Robust Kernel Density Estimation with Median-of-Means principle

Pierre Humbert [*1] **Batiste Le Bars** [*2] **Ludovic Minvielle** [*3]

Abstract

In this paper, we introduce a robust nonparametric density estimator combining the popular Kernel Density Estimation method and the Median-of-Means principle (MoM-KDE). This estimator is shown to achieve robustness for a large class of anomalous data, potentially adversarial. While previous works only prove consistency results under very specific contamination models, this work provides finite-sample high-probability error-bounds without any prior knowledge on the outliers. To highlight the robustness of our method, we introduce an influence function adapted to the considered $\mathcal{O} \cup \mathcal{I}$ framework. Finally, we show that MoM-KDE achieves competitive results when compared with other robust kernel estimators, while having significantly lower computational complexity.

1. Introduction

Over the past years, the task of learning in the presence of outliers has become an increasingly important objective in both statistics and machine learning. Indeed, in many situations, training data can be contaminated by undesired samples, which may badly affect the resulting learning task, especially in adversarial settings. Therefore, building robust estimators and algorithms that are resilient to outliers is becoming crucial in many learning procedures. In particular, the inference of a probability density function from a contaminated random sample is of major concern.

Density estimation methods are mostly divided into parametric and nonparametric techniques. Among the nonparametric family, the Kernel Density Estimator (KDE) is arguably the most known and used for both univariate and multivariate densities (Parzen, 1962; Silverman, 1986; Scott, 2015), but it is also known to be sensitive to datasets contaminated by outliers (Kim and Scott, 2011; 2012; Vandermeulen and Scott, 2014). The construction of a robust KDE is therefore an important area of research that can have useful applications, such as anomaly detection and resilience to adversarial data corruption. Yet, only few works have proposed such a robust estimator.

Kim and Scott (2012) proposed to combine KDE with ideas from M-estimation to construct the so-called Robust Kernel Density Estimator (RKDE). However, no consistency results were provided and robustness was rather shown experimentally. Later, RKDE was proven to converge to the true density, however at the condition that the dataset remains uncorrupted (Vandermeulen and Scott, 2013). More recently, Vandermeulen and Scott (2014) proposed another robust estimator, called Scaled and Projected KDE (SPKDE). Authors proved the L_1-consistency of SPKDE under a variant of the Huber's ε-contamination model (Huber, 1992) where two strong assumptions are made. First, the contamination parameter ε is assumed to be known, and second, the outliers must be uniform over the support of the true density. Unfortunately, as they did not provide rates of convergence, it still remains unclear at which speed SPKDE converges to the true density. Finally, both RKDE and SPKDE require iterative algorithms to compute their estimators, thus increasing the overall complexity of their construction. For theoretical findings, the recent work of Liu and Gao (2019) managed to obtain minimax optimal rates for kernel density estimates, but in the quite restrictive Huber's model (see Sec. 2.1).

In statistical analysis, another idea to construct robust estimators is to use the Median-of-Means principle (MoM). Introduced by Nemirovsky and Yudin (1983), Jerrum et al. (1986), and Alon et al. (1999), the MoM was first designed to estimate the mean of a real random variable. It relies on the simple idea that rather than taking the average of all the observations, the sample is split in several non-overlapping blocks over which the mean is computed. The MoM estimator is then defined as the median of these means. Being easy to compute, the MoM properties have been studied by

[*]Authors in alphabetical order
[1]Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.
[2]Université Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille.
[3]Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190, Gif-sur-Yvette, France. Correspondence to: Batiste Le Bars <batiste.le-bars@inria.fr>, Pierre Humbert <pierre.humbert@universite-paris-saclay.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Minsker et al. (2015) and Devroye et al. (2016) to estimate the means of heavy-tailed distributions. Furthermore, due to its robustness to outliers, MoM-based estimators have recently gained a renewed interest in the machine learning community (Lecué et al., 2020b;a; Laforgue et al., 2020).

Contributions. In this paper, we propose a new robust nonparametric density estimator based on the combination of the Kernel Density Estimation method and the Median-of-Means principle (MoM-KDE). We place ourselves in a more general framework than the classical Huber contamination model, called $\mathcal{O} \cup \mathcal{I}$, which gets rid of any assumption on the outliers. We demonstrate the statistical performance of the estimator through finite-sample high-confidence error bounds in the L_∞-norm and show that MoM-KDE is consistent under the condition that the outlier proportion tends to 0 – a necessary condition in robust kernel density estimation (Liu and Gao, 2019). Additionally, we prove the consistency in the L_1-norm, which is known to reflect the global performance of the estimate (Devroye and Gyorfi, 1985). To the best of our knowledge, this is the first work that presents such results in the context of robust kernel density estimation, under the $\mathcal{O} \cup \mathcal{I}$ framework. As a measure of robustness, we also introduce an influence function adapted to the $\mathcal{O} \cup \mathcal{I}$ framework. It allows us to find the number of outliers above which the MoM-KDE is less sensitive to outliers than the KDE. Finally, we demonstrate the empirical performance of MoM-KDE on both synthetic and real data and show the practical interest of such estimator as it has a lower computational complexity than the baseline RKDE and SPKDE.

2. Median-of-Means Kernel Density Estimation

We first recall the classical kernel density estimator. Let $X_1, \cdots, X_n$ be independent and identically distributed (i.i.d.) random variables that have a probability density function (pdf) $f(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^d$. The Kernel Density Estimate of f (KDE), also called the *Parzen–Rosenblatt estimator*, is a nonparametric estimator given by

$$\hat{f}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) , \qquad (1)$$

where $h > 0$ and $K : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ is an integrable function satisfying $\int K(u)du = 1$ (Tsybakov, 2008). Such a function $K(\cdot)$ is called a *kernel* and the parameter h is called the *bandwidth* of the estimator. The bandwidth is a smoothing parameter that controls the bias-variance tradeoff of $\hat{f}_n(\cdot)$ with respect to the input data.

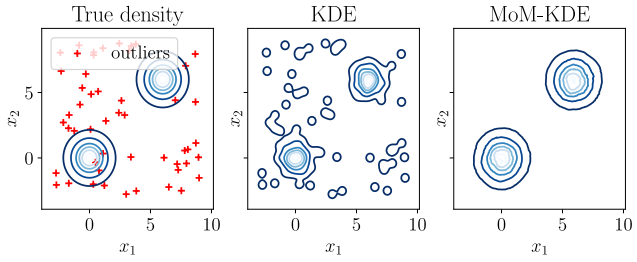


Figure 1. From left to right: True density and outliers from a uniform density, estimation with KDE, and with MoM-KDE.

While this estimator is central in statistics, a major drawback is its weakness against outliers (Kim and Scott, 2008; 2011; 2012). Indeed, as it assigns uniform weights $1/n$ to every $K(\cdot)$ regardless of whether X_i is an outlier or not, inliers and outliers contribute equally in the construction of the KDE, which results in undesired "bumps" over outlier locations in the final estimated density (see Figure 1).

In the following, we propose a KDE-based density estimator robust to the presence of outliers in the sample set. These outliers are considered in a general framework described in the next section.

2.1. Outlier setup

Throughout the paper, we consider the $\mathcal{O} \cup \mathcal{I}$ framework introduced by Lecué and Lerasle (2019). This very general framework allows the presence of outliers in the dataset and relax the standard i.i.d. assumption on each observation. We therefore assume that the n random variables are partitioned into two (unknown) groups: a subset $\{X_i \mid i \in \mathcal{I}\}$ made of inliers, and another subset $\{X_i \mid i \in \mathcal{O}\}$ made of outliers such that $\mathcal{O} \cap \mathcal{I} = \emptyset$ and $\mathcal{O} \cup \mathcal{I} = \{1, \ldots, n\}$. While we suppose the $X_{i \in \mathcal{I}}$ are i.i.d. from a distribution that admits a density f with respect to the Lebesgue measure, no assumption is made on the outliers $X_{i \in \mathcal{O}}$. Hence, these outlying points can be dependent or adversarial.

The $\mathcal{O} \cup \mathcal{I}$ framework is related to the well-known Huber's ε-contamination model (Huber, 1992) where it is assumed that data are i.i.d. with distribution $g = \varepsilon f_{\mathcal{I}} + (1 - \varepsilon) f_{\mathcal{O}}$, and $\varepsilon \in [0, 1)$; the distribution $f_{\mathcal{I}}$ being related to the inliers and $f_{\mathcal{O}}$ to the outliers. However, there are several important differences. First, in the $\mathcal{O} \cup \mathcal{I}$ the proportion of outliers is fixed and equals $|\mathcal{O}|/n$, whereas it is random in the Huber's ε-contamination model (Lerasle, 2019). Second, the $\mathcal{O} \cup \mathcal{I}$ is less restrictive. Indeed, contrary to Huber's model which considers that inliers and outliers are respectively i.i.d from the same distributions, $\mathcal{O} \cup \mathcal{I}$ does not make any assumption about the outliers. This framework should not be confused with the ε-corruption model where an adversary is allowed to directly modify the inliers set (Diakonikolas et al., 2019).

2.2. MoM-KDE

We now present our main contribution, a robust kernel density estimator based on the MoM. This estimator is essentially motivated by the fact that the classical kernel density estimation at one point corresponds to an empirical average (see Equation (1)). Therefore, the MoM principle appears to be an intuitive solution to build a robust version of the KDE. A formal definition of MoM-KDE is given below.

Definition 1. *(MoM Kernel Density Estimator) Let* $1 \leq S \leq n$*, and let* $B_1, \cdots, B_S$ *be a random partition of* $\{1, \cdots, n\}$ *into* S *non-overlapping blocks* B_s *of equal size* $n_s \triangleq n/S$*.*

The MoM Kernel Density Estimator (MoM-KDE) of f *at* x_0 *is given by*

$$\hat{f}_{MoM}(x_0) \propto \text{Median}\left(\hat{f}_{n_1}(x_0), \cdots, \hat{f}_{n_S}(x_0)\right) , \quad (2)$$

where $\hat{f}_{n_s}(x_0)$ *is the value of the standard kernel density estimator at* x_0 *obtained via the samples of the* s*-th block* B_s*. Note that* $\hat{f}_{MoM}(\cdot)$ *do not necessarily integrates to* 1*. However, as suggested by Devroye and Lugosi (2012) (section 5.6), it can always be normalized by its integral.*

Broadly speaking, MoM estimators appear to be a good tradeoff between the unbiased but non robust empirical mean and the robust but biased median (Lecué et al., 2020b). Furthermore, we remark that, when $S = 1$ the standard KDE is recovered.

2.3. Time complexity

The complexity of MoM-KDE to evaluate one point is the same as the standard KDE, $\mathcal{O}(n)$; $\mathcal{O}(S \cdot \frac{n}{S})$ for the block-wise evaluation and $\mathcal{O}(S)$ to compute the median with the *median-of-medians algorithm* (Blum et al., 1973). Since RKDE and SPKDE are KDEs with modified weights, they also perform the evaluation step in $\mathcal{O}(n)$ time. However, these weights need to be learnt, thus requiring an additional non-negligible computing capacity. Indeed, each one of them rely on an iterative method – respectively the iteratively reweighted least squares algorithm and the projected gradient descent algorithm, that both have a complexity of $\mathcal{O}(n_{iter} \cdot n^2)$, where n_{iter} is the number of needed iterations to reach a reasonable accuracy. MoM-KDE on the other hand does not require any learning procedure. Depending on the application that we have, MoM-KDE may require a normalization step. When using a Monte Carlo method, the complexity of such a normalization step at a given precision of $1/\sqrt{n_{norm}}$ is $\mathcal{O}(n_{norm})$ (Weinzierl, 2000). Note that the evaluation step can be accelerated through several ways, hence potentially reducing computational time of all these competing methods (Gray and Moore, 2003a;b; Wang and Scott, 2019; Backurs et al., 2019). Theoretical time complexities are gathered in Table 1.

Method	Learning	Evaluation	Iterative method
KDE	–	$\mathcal{O}(n)$	no
RKDE	$\mathcal{O}(n_{iter} \cdot n^2)$	$\mathcal{O}(n)$	yes
SPKDE	$\mathcal{O}(n_{iter} \cdot n^2)$	$\mathcal{O}(n)$	yes
MoM-KDE	–	$\mathcal{O}(n + n_{norm})$	no

Table 1. Computational complexity.

3. Theoretical analysis

In this section, we give a finite-sample high-probability error bound in the L_∞-norm for MoM-KDE under the $\mathcal{O} \cup \mathcal{I}$ framework. To our knowledge, this work is the first to provide such error bounds in robust kernel density estimation under this framework. In order to build this high-probability error bound, it is assumed, among other standard hypotheses, that the true density is Hölder-continuous, a smoothness property usually considered in KDE analysis (Tsybakov, 2008; Jiang, 2017; Wang et al., 2019). In addition, we show the consistency in the L_1-norm. In this last result, we will see that the aforementioned assumptions are not necessary to obtain the consistency. In the following, we give the necessary definitions and assumptions to perform our non-asymptotic analysis.

3.1. Setup and assumptions

Let us first list the usual assumptions, notably on the considered kernel function, that will allow us to derive our results. They are standard in KDE analysis, and are chosen for their simplicity of comprehension (Tsybakov, 2008; Jiang, 2017). More general hypotheses could be made in order to obtain the same results, notably assuming kernel of order ℓ (see for example the works of Tsybakov (2008) and Wang et al. (2019)).

Assumption 1. (Bounded density) $\|f\|_\infty < \infty$.

We make the following assumptions on the kernel K which are fairly standard.

Assumption 2. The kernel $K : \mathbb{R}^d \longrightarrow \mathbb{R}$ is bounded by a positive constant C_ρ and integrates to 1.

Assumption 3. Let $\mathcal{F}$ be the class of functions of the form $z \in \mathbb{R}^d \mapsto K(x - z)$. Then, $\mathcal{F}$ is a uniformly bounded VC class (Wang et al., 2019).

All the above assumptions are respected by most of the popular kernels, in particular the Gaussian, Exponential, Uniform, Triangular, Cosine kernel, etc. These are key properties to provide the bounds presented in the next section.

Before stating our main results, we recall the definition of

the Hölder class of functions.

Definition 2. *(Hölder class) Let T be an interval of $\mathbb{R}^d$, and $0 < \alpha \leq 1$ and $L > 0$ be two constants. We say that a function $f : T \to \mathbb{R}$ belongs to the Hölder class $\Sigma(L, \alpha)$ if it satisfies*

$$\forall x, x' \in T, \quad |f(x) - f(x')| \leq L\|x - x'\|^{\alpha} . \tag{3}$$

This definition implies a notion of smoothness on the function f, and is a convenient property to bound the bias of KDE-based estimators.

3.2. L_∞ and L_1 consistencies of MoM-KDE

This section states our central finding, a L_∞ finite-sample error bound for MoM-KDE that proves its consistency under the $\mathcal{O} \cup \mathcal{I}$ framework.

Lemma 1. *(L_∞ error-bound of the KDE **without outliers** - (Wang et al., 2019)) Suppose that f belongs to the class of densities $\mathcal{P}(\alpha, L)$ defined as*

$$\mathcal{P}(\alpha, L) \triangleq \left\{ f \mid f \geq 0, \int f(x)dx = 1, \right. \tag{4}$$
$$\left. \text{and } f \in \Sigma(\alpha, L) \right\} ,$$

where $\Sigma(\alpha, L)$ is the Hölder class of functions on $\mathbb{R}^d$ (Definition 2). Grant assumptions 1 to 3 and let $h \in (0, 1)$, $\gamma > 0$, n large enough, and $nh^d \geq 1$. Then with probability at least $1 - \exp(-\gamma)$, we have

$$\|\hat{f}_n - f\|_\infty \leq C_1 \sqrt{\frac{\gamma + \log(1/h)}{nh^d}} + C_2 h^{\alpha} , \tag{5}$$

where $C_2 = L \int \|u\|^{\alpha} K(u)du < \infty$ and C_1 is a constant that only depends on $\|f\|_\infty$, the dimension d, and the kernel properties.

This lemma, which is verified several times in the literature (see e.g. (Giné and Guillou, 2002; Jiang, 2017; Wang et al., 2019)), comes from the well-known bias-variance decomposition, where we separately bound the variance (see e.g. Wang et al. (2019); Kim et al. (2019)) and the bias (see e.g. Tsybakov (2008); Rigollet and Vert (2009)). It shows the consistency of KDE without outliers, as soon as $h \to 0$ and $nh^d \to \infty$.

We now present our main result. Its objective is to show that even under the $\mathcal{O} \cup \mathcal{I}$ framework, we do not need any critical additional hypothesis – besides the ones of the previous lemma – to show that MoM-KDE is consistent.

Proposition 1. (L_∞ error-bound of the MoM-KDE under the $\mathcal{O} \cup \mathcal{I}$) Suppose that f belongs to the class of densities $\mathcal{P}(\alpha, L)$ and grant assumptions 1 to 3. Let S be the number of blocks such that $S \geq 2|\mathcal{O}| + 1$. Then, for any $h \in (0, 1)$, $\gamma > 0$, n/S large enough, and $nh^d \geq S$, we have with probability at least $1 - \exp(-\gamma)$,

$$\|\hat{f}_{MoM} - f\|_\infty \leq C_1 \sqrt{\frac{S(\log(S) + \gamma + \log(1/h))}{nh^d}} + C_2 h^{\alpha} ,$$

where $C_2 = L \int \|u\|^{\alpha} K(u)du < \infty$, and C_1 is a constant that only depends on $\|f\|_\infty$, the dimension d, and the kernel properties.

The proof of this proposition is given in the Appendix. In addition to $S \geq 2|\mathcal{O}| + 1$ in Proposition 1, the other conditions come from those in Lemma 1, which are necessary to obtain the rate in the uncorrupted scenario. The upper bound in the probability is minimal for $S = 2|\mathcal{O}| + 1$. Note also that when $|\mathcal{O}| = 0$ and $S = 1$, we exactly recover the result of Lemma 1 i.e. the rate for the KDE without outliers. This was expected as for $S = 1$, the MoM-KDE is exactly the KDE.

Corollary 1. (Rate of convergence) *Consider the assumptions of Proposition 1 with $S = 2|\mathcal{O}| + 1$, $\gamma = \log(n)$ and let*

$$h \asymp \left(\frac{S \log(n)}{n} \right)^{1/(2\alpha+d)} .$$

With probability higher than $1 - \frac{1}{n}$, we have

$$\|\hat{f}_{MoM} - f\|_\infty \lesssim \left(\frac{|\mathcal{O}|}{n} \log(n) \right)^{\alpha/(2\alpha+d)} + \left(\frac{\log(n)}{n} \right)^{\alpha/(2\alpha+d)} .$$

Corollary 1, proved in Appendix, shows that the rate is split in two terms. On the right, we have the classical minimax optimal rate of the KDE without outliers. On the left, we have a rate that depends on the outlier proportion $|\mathcal{O}|/n$. It also shows that a sufficient condition to obtain consistency (error tending to 0) is to have $n \to \infty$ (standard) and the proportion of outliers $|\mathcal{O}|/n$ tending to 0. This latter condition was also expected since the minimax analysis done for the specific case of the Huber's model already captured it as a necessary condition (Liu and Gao, 2019). Note further that, when there is no outlier, i.e. $|\mathcal{O}| = 0$, we recover the standard KDE minimax optimal rate (Wang et al., 2019).

Last but not least, our two main results illustrate that the convergence of the MoM-KDE only depends on the number of outliers in the dataset, and not on their "type". This estimator is therefore robust in a wide range of scenarios.

In particular, the proof of Proposition 1 and Corollary 1 are also valid under the adversarial scenario (Diakonikolas et al., 2019; Depersin and Lecué, 2021) which is a more complex case than the $\mathcal{O} \cup \mathcal{I}$ framework. For instance, this framework allows the inliers to be highly correlated because of the corruption step.

We now give a L_1-consistency result under mild hypotheses, which is known to reflect the global performance of the estimate. Indeed, small L_1 error leads to accurate probability estimation (Devroye and Gyorfi, 1985).

Proposition 2. *(L_1-consistency in probability) If $n/S \to \infty$, $h \to 0$, $nh^d \to \infty$, $S/\sqrt{nh^d} \to 0$, and $S \geq 2|\mathcal{O}| + 1$, then*

$$\|\hat{f}_{MoM} - f\|_1 \xrightarrow[n\to\infty]{\mathcal{P}} 0 \, .$$

This result is obtained by bounding the left-hand part by the errors in the healthy blocks only, i.e. those without anomalies. Under the hypothesis of the proposition, these errors are known to converge to 0 in probability (Wang et al., 2019). The complete proof is given in the Appendix. Contrary to SPKDE (Vandermeulen and Scott, 2014), no assumption on the outliers generation process is necessary to obtain this consistency result. Moreover, while we need to assume that the proportion of outliers is perfectly known to prove the convergence of SPKDE, the MoM-KDE converges whenever the number of outliers is overestimated.

Finally, note that again in Proposition 2, a condition to obtain consistency is that the fraction of outliers $|\mathcal{O}|/n$ tends to zero. As explained upper, this assumption is natural since the fraction of outliers tending to zero has been shown to be a necessary condition for consistency in the Huber framework (Liu and Gao, 2019).

3.3. Influence function in the $\mathcal{O} \cup \mathcal{I}$ framework

As a measure of robustness, we now introduce an Influence Function (IF) (or sensitivity curve) adapted to the $\mathcal{O} \cup \mathcal{I}$ framework. It is inspired from the classical IF, first proposed by Hampel (1974), which measures how an estimator changes when the initial distribution is modified by adding a small amount of contamination at a point x'. Therefore, it provides a notion of stability in the Huber model framework (Andrews, 1986; Debruyne et al., 2008).

Generalizing the IF from the Huber model to the $\mathcal{O} \cup \mathcal{I}$ framework is not that straightforward. Indeed, the definition of the IF for function estimate (introduced for the Huber model in (Kim and Scott, 2011) ; Definition 1) is

$$\text{IF}_{\text{Huber}}(x, x'; T, F) = \lim_{s \to 0} \frac{T(x; F_s) - T(x; F)}{s} \, ,$$

where $F_s = (1-s)F + s\delta_{x'}$ and T is a function estimate based on F evaluated at x. This formulation is adapted for the Huber model as it consider all samples drawn from a mixture of a true density (here F) and an outlying one (δ_x). This is unfortunately not adapted to the $\mathcal{O} \cup \mathcal{I}$ framework. To highlight the robustness of our estimator using an "IF-based idea", we adapt it to this framework and proposed the following definition.

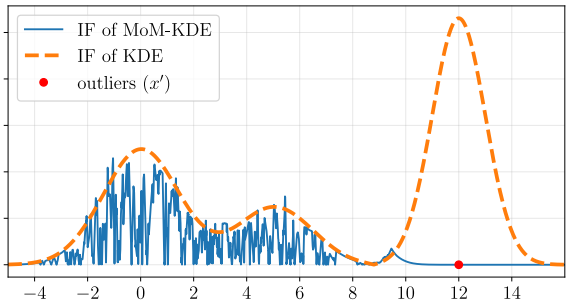


Figure 2. Influence function for MoM-KDE and KDE where $n = 1000$, $m = 40$ outliers are placed at $x' = 12$, and where the samples in $\mathcal{I}_n$ are drawn from the true density: $2/3 \cdot \mathcal{N}(0, 1) + 1/3 \cdot \mathcal{N}(5, 1)$.

Definition 3. *($\text{IF}_{\mathcal{O}\cup\mathcal{I}}$) Let $T_n(x_0; \mathcal{I}_n)$ be a density estimator evaluated at x_0 and learned with a healthy data set $\mathcal{I}_n = \{X_i\}_{i=1}^n$. Let $m \in \mathbb{N}$ and $x' \in \mathbb{R}^d$. The $\text{IF}_{\mathcal{O}\cup\mathcal{I}}$ is defined as:*

$$\text{IF}_{\mathcal{O}\cup\mathcal{I}}(x_0, x', m; \mathcal{I}_n, T_n) \triangleq |T_n(x_0; \mathcal{I}_n) - T_n(x_0; \mathcal{I}_n \cup \{x'\}_{i=1}^m)| \, ,$$

where by healthy points we mean inliers i.e. samples that are independently drawn from the true density function.

Given this definition, $\text{IF}_{\mathcal{O}\cup\mathcal{I}}$ quantifies how much the value at x_0 of an estimated density function changes whenever the healthy dataset is increased by m points located at x'. Therefore, the link with the notion of stability is made obvious: the smaller $\text{IF}_{\mathcal{O}\cup\mathcal{I}}$ is, the more stable and thus robust the estimator is. An illustration of the IF for MoM-KDE and KDE is given in Figure 2. It shows that the IF for MoM-KDE is lower than the one of KDE, especially near the outlying points $\{x'\}$. Note that the variability observed in the case of MoM-KDE is due to the block splitting procedure. See Section C of the Appendix for more results on the IF.

In the next proposition, we provide a lower bound on the number of added samples m over which the $\text{IF}_{\mathcal{O}\cup\mathcal{I}}$ of the MoM-KDE is lower that the one of KDE with high probability.

Proposition 3. *Let $x', x_0 \in \mathbb{R}^d$ and $\mathcal{I}_n$ be a healthy data*

set. Grant assumptions 1 *to* 3 *and denote*

$$a \triangleq \sum_{i \in \mathcal{I}_n} K\left(\frac{X_i - x_0}{h}\right), \quad b \triangleq K\left(\frac{x' - x_0}{h}\right) .$$

Let $S \geq 2m+1$ *with* $m \in [\![0, \frac{n}{2}[\![$ *the number of added samples and* $\delta > 0$ *such that* $|b - a/n| > C_\rho\sqrt{2\delta S/n}$. *If* $m \geq \frac{C_\rho\sqrt{2n\delta S}}{|b-a/n|-C_\rho\sqrt{2\delta S/n}}$, *then with probability higher than* $1 - 4\exp(-\delta)$ *we have:*

$$\mathrm{IF}_{\mathcal{O}\cup\mathcal{I}}(x_0, x', m; \mathcal{I}_n, \hat{f}_{MoM}) \leq \mathrm{IF}_{\mathcal{O}\cup\mathcal{I}}(x_0, x', m; \mathcal{I}_n, \hat{f}_{KDE}) .$$

The proof of this proposition is given in the Appendix.

Given the previous proposition, the lower bound on m over which the MoM-KDE is better than KDE is not necessarily easy to interpret. When everything is fixed except x_0 and x', we see that the bound is low whenever $|b - a/n| = |K(\frac{x_0 - x'}{h}) - \frac{1}{n}\sum K(\frac{x_0 - X_i}{h})|$ is large. A sufficient condition for this is to take x' far from the sampling set $\mathcal{I}_n$, i.e take x' as an outlier. Under this condition, the bound will get even lower whenever x_0 gets closer to x'.

4. Numerical experiments

In this section, we display numerical results supporting the relevance of MoM-KDE. All experiments were run on a personal laptop computer using `Python`. The code of MoM-KDE is made available online.[1].

Comparative methods. In the following experiments, we propose to compare MoM-KDE to the classical KDE and two robust versions of KDE, called RKDE (Kim and Scott, 2012) and SPKDE (Vandermeulen and Scott, 2014).

As previously explained, RKDE takes the ideas of robust M-estimation and translate it to kernel density estimation. The authors point out that classical KDE estimator can be seen as the minimizer of a squared error loss in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ corresponding to the chosen kernel. Instead of minimizing this loss, they propose to minimize a robust version of it:

$$\hat{f}_{\mathrm{RKDE}} = \operatorname*{argmin}_{g \in \mathcal{H}} \sum_{i=1}^{n} \rho(\|\phi(X_i) - g\|_{\mathcal{H}}),$$

where ϕ is the canonical feature map associated to the RKHS and $\rho(\cdot)$ is taken to be either the Huber or the Hampel function, both known to bring robustness in M-estimations. The solution of the newly expressed problem is then found using the iteratively reweighted least squares algorithm. Note that this RKHS approach has also been combined with the MoM principle for robust estimation in (Lerasle et al., 2019). However, no algorithm was proposed to build this estimator, which is why no comparison is made here.

SPKDE proposes to scale and project the standard KDE in a way that it decontaminates the dataset. Recall that $\hat{f}_n$ is the classical KDE estimator of Equation (1), this procedure is done by finding

$$\hat{f}_{\mathrm{SPKDE}} = \operatorname*{argmin}_{g \in \Delta_n} \sum_{i=1}^{n} \|\beta\hat{f}_n - g\|_2,$$

where Δ_n corresponds to the convex hull of $\{K(\frac{X_i - \cdot}{h})\}_{i=1}^n$ and β is an hyperparameter that controls the robustness. The minimization is shown to be equivalent to a quadratic program over the simplex, solved via projected gradient descent.

Metrics. The performance of the MoM-KDE is measured through three metrics. Two are used to measure the similarity between the estimated and the true density. One describes performances of an anomaly detector based on the estimated density. The first one is the Kullback-Leibler divergence (Kullback and Leibler, 1951) which is the most used in robust KDE (Kim and Scott, 2008; 2011; 2012; Vandermeulen and Scott, 2014). Used to measure the similarity between distributions, it is defined as

$$D_{\mathrm{KL}}(\hat{f}\|f) = \int \hat{f}(x) \log\left(\frac{\hat{f}(x)}{f(x)}\right) dx .$$

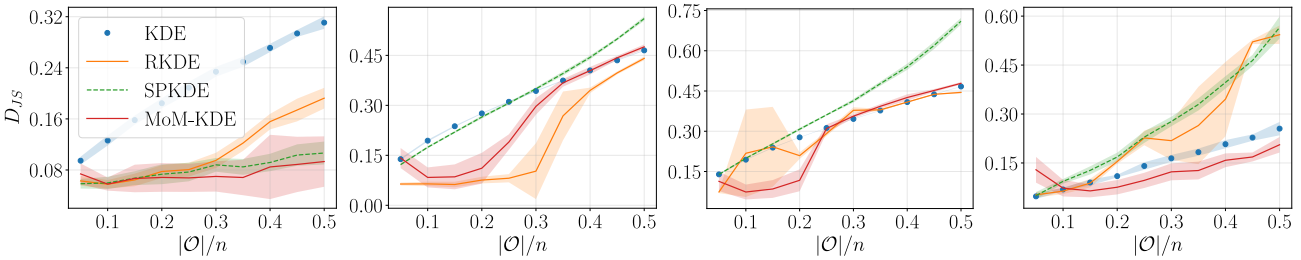
As the Kullback-Leibler divergence is non-symmetric and may have infinite values when distributions do not share the same support, we also consider the Jensen-Shannon divergence (Endres and Schindelin, 2003; Liese and Vajda, 2006). It is a symmetrized version of D_{KL}, with positive values, bounded by 1 (when the logarithm is used in base 2), and has found applications in many fields, such as deep learning (Goodfellow et al., 2014) or transfer learning (Segev et al., 2017). It is defined as

$$D_{\mathrm{JS}}(\hat{f}\|f) = \frac{1}{2}\left(D_{\mathrm{KL}}(\hat{f}\|g) + D_{\mathrm{KL}}(f\|g)\right),$$

with $g = \frac{1}{2}(\hat{f} + f)$.

Motivated by real-world application, the third metric is not related to the true density, which is usually not available in practical cases. Instead, we quantify the capacity of the learnt density to detect anomalies using the well-known Area Under the ROC Curve criterion (AUC) (Bradley, 1997). An input point x_0 is considered abnormal whenever $\hat{f}(x_0)$ is below a given threshold.

[1] https://github.com/lminvielle/mom-kde. For the sake of comparison, we also implemented RKDE and SPKDE.



(a) Uniform (b) Regular Gaussian (c) Thin Gaussian (d) Advsersarial Thin Gaussian

Figure 3. Density estimation with synthetic data. The displayed metric is the Jensen-Shannon divergence. A lower score means a better estimation of the true density.

Hyperparameters. All estimators are built using the Gaussian kernel. The number of blocks S in MoM-KDE is selected on a regular grid of 20 values between 1 and $2|\mathcal{O}|+1$ in order to obtain the lowest D_{JS}. Recall that from a theoretical point of view, we showed that if $S \geq 2|\mathcal{O}|+1$, i.e. at least half of the blocks does not contain anomalies, consistency results and rates can be obtained (Propositions 1 and 2). However, if the proportion of anomalies gets large, $S \geq 2|\mathcal{O}|+1$ implies a small number of samples in each block which may lead to bad estimates. Thus, it seems that $S \geq 2|\mathcal{O}|+1$ is not a necessary condition to obtain good results. In practice, this suggests that S could be lower than $2|\mathcal{O}|+1$, reason why we use this grid search strategy. The bandwidth h is chosen for KDE via the pseudo-likelihood k-cross-validation method (Turlach, 1993), and used for all estimators. The construction of RKDE follows exactly the indications of its authors (Kim and Scott, 2012) and $\rho(\cdot)$ is taken to be the Hampel function as it empirically showed to be the most robust. For SPKDE, the true ratio of anomalies is given as an input parameter.

4.1. Results on synthetic data

To evaluate the efficiency of the MoM-KDE against KDE and its robust competitors, we set up several outlier situations. In all theses situations, we draw $N = 1000$ inliers from an equally weighted mixture of two normal distribution $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$ with $\mu_1 = 0$, $\mu_2 = 6$, and $\sigma_1 = \sigma_2 = 0.5$. The outliers however are sampled through various schemes:

(a) **Uniform.** A uniform distribution $U([\mu_1 - 3, \mu_2 + 3])$ which is the classical setting used for outlier simulation.

(b) **Regular Gaussian.** A *similar*-variance normal distribution $\mathcal{N}(3, 0.5)$ located between the two inlier clusters.

(c) **Thin Gaussian.** A *low*-variance normal distribution $\mathcal{N}(3, 0.01)$ located between the two inliers clusters.

(d) **"Adversarial" Thin Gaussian.** A low variance normal distribution $\mathcal{N}(0, 0.01)$ located on one of the inliers' Gaussian mode. This scenario can be seen as adversarial as an ill-intentioned agent may hide wrong points in region of high density. It is the most challenging setting for standard robust estimators as they are in general robust to outliers located outside the support of the density we wish to estimate.

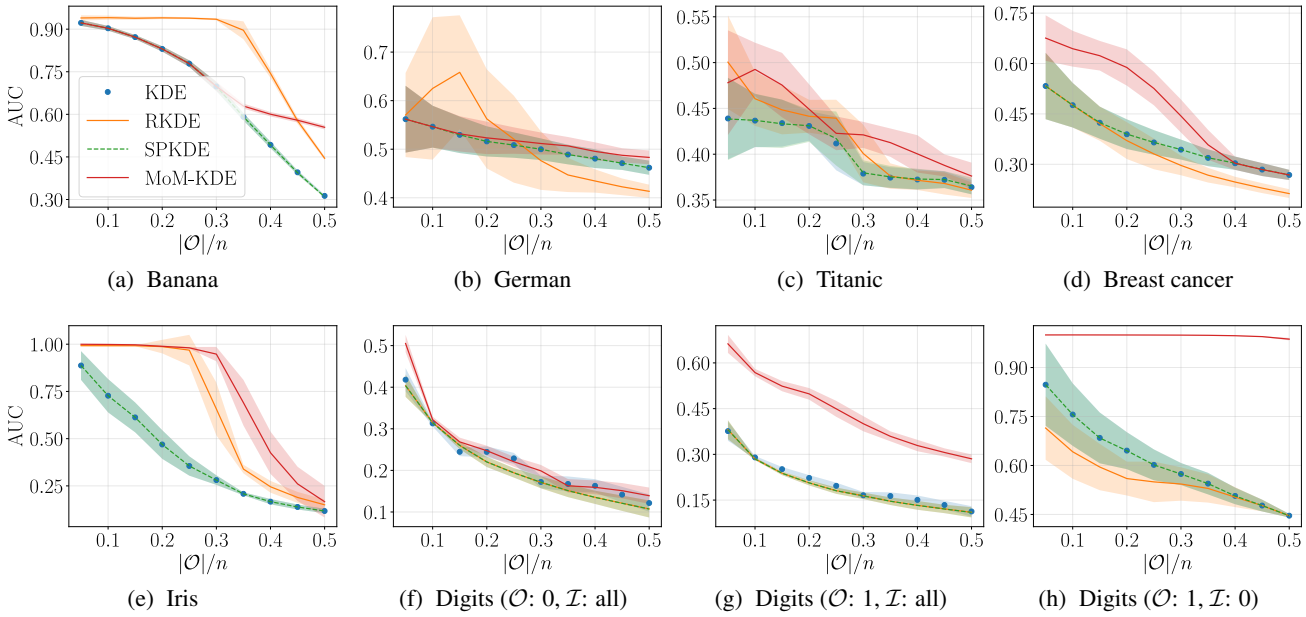
For all situations, we consider several ratios of contamination and set the number of outliers $|\mathcal{O}|$ in order to obtain a ratio $|\mathcal{O}|/n$ ranging from 0.05 to 0.5 with 0.05-wide steps. Finally, to evaluate the pertinence of our results, for each set of parameters, data are generated 10 times.

We display in Figure 3 the results over synthetic data using the D_{JS} score. The average scores and standard deviations over the 10 experiments are represented for each outlier scheme and ratio. Overall, the results show the good performance of MoM-KDE in all the considered situations. Furthermore, they highlight the dependency of the two competitors to the type of outliers. Indeed, as SPKDE is designed to handle uniformly distributed outliers, the algorithm struggles when confronted with differently distributed outliers (see Figure 3 b, c, d). RKDE performs generally better, but fails against adversarial contamination, which may be explained by its tendency to down-weight points located in low-density regions, which for this particular case correspond to the inliers. Results for D_{KL} and AUC are reported in the Appendix C. Generally, they show similar results and the same conclusions on the good performance of MoM-KDE can be made.

In the Appendix C, we also add comparisons with methods that first proceed to an anomaly detection step before fitting a classical KDE. However, their results are not as good as those of the robust estimators.

4.2. Results on real data

Experiments are also conducted over six classification datasets: Banana, German, Titanic, Breast-cancer,



(a) Banana (b) German (c) Titanic (d) Breast cancer

(e) Iris (f) Digits ($\mathcal{O}$: 0, $\mathcal{I}$: all) (g) Digits ($\mathcal{O}$: 1, $\mathcal{I}$: all) (h) Digits ($\mathcal{O}$: 1, $\mathcal{I}$: 0)

Figure 4. Anomaly detection with real datasets, measured with AUC over varying outlier proportion. A higher score means a better detection of the outliers. For Digits, we specify which classes are chosen to be inliers ($\mathcal{I}$) and outliers ($\mathcal{O}$).

Iris, and Digits. They contain respectively $n = 5300, 1000, 2201, 569, 150$ and 1797 data points having $d = 2, 20, 3, 30, 4$ and 64 input dimensions. They are all publicly available either from open repositories at `http://www.raetschlab.org/Members/raetsch/benchmark/` (for the first three) or directly from the Scikit-learn package (for the last three) (Pedregosa et al., 2011). We follow the approach of Kim and Scott (2012) that consists in setting the class labeled 0 as outliers and the rest as inliers. To artificially control the outlier proportion, we randomly downsample the abnormal class to reach a ratio $|\mathcal{O}|/n$ ranging from 0.05 to 0.5 with 0.05-wide steps. When a dataset does not contain enough outliers to reach a given ratio, we similarly downsample the inliers. For each dataset and each ratio, the experiments are performed 50 times, the random downsampling resulting in different learning datasets. The empirical performance is evaluated through the capacity of each estimator to detect anomalies, which we measure with the AUC.

Results are displayed in Figure 4. With the Digits dataset, we also explore additional scenarios with changing inlier and outlier classes (specified in the figure captions). Overall, results are in line with performances observed over synthetic experiments, achieving good results in comparison to its competitors. Note that even in the highest dimensional scenarios, i.e. Digits and Breast cancer ($d = 64$ and $d = 30$), MoM-KDE still behaves well, outperforming its competitors. This behavior for high-dimensionality problems was also notice for robust KDE-based methods in a recent anomaly detection review (Domingues et al., 2018). Additional results are reported in the the Section C of the Appendix.

5. Conclusion

The present paper introduced MoM-KDE, a new efficient way to perform robust kernel density estimation. The method has been shown to be consistent in both L_∞ and L_1 error-norm in presence of very generic outliers. MoM-KDE achieved good empirical results in various situations while having a lower computational complexity than its competitors.

While the present work uses the coordinate-wise median to construct its robust estimator, it might be interesting to investigate the use of other generalizations of the median in high dimension, e.g. the geometric median. Another possible extension of the proposed method is to consider a bootstrap version where several random splits are performed and thus several MoM-KDE are aggregated. In that case, one can expect the final output to be smoother and the bound to be less dependent on the number of blocks S. On the other hand, with such approach the complexity of the method increases. First empirical visualizations of this method can be found in the Appendix.

Acknowledgments

Part of this research was funded by the IdAML chair of Centre Borelli from ENS Paris Saclay.

References

Alon, N., Matias, Y., and Szegedy, M. (1999). The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147.

Andrews, D. W. (1986). Stability comparison of estimators. *Econometrica: Journal of the Econometric Society*, pages 1207–1235.

Backurs, A., Indyk, P., and Wagner, T. (2019). Space and time efficient kernel density estimation in high dimensions. In *Advances in Neural Information Processing Systems*, pages 15773–15782.

Blum, M., Floyd, R. W., Pratt, V. R., Rivest, R. L., and Tarjan, R. E. (1973). Time bounds for selection. *Journal of Computer and System Sciences*, 7:448–461.

Boucheron, S. and Thomas, M. (2012). Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17:1–12.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Debruyne, M., Christmann, A., Hubert, M., and Suykens, J. A. (2008). Robustness and stability of reweighted kernel based regression. Technical report.

Depersin, J. and Lecué, G. (2021). Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *arXiv preprint arXiv:2102.00995*.

Devroye, L. and Gyorfi, L. (1985). *Nonparametric density estimation: the L*1 *view*. New York: John Wiley & Sons.

Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I., et al. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725.

Devroye, L. and Lugosi, G. (2012). *Combinatorial methods in density estimation*. Springer Science & Business Media.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864.

Domingues, R., Filippone, M., Michiardi, P., and Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: experiments and analyses. *Pattern Recognition*, 74:406–421.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Gray, A. G. and Moore, A. W. (2003a). Nonparametric density estimation: toward computational tractability. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 203–211. SIAM.

Gray, A. G. and Moore, A. W. (2003b). Rapid evaluation of multiple density models. In *International Conference on Artificial Intelligence and Statistics*.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Jerrum, M. R., Valiant, L. G., and Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188.

Jiang, H. (2017). Uniform convergence rates for kernel density estimation. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1694–1703. JMLR. org.

Kim, J. and Scott, C. (2008). Robust kernel density estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3381–3384. IEEE.

Kim, J. and Scott, C. (2011). On the robustness of kernel density M-estimators. In *Proceedings of the 28th International Conference on Machine Learning*, pages 697–704. Citeseer.

Kim, J. and Scott, C. D. (2012). Robust kernel density estimation. *The Journal of Machine Learning Research*, 13(Sep):2529–2565.

Kim, J., Shin, J., Rinaldo, A., and Wasserman, L. (2019). Uniform convergence of the kernel density estimator adaptive to intrinsic volume dimension. In *ICML 2019-36th International Conference on Machine Learning*, volume 97.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Laforgue, P., Staerman, G., and Clémençon, S. (2020). How robust is the median-of-means? concentration bounds in presence of outliers. *arXiv preprint arXiv:2006.05240*.

Lecué, G. and Lerasle, M. (2019). Learning from MOM's principles: Le Cam's approach. *Stochastic Processes and their applications*, 129(11):4385–4410.

Lecué, G., Lerasle, M., et al. (2020a). Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*, 48(2):906–931.

Lecué, G., Lerasle, M., and Mathieu, T. (2020b). Robust classification via MOM minimization. *Machine Learning*.

Lerasle, M. (2019). Lecture notes: selected topics on robust statistical learning theory. *arXiv preprint arXiv:1908.10761*.

Lerasle, M., Szabó, Z., Mathieu, T., and Lecué, G. (2019). MONK outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning*, pages 3782–3793. PMLR.

Liese, F. and Vajda, I. (2006). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Liu, H. and Gao, C. (2019). Density estimation with contamination: minimax rates and theory of adaptation. *Electronic Journal of Statistics*, 13(2):3613–3653.

Minsker, S. et al. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335.

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley Interscience, New-York.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Rigollet, P. and Vert, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Segev, N., Harel, M., Mannor, S., Crammer, K., and El-Yaniv, R. (2017). Learn on source, refine on target: a model transfer learning framework with random forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1811–1824.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Sriperumbudur, B. and Steinwart, I. (2012). Consistency and rates for clustering with DBSCAN. In *International Conference on Artificial Intelligence and Statistics*, pages 1090–1098.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Turlach, B. A. (1993). Bandwidth selection in kernel density estimation: a review. In *CORE and Institut de Statistique*. Citeseer.

Vandermeulen, R. and Scott, C. (2013). Consistency of robust kernel density estimators. In *Conference on Learning Theory*, pages 568–591.

Vandermeulen, R. A. and Scott, C. (2014). Robust kernel density estimation by scaling and projection in Hilbert space. In *Advances in Neural Information Processing Systems*, pages 433–441.

Wang, D., Lu, X., and Rinaldo, A. (2019). DBSCAN: optimal rates for density-based cluster estimation. *The Journal of Machine Learning Research*, 20(170):1–50.

Wang, Z. and Scott, D. W. (2019). Nonparametric density estimation for high-dimensional data-algorithms and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(4):e1461.

Weinzierl, S. (2000). Introduction to monte carlo methods. *arXiv preprint hep-ph/0006269*.

Appendix

A. Additional comments

We begin this section with a simple example highlighting the robustness of MoM-KDE compared with KDE (see also Figure 1 for a different visual example).

Example 1. *(MoM-KDE v.s. Uniform KDE) Let the inliers be i.i.d. samples from a uniform distribution on the interval* $[-1,1]$ *and the outliers be i.i.d. samples from another uniform distribution on* $[-3,3]$*. Let the kernel function be the uniform kernel,* $x_0 = 2$ *and* $h \in (0,1)$*. Then if* $S > 2|\mathcal{O}|$*, we obtain*

$$|\hat{f}_{MoM}(x_0) - f(x_0)| = 0 \quad a.s.$$
$$\textit{and} \quad \mathbb{P}\left(|\hat{f}_n(x_0) - f(x_0)| = 0\right) = (1 - h/3)^{|\mathcal{O}|} \neq 1 \ .$$

Proof. Since the inliers are uniform on $[-1,1]$, when $x_0 = 2$, we have $f(x_0) = 0$. Recall that the classical KDE at $x_0 = 2$ with the uniform kernel and $h \in (0,1)$ is:

$$\hat{f}_n(x_0) = \frac{1}{2nh} \sum_{i=1}^{n} I(x_0 - h \leq X_i \leq x_0 + h) \ .$$

The support of the inliers is $[-1,1]$, hence, $\forall i \in \mathcal{I}$, $I(x_0 - h \leq X_i \leq x_0 + h) = 0$ a.s. and without outliers we would have $\hat{f}_n(x_0) = 0$, resulting in no errors at $x_0 = 2$.

To prove the first equality, it suffices to observe that $S > 2|\mathcal{O}|$ implies that more than half the blocks do not contain outliers. Over each one of these block, the KDE at x_0 is thus equal to 0 a.s. making the MoM estimate equal to 0 as well. Finally it results in $|\hat{f}_{MoM}(x_0) - f(x_0)| = \hat{f}_{MoM}(x_0) = 0$, almost surely.

Let us prove the second equality:

$$\begin{aligned}
\mathbb{P}\left(|\hat{f}_n(x_0) - f(x_0)| = 0\right) &= \mathbb{P}\left(\hat{f}_n(x_0) = 0\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} I(x_0 - h \leq X_i \leq x_0 + h) = 0\right) \\
&= \mathbb{P}\left(\sum_{i \in \mathcal{O}} I(x_0 - h \leq X_i \leq x_0 + h) = 0\right) \\
&= \mathbb{P}\Big(\forall i \in \mathcal{O}, X_i \notin [x_0 - h, x_0 + h]\Big) \\
&= (1 - h/3)^{|\mathcal{O}|} \ .
\end{aligned}$$

□

This result shows that the MoM-KDE makes (almost surely) no error at the point x_0. On the contrary, the KDE here has a non-negligible probability to make an error.

We now make three comments on the behavior of the MoM-KDE.

Optimality. Characterizing the optimality i.e. obtaining lower bound/minimax rates, for our framework is a difficult task. This is why, at the moment, we characterize the optimality by comparing our rate with the one of the standard KDE, known to be optimal without outliers (a strategy already used in Lecué et al. (2020a) for example).
Note that Liu and Gao (2019) study minimax rates of the KDE. Nevertheless, their framework is different. Indeed, they consider the Huber model and study the optimality under L^2-convergence. Actually, to be consistent with Liu and Gao (2019), we have also been interested in finding error rates with L^2-convergence. However, the median operator makes the analysis a lot more complicated, as it would require well-suited concentration inequalities for order statistics in the manner of Boucheron and Thomas (2012) but for sub-gaussian random variables.

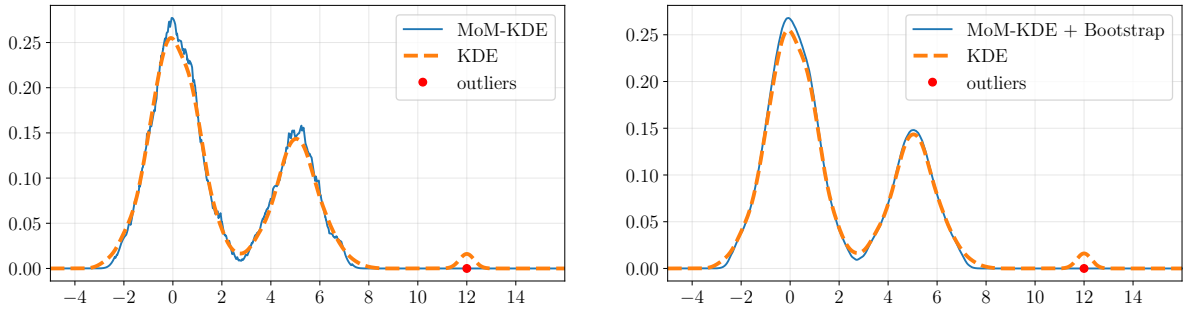


Figure 5. Illustration of MoM-KDE + Bootstrap.

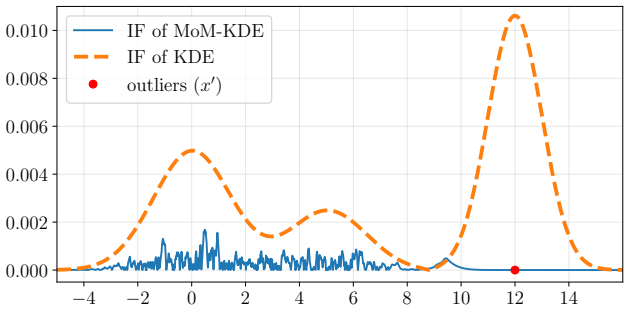


Figure 6. Influence function for the 10-randomized version of MoM-KDE with $S = 2m + 1$ and KDE where the m outliers are placed at $x' = 12$. The true density is $2/3 \cdot \mathcal{N}(0, 1) + 1/3 \cdot \mathcal{N}(5, 1)$.

MoM is dependent from the partition. In Figure 2, we see that the MoM-IF is very erratic. This is essentially due to the partitioning. We actually have several ideas to mitigate this behavior. One natural idea is to perform several partitions/estimations and then aggregate them. However this would come at the price of additional computations and the theoretical analysis would become quickly intractable.

Randomization to reduce the variance. During the process of the MoM-KDE, the split of the dataset is arbitrary. One natural idea, is to consider different splits in order to learn a single MoM-KDE per split, and then average them. The idea being to learn a smoother density estimator. We give an illustration of this procedure in Figure 5. We clearly see that MoM-KDE together with Bootstrap is smoother than just one MoM-KDE. The impact of this strategy is also illustrated in Figures 6 regarding the IF.

B. Technical proofs

Lemma 1 *(L_∞ error-bound of the KDE without outliers - (Wang et al., 2019)) Suppose that f belongs to the class of densities $\mathcal{P}(\alpha, L)$ defined as*

$$\mathcal{P}(\alpha, L) \triangleq \left\{ f \mid f \geq 0, \int f(x)dx = 1, \text{ and } f \in \Sigma(\alpha, L) \right\} ,$$

where $\Sigma(\alpha, L)$ is the Hölder class of function on $\mathbb{R}^d$. Grant assumptions 1 to 3 and let $h \in (0, 1)$, $\gamma > 0$, n large enough, and $nh^d \geq 1$. Then with probability at least $1 - \exp(-\gamma)$, we have

$$\|\hat{f}_n - f\|_\infty \leq C_1 \sqrt{\frac{\gamma + \log(1/h)}{nh^d}} + C_2 h^\alpha ,$$

where $C_2 = L \int \|u\|^\alpha K(u)du < \infty$ and C_1 is a constant that only depends on $\|f\|_\infty$, the dimension d, and the kernel properties.

Proposition 1 *(L_∞ error-bound of the MoM-KDE under the $\mathcal{O} \cup \mathcal{I}$) Suppose that f belongs to the class of densities $\mathcal{P}(\alpha, L)$ and grant assumptions 1 to 3. Let S be the number of blocks such that $S \geq 2|\mathcal{O}| + 1$. Then, for any $h \in (0,1)$, $\gamma > 0$, n/S large enough, and $nh^d \geq S$, we have with probability at least $1 - \exp(-\gamma)$,*

$$\|\hat{f}_{MoM} - f\|_\infty \leq C_1 \sqrt{\frac{S(\log(S) + \gamma + \log(1/h))}{nh^d}} + C_2 h^\alpha \ ,$$

where $C_2 = L \int \|u\|^\alpha K(u) du < \infty$ and C_1 is a constant that only depends on $\|f\|_\infty$, the dimension d, and the kernel properties.

Proof. From the definition of the MoM-KDE, we have the following implication (Lecué et al., 2020b)

$$\left\{ \sup_x \left| \hat{f}_{MoM}(x) - f(x) \right| \geq \varepsilon \right\} \Longrightarrow \left\{ \sup_x \sum_{s=1}^{S} I\left(\left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \geq S/2 \right\} \ .$$

Thus to upper-bound the probability of the left-hand event, it suffices to upper-bound the probability of the right-hand event. Moreover, we have

$$\begin{aligned}
& \left| \hat{f}_{n_s}(x) - f(x) \right| \leq \sup_x \left| \hat{f}_{n_s}(x) - f(x) \right| \\
\Longrightarrow \quad & I\left(\left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \leq I\left(\sup_x \left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \\
\Longrightarrow \quad & \sum_{s=1}^{S} I\left(\left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \leq \sum_{s=1}^{S} I\left(\sup_x \left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \\
\Longrightarrow \quad & \sup_x \sum_{s=1}^{S} I\left(\left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \leq \sum_{s=1}^{S} I\left(\sup_x \left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \ ,
\end{aligned}$$

which implies that

$$\mathbb{P}\left(\sup_x \sum_{s=1}^{S} I\left(\left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \geq S/2 \right) \leq \mathbb{P}\left(\sum_{s=1}^{S} I\left(\sup_x \left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) \geq S/2 \right) \ .$$

Let $Z_s = I\left(\sup_x \left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right)$ and let $\mathcal{S} = \left\{ s \in \{1, \cdots, S\} \mid B_s \cap \mathcal{O} = \emptyset \right\}$ i.e. the set of indices s such that the block B_s does not contain any outliers. Since $\sum_{s \in \mathcal{S}^C} I(\cdot)$ is bounded by $|\mathcal{O}|$, almost surely, the following holds.

$$\sum_{s=1}^{S} I\left(\sup_x \left| \hat{f}_{n_s}(x) - f(x) \right| > \varepsilon \right) = \sum_{s=1}^{S} Z_s = \sum_{s \in \mathcal{S}} Z_s + \sum_{s \in \mathcal{S}^C} Z_s \leq \sum_{s \in \mathcal{S}} Z_s + |\mathcal{O}| \ . \tag{6}$$

Note that $\mathcal{S}$ is never empty thanks to the hypothesis $S \geq 2|\mathcal{O}| + 1$.

For n large enough e.g. $(n/S)h^d > \gamma$ and $(n/S)h^d > |\log(\gamma)|$ (Sriperumbudur and Steinwart (2012) Thm 3.1), Lemma 1 with $\varepsilon = C_1 \sqrt{\frac{S(\gamma + \log(1/h))}{nh^d}} + C_2 h^\alpha$, leads to

$$p_\varepsilon := \mathbb{P}\left(\sup_x \left| \hat{f}_{n_1}(x) - f(x) \right| > \varepsilon \right) \leq e^{-\gamma} \ .$$

Combining this last inequality with equation (6) gives

$$\begin{aligned}\mathbb{P}\left(\sum_{s=1}^{S} I\left(\sup_x \left|\hat{f}_{n_s}(x) - f(x)\right| > \varepsilon\right) \geq S/2\right) &\leq \mathbb{P}\left(\sum_{s\in\mathcal{S}} Z_s + |\mathcal{O}| \geq S/2\right) \\ &\leq \mathbb{P}\left(\sum_{s\in\mathcal{S}} Z_s \geq S/2 - |\mathcal{O}|\right) \\ &\leq \mathbb{P}\left(\sum_{s\in\mathcal{S}} Z_s \geq 1/2\right) = 1 - \mathbb{P}\left(\sum_{s\in\mathcal{S}} Z_s = 0\right) \\ &\leq 1 - (1-p_\varepsilon)^{|\mathcal{S}|} \leq 1 - (1-p_\varepsilon)^S \\ &\leq 1 - (1-e^{-\gamma})^S \ .\end{aligned}$$

Finally, using the fact that $1 - y \cdot e^{-x} \leq (1-e^{-x})^y$, we obtain

$$\mathbb{P}\left(\|\hat{f}_{MoM} - f\|_\infty \leq C_1\sqrt{\frac{S(\gamma + \log(1/h))}{nh^d}} + C_2 h^\alpha\right) \geq 1 - S\cdot e^{-\gamma} \ .$$

Replacing γ by $\log(S) + \gamma > 0$ gives

$$\mathbb{P}\left(\|\hat{f}_{MoM} - f\|_\infty \leq C_1\sqrt{\frac{S(\log(S) + \gamma + \log(1/h))}{nh^d}} + C_2 h^\alpha\right) \geq 1 - e^{-\gamma} \ .$$

Note that we can easily extend this proof to the adversarial framework (Depersin and Lecué, 2021). Indeed, suppose $Z_1, \cdots, Z_S$ defined above equation (6) are corrupted and $Z'_1, \cdots, Z'_S$ are the corresponding uncorrupted version of them. We have that $Z_1, \cdots, Z_S$ are potentially not independent but $Z'_1, \cdots, Z'_S$ are independent and $Z_i = Z'_i$ for any $i \in \mathcal{S}$ (this is the adversarial framework). Then,

$$\sum_{s=1}^{S} Z_s = \sum_{s=1}^{S} Z'_s + \sum_{s=1}^{S}(Z_s - Z'_s) \leq \sum_{s=1}^{S} Z'_s + |\mathcal{O}| \ .$$

The right-hand side is now a sum of i.i.d. bounded random variables and restarting from equation (6) we can conclude.

□

Corollary 1 *(Rate of convergence)* Consider the assumptions of Proposition 1 with $S = 2|\mathcal{O}| + 1$, $\gamma = \log(n)$ and $h \asymp \left(\frac{S\log(n)}{n}\right)^{1/(2\alpha+d)}$. Thus, with probability higher than $1 - \frac{1}{n}$, we have

$$\|\hat{f}_{MoM} - f\|_\infty \lesssim \left(\frac{|\mathcal{O}|}{n}\log(n)\right)^{\alpha/(2\alpha+d)} + \left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)} \ .$$

Proof. From the previous proposition we know that with probabability $1 - 1/n$, we have

$$\begin{aligned}\|\hat{f}_{MoM} - f\|_\infty &\leq C_1\sqrt{\frac{S(\log(S) + \log(n) + \log(1/h))}{nh^d}} + C_2 h^\alpha \\ &\leq C_1\sqrt{\frac{3S\log(n)}{nh^d}} + C_2 h^\alpha \ .\end{aligned}$$

Replacing h by $\left(\frac{S\log(n)}{n}\right)^{1/(2\alpha+d)}$ in this last equation gives us that

$$\|\hat{f}_{MoM} - f\|_\infty \lesssim \left(\frac{S\log(n)}{n}\right)^{\alpha/(2\alpha+d)} \leq \left(\frac{|\mathcal{O}|}{n}\log(n)\right)^{\alpha/(2\alpha+d)} + \left(\frac{\log(n)}{n}\right)^{\alpha/(2\alpha+d)} .$$

□

Proposition 2. *(L_1-consistency in probability) If $n/S \to \infty$, $h \to 0$, $nh^d \to \infty$, $S/\sqrt{nh^d} \to 0$, and $S \geq 2|\mathcal{O}|+1$, then*

$$\|\hat{f}_{MoM} - f\|_1 \xrightarrow[n\to\infty]{\mathcal{P}} 0 .$$

Proof. We first rewrite the MoM-KDE as

$$\hat{f}_{MoM}(x) = \sum_{s=1}^{S} \hat{f}_{n_s}(x) I_{A_s}(x) ,$$

where $A_s = \left\{x \mid \hat{f}_{MoM}(x) = \hat{f}_{n_s}(x)\right\}$. Without loss of generality, we assume that

$$A_k \bigcap_{s\neq\ell}^{S} A_\ell = \emptyset, \quad \bigcup_{s=1}^{S} A_s = \mathbb{R}^d, \quad \text{and} \quad \sum_{s=1}^{S} I_{A_s}(x) = 1 .$$

$$\begin{aligned}
\int \left|\hat{f}_{MoM}(x) - f(x)\right| dx &= \int \left|\sum_{s=1}^{S} \hat{f}_{n_s}(x) I_{A_s}(x) - f(x)\right| dx \\
&= \int \left|\sum_{s=1}^{S} \left(\hat{f}_{n_s}(x) - f(x)\right) I_{A_s}(x)\right| dx \\
&\leq \int \sum_{s=1}^{S} \left|\hat{f}_{n_s}(x) - f(x)\right| I_{A_s}(x) dx \\
&= \sum_{s=1}^{S} \int_{A_s} \left|\hat{f}_{n_s}(x) - f(x)\right| dx \\
&= \sum_{s\in\mathcal{S}} \int_{A_s} \left|\hat{f}_{n_s}(x) - f(x)\right| dx + \sum_{s\in\mathcal{S}^C} \int_{A_s} \left|\hat{f}_{n_s}(x) - f(x)\right| dx . \qquad (7)
\end{aligned}$$

From the L_1-consistency of the KDE in probability, if the number of anomalies grows at a small enough speed (Devroye and Gyorfi, 1985), the left part is bounded, i.e.

$$\sum_{s\in\mathcal{S}} \int_{A_s} \left|\hat{f}_{n_s}(x) - f(x)\right| dx \leq \sum_{s\in\mathcal{S}} \int \left|\hat{f}_{n_s}(x) - f(x)\right| dx \xrightarrow[n\to\infty]{\mathcal{P}} 0 . \qquad (8)$$

To be more specific, since $R_n = \sqrt{nh^d}$, the rate given in Thm 5.1 of Devroye and Gyorfi (1985), the convergence is guarantee whenever S/R_n tends to 0 e.g. with $S = \mathcal{O}(R_n^\tau)$ and any $\tau \in (0,1)$. We now upper-bound the right part of equation (7). Let consider a particular block A_s where $s \in \mathcal{S}^C$. In this block, the estimator f_{n_s} is selected and is calculated with samples containing anomalies. As $\forall x \in A_s$, $f_{n_s}(x)$ is the median (by definition), if $S > 2|\mathcal{O}|$, we can always find a $s' \in \mathcal{S}$ such that $f_{n_s}(x) \leq f_{n_{s'}}(x)$ or $f_{n_s}(x) \geq f_{n_{s'}}(x)$.

Now let denote by $A_s^+ = \left\{x \in A_s \mid \hat{f}_{n_s}(x) \geq f(x)\right\}$ and $A_s^- = \left\{x \in A_s \mid \hat{f}_{n_s}(x) < f(x)\right\}$. We have $A_s^+ \cup A_s^- = A_s$ and each one of these blocks can be decomposed respectively into $|\mathcal{S}|$ sub-blocks (not necessarily disjoint) $\{A_s^{s',+}\}_{s'\in\mathcal{S}}$ and $\{A_s^{s',-}\}_{s'\in\mathcal{S}}$ such that $\forall s' \in \mathcal{S}$,

$A_s^{s',+} = \left\{ x \in A_s \mid \hat{f}_{n_{s'}}(x) \geq \hat{f}_{n_s}(x) \geq f(x) \right\}$ and $A_s^{s',-} = \left\{ x \in A_s \mid \hat{f}_{n_{s'}}(x) \leq \hat{f}_{n_s}(x) < f(x) \right\}$.

Finally, the right-hand term of equation (7) can be upper-bounded by

$$\begin{aligned}
\sum_{s \in \mathcal{S}^C} \int_{A_s} \left| \hat{f}_{n_s}(x) - f(x) \right| dx &\leq \sum_{s \in \mathcal{S}^C} \int_{A_s^+} \left| \hat{f}_{n_s}(x) - f(x) \right| dx + \int_{A_s^-} \left| \hat{f}_{n_s}(x) - f(x) \right| dx \\
&\leq \sum_{s \in \mathcal{S}^C} \sum_{s' \in \mathcal{S}} \int_{A_s^{s',+}} \left| \hat{f}_{n_s}(x) - f(x) \right| dx + \int_{A_s^{s',-}} \left| \hat{f}_{n_s}(x) - f(x) \right| dx \\
&\leq \sum_{s \in \mathcal{S}^C} \sum_{s' \in \mathcal{S}} \int_{A_s^{s',+}} \left| \hat{f}_{n_{s'}}(x) - f(x) \right| dx + \int_{A_s^{s',-}} \left| \hat{f}_{n_{s'}}(x) - f(x) \right| dx \\
&\leq \sum_{s \in \mathcal{S}^C} \sum_{s' \in \mathcal{S}} \int \left| \hat{f}_{n_{s'}}(x) - f(x) \right| dx + \int \left| \hat{f}_{n_{s'}}(x) - f(x) \right| dx
\end{aligned}$$

Since $\forall s' \in \mathcal{S}$ we have $\int |\hat{f}_{n_{s'}}(x) - f(x)| dx \xrightarrow[n \to \infty]{\mathcal{P}} 0$, we can conclude using similar arguments as those used for (8) that $\sum_{s \in \mathcal{S}^C} \int_{A_s} |\hat{f}_{n_s}(x) - f(x)| dx \xrightarrow[n \to \infty]{\mathcal{P}} 0$, which concludes the proof.

□

Proposition 3. *Let $x', x_0 \in \mathbb{R}^d$ and $\mathcal{I}_n$ be an healthy data set. Grant assumptions 1 to 3 and denote*

$$a \triangleq \sum_{i \in \mathcal{I}} K\left(\frac{X_i - x_0}{h} \right), \quad b \triangleq K\left(\frac{x' - x_0}{h} \right) .$$

Let $S \geq 2m + 1$ with $m \in [\![0, \frac{n}{2}[\![$ the number of added samples and $\delta > 0$ such that $|b - a/n| > C_\rho \sqrt{2\delta S/n}$.
If $m \geq \frac{C_\rho \sqrt{2n\delta S}}{|b - a/n| - C_\rho \sqrt{2\delta S/n}}$, then with probability higher than $1 - 4\exp(-\delta)$ we have:

$$\text{IF}_{\mathcal{O} \cup \mathcal{I}}(x_0, x', m; \mathcal{I}_n, \hat{f}_{MoM}) \leq \text{IF}_{\mathcal{O} \cup \mathcal{I}}(x_0, x', m; \mathcal{I}_n, \hat{f}_{KDE}) .$$

Proof.

– The $\text{IF}_{\mathcal{O} \cup \mathcal{I}}$ for the KDE is

$$\begin{aligned}
\text{IF}_{\mathcal{O} \cup \mathcal{I}}(x_0, x', m, \mathcal{I}_n; \hat{f}_{KDE}) &= \left| \frac{1}{(n+m)h^d} \left(\sum_{i=1}^n K\left(\frac{X_i - x_0}{h} \right) + \sum_{i=1}^m K\left(\frac{x' - x_0}{h} \right) \right) - \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h} \right) \right| \\
&= \frac{1}{h^d} \left| \frac{1}{n+m} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h} \right) - \frac{1}{n} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h} \right) + \frac{m}{n+m} K\left(\frac{x' - x_0}{h} \right) \right| \\
&= \frac{1}{h^d} \left| \left(\frac{1}{n+m} - \frac{1}{n} \right) \sum_{i=1}^n K\left(\frac{X_i - x_0}{h} \right) + \frac{m}{n+m} K\left(\frac{x' - x_0}{h} \right) \right| \\
&= \frac{1}{h^d} \left| \left(\frac{1}{n+m} - \frac{1}{n} \right) a + \left(\frac{m}{n+m} \right) b \right| = \frac{1}{h^d} \left| \frac{nb - a}{n^2/m + n} \right| ,
\end{aligned}$$

with $a \triangleq \sum_{i=1}^n K\left(\frac{X_i - x_0}{h} \right)$, and $b \triangleq K\left(\frac{x' - x_0}{h} \right)$.

– $\text{IF}_{\mathcal{O} \cup \mathcal{I}}$ for the MoM-KDE

Let $S > 2m$ be the number of blocks in the MoM-KDE, $\{B_s\}_{s=1}^S$ and $\{\tilde{B}_{s'}\}_{s'=1}^S$ be respectively the blocks of the contaminated data set $\mathcal{I}_n \cup \{x'\}^m$ and the healthy data set. We have:

$$\begin{aligned}\text{IF}_{\mathcal{O}\cup\mathcal{I}}(x_0, x', m, \mathcal{I}_n; \hat{f}_{MoM}) = \frac{1}{h^d}\Bigg|&\text{Median}\left\{\frac{S}{n+m}\left(\sum_{i\in\mathcal{I}_n\cap B_s} K\left(\frac{X_i - x_0}{h}\right) + \sum_{i\in\{x'\}\cap B_s} K\left(\frac{x'-x_0}{h}\right)\right)\right\}_{s=1}^S \\ &- \text{Median}\left\{\frac{S}{n}\sum_{i\in\tilde{B}_{s'}} K\left(\frac{X_i - x_0}{h}\right)\right\}_{s'=1}^S\Bigg| \\ \leq\ & \frac{1}{h^d}\left|\frac{S}{n+m}\sum_{i\in B_s} K\left(\frac{X_i - x_0}{h}\right) - \frac{S}{n}\sum_{i\in\tilde{B}_{s'}} K\left(\frac{X_i - x_0}{h}\right)\right|,\end{aligned}$$

where $\tilde{B}_{s'}$ is the block selected by the median for the healthy MoM-KDE. The inequality is obtained by noticing that, with $S > 2m$, there always exists an healthy block B_s that makes it true.

Finally, denoting

$$\left|\sum_{i\in B_s}\frac{S}{n+m}K\left(\frac{X_i - x_0}{h}\right) - \sum_{i\in\tilde{B}_{s'}}\frac{S}{n}K\left(\frac{X_i - x_0}{h}\right)\right| = \left|Z^{(s)} - Z^{(s')}\right|,$$

and using both the triangular and Hoeffding inequalities, and the fact that $\mathbb{E}(Z^{(s)}) = \mathbb{E}(Z^{(s')})$, we have

$$\begin{aligned}h^d \cdot \text{IF}_{\mathcal{O}\cup\mathcal{I}}(x_0, x', m, \mathcal{I}_n; \hat{f}_{MoM}) &\leq \left|Z^{(s)} - Z^{(s')} - \mathbb{E}(Z^{(s)}) + \mathbb{E}(Z^{(s')})\right| \\ &\leq \left|Z^{(s)} - \mathbb{E}(Z^{(s)})\right| + \left|Z^{(s')} - \mathbb{E}(Z^{(s')})\right|,\end{aligned}$$

and for $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|Z^{(s)} - \mathbb{E}(Z^{(s)})\right| \geq t\right) &\leq 2\exp\left(-\frac{2t^2(n+m))}{SC_\rho^2}\right) \\ \mathbb{P}\left(\left|Z^{(s')} - \mathbb{E}(Z^{(s')})\right| \geq t\right) &\leq 2\exp\left(-\frac{2t^2 n}{SC_\rho^2}\right) = 2\exp\left(-\frac{2n_s t^2}{C_\rho^2}\right),\end{aligned}$$

where $n_s = n/S$. Furthermore, given two real-valued random variables X, Y, we know that for $t > 0$,

$$\mathbb{P}\left(|X| + |Y| \geq 2t\right) \leq \mathbb{P}\left(|X| \geq t\right) + \mathbb{P}\left(|Y| \geq t\right).$$

Therefore, we have

$$\mathbb{P}\left(\left|Z^{(s)} - \mathbb{E}(Z^{(s)})\right| + \left|Z^{(s')} - \mathbb{E}(Z^{(s')})\right| \geq t\right) \leq 4\exp\left(-\frac{n_s t^2}{2C_\rho^2}\right).$$

Setting $t = \frac{C_\rho\sqrt{2\delta}}{\sqrt{n_s}}$ with $\delta > 0$, finally gives $\mathbb{P}\left(\left|Z^{(s)} - Z^{(s')}\right| < t\right) \geq 1 - 4\exp(-\delta)$. We now know that with probability $1 - 4\exp(-\delta)$, $\text{IF}_{\mathcal{O}\cup\mathcal{I}}(x_0, x', m, \mathcal{I}_n; \hat{f}_{MoM}) < \frac{1}{h^d}\frac{C_\rho\sqrt{2\delta}}{\sqrt{n_s}}$ and we seek for which value of m, this value is smaller than

the $\mathrm{IF}_{\mathcal{O}\cup\mathcal{I}}$ of the KDE i.e.

$$\begin{aligned}
\frac{C_\rho\sqrt{2\delta}}{\sqrt{n_s}} \leq \left|\frac{nb-a}{n^2/m+n}\right| &\iff n^2/m+n \leq \frac{\sqrt{n_s}|nb-a|}{C_\rho\sqrt{2\delta}} \\
&\iff 1/m \leq \left(\frac{\sqrt{n_s}|nb-a|}{C_\rho\sqrt{2\delta}} - n\right)/n^2 \\
&\iff m \geq \frac{n^2}{\frac{\sqrt{n_s}|nb-a|}{C_\rho\sqrt{2\delta}} - n} \\
&\iff m \geq \frac{n^2}{\frac{\sqrt{n}|nb-a|}{C_\rho\sqrt{2\delta S}} - n} \\
&\iff m \geq \frac{n}{\frac{\sqrt{n}|b-a/n|}{C_\rho\sqrt{2\delta S}} - 1}\,.
\end{aligned}$$

□

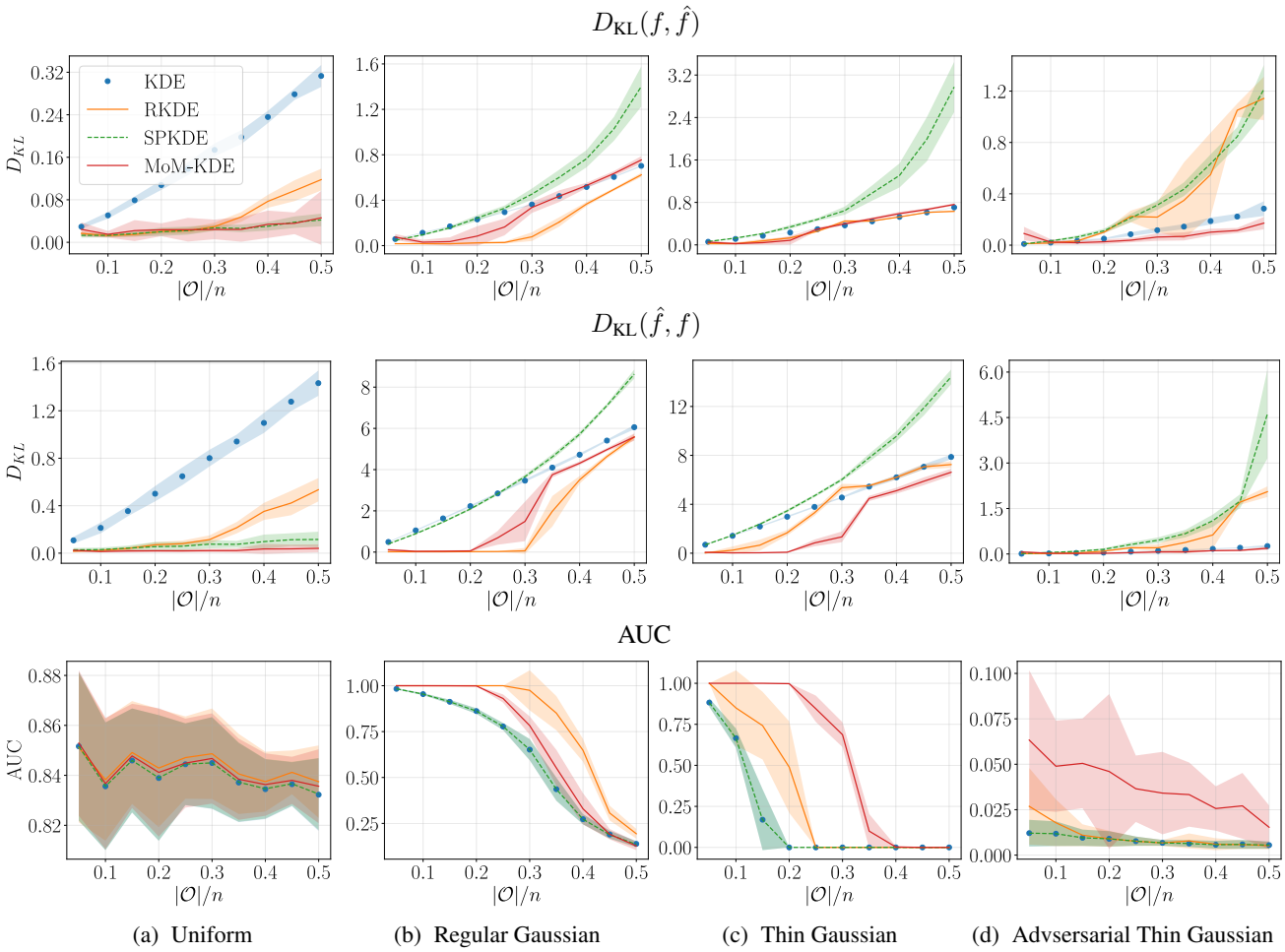
C. Additional results

As stated in the main paper, we display here the additional results containing:

- For synthetic data, the Kullback-Leibler divergence in both directions, i.e. $D_{\mathrm{KL}}(\hat{f}, f)$ and $D_{\mathrm{KL}}(f, \hat{f})$, and the ROC AUC measuring the performance of an anomaly detector based on $\hat{f}$. Results are displayed on Figure 7.
- For synthetic data, a benchmark of Outlier Detection (OD) + KDE. Instead of applying directly a robust estimator, we first proceed to an outlier detection step after which we fit a standard KDE. Results are displayed on Figure 8.
- For Digits data, the ROC AUC measuring the performance of an anomaly detector based on $\hat{f}$. As stated in the main paper, this is done under multiple scenarios, where outliers and inliers can be chosen among the nine available classes. Here we show the AUC when the outliers are set as one class (class 2 to class 9), and inliers are set as "the rest" of all classes. Results are displayed on Figure 9.
- Additional visualizations highlighting the behavior of the Influence Function of KDE and MoM-KDE (Figures 10 and 11).

KL, ROC and AUC for Synthetic data. When considering the Kullback-Leibler divergence, results lead to a very similar conclusion as previously stated, that is, an overall good performance of MoM-KDE while its competitors, notably SPKDE, are more data-dependent. When the density estimate $\hat{f}$ is used in a simple anomaly detector, results are quite different. Indeed, when outliers are uniformly distributed, even if MoM-KDE seems to better estimate the true density (according to D_{JS} and D_{KL}), this doesn't make $\hat{f}_{MoM}$ a better anomaly detector. It seems that in this case, the outliers are either easily detected because distant from the density estimate, or located in dense regions, thus making them impossible to identify, and this for all density estimates provided by competitors. In the case of adversarial contamination, the conclusion is quite similar. Although MoM-KDE better fits the true density, the situation is extremely difficult for anomaly detection, hence making all competitors yield very poor results. In the two other cases – Gaussian outlier, anomaly detection results follow the density estimation.

Benchmark of Outlier Detection (OD) + KDE: We proceed to the OD step using two different methods, the KDE and the Isolation Forest (IF) (Liu et al., 2008). For the first method, we fit a KDE estimate using all data and we erase the proportion $\varepsilon = |\mathcal{O}|/n$ of samples that have the lower density. For the IF, we proceed similarly by removing a proportion ε of outliers. Finally, we fit a KDE estimate on the new samples. The results are given in Figure 8 and show that OD + KDE (purple curve) or IF + KDE (brown) is not as efficient as RKDE or MoM-KDE.

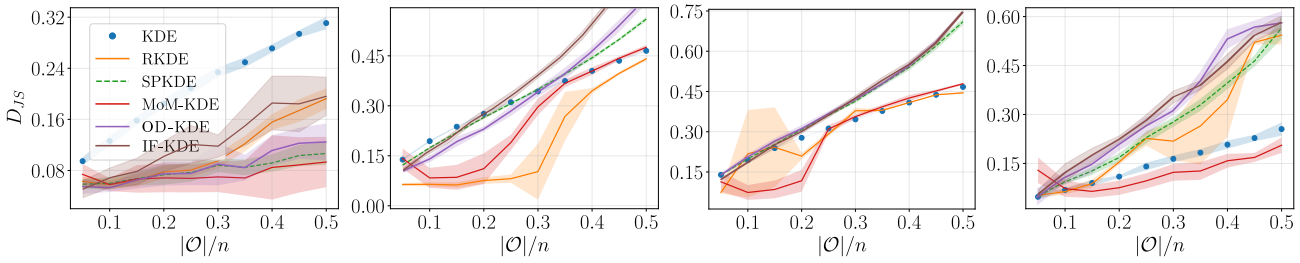


(a) Uniform (b) Regular Gaussian (c) Thin Gaussian (d) Advsersarial Thin Gaussian

Figure 7. Density estimation with synthetic data. The displayed metrics are the Kullback-Leibler divergence (a lower score means a better estimation of the true density) and the AUC (a higher score means a better detection of the outliers).

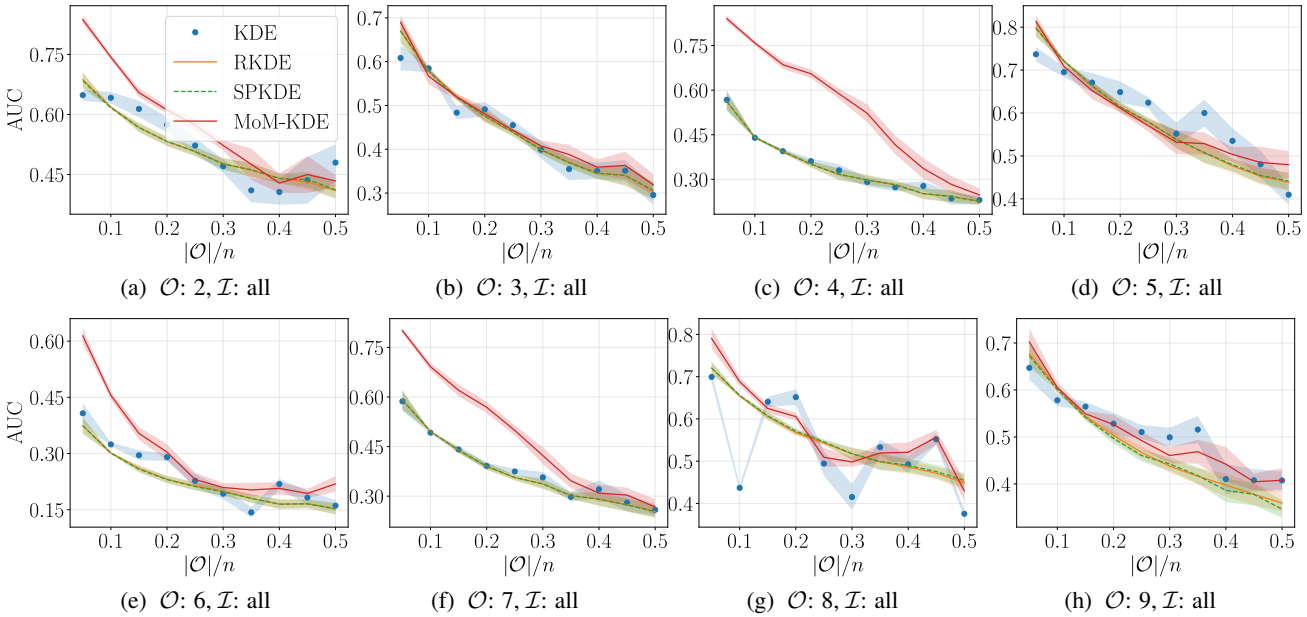
Digits data. Results over Digits scenarios are inline with main conclusions over real data. Although from one scenario to another, all methods have varied results, the overall observation is that MoM-KDE is either similar or better than its competitors.

Influence Function. Illustrations of the behavior of the Influence Function (IF) for MoM-KDE and KDE are given in Figures 10 and 11. They show that the IF for MoM-KDE is lower than the one of KDE when the m outlying points are all placed at a low density point x' (see Figure 10). On the other hand, Figure 11 shows that whenever the outlying points are all placed at a high density point x', the two IF exhibit similar behavior.



(a) Uniform (b) Regular Gaussian (c) Thin Gaussian (d) Advsersarial Thin Gaussian

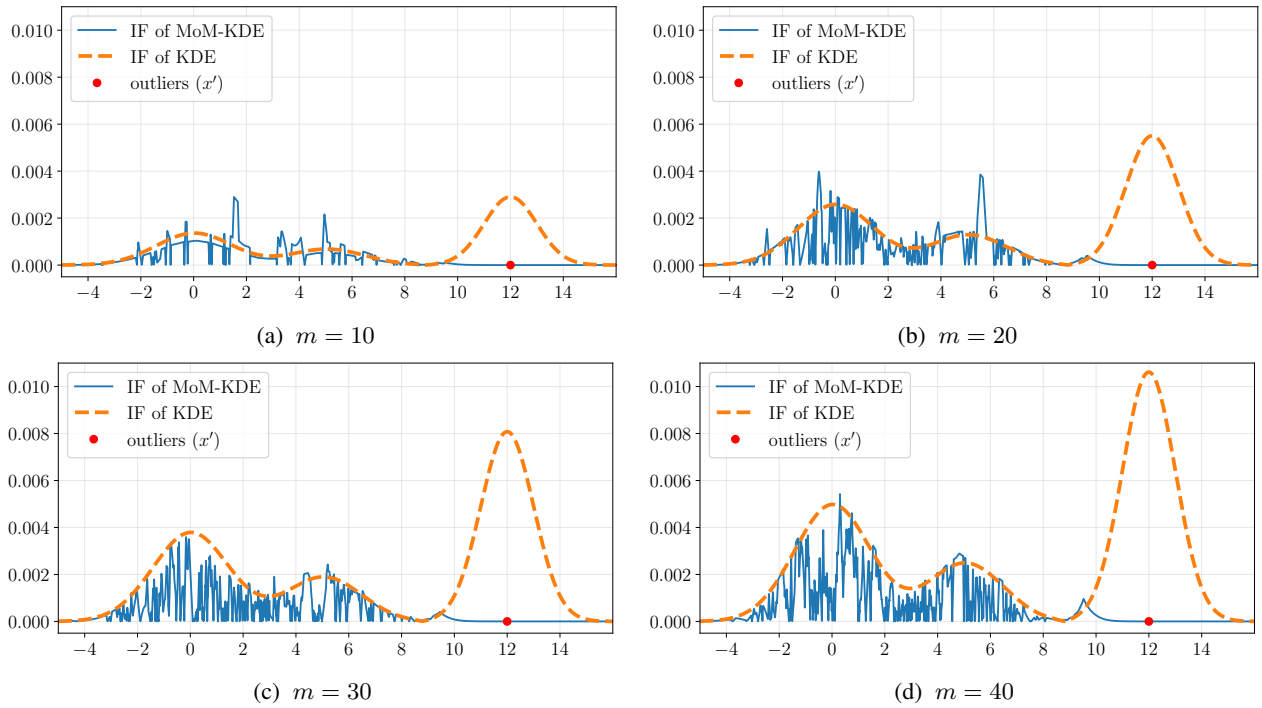
Figure 8. Density estimation with synthetic data. The displayed metric is the Jensen-Shannon divergence. A lower score means a better estimation of the true density.



(a) $\mathcal{O}$: 2, $\mathcal{I}$: all (b) $\mathcal{O}$: 3, $\mathcal{I}$: all (c) $\mathcal{O}$: 4, $\mathcal{I}$: all (d) $\mathcal{O}$: 5, $\mathcal{I}$: all

(e) $\mathcal{O}$: 6, $\mathcal{I}$: all (f) $\mathcal{O}$: 7, $\mathcal{I}$: all (g) $\mathcal{O}$: 8, $\mathcal{I}$: all (h) $\mathcal{O}$: 9, $\mathcal{I}$: all

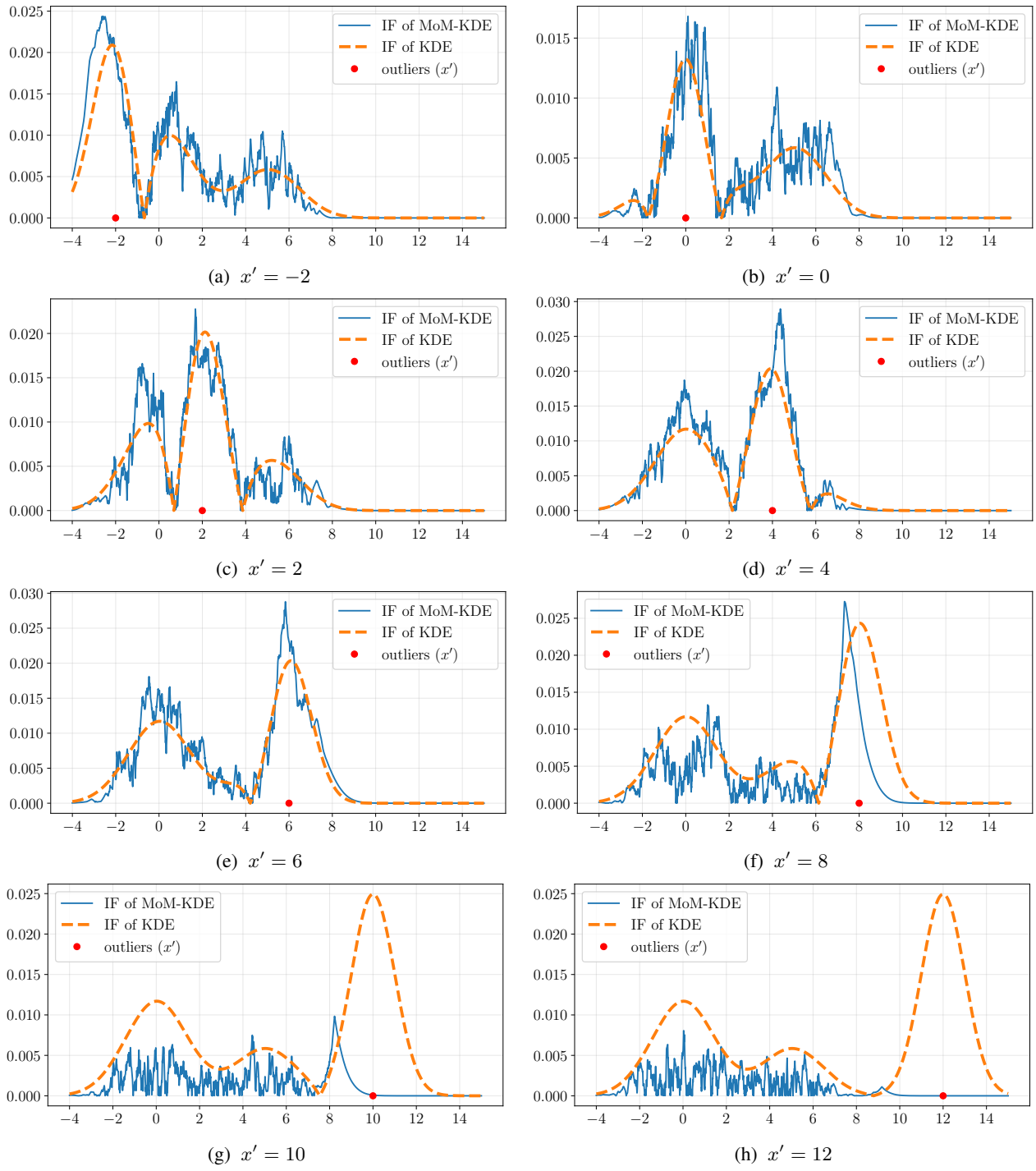
Figure 9. Anomaly detection with Digits data, measured with AUC over varying outlier proportion. A higher score means a better detection of the outliers. We specify which classes are chosen to be inliers ($\mathcal{I}$) and outliers ($\mathcal{O}$).



(a) $m = 10$ (b) $m = 20$

(c) $m = 30$ (d) $m = 40$

Figure 10. Influence function for the 10-randomized version of MoM-KDE with $S = 2m + 1$ and KDE where the m outliers are placed at $x' = 12$. The true density is $2/3 \cdot \mathcal{N}(0, 1) + 1/3 \cdot \mathcal{N}(5, 1)$.



(a) $x' = -2$ (b) $x' = 0$

(c) $x' = 2$ (d) $x' = 4$

(e) $x' = 6$ (f) $x' = 8$

(g) $x' = 10$ (h) $x' = 12$

Figure 11. Influence function for MoM-KDE with $S = 2m + 1$ and KDE where the $m = 100$ outliers are placed at x', and where the samples in $\mathcal{I}_n$ are drawn from the true density: $2/3 \cdot \mathcal{N}(0, 1) + 1/3 \cdot \mathcal{N}(5, 1)$.