x)]^2 dx = \frac{3}{8}\pi^{-\frac{1}{2}}\sigma^{-5}. \tag{18}$$

After some algebraic manipulation, the last integral in (14) is computed to be zero. Now if we consider $\kappa(\cdot)$ to be the one given by (3), then

$$\mu_1(\kappa) = \sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}, \qquad \mu_2(\kappa) = 1. \tag{19}$$

In addition, we have (Salehi and Doostparast, 2015)

$$\begin{aligned}\int \kappa^2(t)dt &= E\left(\phi_{SN}(Z_\lambda;\lambda)\right)\\ &= 2\pi^{-\frac{3}{2}}\arctan\sqrt{1+\lambda^2},\end{aligned} \tag{20}$$

where $Z_\lambda \sim SN(\lambda)$. Substituting (17), (19) and (20) in (15) yields

$$h_{\text{opt}}(\lambda) = c_\lambda \sigma n^{-\frac{1}{3}}, \tag{21}$$

where

$$c_\lambda = \left\{2\left(1+\lambda^{-2}\right)\arctan\sqrt{1+\lambda^2}\right\}^{\frac{1}{3}}. \tag{22}$$

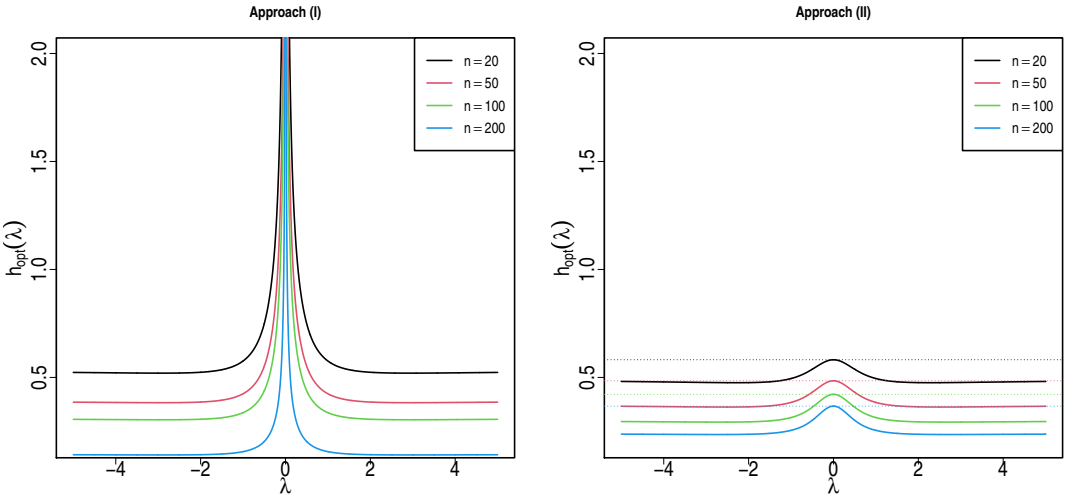

Figure 1: The plots of h_{opt} versus λ for various sample sizes and $\sigma = 1$ based on approaches (I) & (II) given by (21) and (23), respectively. The dotted lines appearing in the right panel correspond to the optimal bandwidths given by (25) which is obtained under the normality assumption.

Now, we give the details of Approach (II). By substituting (17)-(20) in (14), the h_{opt} is obtained to be a real value satisfying the following equation

$$-a + 2bh^3_{\text{opt}}(\lambda) + ch^5_{\text{opt}}(\lambda) = 0, \tag{23}$$

where

$$a := \frac{1}{n} 2\pi^{-3/2} \arctan\sqrt{1+\lambda^2}, \qquad b := \frac{\lambda^2}{2\pi^{\frac{3}{2}}\sigma^3(1+\lambda^2)}, \qquad c := \frac{3}{8\pi^{\frac{1}{2}}\sigma^5}. \tag{24}$$

Figure 1 displays the plots of $h_{\text{opt}}(\lambda)$ obtained from (21) and (23), respectively, when sample size spans from small to large. As it is observed, the optimal bandwidth obtained via Approach (I) diverges as λ tends to zero. So, the density f will be over-smoothed for small values of λ. On the contrary, that of the second approach has a peak around zero according to the right pane of Fig. 1. This peak coincides with the optimal bandwidth under the normal kernel which is known to be (Bowman and Azzalini, 1997, pp 31)

$$h = \left(\frac{4}{3n}\right)^{\frac{1}{5}} \sigma. \tag{25}$$

Hence, it is reasonable to ignore Approach (I) and rather exploit from Approach (II). By this manner, we shrink the bandwidth (slightly towards zero) more than the one obtained under the normal assumption. Azzalini and Torelli (2007) empirically showed that this property is often desirable when a KDE is applied in their density-based clustering.

3.1.1 Boundary bias comparison

For illustrating the impact of the use of the SN and the ST as asymmetric kernels in reducing the boundary bias, we have generated 1000 samples of size $n = 150$ from symmetric/asymmetric distributions whose densities are bounded. The average of the usual non-parametric KDE (with Gaussian kernel and h given by (25)) and the average of the proposed semi-parametric KDE with SN and ST kernels and h given by (23) are displayed in Figure 2. This figure shows that the proposed estimators are smaller than the classical KDE with the Gaussian kernel at the outside of the support and near the boundary points for the asymmetrical bounded distributions. However, their performance is the same for the symmetrical bounded distributions (see the bottom-left and the center-right panes in Figure 2). It has been shown that when the density has high concentrations close to the boundaries, the problem of boundary bias is more severe below the lower bound and above the upper bound (Rattihalli and Patil, 2019). The SN kernel outperforms the ST one under the mentioned circumstances (see the top-left case in Figure 2). In addition, the KDEs with SN and ST kernels are much more closer to the true density than the Gaussian one even in the central part.

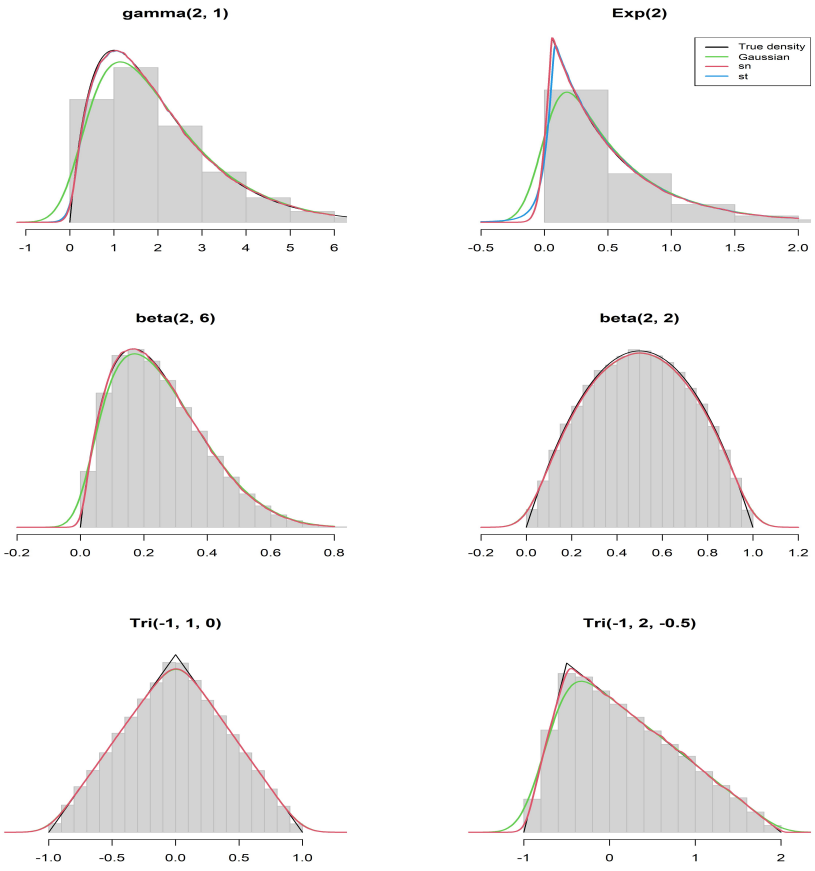

Figure 2: The plots of the true density, the average of the semi-parametric KDEs with the SN and the ST kernels and the classical KDE with the Gaussian kernel for various bounded distributions.

3.2 Multi-dimensional case

Assume $f(\mathbf{x})$ is a d-variate density. Then, the kernel density estimator in its most general form is defined as (Wand and Jones, 1995)

$$\hat{f}(\mathbf{x}) = n^{-1}|H|^{-\frac{1}{2}} \sum_{i=1}^{n} K\left(H^{-\frac{1}{2}}(\mathbf{x} - \mathbf{X}_i)\right), \tag{26}$$

where $\mathbf{x} := (x_1, \ldots, x_d)^\top$, $\mathbf{X}_i := (X_{i1}, \ldots, X_{id})^\top$, $i = 1, \ldots, n$, is the ith observation, H is the so-called bandwidth matrix (a symmetric positive definite $d \times d$ matrix) and $K(\cdot)$ is a d-variate kernel density. In the present paper, we restrict ourselves to a special case of this general form given by (2). Its bandwidth indeed belongs to the class of diagonal bandwidth matrices, namely $H = \mathrm{Diag}(h_1^2, \ldots, h_d^2)$. Notice that in (2), the marginal kernels $\kappa_j(\cdot)$ are allowed to vary while they are typically considered to be identical in the literature. Now, let $\mathcal{H}_f(\mathbf{x}) = [f''_{(\ell,j)}]$ and $\mathcal{D}(\mathbf{x}) = [f'_{(j)}]$ denote the Hessian matrix and the vector of the first derivatives of $f(\mathbf{x})$, respectively. More specifically

$$\begin{aligned} f'_{(j)} &= \frac{\partial}{\partial x_j} f(\mathbf{x}), \\ f''_{(\ell,j)} &= \frac{\partial^2}{\partial x_\ell \partial x_j} f(\mathbf{x}), \qquad \ell, j = 1, \ldots, d. \end{aligned} \tag{27}$$

Suppose further that

(i) All $f''_{(\ell,j)}$'s are piecewise continuous and square-integrable.

(ii) $h_j = h_{j,n}$, $j = 1, \ldots, d$, is a sequence of bandwidths tending to zero as $n \to \infty$.

(iii) Each component $\kappa_j(\cdot)$ satisfies the conditions (7).

Under the above definition and assumptions, we will obtain the bias, variance and accordingly the AMISE of $\hat{f}(\mathbf{x})$ given by (2) as follows.

Lemma 1 *Given Assumptions (i)-(iv), the bias and the variance of* $\hat{f}(\mathbf{x})$ *in* (2) *are respectively derived as*

$$\begin{aligned}\mathrm{Bias}\hat{f}(\mathbf{x}) &= -\sum_{j=1}^{d}\mu_1(\kappa_j)h_j f'_{(j)} + \frac{1}{2}\sum_{j=1}^{d}\mu_2(\kappa_j)h_j^2 f''_{(j,j)} \\ &\quad + \sum_{j<\ell}\mu_1(\kappa_j)\mu_1(\kappa_\ell)h_j h_\ell f''_{(\ell,j)} + o\left(\sum_{j=1}^{d}h_j^2\right), \end{aligned} \tag{28}$$

and

$$\mathrm{Var}\hat{f}(\mathbf{x}) = \frac{1}{n}f(\mathbf{x})\prod_{j=1}^{d}\frac{1}{h_j}\int \kappa_j^2(t_j)dt_j + O(n^{-1}). \tag{29}$$

Proof 1 *Refer to Appendix.*

As it is observed from Lemma 1, the bias is related to the derivatives of the target density being estimated. So, one can obtain an appropriate value for the bandwidth under the assumption of normality. Specifically, analogues to the preceding subsection, we assume that $f(\mathbf{x}) = \prod_{j=1}^{d}\phi(x_j/\sigma_j)/\sigma_j$. Let us also simplify (28) and (29) by setting $h_j = h$, $j = 1,\ldots,d$. The next remark deals with obtaining the AMISE by imposing these restrictions.

Theorem 1 *Assume* $h_j = h$, $j = 1,\ldots,d$. *Then, the AMISE of* $\hat{f}(\mathbf{x})$ *has the form*

$$\mathrm{AMISE}_h f(\mathbf{x}) = \frac{1}{nh^d}\prod_{j=1}^{d}\int \kappa_j^2(t)dt + A\left(Bh^2 + Ch^4\right), \tag{30}$$

where

$$\begin{aligned} A &:= \frac{(2\sqrt{\pi})^{-d}}{16\prod_{j=1}^{d}\sigma_j}, \qquad B := 8\sum_{j=1}^{d}\mu_1^2(\kappa_j)\sigma_j^{-2}, \\ C &:= 3\sum_{j=1}^{d}\mu_2^2(\kappa_j)\sigma_j^{-4} + 2\sum_{j<\ell}\left(\mu_2(\kappa_j)\mu_2(\kappa_\ell) + \mu_1^2(\kappa_j)\mu_1^2(\kappa_\ell)\right)\sigma_j^{-2}\sigma_\ell^{-2}. \end{aligned} \tag{31}$$

Proof 2 *Refer to Appendix.*

Remark 1 *As a special case, if* $\kappa_j(\cdot) = \phi(\cdot)$, $j = 1, \ldots, d$, *then the optimal bandwidth will be of the form*

$$h = \left\{ \frac{4 \prod_{j=1}^{d} \sigma_j}{3 \sum_{j=1}^{d} \sigma_j^{-4} + 2 \sum_{j<\ell} \sigma_j^{-2} \sigma_\ell^{-2}} \frac{d}{n} \right\}^{\frac{1}{d+4}}, \tag{32}$$

and will reduce to (25) *as* $d = 1$. *These formulae also coincide with Eqs (4.12) and (3.28) of Silverman (1986), respectively.*

Now, if we consider $\kappa_j(\cdot)$ to be either skew-symmetric cases (3) or (4), then, a quick conclusion from Theorem 1 is made as follows.

Remark 2 *The optimal bandwidth following Approach (I) is derived as*

$$h_{\text{opt}} = \left\{ \frac{d\,(2\sqrt{\pi})^d \prod_{j=1}^{d} \sigma_j \int \kappa_j^2(t)dt}{n \sum_{j=1}^{d} \mu_1^2(\kappa_j) \sigma_j^{-2}} \right\}^{\frac{1}{d+2}}, \tag{33}$$

while Approach (II) leads us to find the optimal bandwidth via solving a non-linear equation of the form

$$-\frac{d}{n} \prod_{j=1}^{d} \int \kappa_j^2(t)dt + 2ABh_{opt}^{d+2} + 4ACh_{opt}^{d+4} = 0, \tag{34}$$

where constants A, B *and* C *are given by* (31). *For the SN case, the integral involved in* (33) *and* (34) *is given by* (20) *considering* λ_j *rather than* λ. *The unknown parameters* σ_j *can be substituted by* s_j, *the sample standard deviation of the* j*th variable already defined in* (5), *and the* λ_j *can be approximated by its MPLE discussed in Section 2.*

It can be easily shown that by setting $d = 1$ in (33) and (34), they simplify to their univariate versions (21) and (23), respectively. Figure 3 shows the surface plots of both optimal bandwidths given by Remark 2 for $d = 2$ and different sample sizes. It is observed that the behaviour of the optimal bandwidths in the bivariate case are similar to those of the univariate ones exhibited in Figure 1. In this case also Approach (II) is more appropriate than Approach (I) due to the same reasons.

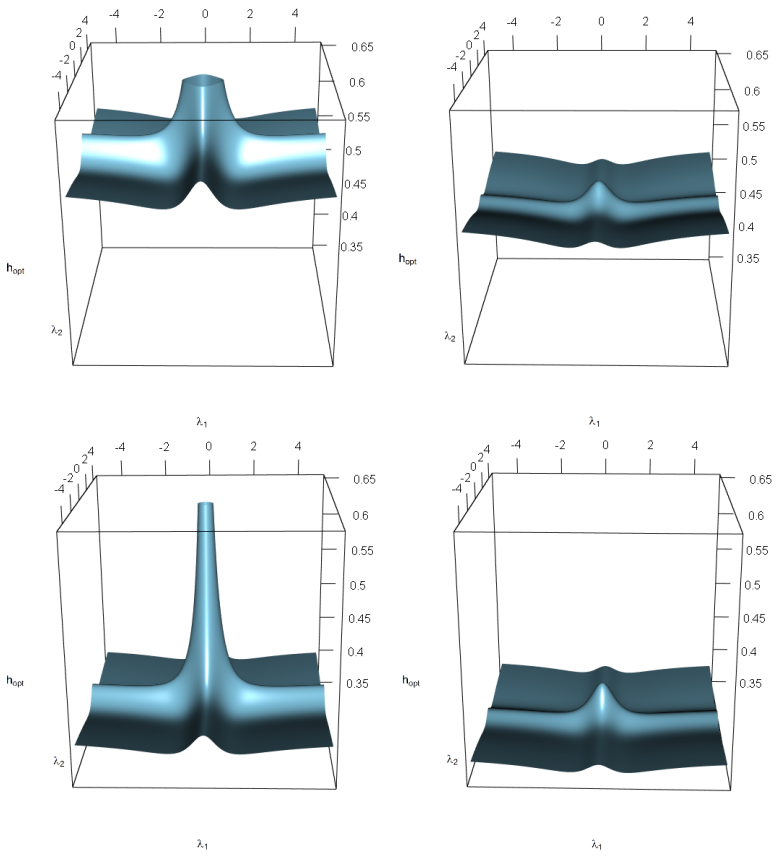

Figure 3: The plots of h_{opt} versus the pair of (λ_1, λ_2) for $\sigma_1 = \sigma_2 = 1$, $n = 50$ (top panels) and $n = 200$ (bottom panels), and based on approaches (I) (right panels) and (II) (left panels) given by (33) and (34), respectively.

4 Clustering via the semiparametric KDE

The clustering algorithm `pdfCluster`, which was briefly described in the introduction, consists of two main stages summarized by Algorithm 2. The first step of `pdfCluster` is

Algorithm 2: Clustering via KDE

1 Constructing a cluster tree and then obtaining the initial clusters.

2 Allocating possible unlabelled points to the clusters formed by Azzalini's Step 1.

done using "spatial tessellation" when the dimension is less than 6 (Azzalini and Menardi, 2014, Section 2.2), while it is carried out using "pairwise connections" for higher dimensions (Azzalini and Menardi, 2014, Section 2.3). In either cases, $\hat{f}(\mathbf{x})$ in (1) has been used with normal (or t_7) kernel. As it is mentioned in Section 1, the number of clusters in `pdfCluster` is not determined by the user while it is automatically specified by the procedure (in its Step 1) as the number of the modes of the estimated density. But this step sometimes leaves some points with no label. In the second step of `pdfCluster`, the unlabelled points are assigned to the existing clusters using a classification algorithm which works based on a nonparametric KDE borrowing $\hat{f}(\mathbf{x})$ in (1) once again. This approach and its improved version are explained by Azzalini and Torelli (2007) and Azzalini and Menardi (2014) (Section 2.2), respectively. The implementation of this clustering method is carried out by the R function `pdfCluster`() in the Package `pdfCluster`.

In this section, we intend to plug-in the $\hat{f}(\mathbf{x})$ which was obtained through a semi-parametric approach described by Algorithm 1 in both steps of `pdfCluster` presented by Algorithm 2. Thus, the resulting procedure will be a semi-parametric KDE clustering with skew kernels. In the sequel, we demonstrate this procedure by using some real data sets as well as some simulated samples. We apply two indexes for assessing the clusters obtained:

the ARI criterion proposed by Hubert and Arabie (1985), as an external validity metric, and the density-based Silhouette (DBS) information proposed by Menardi (2011) as an internal validity index. In contrast to the ordinary Rand index which only varies on the interval [0, 1], the ARI can also yield negative values and has an expected value of zero under random classification. However, the higher values of the ARI correspond to the better model. It should be noted that clustering techniques fall within the unsupervised learning algorithms where auxiliary information is not always attached to the observations. In such situations, a common approach is to use an internal validity measure (Ingrassia and Punzo, 2020), e.g. the Silhouette information. The DBS is an adaptation of the Silhouette information which is suitable for density-based clustering procedures. The interpretation of the DBS is the same as the existing Silhouette information, i.e. large values of the DBS are an indication of a well clustered data point while small values of the DBS mean low confidence in the clustering. Negative values of the DBS can also occur and it is usually evidence of an incorrect allocation of the observation.

4.1 Real data analysis

We present here the analysis of three real data sets whose dimensions are five, two and one, respectively. The first real data set which is called *olive oil* data, was originally presented by Forina et al (1983) and then used by other researchers for illustrating various clustering techniques. Among of them, we refer you to Azzalini and Torelli (2007) and Menardi (2011) who used this data set for illustrating the non-parametric `pdfCluster` procedure. The data set collected $n = 572$ species of olive oil with eight chemical measurements produced in various areas of Italy. Azzalini and Torelli (2007) carried out some manipulations on the data to reduce the data dimensionality and used only the first five principal components.

Here, we use the same data for all clustering methods being used. The scatter plot of the first two components of the *olive oil* data colored based on the actual labels are displayed in Figure 4. As it is observed from Figure 4, the green cluster can be potentially challenging to be correctly detected by a clustering method.

The next example considers the *geyser* data set which is available in the R package `sm` (Bowman and Azzalini, 2018). It consists of $n = 299$ observations representing the features eruption time in minutes (*duration*) and the waiting time before the eruption for the Old Faithful *geyser* in Yellowstone National Park, Wyoming. A version of this data set including $n = 272$ observations of eruption length has been analyzed by Bagnato et al (2017) from a clustering perspective. Unlike the preceding example, we are not able to compare the clustering results with true groupings, since it does not explicitly exist. Thus, in this example, we can only employ an internal validity measure e.g. the DBS, for comparing various clustering methods.

In the third example, we aim to examine the ability of the proposed clustering algorithm on a univariate data set. To this end, we work with the *duration* variable given by the *geyser* data explained in the previous example. Of course, in this example, we do not also have the actual groups, however, it seems that the data contain two groups, each located around a mode (see Figure 9). For the Gaussian and the t_7 kernels, we use the optimal bandwidth under the normality assumption with the setting `hmult` $=$ `1`, a scalar multiplied by the bandwidth (Azzalini and Menardi, 2014, Section 3.2), in the R function `pdfCluster` while for the asymmetric kernels SN and ST we use the ones obtained in the preceding section.

The distribution of the clusters as well as the DBS plots of some selected clustering methods implemented on the data sets are exhibited in Figures 5-10. The latter plots

display the DBS obtained for each observation and also return their median for each cluster. But, as overall accuracy measures for each clustering method, some specific quartiles of the DBSs of all observations are also reported by Table 1. From this table, one can pick the median as the final accuracy measure of a given clustering method. Although the mean is not a good candidate for this purpose in all scenarios due to the lack of robustness, it is also attached to Table 1. Figures 4 and 5 reveals that heavy tails kernels (t_7 and ST) have been more successful than the Gaussian and the SN kernels in detecting the problematic green cluster of the *olive oil* data, but, it is observed from the numerical metrics given by Table 1 that they have not been the best choice in overall. Moreover, among the Gaussian and the SN kernels, the `pdfCluster` with the latter kernel has recorded the best result according to the DBS information given by Table 1 and Figure 6.

Concerning the *geyser* data, the superiority of the fitting of the semi-parametric KDE with SN and ST kernels comparing to that of the non-parametric KDE with the symmetric kernels Gaussian and t_7 is revealed at the first glance on Figure 7. In other words, Figure 7 confirms the simulation results of the boundary bias problem given by Section 3.1.1. As for the clustering results on this set of data, Table 1 and Figure 8 show that the SN is the best kernel for the `pdfCluster` algorithm. With regard to the univariate data set *duration*, intuition from Table 1 and Figure 10 suggests that the Gaussian is the last choice as the kernel of the `pdfCluster` method. Figure 9 shows once again that using the asymmetric kernels SN and ST reduces the weights outside the boundary of the support (refer to Figure 2 as well).

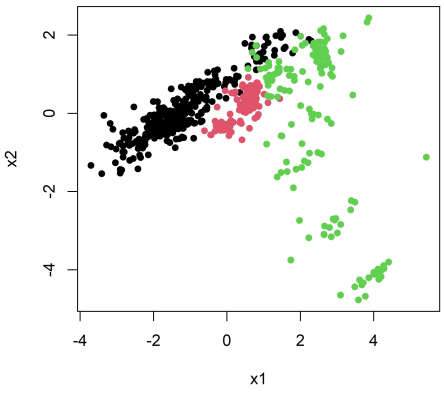

Figure 4: The distribution of the actual clusters in the space of the first two principal components of the *olive oil* data.

Table 1: Summary of DBS results of clusters obtained by the `pdfCluster` method with various kernels on three real data sets.

		DBS						ARI
Data	kernel	min	Q1	median	mean	Q3	max	
olive oil	Gaussian	-0.019	0.258	0.321	0.350	0.416	1.000	0.885
	t_7	-0.007	0.059	0.077	0.134	0.144	1.000	0.825
	SN	0.061	0.266	0.359	0.371	0.448	1.000	0.700
	ST	-0.025	0.057	0.072	0.118	0.109	1.000	0.676
geyser	Gaussian	0.041	0.097	0.175	0.205	0.303	1.000	
	t_7	0.041	0.097	0.175	0.205	0.303	1.000	
	SN	-0.008	0.159	0.260	0.266	0.342	1.000	
	ST	0.078	0.173	0.210	0.250	0.297	1.000	
duration	Gaussian	-0.005	0.076	0.130	0.158	0.198	1.000	
	t_7	0.029	0.088	0.143	0.167	0.214	1.000	
	SN	0.029	0.088	0.143	0.167	0.214	1.000	
	ST	0.029	0.088	0.143	0.167	0.214	1.000	

4.2 Simulation study

In this section, we are intending to carry out an extensive simulation in order to investigate the impact of using asymmetric kernels on the performance of the `pdfCluster` algorithm. The aim of this section is twofold: 1) to use the Monte Carlo simulation under some sce-

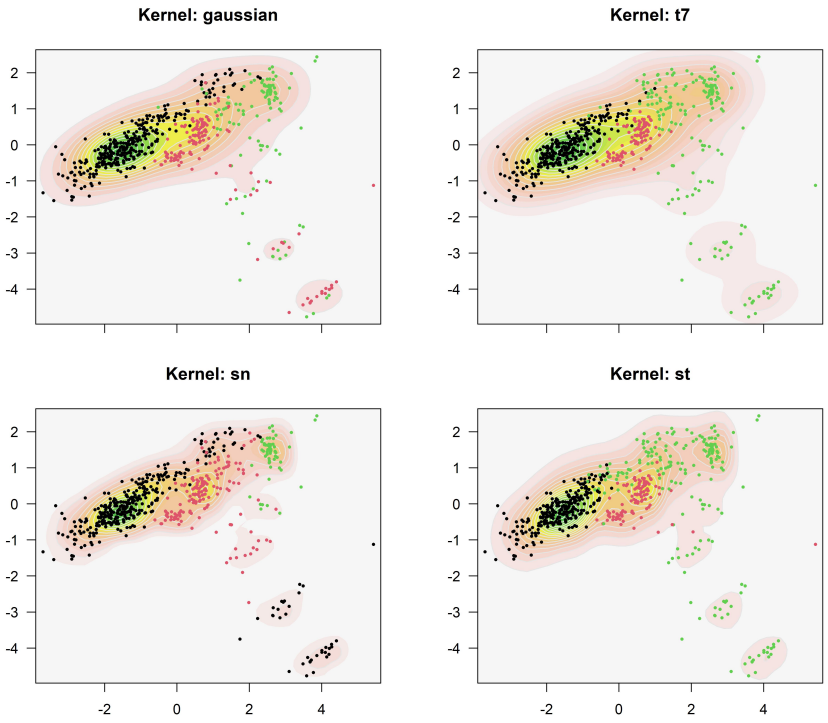

Figure 5: Clustering results of *olive oil* data as well as contour levels of the various fitted kernel estimates.

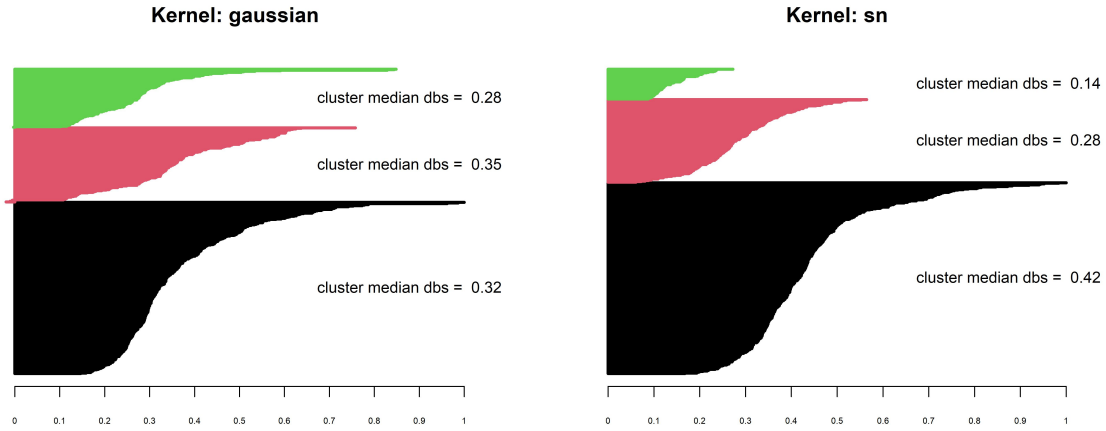

Figure 6: dbs plots of three clusters obtained by the `pdfCluster` method with Gaussian and SN kernels on the *olive oil* data.

narios and 2) to check the sensitivity of the chosen sample in the real data sets using the bootstrap re-sampling method. For both mechanisms, the number of replications is considered to be 1000. For each replication, the median of the DBS of the observations is recorded

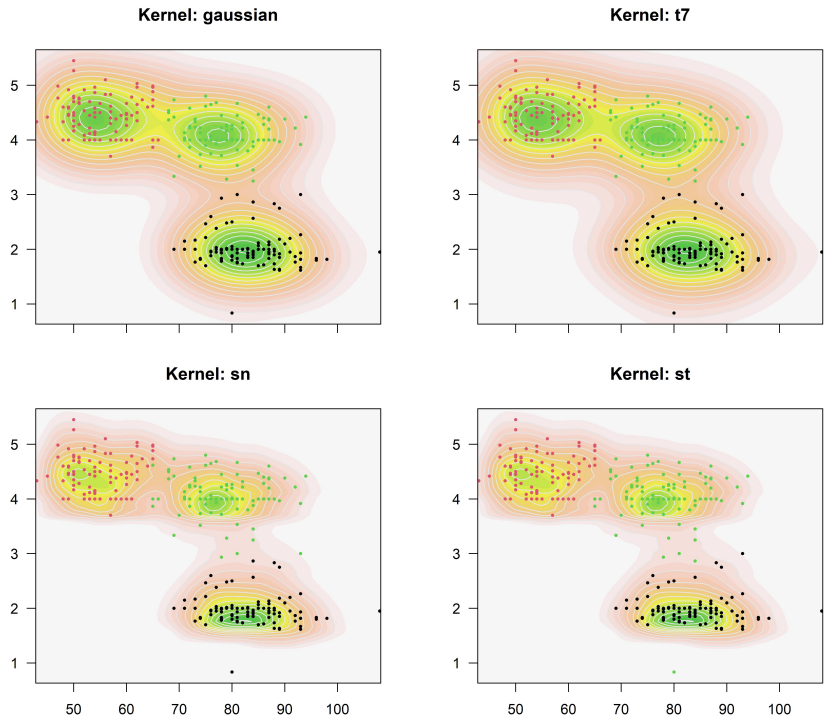

Figure 7: Clustering results of *geyser* data as well as contour levels of the various fitted kernel estimates.

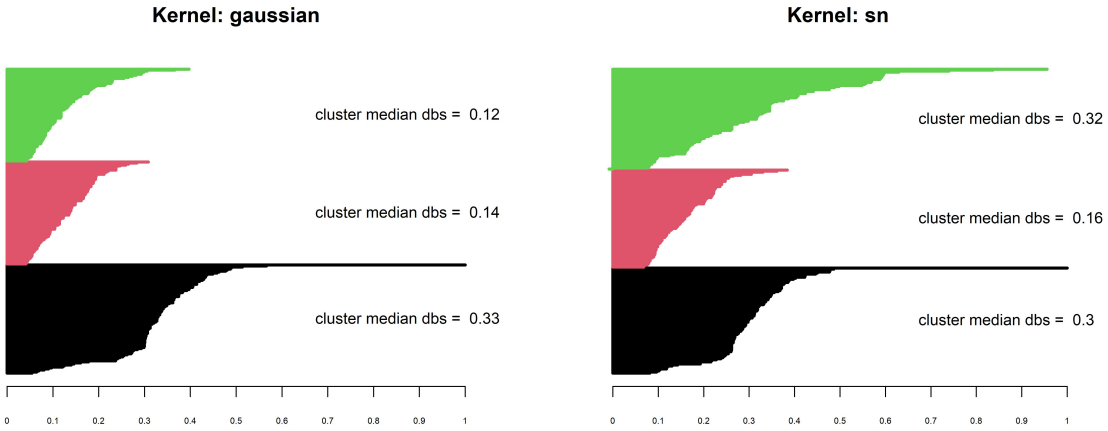

Figure 8: dbs plots of three clusters obtained by the `pdfCluster` method with Gaussian and SN kernels on the *geyser* data.

as the representative candidate of this internal criterion. To visualize the distributions of the ARI and the DBS, the adjusted boxplot is utilized (Hubert and Vandervieren, 2008). It is to be noted that for the *kmeans* clustering we only compute the ARI since it belongs

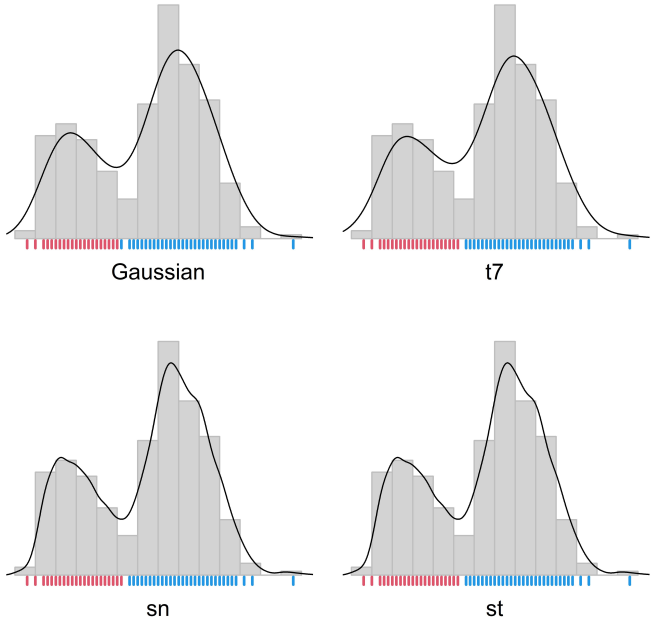

Figure 9: Clustering results of *duration oil* data as well as contour levels of the various fitted kernel estimates.

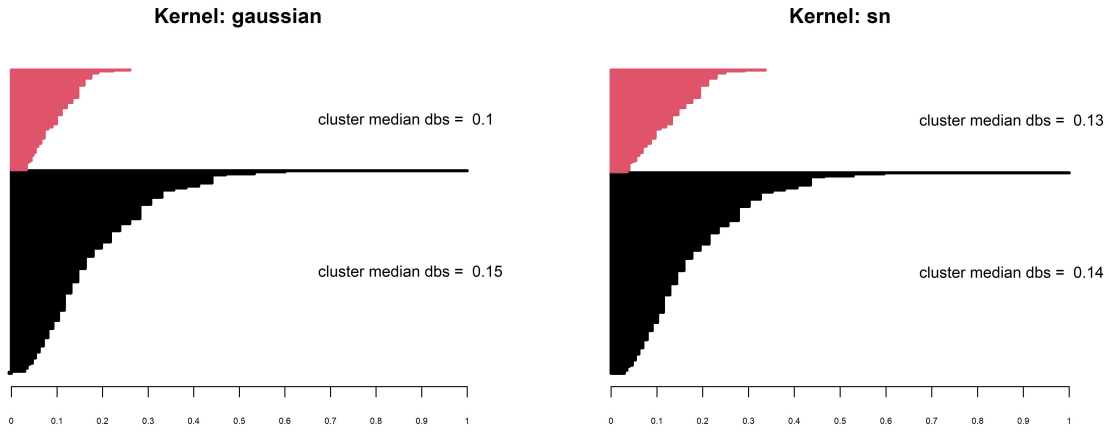

Figure 10: dbs plots of two clusters obtained by the `pdfCluster` method with Gaussian and SN kernels on the *duration* data.

to the distance-based algorithms, not the density-based ones.

4.2.1 Monte Carlo simulation

We consider the following three simulation scenarios for evaluating the `pdfCluster` procedure with the mentioned symmetric and asymmetric kernels.

- *Scenario 1*: Our first scenario deals with a sample of size 180 of simulated data from two sub-populations whose sizes are 100 and 80, respectively. More precisely, both portions of the data set come from bivariate skew-t distribution with location vectors $(10, 11)^\top$ and $(5, 3)^\top$, dispersion matrices I_2 and $I_2 + \mathbf{1}$, slant vectors $(9, 5)^\top$ and $(5, 3)^\top$ and degrees of freedoms $\nu_1 = 2$ and $\nu_2 = 5$, respectively.

- *Scenario 2*: The second scenario uses a sample of size 300 of simulated bivariate data from three sub-populations. More specifically, this data set consists of two variables which are independent and identically distributed observations. Each column gathers a sample of size 150 from the standard SN distribution (3) with $\lambda = 2$, a sample of size 100 from $N(3, \sigma = 0.3)$ and a sample of size 50 from the SN distribution with the location parameter $\xi = 8$, the unit scale parameter, and $\lambda = 2$.

- *Scenario 3*: In the third scenario, we apply our method on a sample of size 600 of simulated data from three sub-populations with sample sizes 200, 300 and 100, respectively. All of the populations follow from bivariate skew-t distribution with location vectors $(10, 11)^\top$, $(5, 3)^\top$ and $(2, 20)^\top$, dispersion matrices I_2, $I_2 + \mathbf{1}$ and $I_2 + 2 \times \mathbf{1}$, slant vectors $(9, 5)^\top$, $(5, 3)^\top$ and $(0, 3)^\top$ and degrees of freedoms $\nu_1 = 5$, $\nu_2 = 2$ and $\nu_3 = 7$, respectively.

- *Scenario 4*: In this case, a sample of size 180 of simulated data from two sub-populations is drawn. Both groups of the data set come from 5-variate skew-t distribution with location vectors $\mu_1 = (10, 11, 5, 6, -6)^\top$ and $\mu_2 = (5, 3, 0, 0, 1)^\top$,

dispersion matrices $I_5 + 0.5 \times \mathbf{1}$ and $I_5 + \mathbf{1}$, slant vectors $\alpha_1 = (9, 5, 2, 0, 1)^\top$ and $\alpha_2 = (5, 3, 0, 2, 3)^\top$ and degrees of freedoms $\nu_1 = 2$ and $\nu_2 = 5$, respectively.

The Monte Carlo distributions of the DBS and the ARI under the above-mentioned scenarios are given by Table 2 and Figure 11. From the results of the ARI metric it is concluded that it does not matter which kernel is selected, the density-based `pdfCluster` has been reasonably successful to detect the right clusters comparing to the conventional clustering *kmeans* (except the last scenario). But, it is hard to choose between the symmetric and asymmetric kernels by using this criterion. Thus, the DBS information can help us in such circumstances. The DBS results indicate that the `pdfCluster` algorithm supplied by the asymmetric kernel SN always outperforms the other kernels and its accuracy is significantly higher than those of the symmetric kernels in Scenario 2. However, the ST kernel has been the best option in the mentioned scenario but is not considerably better than the SN.

4.2.2 More explorations on real-life data

In Section 4.1, we analyzed three real data sets and observed that the semi-parametric `pdfCluster` with asymmetric kernel SN outperformed the other competitors according to the internal validity measure of DBS. In this section, we are seeking to present more distributional information on the DBS metric computed for the `pdfCluster` algorithm. To this end, we employ the non-parametric bootstrap re-sampling method. The results of the bootstrap are presented by Table 3 and Figure 12. These results coincide with the results obtained in Section 4.1 and confirm the appropriateness of the SN kernel on the three real data sets once again.

Table 2: Central tendencies, quartiles and standard deviation of the empirical distribution of the DBS and the ARI of various clustering methods for the scenarios given by the previous section.

		DBS					ARI				
Scenario	method	median	mean	Q1	Q3	sd	median	mean	Q1	Q3	sd
	kmeans	-	-	-	-	-	0.846	0.851	0.798	0.922	0.085
	Gaussian	0.112	0.132	0.075	0.171	0.076	0.978	0.964	0.956	0.988	0.054
1	t_7	0.115	0.137	0.075	0.180	0.081	0.956	0.956	0.934	0.978	0.055
	SN	0.132	0.152	0.087	0.197	0.092	0.956	0.926	0.913	0.978	0.107
	ST	0.112	0.130	0.073	0.171	0.077	0.934	0.933	0.913	0.978	0.071
	kmeans						0.965	0.920	0.954	0.977	0.161
	Gaussian	0.154	0.163	0.123	0.199	0.053	0.988	0.986	0.977	1.000	0.013
2	t_7	0.159	0.163	0.123	0.199	0.054	0.988	0.985	0.977	1.000	0.014
	SN	0.207	0.215	0.166	0.246	0.087	1.000	0.943	0.988	1.000	0.159
	ST	0.209	0.215	0.166	0.249	0.074	1.000	0.966	0.988	1.000	0.121
	kmeans						0.876	0.840	0.857	0.891	0.128
	Gaussian	0.102	0.113	0.064	0.154	0.061	0.872	0.856	0.852	0.888	0.098
3	t_7	0.094	0.108	0.064	0.144	0.061	0.868	0.847	0.849	0.888	0.111
	SN	0.116	0.121	0.063	0.168	0.073	0.869	0.865	0.849	0.888	0.051
	ST	0.065	0.085	0.041	0.107	0.062	0.849	0.849	0.823	0.876	0.055
	kmeans						0.919	0.907	0.851	0.966	0.067
	gaussian	0.169	0.186	0.126	0.234	0.086	0.978	0.966	0.956	1.000	0.050
4	t7	0.167	0.176	0.126	0.213	0.074	0.978	0.971	0.956	1.000	0.033
	sn	0.202	0.216	0.149	0.265	0.097	0.966	0.939	0.932	0.978	0.107
	st	0.174	0.187	0.130	0.226	0.083	0.956	0.927	0.913	0.978	0.110

5 Conclusion and discussion

This paper proposed a semi-parametric KDE with a more flexible family of kernels including SN and ST. Furthermore, the optimal bandwidth under the mentioned kernels was obtained. Through a simulation study, we observed that the proposed estimator not only reduces boundary bias but also it is closer to the true density comparing to that of the usual KDE employing the Gaussian kernel. Then, we applied the proposed semi-parametric KDE in the density-based `pdfCluster` of Azzalini and Torelli (2007) and Azzalini and Menardi (2014) to improve the performance of this clustering method when we are exposed to skewed

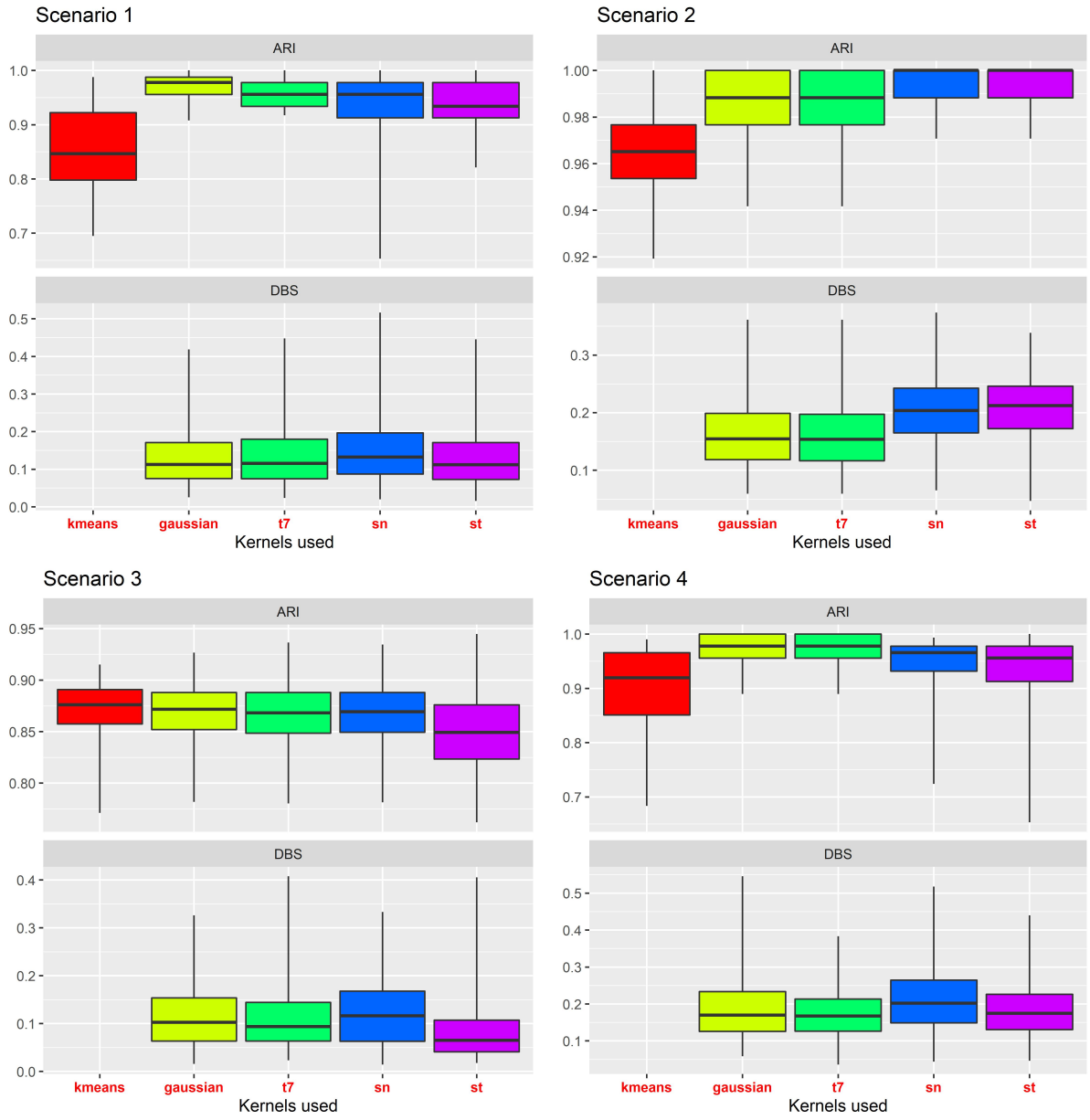

Figure 11: Empirical distribution of the DBS and the ARI of various clustering methods for the various scenarios used in the previous section.

and bounded distributions of data. The application of the semi-parametric `pdfCluster` was presented on three real data sets. Moreover, a relatively comprehensive simulation study was conducted based on three artificial scenarios. For the comparison purpose, we used the ARI and the DBS as external and internal validation criteria. In practice, we found that if either the data supports are bounded or the data are asymmetric in some

Table 3: Central tendencies, quartiles and standard deviation of the empirical distribution of the DBS of the `pdfCluster` method with various kernels on the real data sets.

Data	kernel	median	mean	Q1	Q3	sd
Olive oil	Gaussian	0.201	0.208	0.165	0.230	0.070
	t_7	0.192	0.202	0.172	0.213	0.053
	SN	0.245	0.284	0.177	0.330	0.192
	ST	0.189	0.183	0.086	0.264	0.103
geyser	Gaussian	0.224	0.234	0.173	0.286	0.075
	t_7	0.231	0.235	0.170	0.287	0.075
	SN	0.239	0.238	0.190	0.281	0.068
	ST	0.193	0.204	0.154	0.255	0.076
duration	Gaussian	0.133	0.161	0.116	0.209	0.054
	t_7	0.133	0.159	0.115	0.202	0.052
	SN	0.142	0.157	0.119	0.195	0.062
	ST	0.140	0.155	0.118	0.190	0.069

clusters, then, it would be better to use the semi-parametric `pdfCluster` model proposed. In the contrary, when the underlying data are symmetric in each cluster, no matter which kernel is selected for the `pdfCluster`. Thus, in such a situation, if we are exposed to a big data set with thousands of observations and variables, it would be logical to employ the non-parametric `pdfCluster` rather than its semi-parametric counterpart due to the computation time. To see the computation time of the `pdfCluster` algorithm based on different symmetric and asymmetric kernels, we replicated Scenario 1-3 given by Section 4.2.1 for 1000 times and report the average runtime in Table 4. Although the computation time of the semi-parametric `pdfCluster` algorithm is negligible, the non-parametric `pdfCluster` is more time-efficient than the semi-parametric one since in the latter we should estimate the nuisance parameters involved while in the former we do not have such an extra step.

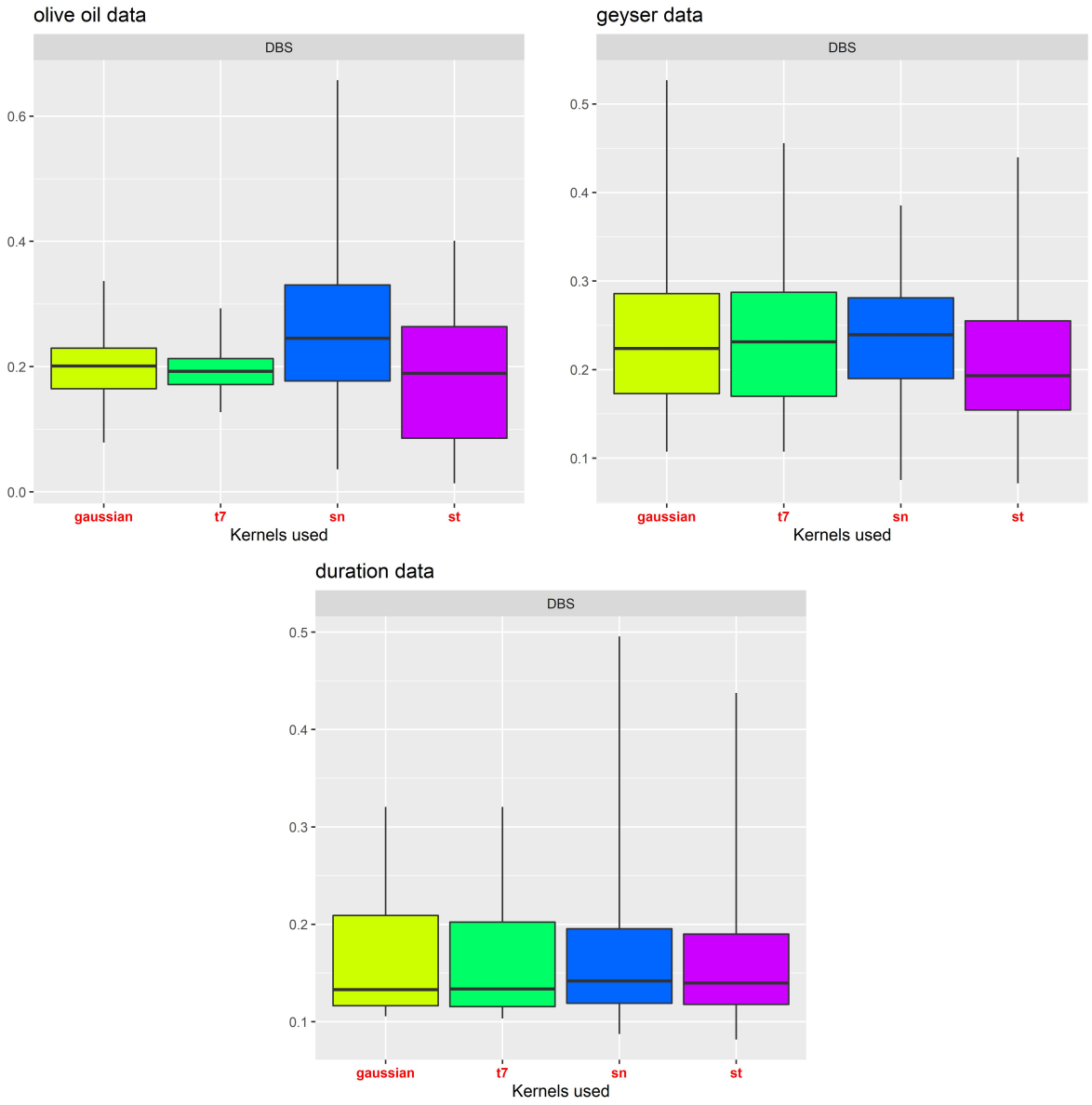

Figure 12: Empirical distributions of the DBS of the `pdfCluster` method with various kernels implemented on the real data sets.

Acknowledgment

The authors are thankful to the reviewers for their constructive comments that significantly improved this paper.

Table 4: The average computation time elapsed (in Sec.) for implementation of the non-parametric and semi-parametric `pdfCluster` algorithms on the data generated under some scenarios given by Section 4.2.1 based on 1000 Monte Carlo replications.

Scenario	Gaussian	t_7	SN	ST
1	0.30	0.31	1.24	1.32
2	0.62	0.61	1.80	2.10
3	0.81	0.76	2.55	3.5

Appendix

Proof of Lemma 1:

Proof 3 *The bias of $\hat{f}$ evaluated at $\mathbf{x}$ is obtained as*

$$\begin{aligned} \text{Bias}\hat{f}(\mathbf{x}) &= \int \prod_{j=1}^{d} \frac{1}{h_j} \kappa_j \left(\frac{x_j - y_j}{h_j} \right) f(\mathbf{y}) d\mathbf{y} - f(\mathbf{x}) \\ &= \int \prod_{j=1}^{d} \kappa_j (t_j) \left\{ f\left(\mathbf{x} - H^{\frac{1}{2}} \mathbf{t} \right) - f(\mathbf{x}) \right\} d\mathbf{t}, \end{aligned} \tag{35}$$

where $H = \text{Diag}(h_1^2, \ldots, h_d^2)$. Now using the multivariate Taylor's theorem (Wand and Jones, 1995, pp. 94) we get

$$\begin{aligned} \text{Bias}\hat{f}(\mathbf{x}) &= \int \prod_{j=1}^{d} \kappa_j (t_j) \{ -\mathbf{t}^\top H^{\frac{1}{2}} \mathcal{D}_f(\mathbf{x}) + \frac{1}{2} \mathbf{t}^\top H^{\frac{1}{2}} \mathcal{H}_f(\mathbf{x}) H^{\frac{1}{2}} \mathbf{t} + \text{o}(\text{Tr}(H)) \} d\mathbf{t} \\ &= -\int \prod_{j=1}^{d} \kappa_j (t_j) \sum_{j=1}^{d} t_j h_j f'_{(j)} d\mathbf{t} \\ &\quad + \frac{1}{2} \int \prod_{j=1}^{d} \kappa_j (t_j) \left(t_1^2 h_1^2 f''_{(1,1)} + t_1 h_1 \sum_{j=2}^{d} t_j h_j f''_{(j,1)} + \cdots \right. \\ &\quad \left. + t_d^2 h_d^2 f''_{(d,d)} + t_d h_d \sum_{j=1}^{d-1} t_j h_j f''_{(j,d)} \right) d\mathbf{t} + \text{o}(\text{Tr}(H)), \end{aligned}$$

where $\mathcal{D}_f(\cdot)$ and $\mathcal{H}_f(\cdot)$ are respectively the vector of first derivatives and the Hessian matrix already defined. So the desired result of the bias follows. As for the variance, essentially the same algebraic manipulations as in the one-dimensional case yield (29).

Proof of Theorem 1:

Proof 4 *Considering* $f(\mathbf{x}) = \prod_{j=1}^{d} \phi(x_j/\sigma_j)/\sigma_j$, *it can be easily seen that*

$$\begin{aligned} &\int f'_{(j)} f'_{(\ell)} d\mathbf{x} = \int f'_{(j)} f''_{(j,j)} d\mathbf{x} = \int f'_{(j)} f''_{(\ell,j)} d\mathbf{x} \\ &= \int f''_{(j,j)} f''_{(\ell,j)} d\mathbf{x} = 0, \qquad \ell, j \in \{1, \dots d\},\ \ell \neq j \end{aligned} \tag{36}$$

and also

$$\begin{aligned} \int f''_{(\ell,j)} f''_{(\ell',j')} d\mathbf{x} = 0, \qquad & \ell', j' \in \{1, \dots d\}, \\ & \ell' \neq j',\ \ell \neq \ell', j \neq j'. \end{aligned} \tag{37}$$

Thus, by using Lemma 1 and assuming $h_j = h$, $j = 1, \dots, d$ *we get*

$$\begin{aligned} \int \mathrm{Bias}^2 f(\mathbf{x}) d\mathbf{x} &= h^2 \sum_{j=1}^{d} \mu_1^2(\kappa_j) \int \left(f'_{(j)}\right)^2 d\mathbf{x} \\ &+ \frac{1}{4} h^4 \sum_{j=1}^{d} \mu_2^2(\kappa_j) \int \left(f''_{(j,j)}\right)^2 d\mathbf{x} \\ &+ \frac{1}{4} h^4 2 \sum_{j<\ell} \mu_2(\kappa_j) \mu_2(\kappa_\ell) \int f''_{(j,j)} f''_{(\ell,\ell)} d\mathbf{x} \\ &+ \frac{1}{4} h^4 2 \sum_{j<\ell} \mu_1^2(\kappa_j) \mu_1^2(\kappa_\ell) \int \left(f''_{(\ell,j)}\right)^2 d\mathbf{x}, \end{aligned} \tag{38}$$

where

$$\begin{aligned} \int \left(f'_{(j)}\right)^2 d\mathbf{x} &= \int \frac{x_j^2}{\sigma_j^4} \prod_{\ell=1}^{d} \frac{1}{\sigma_\ell^2} \phi^2\left(\frac{x_\ell}{\sigma_\ell}\right) d\mathbf{x} \\ &= \frac{\sigma_j^{-2}}{\prod_{j=1}^{d} \sigma_j} \int t_j^2 \phi^2(t_j) dt_j \prod_{\substack{\ell=1 \\ \ell \neq j}}^{d} \int \phi^2(t_\ell) dt_\ell \\ &= \frac{\sigma_j^{-2}}{\prod_{j=1}^{d} \sigma_j} E\left(Z^2 \phi(Z)\right) \left\{E\left(\phi(Z)\right)\right\}^{d-1} \\ &= \frac{(2\sqrt{\pi})^{-d}}{2 \prod_{j=1}^{d} \sigma_j} \sigma_j^{-2}, \end{aligned} \tag{39}$$

$$\begin{aligned}\int \left(f''_{(j,j)}\right)^2 d\mathbf{x} &= \int \left\{\frac{1}{\sigma_j^3}\phi\left(\frac{x_j}{\sigma_j}\right) - \frac{x_j^2}{\sigma_j^5}\phi\left(\frac{x_j}{\sigma_j}\right)\right\}^2 dx_j \prod_{\substack{\ell=1\\ \ell\neq j}}^{d} \int \frac{1}{\sigma_\ell^2}\phi^2\left(\frac{x_\ell}{\sigma_\ell}\right) dx_\ell \\ &= \sigma_j^{-5}\left\{E\left(\phi(Z)\right) + E\left(Z^4\phi(Z)\right) - 2E\left(Z^2\phi(Z)\right)\right\} \\ &\quad \times \left\{E(\phi(Z))\right\}^{d-1} \prod_{\substack{\ell=1\\ \ell\neq j}}^{d} \frac{1}{\sigma_\ell} \\ &= \frac{3(2\sqrt{\pi})^{-d}}{4\prod_{j=1}^{d}\sigma_j}\sigma_j^{-4}, \end{aligned} \tag{40}$$

$$\begin{aligned}\int f''_{(j,j)} f''_{(\ell,\ell)} d\mathbf{x} &= \int \left\{\frac{1}{\sigma_j^3}\phi\left(\frac{x_j}{\sigma_j}\right) - \frac{x_j^2}{\sigma_j^5}\phi\left(\frac{x_j}{\sigma_j}\right)\right\} \frac{1}{\sigma_\ell}\phi\left(\frac{x_\ell}{\sigma_\ell}\right) \\ &\quad \times \left\{\frac{1}{\sigma_\ell^3}\phi\left(\frac{x_\ell}{\sigma_\ell}\right) - \frac{x_\ell^2}{\sigma_\ell^5}\phi\left(\frac{x_\ell}{\sigma_\ell}\right)\right\} \frac{1}{\sigma_j}\phi\left(\frac{x_j}{\sigma_j}\right) dx_\ell dx_j \\ &\quad \times \prod_{\substack{m=1\\ m\neq j,\ell}}^{d} \int \frac{1}{\sigma_m^2}\phi^2\left(\frac{x_m}{\sigma_m}\right) dx_m \\ &= \sigma_j^{-3}\sigma_\ell^{-3}\left\{E^2\left(\phi(Z)\right) + E^2\left(Z^2\phi(Z)\right) - 2E\left(Z^2\phi(Z)\right)E\left(\phi(Z)\right)\right\} \\ &\quad \times \left\{E(\phi(Z))\right\}^{d-2} \prod_{\substack{m=1\\ m\neq j,\ell}}^{d} \frac{1}{\sigma_m} \\ &= \frac{(2\sqrt{\pi})^{-d}}{4\prod_{j=1}^{d}\sigma_j}\sigma_j^{-2}\sigma_\ell^{-2}, \end{aligned} \tag{41}$$

and similarly

$$\begin{aligned}\int \left(f''_{(j,\ell)}\right)^2 d\mathbf{x} &= \frac{\sigma_j^{-2}\sigma_\ell^{-2}}{\prod_{j=1}^{d}\sigma_j} E^2\left(Z^2\phi(Z)\right)\left\{E\left(\phi(Z)\right)\right\}^{d-2} \\ &= \frac{(2\sqrt{\pi})^{-d}}{4\prod_{j=1}^{d}\sigma_j}\sigma_j^{-2}\sigma_\ell^{-2}. \end{aligned} \tag{42}$$

Finally, from (29) *the desired result follows.*

Compliance with ethical standards

Funding: This work was based upon research supported by the South African National Research Foundation (SRUG190308422768 nr. 120839) SARChI Research Chair in Computational and Methodological Statistics (UID: 71199), STATOMET at the Department

of Statistics at the University of Pretoria.

Conflict of interest: The authors declare that they have no conflict of interest.

Ethical Conduct: Not applicable.

Data Availability Statements: The *geyser* and the *olive oil* data sets are available in the R packages `sm` and `pdfCluster`, respectively.

References

Abadir, K.M., Lawford, S.: Optimal asymmetric kernels, Economics Letters. **83**, 61–68 (2004)

Azzalini, A. and Arellano-Valle, R.B.: Maximum penalized likelihood estimation for skew-normal and skew-t distributions, Journal of statistical planning and inference, **143**, 419–433 (2013)

Azzalini A, Torelli N: Clustering Via Nonparametric Density Estimation. Statistics and Computing. **17**, 71–80 (2007)

Azzalini, A., Menardi, G.: Clustering via Nonparametric Density Estimation: The R Package pdfCluster. Journal of Statistical Software. **57**, 1–26, (2014)

Azzalini A., Salehi M.: Some Computational Aspects of Maximum Likelihood Estimation of the Skew-t Distribution. In: Bekker A., Chen G., Ferreira J. (eds) Computational and Methodological Statistics and Biostatistics. Emerging Topics in Statistics and Biostatistics. Springer, Cham. (2020) http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-030-42196-0_1

Bagnato, L., Punzo, A. and Zoia, M. G.: The multivariate leptokurtic-normal distribution and its application in model-based clustering. Canadian Journal of Statistics, **45**, 95–119 (2017)

Bowman, A. W. and Azzalini, A.: R package 'sm': nonparametric smoothing methods (version 2.2-5.6) (2018) URL http://www.stats.gla.ac.uk/ adrian/sm

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E.: Model-based clustering and classification for data science: with applications in R (Vol. 50). Cambridge University Press.

Bouezmarni, T. and Scaillet, O.: Consistency of Asymmetric Kernel Density Estimators and Smoothed Histograms with Application to Income Data. Econometric Theory. **21**, 390–412 (2005)

Bowman A.W. and Azzalini, A.: Applied Smoothing Techniques for Data Analysis. Claredon Press, Oxford (1997)

Chacon, Jose E.: A Population Background for Nonparametric Density-Based Clustering. Statist. Sci. **30**, 518–532 (2015)

Chen, S.: Beta kernel estimators for density functions. Comput Stat Data Anal, **31**, 131–145 (1999)

Chen, S.X.: Probability Density Function Estimation Using Gamma Kernels. Annals of the Institute of Statistical Mathematics. **52**, 471–480 (2000)

Fernandez, M. and Monteiro, P. K.: Central limit theorem for asymmetric kernel functionals, Ann. Inst. Statist. Math., **57** 425–442 (2005)

Forina, M., Armanino, C., Lanteri, S., Tiscornia, E.: Classification of olive oils from their fatty acid composition. In: Martens, M., Russwurm, H.J. (eds.) Food Research and Data Analysis, 189–214. Appl. Sci., London (1983)

Fraley, C., and Raftery, A. E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association, **97**(458), 611–631 (2002)

Hjort, N.L. and Glad, I.K.: Nonparametric Density Estimation with a Parametric Start. Ann. Statist. **23**, 882–904 (1995)

Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification. **2**, 193–218 (1985)

Hubert, M. and Vandervieren, E.: An adjusted boxplot for skewed distributions, Computational Statistics and Data Analysis. **52**, 5186–5201 (2008)

Ingrassia, S., Punzo, A. Cluster Validation for Mixtures of Regressions via the Total Sum of Squares Decomposition. J Classif **37**, 526–547 (2020)

Kuruwita,C.N., Kulasekera,K.B., Padgett,W.J.: Density estimation using asymmetric kernels and Bayes bandwidths with censored data, Journal of Statistical Planning and Inference. **140**, 1765–1774 (2010)

Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B.: Identifying Mixtures of Mixtures Using Bayesian Estimation, Journal of Computational and Graphical Statistics, **26**, 285–295 (2017)

Marron, J. S., Ruppert, D.: Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society: Series B (Methodological). **56**, 653–671 (1994)

Mazza, A. and Punzo, A.: Discrete Beta Kernel Graduation of Age-Specific Demographic Indicators. In: Ingrassia S., Rocci R., Vichi M. (Eds.), New Perspectives in Statistical Modeling and Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 127—134, Berlin Heidelberg: Springer-Verlag (2011)

Mazza, A. and Punzo, A.: Using the Variation Coefficient for Adaptive Discrete Beta Kernel Graduation. In: Giudici P., Ingrassia S., Vichi M. (Eds.), Statistical Models for Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 225-–232, Switzerland: Springer International Publishing (2013a)

Mazza, A. and Punzo, A.: Graduation by Adaptive Discrete Beta Kernels. In: Giusti A., Ritter G., Vichi M. (Eds.), Classification and Data Mining, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 243–250, Berlin Heidelberg: Springer-Verlag (2013b)

Mazza, A. and Punzo, A.: DBKGrad: An R package for mortality rates graduation by fixed and adaptive discrete beta kernel techniques. Journal of Statistical Software, **57**, 1–18 (2014)

Mazza, A. and Punzo, A.: Bivariate discrete beta kernel graduation of mortality data. Lifetime Data Analysis, **21**, 419—433 (2015)

McNicholas, P. D.: Mixture model-based classification. CRC press (2016)

Menardi G.: Density based Silhouette diagnostics for clustering methods. Statistics and Computing. **21**, 295–308 (2011)

Menardi G., Azzalini, A.: An advancement in clustering via nonparametric density estimation. Statistics and Computing. **24**, 753—767 (2014)

Millard, S.: Contributions to Mixture Regression Modelling with Applications in Industry. PhD Thesis, University of Pretoria (2019)

Moss, J., Tveten, M.: kdensity: An R package for kernel density estimation with parametric starts and asymmetric kernels. Journal of Open Source Software, **4**, 1566 (2019)

Lee, S., McLachlan, G.J.: Finite mixtures of multivariate skew t-distributions: some recent and new results. Stat Comput. **24**, 181–202 (2014)

Lin, TI., Lee, JC., Hsieh, WJ.: Robust Mixture Modelling Using the Skew-t Distribution. Statistics and Computing. **17**, 81–92 (2007)

Loperfido, N.: Finite Mixtures, Projection Pursuit and Tensor Rank: a Triangulation. Advances in Data Analysis and Classification **31**, 145-173 (2019)

Punzo, A.: Discrete Beta-Type Models. In: Locarek-Junge H., Weihs C. (Eds.), Classification as a Tool for Research, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 253–261, Berlin Heidelberg: Springer-Verlag (2010)

Rattihalli, R.N., Patil, S.B.: Data Dependent Asymmetric Kernels for Estimating the Density Function. Sankhya A. **83**, 155—186 (2021)

Salehi, M., Azzalini, A.: On application of the univariate Kotz distribution and some of its extensions, METRON. **76**, 177–201 (2018)

Salehi, M., Doostparast, M.: Expressions for moments of order statistics and records from the skew-normal distribution in terms of multivariate normal orthant probabilities, Stat. Methods Appl. **24**, 547–568 (2015)

Saulo, H., Leiva, V., Ziegelmann, F.A. et al: A nonparametric method for estimating asymmetric densities based on skewed Birnbaum-Saunders distributions applied to environmental data. Stoch Environ Res Risk Assess. **27**, 1479–1491 (2013)

Silverman, BW.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)

Tomarchio, S. D. and Punzo, A.: Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. Journal of the Royal Statistical Society: Series A, **182**, 1247—1266 (2019)

Vrbik, I. and McNicholas, P.D.: Analytic calculations for the EM algorithm for multivariate skew-t mixture models, Statistics & Probability Letters, **82**, 1169–1174 (2012)

Wand, M.P. and Jones, M.C.: Kernel Smoothing. Chapman & Hall, London (1995)