On the Robustness of Kernel Density M-Estimators

JooSeuk Kim STANNUM@UMICH.EDU
Clayton D. Scott CLAYSCOT@UMICH.EDU
Department of Electrical Engineering and Computer Science, University of Michigan
1301 Beal Avenue, Ann Arbor, MI, 48109-2122 USA

Abstract

We analyze a method for nonparametric density estimation that exhibits robustness to contamination of the training sample. This method achieves robustness by combining a traditional kernel density estimator (KDE) with ideas from classical M-estimation. The KDE based on a Gaussian kernel is interpreted as a sample mean in the associated reproducing kernel Hilbert space (RKHS). This mean is estimated robustly through the use of a robust loss, yielding the so-called robust kernel density estimator (RKDE). This robust sample mean can be found via a kernelized iteratively re-weighted least squares (IRWLS) algorithm. Our contributions are summarized as follows. First, we present a representer theorem for the RKDE, which gives an insight into the robustness of the RKDE. Second, we provide necessary and sufficient conditions for kernel IRWLS to converge to the global minimizer, in the Gaussian RKHS, of the objective function defining the RKDE. Third, characterize and provide a method for computing the influence function associated with the RKDE. Fourth, we illustrate the robustness of the RKDE through experiments on several data sets.

1. Introduction

This paper addresses a method of nonparametric density estimation that exhibits robustness to contamination of the training sample, meaning the training sample consists of some realizations that are not from the density being estimated. Such robust density estimators are motivated, for example, by the problem of anomaly detection. When labeled examples of anomalies are unavailable, it is common to define an anomaly detector by thresholding a density estimate based on non-anomalous data. In applications where it is difficult or impossible to obtain a pure sample (containing no anomalies), robust density estimation can mitigate the impact of contamination.

We analyze a method for robust nonparametric density estimation described by Kim & Scott (2008). This method achieves robustness by combining a traditional kernel density estimator (KDE) with ideas from M-estimation (Huber, 1964; Hampel, 1974). The KDE based on a Gaussian kernel is interpreted as a sample mean in the reproducing kernel Hilbert space (RKHS) associated with the kernel. The sample mean is estimated robustly through the use of a robust loss, yielding the so-called robust kernel density estimator (RKDE). To implement the RKDE, Kim & Scott (2008) introduce a kernelized form of iterative re-weighted least squares (IRWLS). The algorithm is evaluated on 1 and 2 dimensional synthetic datasets.

We make four contributions to the understanding of the RKDE. First, we present a representer theorem for the RKDE, and based on which we give an explanation why the RKDE is robust to outliers. Second, we provide necessary and sufficient conditions for kernel IRWLS to converge to the global minimizer, in the Gaussian RKHS, of the objective function defining the RKDE. Third, we define, characterize, and provide a method for computing the influence function associated with the RKDE. The influence function quantifies the impact on the density estimator of perturbing the random sample with a new data point. Fourth, we conduct experiments on several synthetic and real data sets to illustrate the robustness of the RKDE. In particular, we demonstrate robustness through an empirical investigation of both influence functions and of anomaly detectors based on contaminated data.

Previous work combining robust estimation and kernel methods has focused primarily on supervised learning

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

problems. M-estimation applied to kernel regression has been studied by various authors (see Brabanter et al. (2009) and references within). Robust surrogate losses for kernel-based classifiers have also been studied (Xu et al., 2006). To our knowledge, the RKDE is the first application of M-estimation ideas in kernel density estimation. The problem of nonparametric density estimation with contaminated data, in the sense considered here, has also received little attention. Several papers have considered nonparametric density estimation in the case where data are corrupted by additive noise having known distribution (see, for example, Devroye (1989)). In contrast, we suppose that most of the data come from the target distribution, but a small portion come from some alternative distribution. We begin in Section 2 with a review of the RKDE and the IRWLS algorithm of Kim & Scott (2008). In Section 3 we provide a representer theorem for the RKDE, and necessary and sufficient conditions for the convergence of IRWLS algorithm. The influence function is developed in Section 4, and experimental results are reported in Section 5. Complete proofs can be found at http://www-personal.umich.edu/~stannum/rkde-supple.pdf.

2. Kernel Density M-Estimation

Let $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$ be a random sample from a distribution F with a density f. The kernel density estimate of f, also called the Parzen window estimate, is a nonparametric estimate given by

$$\widehat{f}_{KDE}\left(\mathbf{x}\right) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma\left(\mathbf{x}, \mathbf{X}_i\right),$$

where $k_\sigma\left(\mathbf{x}, \mathbf{X}_i\right)$ is a kernel function. A commonly used kernel function, which we will work with from now on, is the Gaussian kernel

$$k_\sigma\left(\mathbf{x}, \mathbf{X}_i\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(-\frac{\|\mathbf{x}-\mathbf{X}_i\|^2}{2\sigma^2}\right).$$

For the Gaussian kernel, there exists a mapping $\Phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is an infinite dimensional Hilbert space, such that $k_\sigma\left(\mathbf{x}, \mathbf{x}'\right) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}')\rangle$. We will assume that $\Phi(\mathbf{x})$ is the canonical feature map, $\Phi(\mathbf{x}) = k_\sigma(\cdot, \mathbf{x})$. We also recall the reproducing property, which states that for all $g \in \mathcal{H}$, $g(\mathbf{x}) = \langle \Phi(\mathbf{x}), g\rangle$ (Steinwart & Christmann, 2008).

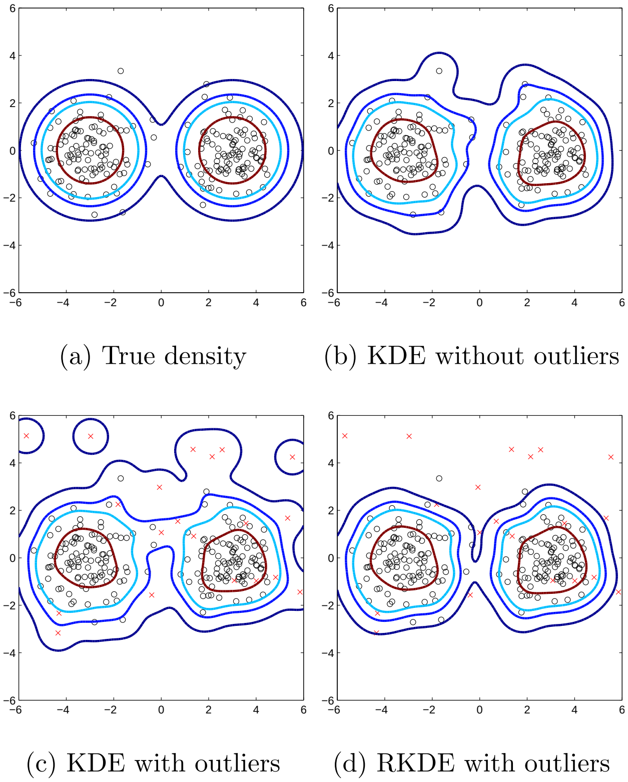


Figure 1. Contours of true density and kernel density estimates along with data samples from true density (o) and outliers (x). 200 data samples are from the true distribution and 20 outliers are from a uniform distribution.

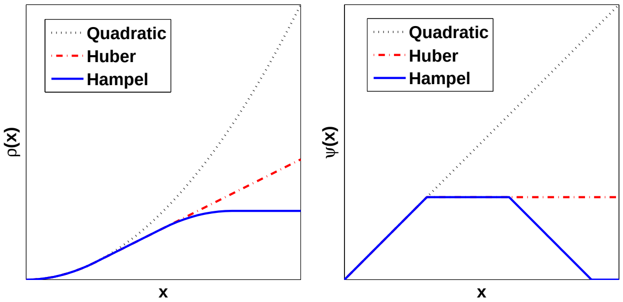
From this point of view, the KDE can be expressed as

$$\begin{aligned}\widehat{f}_{KDE}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n} \langle \Phi(\mathbf{x}), \Phi(\mathbf{X}_i)\rangle \\ &= \left\langle \Phi(\mathbf{x}), \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{X}_i)\right\rangle.\end{aligned}$$

By the reproducing property of $\Phi(\mathbf{x})$, $\widehat{f}_{KDE} \in \mathcal{H}$ can be seen as $\frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{X}_i)$, the sample mean of $\Phi(\mathbf{X}_i)$'s, or equivalently, the solution of

$$\min_{g\in\mathcal{H}} \sum_{i=1}^{n} \|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}^2.$$

Consider the case where the training sample is contaminated by outliers, i.e., some of $\mathbf{X}_1, \cdots, \mathbf{X}_n \in \mathbb{R}^d$ are not from F. As we can see in Figure 1 (c), the KDE is affected by outliers such that the density estimate has small bumps over the regions where the outliers exist. This is because it assigns uniform weights $1/n$ to every $\Phi(\mathbf{X}_i)$ regardless of whether $\mathbf{X}_i$ is an outlier or not, which, in turn, comes from the use of the quadratic loss of $\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}$.



(a) ρ functions (b) ψ functions

Figure 2. The comparison between three different ρ and ψ functions: quadratic, Huber's, and Hampel's.

Kim & Scott (2008) proposed the robust kernel density estimate, a robust version of the kernel density estimate. They extend the notion of M-estimator previously used in Euclidean space to the Hilbert space $\mathcal{H}$ in order to find a robust sample mean of the $\Phi(\mathbf{X}_i)$'s. For a robust loss function $\rho(x)$ on $x \geq 0$, the robust kernel density estimate is defined as

$$\widehat{f}_{RKDE} = \arg\min_{g\in\mathcal{H}} \sum_{i=1}^{n} \rho\big(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}\big).$$

Well-known examples of robust loss functions are Huber's or Hampel's ρ. Unlike the quadratic loss, these loss functions have the property that $\psi \triangleq \rho'$ is bounded. For Huber's ρ, ψ is given by

$$\psi\,(x) = \begin{cases} x & ,\ 0 \leq x \leq a \\ a & ,\ a < x. \end{cases} \tag{1}$$

and for Hampels' ψ,

$$\psi(x) = \begin{cases} x & ,0 \leq x < a \\ a & ,a \leq x < b \\ a\cdot(c-x)/(c-b) & ,b \leq x < c \\ 0 & ,c \leq x. \end{cases} \tag{2}$$

These functions are plotted in Figure 2.

Kim & Scott (2008) also propose a *kernelized* iteratively re-weighted least squares (IRWLS) algorithm for computing $\widehat{f}_{RKDE}$. Starting with initial $w_i^{(0)} \in \mathbb{R}$, $i = 1,\ldots,n$, the algorithm generates a sequence $\{f^{(k)}\}$ by iterating on the following procedure:

$$f^{(k)} = \sum_{i=1}^{n} w_i^{(k-1)}\Phi(\mathbf{X}_i) = \sum_{i=1}^{n} w_i^{(k-1)}k(\cdot,\mathbf{X}_i),$$

$$w_i^{(k)} = \frac{\varphi(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathcal{H}})}{\sum_{j=1}^{n}\varphi(\|\Phi(\mathbf{X}_j) - f^{(k)}\|_{\mathcal{H}})},$$

where $\varphi(x) = \psi(x)/x$. It was shown that $\|\Phi(\mathbf{X}_j) - f^{(k)}\|_{\mathcal{H}}$ can be computed using the *kernel trick.*

3. Representer Theorem and IRWLS Convergence

For greater generality that will be needed in Section 4, we define M-estimates in $\mathcal{H}$ with respect to a general probability distribution μ. Given μ, we define the kernel density M-estimate $f_\mu \in \mathcal{H}$ as a minimizer of $J_\mu(g)$, where

$$J_\mu(g) = \int \rho\big(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}\big)\,d\mu(\mathbf{x}). \tag{3}$$

If μ is the empirical distribution $F_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{X}_i}$, then

$$J_{F_n}(g) = \frac{1}{n}\sum_{i=1}^{n}\rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}),$$

and thus $f_{F_n} = \widehat{f}_{RKDE}$.

We consider the following assumptions on ρ and ψ:

(A1) ρ is non-decreasing, $\rho(0) = 0$, and $\rho(x)/x \to 0$ as $x \to 0$

(A2) $\psi(x)$ and $\psi(x)/x$ are continuous

(A3) $\psi(x)$ and $\psi(x)/x$ are bounded

which hold for Huber or Hampel's ψ.

3.1. Representer Theorem

In this section, we will describe how $\widehat{f}_{RKDE}$ can be expressed as a weighted combination of the $k_\sigma(\mathbf{x},\mathbf{X}_i)$'s, where the weights offer insight into the robustness of the RKDE. Let $V_\mu : \mathcal{H} \to \mathcal{H}$ be given by

$$V_\mu(g) = \int \frac{\psi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}})}{\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}}\cdot(\Phi(\mathbf{x}) - g)\,d\mu(\mathbf{x})$$

and $\mathcal{D} \subset \mathcal{H}$ be the convex set defined as

$$\mathcal{D} = \left\{ g \,\middle|\, g = \int \Phi(\mathbf{y})d\mu'(\mathbf{y}), \mu' \in \Lambda \right\}$$

where Λ is the set of probability distributions on $\mathbb{R}^d$. (For the integral of $\mathcal{H}$-valued functions, see Berlinet & Thomas-Agnan (2004).) $V_\mu(g)$ is related to the Gateaux differential of $J_\mu(g)$ in that for $h \in \mathcal{H}$,

$$\delta J_\mu(g;h) = -\big\langle V_\mu(g), h\big\rangle$$

where $\delta T(x,h)$ is the Gateaux differential of T at x with incremental h (Luenberger, 1997).

Lemma 1. *Suppose assumptions (A1)-(A3) are satisfied. Then,*

(a) f_μ *satisfies* $V_\mu(f_\mu) = \mathbf{0}$

(b) $f_\mu \in \mathcal{D}$

(c) f_μ *is a density*

Proof sketch. First, (a) directly comes from the fact that a minimizer f_μ of J_μ has to satisfy $\delta J_\mu(f_\mu; h) = \mathbf{0}$ for $\forall h \in \mathcal{H}$. By expressing $V_\mu(f_\mu) = \mathbf{0}$ in terms of f_μ, we obtain $f_\mu(\mathbf{x}) = \int w(\mathbf{y}) k_\sigma(\mathbf{x}, \mathbf{y})\, d\mu(\mathbf{y})$ for some $w \in L_1(\mu)$ such that $w \geq 0$ and $\|w\|_{L_1} = 1$. This establishes (b) and (c). □

From the above lemma with $\mu = F_n$, we have the following representer theorem for $f_{F_n} = \widehat{f}_{RKDE}$, similar to those known for supervised kernel methods.

Theorem 1 (Representer Theorem)**.** *Suppose assumptions (A1)-(A3) are satisfied. Then,*

$$\widehat{f}_{RKDE}(\mathbf{x}) = \sum_{i=1}^{n} w_i k_\sigma(\mathbf{x}, \mathbf{X}_i) \tag{4}$$

where $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 1$. *Furthermore,*

$$w_i \propto \varphi(\|\Phi(\mathbf{X}_i) - \widehat{f}_{RKDE}\|_{\mathcal{H}}).$$

Note that while $\varphi(x)$ is constant for the quadratic loss, the Huber or Hampel's counterparts decrease as x increases. If φ is decreasing, w_i will be small when $\|\Phi(\mathbf{X}_i) - \widehat{f}_{RKDE}\|_{\mathcal{H}}$ is large. Now for any $g \in \mathcal{H}$,

$$\begin{aligned}\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}^2 &= \langle \Phi(\mathbf{X}_i) - g, \Phi(\mathbf{X}_i) - g\rangle \\ &= \|\Phi(\mathbf{X}_i)\|_{\mathcal{H}}^2 - 2\langle \Phi(\mathbf{X}_i), g\rangle + \|g\|_{\mathcal{H}}^2 \\ &= \left(\sqrt{2\pi}\sigma\right)^{-d} - 2g(\mathbf{X}_i) + \|g\|_{\mathcal{H}}^2,\end{aligned}$$

Taking $g = \widehat{f}_{RKDE}$, we see that w_i is small when $\widehat{f}_{RKDE}(\mathbf{X}_i)$ is small. Therefore, the RKDE is robust in the sense that it down-weights outlying points.

Lemma 1. (a) provides a necessary condition for f_μ to be the minimizer of (3). With additional assumptions on ρ and/or μ, this is also sufficient.

Theorem 2. *Suppose assumptions (A1)-(A3) are satisfied. J_μ is strictly convex provided either (1) ρ is strictly convex, or (2) ρ is convex, strictly increasing, and μ is not a discrete measure having only* 1 *or* 2 *atoms. If J_μ is strictly convex, then $V_\mu(g) = \mathbf{0}$ is sufficient for $g = f_\mu$.*

3.2. Convergence of IRWLS Method in Hilbert Space

In general, the equation $V_{F_n}(g) = \mathbf{0}$ does not have a closed form solution for f_{F_n}. The kernelized IRWLS algorithm explained in Section 2 has been proposed to find f_{F_n} in an iterative way. In fact, the kernelized IRWLS can be viewed as a kind of optimization transfer/majorize-minimize (MM) algorithm (Lange & Yang, 2000; Jacobson & Fessler, 2007) with a quadratic surrogate for ρ.

The convergence to some value, not necessarily optimal, of $\{J_{F_n}(f^{(k)})\}_{k=1}^{\infty}$ is proven in Kim & Scott (2008), but the convergence of $\{f^{(k)}\}_{k=1}^{\infty}$ is still in question. The next theorem characterizes the convergence of this sequence.

Theorem 3. *Suppose assumptions (A1)-(A3) are satisfied, and $\varphi(x)$ is nonincreasing. Let*

$$\mathcal{S} = \left\{g \in \mathcal{H} \,\middle|\, V_{F_n}(g) = \mathbf{0}\right\}$$

and $\{f^{(k)}\}_{k=1}^{\infty}$ be the sequence produced by the kernelized IRWLS algorithm. Then, $\mathcal{S} \neq \emptyset$ and

$$\|f^{(k)} - \mathcal{S}\|_{\mathcal{H}} \triangleq \inf_{g \in \mathcal{S}} \|f^{(k)} - g\|_{\mathcal{H}} \to 0$$

as $k \to \infty$.

Proof Sketch. Proof by contradiction. Suppose $\|f^{(k)} - \mathcal{S}\|_{\mathcal{H}} \nrightarrow 0$. Then, there exist $\epsilon > 0$ such that we can construct a subsequence $\{f^{(k_l)}\}_{l=1}^{\infty}$ with $\|f^{(k_l)} - \mathcal{S}\|_{\mathcal{H}} \geq \epsilon$ for $l = 1, 2, \ldots$. Since $\{f^{(k_l)}\}_{k=1}^{\infty}$ lies in a compact set, it has a convergent subsequence with limit $f^{\dagger} \in \mathcal{S}$. Thus, we can choose j such that $\|f^{(k_j)} - f^{\dagger}\|_{\mathcal{H}} \leq \epsilon/2$. This is a contradiction because

$$\epsilon \leq \inf_{g \in \mathcal{S}} \|f^{(k_j)} - g\|_{\mathcal{H}} \leq \|f^{(k_j)} - f^{\dagger}\|_{\mathcal{H}} \leq \epsilon/2.$$

□

In words, if the number of iterations grows, $f^{(k)}$ becomes arbitrarily close to the set of the stationary points of J_{F_n}, points $g \in \mathcal{H}$ satisfying $\delta J_{F_n}(g; h) = 0 \quad \forall h \in \mathcal{H}$.

Corollary 1. *Suppose that the assumptions in Theorem 3 hold. In addition, assume that ρ is convex and strictly increasing, and $\{\mathbf{X}_i\}_{i=1}^{n}$ contains at least three distince $\mathbf{X}_i$'s. Then, $\{f^{(k)}\}_{k=1}^{\infty}$ converges to the unique global minimizer of (3).*

4. Influence Function for Robust KDE

To quantify the robustness of the RKDE, we introduce the influence function. First, we recall the traditional influence function from robust statistics. Let $T(\mu)$ be an estimator based on μ. As a measure of robustness of T, the influence function was proposed by Hampel (1974). The influence function (IF) for T at F is defined as

$$IF(x'; T, F) = \lim_{s \to 0} \frac{T((1-s)F + s\delta_{x'}) - T(F)}{s}.$$

Basically, $IF(x';T,F)$ represents how $T(F)$ changes when the distribution F is contaminated with infinitesimal probability mass at x'. One robustness measure of T is whether the corresponding IF is bounded or not.

For example, the maximum likelihood estimator for the unknown mean θ of Gaussian distribution is the sample mean $T(\mu)$,

$$T(\mu) = E_\mu[X] = \int x \, d\mu(x). \tag{5}$$

The influence function for $T(F)$ in (5) is

$$\begin{aligned} IF(x';T,F) &= \lim_{s\to 0} \frac{T((1-s)F + s\delta_{x'}) - T(F)}{s} \\ &= x' - E_F[X]. \end{aligned}$$

Since $|IF(x';T,F)|$ increases without bound as x' goes to $\pm\infty$, the estimator is considered as not robust.

Now, we define a similar concept for a function estimate. Since the estimate is a function, not a scalar, we should be able to express the change of the function value at every $\mathbf{x}$.

Definition 1 (IF for function estimate)**.** *Let $T(\mathbf{x};\mu)$ be a function estimate based on μ, evaluated at $\mathbf{x}$. Then, we define the influence function for $T(\mathbf{x};F)$ as*

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

where $F_s = (1-s)F + s\delta_{\mathbf{x}'}$.

$IF(\mathbf{x},\mathbf{x}';T,F)$ represents the change of the estimated function T at $\mathbf{x}$ when we add infinitesimal probability mass at $\mathbf{x}'$ to F.

For example, the standard KDE is $T(\mathbf{x};F) = \widehat{f}_{KDE}(\mathbf{x};F) = \int k_\sigma(\mathbf{x},\mathbf{y})dF(\mathbf{y}) = E_F[k_\sigma(\mathbf{x},\mathbf{X})]$ where $\mathbf{X} \sim F$. In this case, the influence function is

$$\begin{aligned} & IF(\mathbf{x},\mathbf{x}';\widehat{f}_{KDE},F) \\ &= \lim_{s\to 0} \frac{\widehat{f}_{KDE}(\mathbf{x};F_s) - \widehat{f}_{KDE}(\mathbf{x};F)}{s} \\ &= \lim_{s\to 0} \frac{E_{F_s}[k_\sigma(\mathbf{x},\mathbf{X})] - E_F[k_\sigma(\mathbf{x},\mathbf{X})]}{s} \\ &= \lim_{s\to 0} \frac{-sE_F[k_\sigma(\mathbf{x},\mathbf{X})] + sE_{\delta_{\mathbf{x}'}}[k_\sigma(\mathbf{x},\mathbf{X})]}{s} \\ &= -E_F[k_\sigma(\mathbf{x},\mathbf{X})] + E_{\delta_{\mathbf{x}'}}[k_\sigma(\mathbf{x},\mathbf{X})] \\ &= -E_F[k_\sigma(\mathbf{x},\mathbf{X})] + k_\sigma(\mathbf{x},\mathbf{x}') \end{aligned} \tag{6}$$

With the empirical distribution F_n,

$$\begin{aligned} & IF(\mathbf{x},\mathbf{x}';\widehat{f}_{KDE},F_n) \\ &= -\frac{1}{n}\sum_{i=1}^n k_\sigma(\mathbf{x},\mathbf{x}_i) + k_\sigma(\mathbf{x},\mathbf{x}'). \end{aligned} \tag{7}$$

For the robust KDE, $T(\mathbf{x},F) = \widehat{f}_{RKDE}(\mathbf{x};F) = \langle \Phi(\mathbf{x}), f_F \rangle$, we have the following characterization of the influence function. Let $q(x) = x\psi'(x) - \psi(x)$.

Theorem 4. *Suppose assumptions (A1)-(A3) are satisfied. In addition, assume that $\varphi(x)$ is nonincreasing and Lipschitz continuous, and $f_{F_s} \to f_F$ as $s \to 0$. If $\dot{f}_F \triangleq \lim_{s\to 0} \frac{f_{F_s} - f_F}{s}$ exists, then*

$$IF(\mathbf{x},\mathbf{x}';\widehat{f}_{RKDE},F) = \langle \Phi(\mathbf{x}), \dot{f}_F \rangle$$

where $\dot{f}_F \in \mathcal{H}$ satisfies

$$\begin{aligned} & \left(\int \varphi(\|\Phi(\mathbf{x}) - f_F\|_{\mathcal{H}}) dF \right) \cdot \dot{f}_F \\ & + \quad \int \Bigg(\frac{\langle \dot{f}_F, \Phi(\mathbf{x}) - f_F \rangle}{\|\Phi(\mathbf{x}) - f_F\|_{\mathcal{H}}^3} \\ & \qquad \cdot q(\|\Phi(\mathbf{x}) - f_F\|_{\mathcal{H}}) \cdot \big(\Phi(\mathbf{x}) - f_F \big) \Bigg) dF(\mathbf{x}) \\ & = \quad (\Phi(\mathbf{x}') - f_F) \cdot \varphi(\|\Phi(\mathbf{x}') - f_F\|_{\mathcal{H}}). \end{aligned} \tag{8}$$

Unfortunately, for Huber or Hampel's ρ function, there is no closed form solution for $\dot{f}_F$ of (8). However, if we work with F_n instead of F, we can find $\dot{f}_{F_n}$ explicitly. Let

$$\begin{aligned} \mathbf{1} &= [1,\ldots,1]^T, \\ \mathbf{k}' &= [k_\sigma(\mathbf{x}',\mathbf{X}_1),\ldots,k_\sigma(\mathbf{x}',\mathbf{X}_n)]^T \end{aligned}$$

and I_n be a $n \times n$ identity matrix, $K := (k_\sigma(\mathbf{X}_i,\mathbf{X}_j))_{i=1,j=1}^n$ be the kernel matrix, Q be a diagonal matrix with $Q_{ii} = q(\|\Phi(\mathbf{X}_i) - f_{F_n}\|_{\mathcal{H}}) / \|\Phi(\mathbf{X}_i) - f_{F_n}\|_{\mathcal{H}}^3$, and

$$\begin{aligned} c &= \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - f_{F_n}\|_{\mathcal{H}}), \\ \mathbf{w} &= [w_1,\ldots,w_n]^T, \end{aligned}$$

where $\mathbf{w}$ gives the RKDE weights as in Theorem 1.

Theorem 5. *Suppose assumptions (A1)-(A3) are satisfied. In addition, assume that $\varphi(x)$ is nonincreasing and Lipschitz continuous, $f_{F_{n,s}} \to f_{F_n}$ as $s \to 0$, and $\{\mathbf{X}_i\}$ are distinct. Then,*

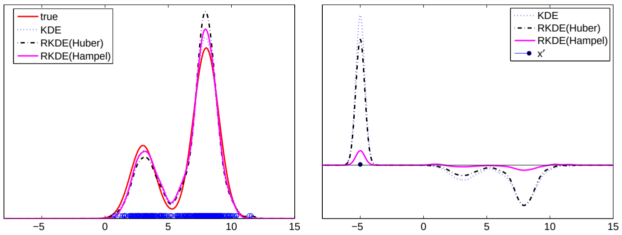
$$IF(\mathbf{x},\mathbf{x}';\widehat{f}_{RKDE},F_n) = \sum_{i=1}^n \alpha_i k_\sigma(\mathbf{x},\mathbf{X}_i) + \alpha' k_\sigma(\mathbf{x},\mathbf{x}')$$

where

$$\alpha' = n \cdot \varphi(\|\Phi(\mathbf{x}') - f_{F_n}\|_{\mathcal{H}})/c$$

and $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_n]^T$ is the solution of the following system of linear equations:

$$\begin{aligned} & \left\{ cI_n + (I_n - \mathbf{1}\cdot\mathbf{w}^T)^T Q (I_n - \mathbf{1}\cdot\mathbf{w}^T) K \right\} \boldsymbol{\alpha} \\ & = \quad -n\varphi(\|\Phi(\mathbf{x}') - f_{F_n}\|_{\mathcal{H}})\mathbf{w} \\ & \qquad -\alpha'(I_n - \mathbf{1}\cdot\mathbf{w}^T)^T Q \cdot (I_n - \mathbf{1}\cdot\mathbf{w}^T) \cdot \mathbf{k}'. \end{aligned}$$



(a) (b)

Figure 3. (a) true density and density estimates. (b) IF as a function of $\mathbf{x}$ when $\mathbf{x}' = -5$

The condition $f_{F_{n,s}} \to f_{F_n}$ is satisfied, for example, when J_{F_n} is strictly convex (see Theorem 2).

As mentioned above, in classical robust statistics, the robustness of an estimator can be determined by the boundedness of the corresponding influence function. However, the influence functions for density estimators are bounded even if $\|\mathbf{x}'\| \to \infty$. Therefore, when we compare the robustness of density estimates, we compare how close the influence functions are to the zero function.

Simulation results are shown in Fig 3. As we can see in (b), for a point $\mathbf{x}'$ in the tails of F, the influence functions for robust KDEs are overall smaller than that of standard KDE in absolute value (especially with Hampel's loss).

We give a possible explanation for this observation. Assume that the parameter a in (1) or (2) is such that for all $\mathbf{X}_i$, $\|\Phi(\mathbf{X}_i) - f_{F_n}\|_{\mathcal{H}} \leq a$. Equivalently, $f_{F_n}(\mathbf{X}_i) \geq \lambda$ for a corresponding λ. This results in $\mathbf{w} = \frac{1}{n}\mathbf{1}$, $c = n$, and $Q = \mathbf{0}$. In this case, the influence function is

$$
\begin{aligned}
& IF(\mathbf{x}, \mathbf{x}'; \widehat{f}_{RKDE}, F_n) \\
&= \kappa \cdot \left(-\frac{1}{n} \sum_{i=1}^{n} k_\sigma(\mathbf{x}, \mathbf{x}_i) + k_\sigma(\mathbf{x}, \mathbf{x}') \right)
\end{aligned}
$$

where $\kappa = \varphi(\|f_{F_n} - \Phi(\mathbf{x}')\|_{\mathcal{H}})$. If we compare the above result with (7), the robust KDE will less affected than the KDE by an outlier $\mathbf{x}'$ with $f_{F_n}(\mathbf{x}') < \lambda$, in which case $\kappa < 1$.

5. Experiments

We demonstrate experimental results on synthetic and real data sets. In each experiment, the parameters a, b, and c in (1) and (2) are set as follows. First, we compute $f^{(1)}$, the RKDE based on $\rho = |\cdot|$, and $d_i = \|\Phi(\mathbf{X}_i) - f^{(1)}\|_{\mathcal{H}}$. Then, a is set to be the median of $\{d_i\}$, b the 95th percentile of $\{d_i\}$, and $c = \max\{d_i\}$. IRWLS is always initialized with uniform weights.

5.1. Synthetic data

First, we demonstrate the robustness of the RKDE on 1, 2, and 5 dimensional synthetic data. For each dimension, the true distribution is a mixture of two normal distributions and the outlying distribution is a single normal distribution. These are summarized in Table 1. The bandwidth σ of the Gaussian kernel is chosen via least square cross validation (LSCV) (Turlach, 1993).

As a quantitative measure of how close the estimated density is to the true density, we compute $\|f - \widehat{f}\|_{L_2}$. For each $\epsilon = 0$, 0.05, 0.10, 0.15, and 0.20, we generate a random sample of size n from the true distribution and add m outliers where $m = \epsilon \cdot n$ (n is given in Table 1). Figure 4 (a) - (c) show the average L_2 error over 100 simulations as a function of ϵ. All three density estimates provide similar L_2 errors when there are no outliers, i.e., $\epsilon = 0$. However, in the presence of outliers, $\epsilon > 0$, we see that RKDEs (especially with Hampel's loss) have smaller L_2 errors than KDEs.

As another measure of robustness, we compare the influence functions for the density estimates given in Theorem 5. We examine $\alpha(\mathbf{x}') = IF(\mathbf{x}', \mathbf{x}'; T, F_n)$ and

$$
\beta(\mathbf{x}') = \left(\int \big(IF(\mathbf{x}, \mathbf{x}'; T, F_n) \big)^2 d\mathbf{x} \right)^{1/2}.
$$

In words, $\alpha(\mathbf{x}')$ is the change of the density estimate value at an added point $\mathbf{x}'$ and $\beta(\mathbf{x}')$ is an overall impact of $\mathbf{x}'$ on the density estimate over $\mathbb{R}^d$. We generate 1000 random samples from the outlying distribution, each of which serves as an $\mathbf{x}'$. This gives us 1000 $\alpha(\mathbf{x}')$'s and $\beta(\mathbf{x}')$'s. The boxplot of these are shown in Figure 4 (d) - (i), from which we can see that RKDEs are less affected by outliers $\mathbf{x}'$ than KDEs.

5.2. Application to Anomaly Detection

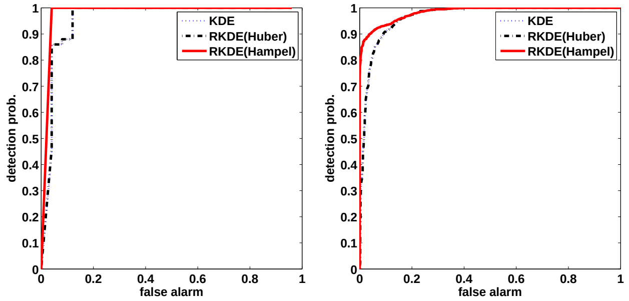
We apply RKDEs in anomaly detection problems with benchmark data sets. Each density estimate serves as an anomaly detector by thresholding the value of the density estimate at a test point. Robustness can be checked by comparing a performance measure, e.g., AUC, of the anomaly detectors, where the density estimates are based on contaminated training data.

We conduct experiments on 15 benchmark data sets (Banana, B. Cancer, Diabetes, F. Solar, German, Heart, Image, Ringnorm, Splice, Thyroid, Twonorm, Waveform, Pima Indian, Iris, MNIST) [1], which were originally used in the task of classification. For each

[1] http://www.fml.tuebingen.mpg.de/Members/ for the first 12 data sets and UCI machine learning repository for the last 3 data sets.

Table 1. Summary of distributions. $\mathcal{N}_d(\mathbf{x};\mathbf{m};C)$ represents a d-dimensional normal distribution with mean $\mathbf{m} \in \mathbb{R}^d$ and $d \times d$ covariance matrix C. n is the number of samples from true distribution.

dimension	# of samples	true distribution $(1-\eta)\cdot\mathcal{N}_d(\mathbf{x};\mathbf{m}_1,\lambda_1 I)+\eta\cdot\mathcal{N}_d(\mathbf{x};\mathbf{m}_2,\lambda_2 I)$	outlying distribution $\mathcal{N}_d(\mathbf{x};\mathbf{m}_0;\lambda_0 I)$
1	n=200	$\eta = 0.4$ $\mathbf{m}_1 = -3,\ \lambda_1 = 1.5$ $\mathbf{m}_2 = +3,\ \lambda_2 = 1.5$	$\mathbf{m}_0 = 10,\ \lambda_0 = 2.25$
2	n=400	$\eta = 0.5$ $\mathbf{m}_1 = [-3,0]^T,\ \lambda_1 = 1$ $\mathbf{m}_2 = [+3,0]^T,\ \lambda_2 = 1$	$\mathbf{m}_0 = [0,3]^T,\ \lambda_0 = 1$
5	n=1000	$\eta = 0.6$ $\mathbf{m}_1 = [-1,1,-1,1,-1]^T,\ \lambda_1 = 0.5$ $\mathbf{m}_2 = [0,\ 0,\ 0,\ 0,\ 0]^T,\ \lambda_2 = 0.5$	$\mathbf{m}_0 = [3,-3,3,-3,3]^T,\ \lambda_0 = 1$



(a) Iris, $\epsilon = 0.1$ (b) MNIST, $\epsilon = 0.2$

Figure 5. Examples of ROC.

Table 2. The comparison of average ranks of the three density estimators, by the Friedman test. The critical difference of the post-hoc Nemenyi test is 0.86 at a significance level of 0.05.

ϵ	KDE	RKDE (Huber)	RKDE (Hampel)	p-value
0.00	2.17	1.90	1.93	0.71
0.05	2.57	2.23	1.20	0.00
0.10	2.67	2.20	1.13	0.00
0.15	2.67	2.20	1.13	0.00
0.20	2.67	2.20	1.13	0.00

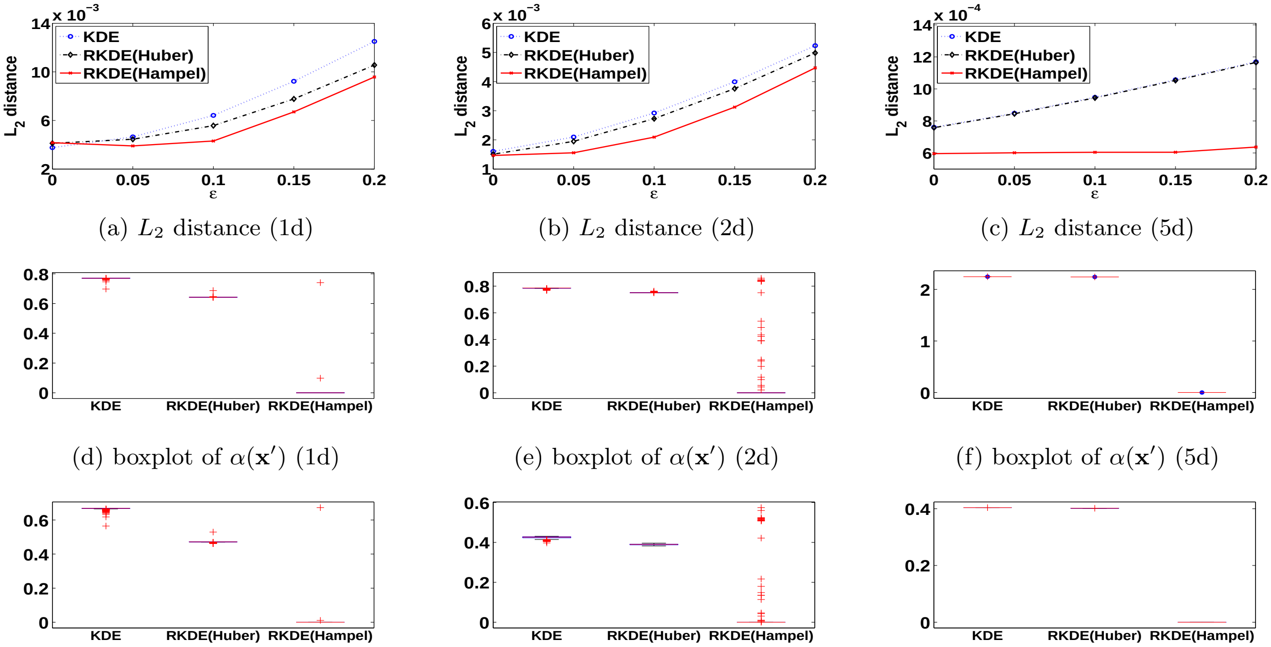
data set with two classes, we take one class as the nominal data and the other class as anomalies. For Iris, there are 3 classes and we take one class as nominal data and the other two as anomalies. For MINST, we choose to use 0 digit as nominal and 1 digit as anomalies. For MNIST, the original dimension 784 is reduced to 8 via kernel PCA using Gaussian kernel with bandwidth 30.

For each data set, the training sample consists of n nominal data points and m outliers, and as mentioned before $m = \epsilon \cdot n$ for $\epsilon = 0,\ 0.05,\ 0.10,\ 0.15$, and 0.20. The bandwidth of the Gaussian kernel is set as the median distance to the nearest neighbor. KDEs and RKDEs are estimated based on these contaminated training data, and ROCs are generated by varying the threshold. Examples of the ROCs are shown in Figure 5. While the ROCs from the RKDE with Huber's loss is fairly close to that of the KDE, the RKDE with Hampel's loss provides better detection probabilities, especially at low false alarm rates. This results in higher AUC.

To compare the density estimators across multiple data sets, we adopt the methodology of Dĕmsar (2006). For each data set and each ϵ, the density estimates are ranked 1 (best) through 3 (worst) based on AUC. For each ϵ, we use the Friedman test in order to determine whether there was a significant difference in the average ranks of the three density estimators across the data sets. The average ranks and p-values are shown in Tables 2. The results indicate that there is a significant difference among the estimators with the exception of $\epsilon = 0$. For three methods on 15 data sets, with a significance level of 0.05, the critical difference (CD) for the Nemenyi test is 0.86. If the average ranks differs by more than the CD, the methods are deemed to be significantly different. This indicates that RKDEs with Hampel's loss are significantly better than KDEs and RKDEs with Huber's loss where $\epsilon > 0$.

6. Conclusions

In this paper, we have investigated the convergence and robustness of the kernel density M-estimators. We derive an influence function for the estimator and give an explanation of why RDKEs are more robust than KDEs through the influence function. The argument



(a) L_2 distance (1d)

(b) L_2 distance (2d)

(c) L_2 distance (5d)

(d) boxplot of $\alpha(\mathbf{x}')$ (1d)

(e) boxplot of $\alpha(\mathbf{x}')$ (2d)

(f) boxplot of $\alpha(\mathbf{x}')$ (5d)

(g) boxplot of $\beta(\mathbf{x}')$ (1d)

(h) boxplot of $\beta(\mathbf{x}')$ (2d)

(i) boxplot of $\beta(\mathbf{x}')$ (5d)

Figure 4. Experimental results on 1, 2, and 5 dimensional synthetic data.

is also supported by experimental results on several data sets.

Acknowledgements

This work was supported in part by NSF Award No. 0830490.

References

Berlinet, A. and Thomas-Agnan, C. Reproducing kernel Hilbert spaces in probability and statistics. 2004.

Brabanter, K. D., Pelckmans, K., and et al. Robustness of kernel based regression: A comparison of iterative weighting schemes. *Proceedings of the 19th International Conference on Artificial Neural Networks (ICANN)*, pp. 100–110, 2009.

Devroye, L. Consistent deconvolution in density estimation. *The Canadian Journal of Statistics*, (2): 235–239, 1989.

Děmsar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

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

Huber, P. J. Robust estimation of a location parameter. *Ann. Math. Statist*, 35:45, 1964.

Jacobson, M. W. and Fessler, J. A. An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms. *IEEE Transactions on Image Processing*, 16(10):2411–2422, October 2007.

Kim, J. and Scott, C. Robust kernel density estimation. *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2008.

Lange, K. and Yang, D. R. Hunterand I. Optimization transfer using surrogate objective functions. *J. Computational and Graphical Stat.*, 9(1):1–20, March 2000.

Luenberger, David G. *Optimization by Vector Space Methods*. Wiley-Interscience, New York, 1997.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer, 2008.

Turlach, B.A. Bandwidth selection in kernel density estimation: A review. *Technical Report 9317, C.O.R.E. and Institut de Statistique, Université Catholique de Louvain*, 1993.

Xu, L., Crammer, K., and Schuurmans, D. Robust support vector machine training via convex outlier ablation. *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.