

Robust Kernel Density Estimation

JooSeuk Kim

Clayton D. Scott*

Electrical Engineering and Computer Science

University of Michigan

Ann Arbor, MI 48109-2122 USA

STANNUM@UMICH.EDU

CLAYSCOT@UMICH.EDU

Editor: Kenji Fukumizu

Abstract

We propose 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. We interpret the KDE based on a positive semi-definite kernel as a sample mean in the associated reproducing kernel Hilbert space. Since the sample mean is sensitive to outliers, we estimate it robustly via M -estimation, yielding a robust kernel density estimator (RKDE).

An RKDE can be computed efficiently via a kernelized iteratively re-weighted least squares (IRWLS) algorithm. Necessary and sufficient conditions are given for kernelized IRWLS to converge to the global minimizer of the M -estimator objective function. The robustness of the RKDE is demonstrated with a representer theorem, the influence function, and experimental results for density estimation and anomaly detection.

Keywords: outlier, reproducing kernel Hilbert space, kernel trick, influence function, M -estimation

1. Introduction

The kernel density estimator (KDE) is a well-known nonparametric estimator of univariate or multivariate densities, and numerous articles have been written on its properties, applications, and extensions (Silverman, 1986; Scott, 1992). However, relatively little work has been done to understand or improve the KDE in situations where the training sample is contaminated. This paper addresses a method of nonparametric density estimation that generalizes the KDE, and exhibits robustness to contamination of the training sample.¹

Consider training data following a contamination model

$$\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{iid}{\sim} (1-p)f_0 + pf_1,$$

where f_0 is the “nominal” density to be estimated, f_1 is the density of the contaminating distribution, and $p < \frac{1}{2}$ is the proportion of contamination. Labels are not available, so that the problem is unsupervised. The objective is to estimate f_0 while making no parametric assumptions about the nominal or contaminating distributions.

*, Also in the Department of Statistics.

1. Shorter versions of this work previously appeared at the International Conference on Acoustics, Speech, and Signal Processing (Kim and Scott, 2008) and the International Conference on Machine Learning (Kim and Scott, 2011).

Consider KDE as a mean & instead of using KDE, use M-estimation to find the mean

Clearly f_0 cannot be recovered if there are *no* assumptions on f_0, f_1 and p . Instead, we will focus on a **set of nonparametric conditions that are reasonable in many practical applications**. In particular, we will assume that, relative to the nominal data, the contaminated data are

- Assumptions about f_0 & f_1**
- (a) *outlying*: the densities f_0 and f_1 have relatively little overlap
 - (b) *diffuse*: f_1 is not too spatially concentrated relative to f_0
 - (c) *not abundant*: a minority of the data come from f_1

Although we will not be stating these conditions more precisely, they capture the intuition behind the quantitative results presented below.

As a motivating application, consider anomaly detection in a computer network. Imagine that several multi-dimensional measurements $\mathbf{X}_1, \dots, \mathbf{X}_n$ are collected. For example, each \mathbf{X}_i may record the volume of traffic along certain links in the network, at a certain instant in time (Chhabra et al., 2008). If each measurement is collected when the network is in a nominal state, these data could be used to construct an anomaly detector by first estimating the density f_0 of nominal measurements, and then thresholding that estimate at some level to obtain decision regions. **Unfortunately, it is often difficult to know that the data are free of anomalies, because assigning labels (nominal vs. anomalous) can be a tedious, labor intensive task. Hence, it is necessary to estimate the nominal density (or a level set thereof) from contaminated data. Furthermore, the distributions of both nominal and anomalous measurements are potentially complex, and it is therefore desirable to avoid parametric models.**

The proposed method achieves robustness by combining a traditional kernel density estimator with ideas from M -estimation (Huber, 1964; Hampel, 1974). The KDE based on a translation invariant, positive semi-definite (PSD) kernel is interpreted as a sample mean in the reproducing kernel Hilbert space (RKHS) associated with the kernel. Since the sample mean is sensitive to outliers, we estimate it robustly via M -estimation, yielding a robust kernel density estimator (RKDE). We describe a kernelized iteratively re-weighted least squares (KIRWLS) algorithm to efficiently compute the RKDE, and provide necessary and sufficient conditions for the convergence of KIRWLS to the RKDE.

We also offer three arguments to support the claim that the RKDE robustly estimates the nominal density and its level sets. **First, we characterize the RKDE by a representer theorem. This theorem shows that the RKDE is a weighted KDE, and the weights are smaller for more outlying data points. Second, we study the influence function of the RKDE, and show through an exact formula and numerical results that the RKDE is less sensitive to contamination by outliers than the KDE. Third, we conduct experiments on several benchmark data sets that demonstrate the improved performance of the RKDE, relative to competing methods, at both density estimation and anomaly detection.**

One motivation for this work is that the traditional kernel density estimator is well-known to be sensitive to outliers. Even without contamination, the standard KDE tends to overestimate the density in regions where the true density is low. This has motivated several authors to consider variable kernel density estimators (VKDEs), which employ a data-dependent bandwidth at each data point (Breiman et al., 1977; Abramson, 1982; Terrell and Scott, 1992). This bandwidth is adapted to be larger where the data are less dense, with the aim of decreasing the aforementioned bias. Such methods have been applied in outlier detection and computer vision applications (Comaniciu et al., 2001; Latecki et al., 2007), and are one possible approach to robust nonparametric density estimation. We compare against these methods in our experimental study.

Density estimation with positive semi-definite kernels has been studied by several authors. Vapnik and Mukherjee (2000) optimize a criterion based on the empirical cumulative distribution function over the class of weighted KDEs based on a PSD kernel. Shawe-Taylor and Dolia (2007) provide a refined theoretical treatment of this approach. Song et al. (2008) adopt a different criterion based on Hilbert space embeddings of probability distributions. Our approach is somewhat similar in that we attempt to match the mean of the empirical distribution in the RKHS, but our criterion is different. These methods were also not designed with contaminated data in mind.

We show that the standard kernel density estimator can be viewed as the solution to a certain least squares problem in the RKHS. The use of quadratic criteria in density estimation has also been previously developed. The aforementioned work of Song et al. optimizes the norm-squared in Hilbert space, whereas Kim (1995), Girolami and He (2003), Kim and Scott (2010) and Mahapatruni and Gray (2011) adopt the integrated squared error. Once again, these methods are not designed for contaminated data.

Previous work combining robust estimation and kernel methods has focused primarily on supervised learning problems. M -estimation applied to kernel regression has been studied by various authors (Christmann and Steinwart, 2007; Debruyne et al., 2008a,b; Zhu et al., 2008; Wibowo, 2009; Brabanter et al., 2009). Robust surrogate losses for kernel-based classifiers have also been studied (Xu et al., 2006). In unsupervised learning, a robust way of doing kernel principal component analysis, called spherical KPCA, has been proposed, which applies PCA to feature vectors projected onto a unit sphere around the spatial median in a kernel feature space (Debruyne et al., 2010). The kernelized spatial depth was also proposed to estimate depth contours nonparametrically (Chen et al., 2009). To our knowledge, the RKDE is the first application of M -estimation ideas in kernel density estimation.

In Section 2 we propose robust kernel density estimation. In Section 3 we present a representer theorem for the RKDE. In Section 4 we describe the KIRWLS algorithm and its convergence. The influence function is developed in Section 5, Section 6 describes a straightforward extension to non-reproducing kernels, and experimental results are reported in Section 7. Conclusions are offered in Section 8. Section 9 contains proofs of theorems. Matlab code implementing our algorithm is available at www.eecs.umich.edu/~cscott.

2. Robust Kernel Density 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

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

where k_{σ} is a kernel function with bandwidth σ . To ensure that $\hat{f}_{KDE}(\mathbf{x})$ is a density, we assume the kernel function satisfies $k_{\sigma}(\cdot, \cdot) \geq 0$ and $\int k_{\sigma}(\mathbf{x}, \cdot) d\mathbf{x} = 1$. We will also assume that $k_{\sigma}(\mathbf{x}, \mathbf{x}')$ is translation invariant, in that $k_{\sigma}(\mathbf{x} - \mathbf{z}, \mathbf{x}' - \mathbf{z}) = k_{\sigma}(\mathbf{x}, \mathbf{x}')$ for all \mathbf{x}, \mathbf{x}' , and \mathbf{z} .

In addition, we require that k_{σ} be positive semi-definite, which means that the matrix $(k_{\sigma}(x_i, x_j))_{1 \leq i, j \leq m}$ is positive semi-definite for all positive integers m and all $x_1, \dots, x_m \in \mathbb{R}^d$. Well-known examples of kernels satisfying all of the above properties are the Gaussian kernel

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

the multivariate Student kernel

$$k_{\sigma}(\mathbf{x}, \mathbf{x}') = \left(\frac{1}{\sqrt{\pi}\sigma} \right)^d \cdot \frac{\Gamma((v+d)/2)}{\Gamma(v/2)} \cdot \left(1 + \frac{1}{v} \cdot \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2} \right)^{-\frac{v+d}{2}},$$

and the Laplacian kernel

$$k_{\sigma}(\mathbf{x}, \mathbf{x}') = \frac{c_d}{\sigma^d} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{\sigma}\right)$$

where c_d is a constant depending on the dimension d that ensures $\int k_{\sigma}(\mathbf{x}, \cdot) d\mathbf{x} = 1$ (Scovel et al., 2010).

Every PSD kernel k_{σ} is associated with a unique Hilbert space of functions called its reproducing kernel Hilbert space (RKHS) which we will denote \mathcal{H} , and k_{σ} is called the reproducing kernel of \mathcal{H} . For every \mathbf{x} , $\Phi(\mathbf{x}) \triangleq k_{\sigma}(\cdot, \mathbf{x})$ is an element of \mathcal{H} , and therefore so too is the KDE. See Steinwart and Christmann (2008) for a thorough treatment of PSD kernels and RKHSs. For our purposes, the critical property of \mathcal{H} is the so-called *reproducing property*. It states that for all $g \in \mathcal{H}$ and all $\mathbf{x} \in \mathbb{R}^d$, $g(\mathbf{x}) = \langle \Phi(\mathbf{x}), g \rangle_{\mathcal{H}}$. As a special case, taking $g = k_{\sigma}(\cdot, \mathbf{x}')$, we obtain

$$k_{\sigma}(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}$$

for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. We also note that, by translation invariance, the functions $\Phi(\mathbf{x})$ have constant norm in \mathcal{H} because

$$\|\Phi(\mathbf{x})\|_{\mathcal{H}}^2 = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle_{\mathcal{H}} = k_{\sigma}(\mathbf{x}, \mathbf{x}) = k_{\sigma}(\mathbf{0}, \mathbf{0}).$$

We will denote $\tau = \|\Phi(\mathbf{x})\|_{\mathcal{H}}$.

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

$$\begin{aligned} \hat{f}_{KDE}(\cdot) &= \frac{1}{n} \sum_{i=1}^n k_{\sigma}(\cdot, \mathbf{X}_i) \\ &= \frac{1}{n} \sum_{i=1}^n \Phi(\mathbf{X}_i), \end{aligned}$$

the sample mean of the $\Phi(\mathbf{X}_i)$'s in \mathcal{H} . Equivalently, $\hat{f}_{KDE} \in \mathcal{H}$ is the solution of

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

Being the solution of a least squares problem, the KDE is sensitive to the presence of outliers among the $\Phi(\mathbf{X}_i)$'s. To reduce the effect of outliers, we propose to use M -estimation (Huber, 1964) 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

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

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. Huber's ρ and ψ are given by

$$\rho(x) = \begin{cases} x^2/2 & , 0 \leq x \leq a \\ ax - a^2/2 & , a < x \end{cases}$$

$$\psi(x) = \begin{cases} x & , 0 \leq x \leq a \\ a & , a < x, \end{cases} \quad (3)$$

and Hampel's ρ and ψ are

$$\rho(x) = \begin{cases} x^2/2 & , 0 \leq x < a \\ ax - a^2/2 & , a \leq x < b \\ a(x-c)^2/2(b-c) + a(b+c-a)/2 & , b \leq x < c \\ a(b+c-a)/2 & , c \leq x \end{cases}$$

$$\psi(x) = \begin{cases} x & , 0 \leq x < a \\ a & , a \leq x < b \\ a \cdot (c-x)/(c-b) & , b \leq x < c \\ 0 & , c \leq x. \end{cases} \quad (4)$$

The functions $\rho(x)$, $\psi(x)$, and $\psi(x)/x$ are plotted in Figure 1, for the quadratic, Huber, and Hampel losses. Note that while $\psi(x)/x$ is constant for the quadratic loss, for Huber's or Hampel's loss, this function is decreasing in x . This is a desirable property for a robust loss function, which will be explained later in detail. While our examples and experiments employ Huber's and Hampel's losses, many other losses can be employed.

We will argue below that \hat{f}_{RKDE} is a valid density, having the form $\sum_{i=1}^n w_i k_{\sigma}(\cdot, \mathbf{X}_i)$ with weights w_i that are nonnegative and sum to one. To illustrate the estimator, Figure 2 (a) shows a contour plot of a Gaussian mixture distribution on \mathbb{R}^2 . Figure 2 (b) depicts a contour plot of a KDE based on a training sample of size 200 from the Gaussian mixture. As we can see in Figure 2 (c) and (d), when 20 contaminating data points are added, the KDE is significantly altered in low density regions, while the RKDE is much less affected.

We remark that the definition of the RKDE does not require that k_{σ} be a reproducing kernel, only that the functions $\Phi(\mathbf{x}) = k_{\sigma}(\cdot, \mathbf{x})$ belong to a Hilbert space. Indeed, almost all of the results presented below hold in this more general setting. However, we restrict our attention to reproducing kernels for two reasons. First, with reproducing kernels, inner products in \mathcal{H} can be easily computed via the kernel, leading to efficient implementation. Second, the reproducing property allows us to interpret the representer theorem and influence function to infer robustness of the RKDE. With non-reproducing kernels, these interpretations are less clear. The extension to non-RKHS Hilbert spaces is discussed in Section 6, with specific examples.

Throughout this paper, we define $\phi(x) \triangleq \psi(x)/x$ and consider the following assumptions on ρ , ψ , and ϕ :

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

(A2) $\phi(0) \triangleq \lim_{x \rightarrow 0} \frac{\psi(x)}{x}$ exists and is finite

(A3) ψ and ϕ are continuous

(A4) ψ and ϕ are bounded

(A5) ϕ is Lipschitz continuous

which hold for Huber's and Hampel's losses, as well as several others.

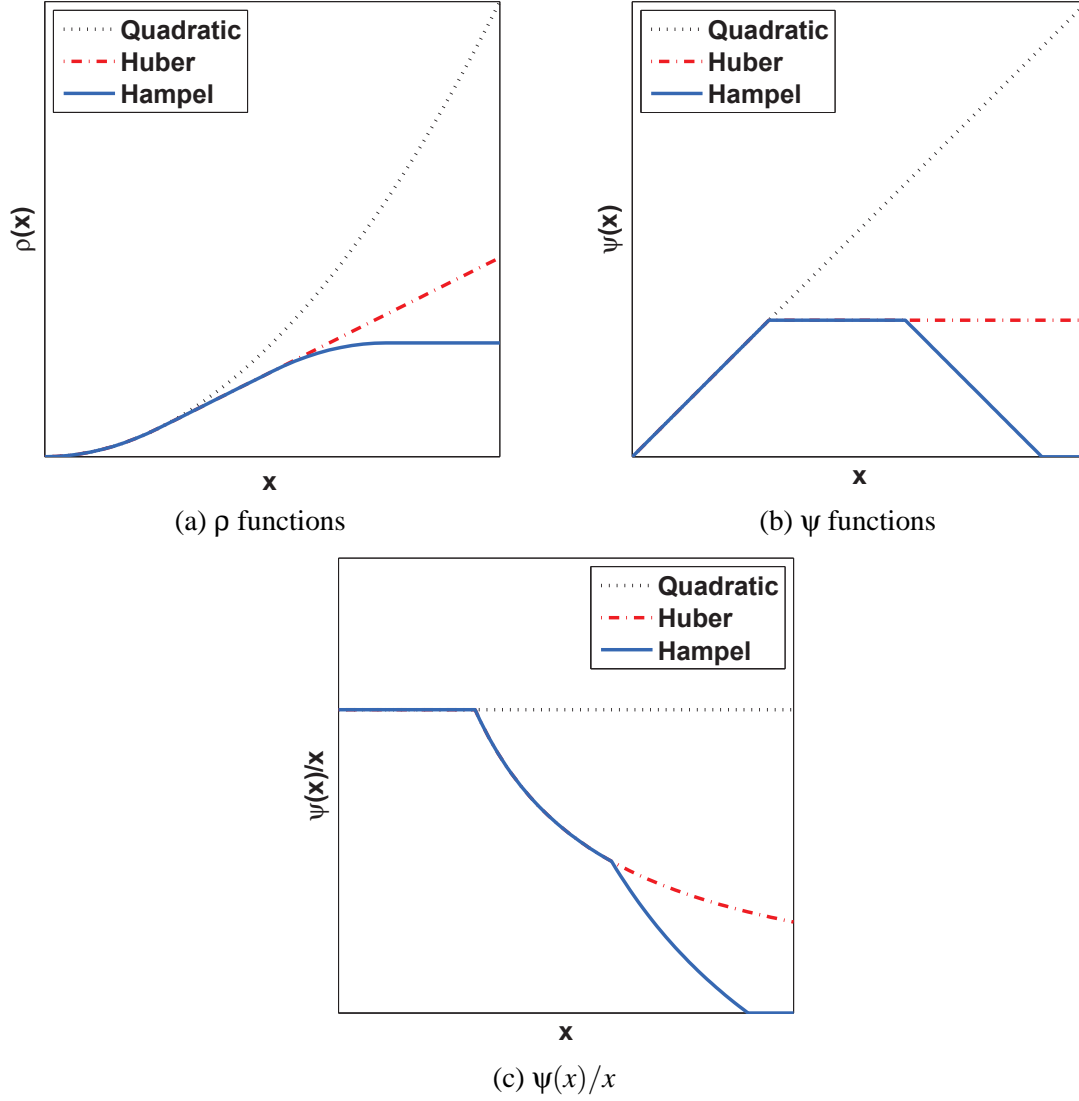


Figure 1: The comparison between three different $\rho(x)$, $\psi(x)$, and $\psi(x)/x$: quadratic, Huber's, and Hampel's.

3. Representer Theorem

In this section, we will describe how \hat{f}_{RKDE} can be expressed as a weighted combination of the $k_{\sigma}(\cdot, \mathbf{X}_i)$'s. A formula for the weights explains how a robust sample mean in \mathcal{H} translates to a robust nonparametric density estimate. We also present necessary and sufficient conditions for a function to be an RKDE. From (2), $\hat{f}_{RKDE} = \arg \min_{g \in \mathcal{H}} J(g)$, where

$$J(g) = \frac{1}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}). \quad (5)$$

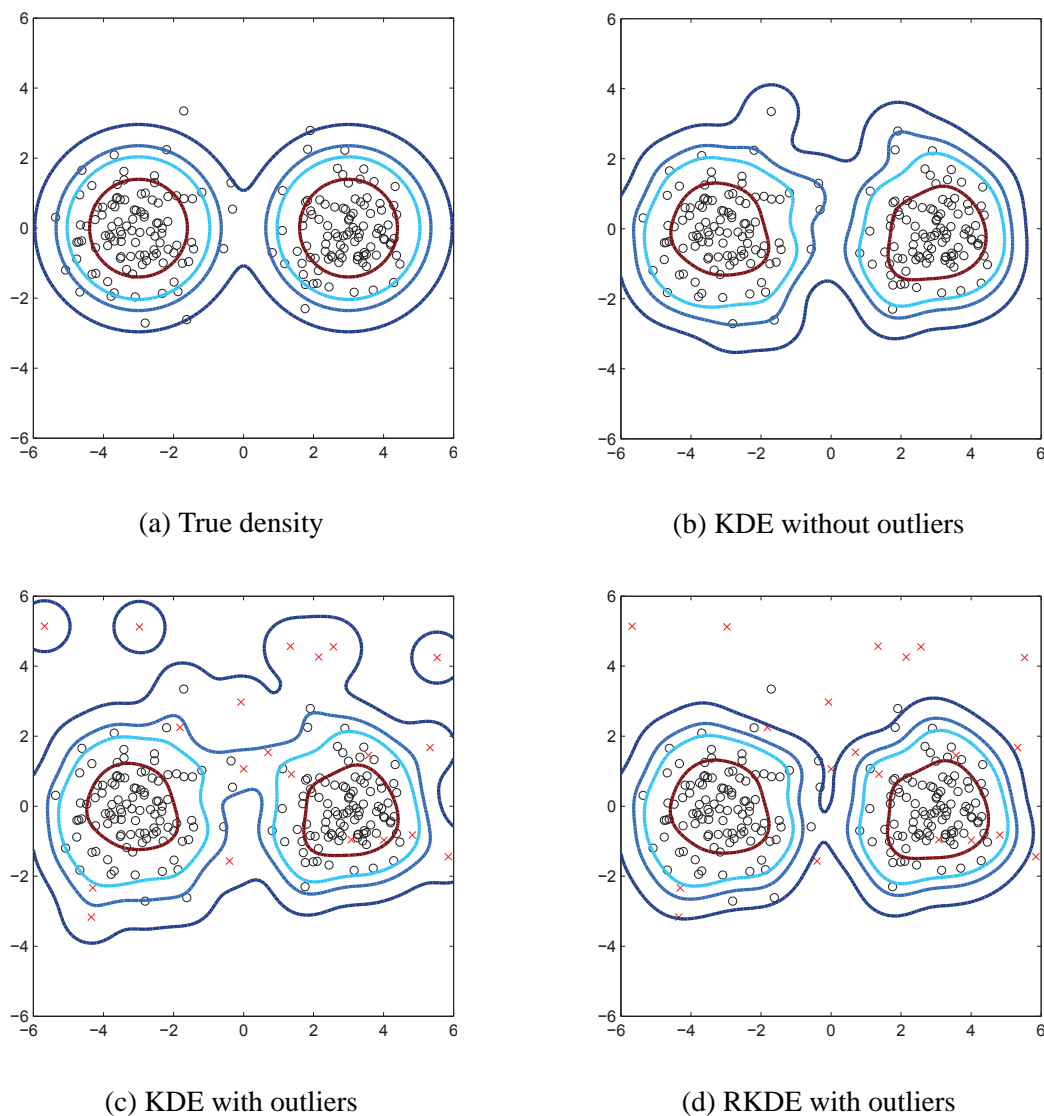


Figure 2: Contours of a nominal density (a) and kernel density estimates (b-d) along with data samples from the nominal density (o) and contaminating density (x). 200 points are from the nominal distribution and 20 contaminating points are from a uniform distribution.

First, let us find necessary conditions for g to be a minimizer of J . Since the space over which we are optimizing J is a Hilbert space, the necessary conditions are characterized through Gateaux differentials of J . Given a vector space \mathcal{X} and a function $T : \mathcal{X} \rightarrow \mathbb{R}$, the Gateaux differential of T at $x \in \mathcal{X}$ with incremental $h \in \mathcal{X}$ is defined as

$$\delta T(x; h) = \lim_{\alpha \rightarrow 0} \frac{T(x + \alpha h) - T(x)}{\alpha}.$$

If $\delta T(x_0; h)$ is defined for all $h \in \mathcal{X}$, a necessary condition for T to have a minimum at x_0 is that $\delta T(x_0; h) = 0$ for all $h \in \mathcal{X}$ (Luenberger, 1997). From this optimality principle, we can establish the following lemma.

Lemma 1 *Suppose assumptions (A1) and (A2) are satisfied. Then the Gateaux differential of J at $g \in \mathcal{H}$ with incremental $h \in \mathcal{H}$ is*

$$\delta J(g; h) = -\langle V(g), h \rangle_{\mathcal{H}}$$

where $V : \mathcal{H} \rightarrow \mathcal{H}$ is given by

$$V(g) = \frac{1}{n} \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{X}_i) - g).$$

A necessary condition for $g = \hat{f}_{RKDE}$ is $V(g) = \mathbf{0}$.

Lemma 1 is used to establish the following representer theorem, so named because \hat{f}_{RKDE} can be represented as a weighted combination of kernels centered at the data points. Similar results are known for supervised kernel methods (Schölkopf et al., 2001).

Theorem 2 *Suppose assumptions (A1) and (A2) are satisfied. Then,*

$$\hat{f}_{RKDE}(\mathbf{x}) = \sum_{i=1}^n w_i k_{\sigma}(\mathbf{x}, \mathbf{X}_i) \quad (6)$$

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

$$w_i \propto \varphi(\|\Phi(\mathbf{X}_i) - \hat{f}_{RKDE}\|_{\mathcal{H}}). \quad (7)$$

It follows that \hat{f}_{RKDE} is a density. The representer theorem also gives the following interpretation of the RKDE. If φ is decreasing, as is the case for a robust loss, then w_i will be small when $\|\Phi(\mathbf{X}_i) - \hat{f}_{RKDE}\|_{\mathcal{H}}$ is large. Now for any $g \in \mathcal{H}$,

$$\begin{aligned} \|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}^2 &= \langle \Phi(\mathbf{X}_i) - g, \Phi(\mathbf{X}_i) - g \rangle_{\mathcal{H}} \\ &= \|\Phi(\mathbf{X}_i)\|_{\mathcal{H}}^2 - 2\langle \Phi(\mathbf{X}_i), g \rangle_{\mathcal{H}} + \|g\|_{\mathcal{H}}^2 \\ &= \tau^2 - 2g(\mathbf{X}_i) + \|g\|_{\mathcal{H}}^2, \end{aligned}$$

where the last step follows from the reproducing property. Taking $g = \hat{f}_{RKDE}$, we see that w_i is small when $\hat{f}_{RKDE}(\mathbf{X}_i)$ is small. Therefore, the RKDE is robust in the sense that it down-weights outlying points.

Theorem 2 provides a necessary condition for \hat{f}_{RKDE} to be the minimizer of (5). With an additional assumption on J , this condition is also sufficient.

Theorem 3 *Suppose that assumptions (A1) and (A2) are satisfied, and J is strictly convex. Then (6), (7), and $\sum_{i=1}^n w_i = 1$ are sufficient for \hat{f}_{RKDE} to be the minimizer of (5).*

Since the previous result assumes J is strictly convex, we give some simple conditions that imply this property.

Lemma 4 J is strictly convex provided either of the following conditions is satisfied:

- (i) ρ is strictly convex and non-decreasing.
- (ii) ρ is convex, strictly increasing, $n \geq 3$, and $K = (k_\sigma(\mathbf{X}_i, \mathbf{X}_j))_{i,j=1}^n$ is positive definite.

The second condition implies that J can be strictly convex even for the Huber loss, which is convex but not strictly convex.

4. KIRWLS Algorithm and Its Convergence

In general, (2) does not have a closed form solution and \hat{f}_{RKDE} has to be found by an iterative algorithm. Fortunately, the iteratively re-weighted least squares (IRWLS) algorithm used in classical M -estimation (Huber, 1964) can be extended to a RKHS using the *kernel trick*. The kernelized iteratively re-weighted least squares (KIRWLS) algorithm starts with initial $w_i^{(0)} \in \mathbb{R}$, $i = 1, \dots, n$ such that $w_i^{(0)} \geq 0$ and $\sum_{i=1}^n w_i^{(0)} = 1$, and generates a sequence $\{f^{(k)}\}$ by iterating on the following procedure:

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

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

Intuitively, this procedure is seeking a fixed point of Equations (6) and (7). The computation of $\|\Phi(\mathbf{X}_j) - f^{(k)}\|_{\mathcal{H}}$ can be done by observing

$$\begin{aligned} \|\Phi(\mathbf{X}_j) - f^{(k)}\|_{\mathcal{H}}^2 &= \langle \Phi(\mathbf{X}_j) - f^{(k)}, \Phi(\mathbf{X}_j) - f^{(k)} \rangle_{\mathcal{H}} \\ &= \langle \Phi(\mathbf{X}_j), \Phi(\mathbf{X}_j) \rangle_{\mathcal{H}} - 2\langle \Phi(\mathbf{X}_j), f^{(k)} \rangle_{\mathcal{H}} + \langle f^{(k)}, f^{(k)} \rangle_{\mathcal{H}}. \end{aligned}$$

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

$$\begin{aligned} \langle \Phi(\mathbf{X}_j), \Phi(\mathbf{X}_j) \rangle_{\mathcal{H}} &= k_\sigma(\mathbf{X}_j, \mathbf{X}_j) \\ \langle \Phi(\mathbf{X}_j), f^{(k)} \rangle_{\mathcal{H}} &= \sum_{i=1}^n w_i^{(k-1)} k_\sigma(\mathbf{X}_j, \mathbf{X}_i) \\ \langle f^{(k)}, f^{(k)} \rangle_{\mathcal{H}} &= \sum_{i=1}^n \sum_{l=1}^n w_i^{(k-1)} w_l^{(k-1)} k_\sigma(\mathbf{X}_i, \mathbf{X}_l). \end{aligned}$$

Recalling that $\Phi(\mathbf{x}) = k_\sigma(\cdot, \mathbf{x})$, after the k th iteration

$$f^{(k)}(\mathbf{x}) = \sum_{i=1}^n w_i^{(k-1)} k_\sigma(\mathbf{x}, \mathbf{X}_i).$$

Therefore, KIRWLS produces a sequence of weighted KDEs. The computational complexity is $O(n^2)$ per iteration. In our experience, the number of iterations needed is typically well below 100. Initialization and termination are discussed in the experimental study below.

KIRWLS can also be viewed as a kind of optimization transfer/majorize-minimize algorithm (Lange et al., 2000; Jacobson and Fessler, 2007) with a quadratic surrogate for ρ . This perspective is used in our analysis in Section 9.4, where $f^{(k)}$ is seen to be the solution of a weighted least squares problem in \mathcal{H} .

The next theorem characterizes the convergence of KIRWLS in terms of $\{J(f^{(k)})\}_{k=1}^{\infty}$ and $\{f^{(k)}\}_{k=1}^{\infty}$.

Theorem 5 *Suppose assumptions (A1) - (A3) are satisfied, and $\phi(x)$ is nonincreasing. Let*

$$\mathcal{S} = \{g \in \mathcal{H} \mid V(g) = \mathbf{0}\}$$

and $\{f^{(k)}\}_{k=1}^{\infty}$ be the sequence produced by the KIRWLS algorithm. Then, $J(f^{(k)})$ monotonically decreases at every iteration and converges. Also, $\mathcal{S} \neq \emptyset$ and

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

as $k \rightarrow \infty$.

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

Corollary 6 *Suppose that the assumptions in Theorem 5 hold and J is strictly convex. Then $\{f^{(k)}\}_{k=1}^{\infty}$ converges to \hat{f}_{RKDE} in the \mathcal{H} -norm and the supremum norm.*

Proof Convergence in the \mathcal{H} -norm follows from the previous result because under strict convexity of J , $|\mathcal{S}| = 1$. Convergence in the supremum norm now follows from the reproducing property and Cauchy-Schwarz because, for any \mathbf{x} ,

$$\begin{aligned} |f^{(k)}(\mathbf{x}) - \hat{f}_{RKDE}(\mathbf{x})| &= |\langle \Phi(\mathbf{x}), f^{(k)} - \hat{f}_{RKDE} \rangle_{\mathcal{H}}| \\ &\leq \tau \|f^{(k)} - \hat{f}_{RKDE}\|_{\mathcal{H}}. \end{aligned}$$

■

5. Influence Function for Robust KDE

To quantify the robustness of the RKDE, we study the influence function. First, we recall the traditional influence function from robust statistics. Let $T(F)$ be an estimator of a scalar parameter based on a distribution F . 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

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

where $\delta_{x'}$ represents a discrete distribution that assigns probability 1 to the point x' . 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 of a Gaussian distribution is the sample mean $T(F)$,

$$T(F) = E_F[X] = \int x dF(x). \quad (8)$$

The influence function for $T(F)$ in (8) is

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

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

Now, consider 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 7 (IF for function estimate) Let $T(\mathbf{x}; F)$ be a function estimate based on F , evaluated at \mathbf{x} . We define the influence function for $T(\mathbf{x}; F)$ as

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

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

$$\begin{aligned} T(\mathbf{x}; F) &= \hat{f}_{KDE}(\mathbf{x}; F) = \int k_\sigma(\mathbf{x}, \mathbf{y}) dF(\mathbf{y}) \\ &= E_F[k_\sigma(\mathbf{x}, \mathbf{X})] \end{aligned}$$

where $\mathbf{X} \sim F$. In this case, the influence function is

$$\begin{aligned} IF(\mathbf{x}, \mathbf{x}'; \hat{f}_{KDE}, F) &= \lim_{s \rightarrow 0} \frac{\hat{f}_{KDE}(\mathbf{x}; F_s) - \hat{f}_{KDE}(\mathbf{x}; F)}{s} \\ &= \lim_{s \rightarrow 0} \frac{E_{F_s}[k_\sigma(\mathbf{x}, \mathbf{X})] - E_F[k_\sigma(\mathbf{x}, \mathbf{X})]}{s} \\ &= \lim_{s \rightarrow 0} \frac{-sE_F[k_\sigma(\mathbf{x}, \mathbf{X})] + sE_{\delta_{\mathbf{x}'}}[k_\sigma(\mathbf{x}, \mathbf{X})]}{s} \\ &= -E_F[k_\sigma(\mathbf{x}, \mathbf{X})] + E_{\delta_{\mathbf{x}'}}[k_\sigma(\mathbf{x}, \mathbf{X})] \\ &= -E_F[k_\sigma(\mathbf{x}, \mathbf{X})] + k_\sigma(\mathbf{x}, \mathbf{x}'). \end{aligned} \quad (9)$$

With the empirical distribution $F_n = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}$,

$$IF(\mathbf{x}, \mathbf{x}'; \hat{f}_{KDE}, F_n) = -\frac{1}{n} \sum_{i=1}^n k_\sigma(\mathbf{x}, \mathbf{x}_i) + k_\sigma(\mathbf{x}, \mathbf{x}'). \quad (10)$$

To investigate the influence function of the RKDE, we generalize its definition to a general distribution μ , writing $\hat{f}_{RKDE}(\cdot; \mu) = f_\mu$ where

$$f_\mu = \arg \min_{g \in \mathcal{H}} \int \rho(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) d\mu(\mathbf{x}).$$

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

Theorem 8 *Suppose assumptions (A1)-(A5) are satisfied. In addition, assume that $f_{F_s} \rightarrow f_F$ as $s \rightarrow 0$. If $\dot{f}_F \triangleq \lim_{s \rightarrow 0} \frac{f_{F_s} - f_F}{s}$ exists, then*

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

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

$$\begin{aligned} & \left(\int \varphi(\|\Phi(\mathbf{x}) - f_F\|_{\mathcal{H}}) dF \right) \cdot \dot{f}_F \\ & + \int \left(\frac{\langle \dot{f}_F, \Phi(\mathbf{x}) - f_F \rangle_{\mathcal{H}}}{\|\Phi(\mathbf{x}) - f_F\|_{\mathcal{H}}^3} \cdot q(\|\Phi(\mathbf{x}) - f_F\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{x}) - f_F) \right) dF(\mathbf{x}) \\ & = (\Phi(\mathbf{x}') - f_F) \cdot \varphi(\|\Phi(\mathbf{x}') - f_F\|_{\mathcal{H}}). \end{aligned} \quad (11)$$

Unfortunately, for Huber or Hampel's ρ , there is no closed form solution for \dot{f}_F of (11). However, if we work with F_n instead of F , we can find \dot{f}_{F_n} explicitly. Let

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

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

$$\gamma = \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - f_{F_n}\|_{\mathcal{H}}),$$

and

$$\mathbf{w} = [w_1, \dots, w_n]^T,$$

where \mathbf{w} gives the RKDE weights as in (6).

Theorem 9 *Suppose assumptions (A1)-(A5) are satisfied. In addition, assume that*

- $f_{F_{n,s}} \rightarrow f_{F_n}$ as $s \rightarrow 0$ (satisfied when J is strictly convex)
- the extended kernel matrix K' based on $\{\mathbf{X}_i\}_{i=1}^n \cup \{\mathbf{x}'\}$ is positive definite.

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

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

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

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

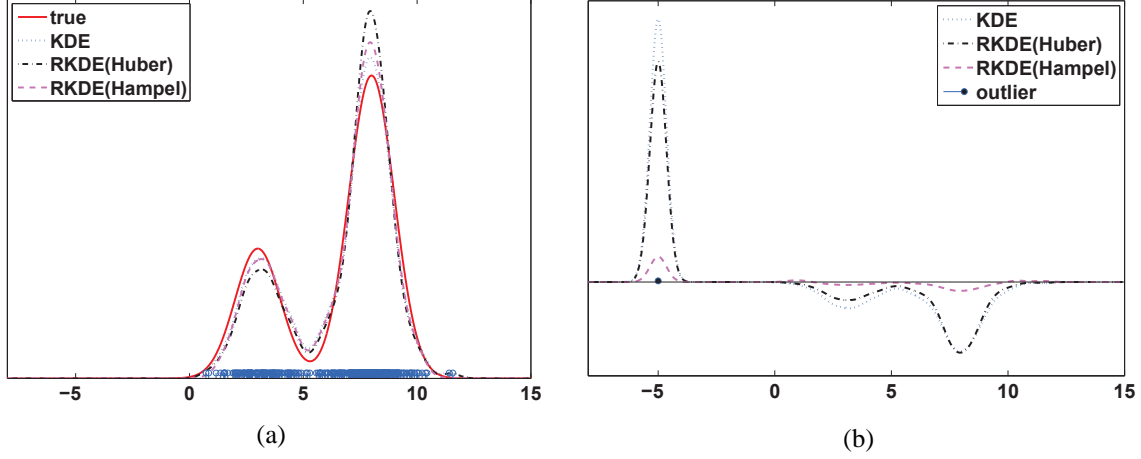


Figure 3: (a) true density and density estimates. (b) IF as a function of \mathbf{x} when $\mathbf{x}' = -5$

Note that α' captures the amount by which the density estimator changes near \mathbf{x}' in response to contamination at \mathbf{x}' . Now α' is given by

$$\alpha' = \frac{\varphi(\|\Phi(\mathbf{x}') - f_{F_n}\|_{\mathcal{H}})}{\frac{1}{n} \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - f_{F_n}\|_{\mathcal{H}})}.$$

For a standard KDE, we have $\varphi \equiv 1$ and $\alpha' = 1$, in agreement with (10). For robust ρ , $\varphi(\|\Phi(\mathbf{x}') - f_{F_n}\|_{\mathcal{H}})$ can be viewed as a measure of “inlyingness”, with more inlying points having larger values. This follows from the discussion just after Theorem 2, which leverages the reproducing property. If the contaminating point \mathbf{x}' is less inlying than the average \mathbf{X}_i , then $\alpha' < 1$. Thus, the RKDE is less sensitive to outlying points than the KDE.

As mentioned above, in classical robust statistics, the robustness of an estimator can be inferred from the boundedness of the corresponding influence function. However, the influence functions for density estimators are bounded even if $\|\mathbf{x}'\| \rightarrow \infty$. 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 Figure 3 for a synthetic univariate distribution. Figure 3 (a) shows the density of the distribution, and three estimates. Figure 3 (b) shows the corresponding influence functions. As we can see in (b), for a point \mathbf{x}' in the tails of F , the influence functions for the robust KDEs are overall smaller, in absolute value, than those of the standard KDE (especially with Hampel’s loss). Additional numerical results are given in Section 7.2.

Finally, it is interesting to note that for any density estimator \hat{f} ,

$$\int IF(\mathbf{x}, \mathbf{x}'; \hat{f}, F) d\mathbf{x} = \lim_{s \rightarrow 0} \frac{\int \hat{f}(\mathbf{x}; F_s) d\mathbf{x} - \int \hat{f}(\mathbf{x}; F) d\mathbf{x}}{s} = 0.$$

Thus $\alpha' = -\sum_{i=1}^n \alpha_i$ for a robust KDE. This suggests that since \hat{f}_{RKDE} has a smaller increase at \mathbf{x}' (compared to the KDE), it will also have a smaller decrease (in absolute value) near the training data. Therefore, the norm of $IF(\mathbf{x}, \mathbf{x}'; \hat{f}_{RKDE}, F_n)$, viewed as a function of \mathbf{x} , should be smaller when \mathbf{x}' is an outlier. We confirm this in our experiments in Section 7.

6. Generalization to Other Hilbert Spaces

So far, we have focused our attention on PSD kernels and viewed the KDE as an element of the RKHS associated with the kernel. However, the RKDE can be defined in a more general setting. In particular, it is only necessary that the functions $\Phi(\mathbf{x}) = k_\sigma(\cdot, \mathbf{x})$ belong to a Hilbert space \mathcal{H} . Then one can still obtain all the previous results, that is, Lemmas 1 and 4, Theorems 2, 3, 5, 8, and 9, and Corollary 6 (except convergence in the supremum norm). (For Theorems 8 and 9 it is necessary to additionally assume that $\|\Phi(\mathbf{x})\|_{\mathcal{H}}$ is bounded independent of \mathbf{x} .) The only necessary change is that inner products of the form $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}$ can no longer be computed as $k_\sigma(\mathbf{x}, \mathbf{x}')$. Thus, K in Lemma 4 (ii), \mathbf{k}', K, K' in Theorem 9, and various terms in the KIRWLS algorithm should now be computed with the inner product on \mathcal{H} .

It is also interesting to note that this generalization gives a representer theorem for non-RKHS Hilbert spaces. This contrasts with standard approaches to supervised learning that minimize an empirical risk plus regularization term. In those cases, a representer theorem may be more difficult to obtain when the function class is not an RKHS.

There are some examples of non-RKHS Hilbert spaces where the inner product can still be computed efficiently. For example, consider $\mathcal{H} = L^2(\mathbb{R}^d)$ and denote

$$\begin{aligned}\tilde{k}_\sigma(\mathbf{x}, \mathbf{x}') &= \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{L^2(\mathbb{R}^d)} \\ &= \int k_\sigma(\mathbf{z}, \mathbf{x}) k_\sigma(\mathbf{z}, \mathbf{x}') d\mathbf{z}.\end{aligned}$$

For the multivariate Gaussian kernel, $\tilde{k}_\sigma = k_{\sqrt{2}\sigma}$. For the multivariate Cauchy kernel (the multivariate Student kernel with $\nu = 1$; see Section 2), it holds that $\tilde{k}_\sigma(\mathbf{x}, \mathbf{x}') = k_{2\sigma}(\mathbf{x}, \mathbf{x}')$ (Berry et al., 1996). For the multivariate Laplacian product kernel,

$$k_\sigma(\mathbf{x}, \mathbf{x}') = \frac{1}{(2\sigma)^d} \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}'\|_1}{\sigma}\right\},$$

it is true that

$$\tilde{k}_\sigma(\mathbf{x}, \mathbf{x}') = \frac{1}{(4\sigma)^d} \prod_{l=1}^d \left(1 + \frac{|\mathbf{x}_l - \mathbf{x}'_l|}{\sigma}\right) \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}'\|_1}{\sigma}\right\}.$$

For kernels without a closed-form expression for \tilde{k}_σ , the inner product may still be calculated numerically. For radially symmetric kernels this entails a simple one-dimensional table, as $\tilde{k}_\sigma(\mathbf{x}, \mathbf{x}')$ depends only on $\|\mathbf{x} - \mathbf{x}'\|$, and these values may be tabulated in advance.

As we noted previously, however, we rely on the reproducing property to deduce robustness of the RKDE from the representer theorem and the influence function. For non-RKHS Hilbert spaces, these arguments are less clear. We have not experimentally investigated non-reproducing kernels, and so cannot comment on the robustness of RKDEs based on such kernels in practice.

7. Experiments

The experimental setup is described in 7.1, and results are presented in 7.2.

7.1 Experimental Setup

Data, methods, and evaluation are now discussed.

7.1.1 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), which were originally used in the task of classification. The data sets are available online: see <http://www.fml.tuebingen.mpg.de/Members/> for the first 12 data sets and the UCI machine learning repository for the last 3 data sets. There are 100 randomly permuted partitions of each data set into “training” and “test” sets (20 for Image, Splice, and MNIST).

Given $\mathbf{X}_1, \dots, \mathbf{X}_n \sim f = (1 - p) \cdot f_0 + p \cdot f_1$, our goal is to estimate f_0 , or the level sets of f_0 . For each data set with two classes, we take one class as the nominal data from f_0 and the other class as contamination from f_1 . For Iris, there are 3 classes and we take one class as nominal data and the other two as contamination. For MNIST, we choose to use digit 0 as nominal and digit 1 as contamination. For MNIST, the original dimension 784 is reduced to 8 via kernel PCA using a Gaussian kernel with bandwidth 30. For each data set, the training sample consists of n_0 nominal data and n_1 contaminating points, where $n_1 = \varepsilon \cdot n_0$ for $\varepsilon = 0, 0.05, 0.10, 0.15, 0.20, 0.25$ and 0.30 . Note that each ε corresponds to an anomaly proportion p such that $p = \frac{\varepsilon}{1+\varepsilon}$. n_0 is always taken to be the full amount of training data for the nominal class.

7.1.2 METHODS

In our experiments, we compare three density estimators: the standard kernel density estimator (KDE), variable kernel density estimator (VKDE), and robust kernel density estimator (RKDE) with Hampel’s loss. For all methods, the Gaussian kernel in (1) is used as the kernel function k_σ and the kernel bandwidth σ is set as the median distance of a training point \mathbf{X}_i to its nearest neighbor.

The VKDE has a variable bandwidth for each data point,

$$\hat{f}_{VKDE}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n k_{\sigma_i}(\mathbf{x}, \mathbf{X}_i),$$

and the bandwidth σ_i is set as

$$\sigma_i = \sigma \cdot \left(\frac{\eta}{\hat{f}_{KDE}(\mathbf{X}_i)} \right)^{1/2}$$

where η is the mean of $\{\hat{f}_{KDE}(\mathbf{X}_i)\}_{i=1}^n$ (Abramson, 1982; Comaniciu et al., 2001). There is another implementation of the VKDE where σ_i is based on the distance to its k -th nearest neighbor (Breiman et al., 1977). However, this version did not perform as well and is therefore omitted.

For the RKDE, the parameters a , b , and c in (4) are set as follows. First, we compute \hat{f}_{med} , which is the RKDE obtained with $\rho = |\cdot|$, and set $d_i = \|\Phi(\mathbf{X}_i) - \hat{f}_{med}\|_{\mathcal{H}}$. Then, a is set to be the median of $\{d_i\}$, b the 75th percentile of $\{d_i\}$, and c the 85th percentile of $\{d_i\}$. After finding these parameters, we initialize $w_i^{(0)}$ such that $f^{(1)} = \hat{f}_{med}$ and terminate KIRWLS when

$$\frac{|J(f^{(k+1)}) - J(f^{(k)})|}{J(f^{(k)})} < 10^{-8}.$$

7.1.3 EVALUATION

We evaluate the performance of the three density estimators in three different settings. First, we use the influence function to study sensitivity to outliers. Second and third, we compare the methods

at the tasks of density estimation and anomaly detection, respectively. In each case, an appropriate performance measure is adopted. These are explained in detail in Section 7.2. To compare a pair of methods across multiple data sets, we adopt the Wilcoxon signed-rank test (Wilcoxon, 1945). Given a performance measure, and given a pair of methods and ϵ , we compute the difference h_i between the performance of two density estimators on the i th data set. The data sets are ranked 1 through 15 according to their absolute values $|h_i|$, with the largest $|h_i|$ corresponding to the rank of 15. Let R_1 be the sum of ranks over these data sets where method 1 beats method 2, and let R_2 be the sum of the ranks for the other data sets. The signed-rank test statistic $T \triangleq \min(R_1, R_2)$ and the corresponding p -value are used to test whether the performances of the two methods are significantly different. For example, the critical value of T for the signed rank test is 25 at a significance level of 0.05. Thus, if $T \leq 25$, the two methods are significantly different at the given significance level, and the larger of R_1 and R_2 determines the method with better performance.

7.2 Experimental Results

We begin by studying influence functions.

7.2.1 SENSITIVITY USING INFLUENCE FUNCTION

As the first measure of robustness, we compare the influence functions for KDEs and RKDEs, given in (10) and Theorem 9, respectively. To our knowledge, there is no formula for the influence function of VKDEs, and therefore VKDEs are excluded in the comparison. We examine $\alpha(\mathbf{x}') = IF(\mathbf{x}', \mathbf{x}'; T, F_n)$ and

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

In words, $\alpha(\mathbf{x}')$ reflects 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 .

In this experiment, ϵ is equal to 0, that is, the density estimators are learned from a pure nominal sample. Then, we take contaminating points from the test sample, each of which serves as an \mathbf{x}' . This gives us multiple $\alpha(\mathbf{x}')$'s and $\beta(\mathbf{x}')$'s. The performance measures are the medians of $\{\alpha(\mathbf{x}')\}$ and $\{\beta(\mathbf{x}')\}$ (smaller means better performance). The results using signed rank statistics are shown in Table 1. The results clearly states that for all data sets, RKDEs are less affected by outliers than KDEs.

7.2.2 KULLBACK-LEIBLER (KL) DIVERGENCE

Second, we present the Kullback-Leibler (KL) divergence between density estimates \hat{f} and f_0 ,

$$D_{KL}(\hat{f} || f_0) = \int \hat{f}(\mathbf{x}) \log \frac{\hat{f}(\mathbf{x})}{f_0(\mathbf{x})} d\mathbf{x}.$$

This KL divergence is large whenever \hat{f} estimates f_0 to have mass where it does not. For contamination characterized by properties (a), (b), and (c) in the Introduction, we expect this performance measure to capture the robustness of a density estimator.

The computation of D_{KL} is done as follows. Since we do not know the nominal f_0 , it is estimated as \tilde{f}_0 , a KDE based on a separate nominal sample, obtained from the test data for each benchmark

method 1	method 2		$\alpha(\mathbf{x}')$	$\beta(\mathbf{x}')$
RKDE	KDE	R_1	120	120
		R_2	0	0
		T	0	0
		p -value	0.00	0.00

Table 1: The signed-rank statistics and p -values of the Wilcoxon signed-rank test using the medians of $\{\alpha(\mathbf{x}')\}$ and $\{\beta(\mathbf{x}')\}$ as a performance measure. If R_1 is larger than R_2 , method 1 is better than method 2.

data set. Then, the integral is approximated by the sample mean, that is,

$$D_{KL}(\hat{f}||f_0) \approx \frac{1}{n'} \sum_{i=1}^{n'} \log \frac{\hat{f}(\mathbf{x}'_i)}{\tilde{f}_0(\mathbf{x}'_i)}$$

where $\{\mathbf{x}'_i\}_{i=1}^{n'}$ is an i.i.d sample from the estimated density \hat{f} with $n' = 2n = 2(n_0 + n_1)$. Note that the estimated KL divergence can have an infinite value when $\tilde{f}_0(\mathbf{y}) = 0$ (to machine precision) and $\hat{f}(\mathbf{y}) > 0$ for some $\mathbf{y} \in \mathbb{R}^d$. The averaged KL divergence over the permutations are used as the performance measure (smaller means better performance). In Table 2, the rank of the three methods are shown for each data set and ϵ .

Table 3 summarizes the results using the Wilcoxon signed-rank test. When comparing RKDEs and KDEs, the results show that KDEs have smaller KL divergence than RKDEs with $\epsilon = 0$. As ϵ increases, however, RKDEs estimate f_0 more accurately than KDEs. The results also demonstrate that VKDEs are the worst in the sense of KL divergence. Note that VKDEs place a total mass of $1/n$ at all \mathbf{X}_i , whereas the RKDE will place a mass $w_i < 1/n$ at outlying points.

Since KL divergence is not symmetric, we also compute KL divergence between f_0 and \hat{f} ,

$$\begin{aligned} D_{KL}(f_0||\hat{f}) &= \int f_0(\mathbf{x}) \log \frac{f_0(\mathbf{x})}{\hat{f}(\mathbf{x})} d\mathbf{x} \\ &= \int f_0(\mathbf{x}) \log f_0(\mathbf{x}) d\mathbf{x} - \int f_0(\mathbf{x}) \log \hat{f}(\mathbf{x}) d\mathbf{x} \end{aligned}$$

This KL divergence is large whenever \hat{f} estimates f_0 not to have mass where it does.

Since f_0 is the same when comparing different estimate \hat{f} , we only have to compare the second term, which is approximated as

$$-\int f_0(\mathbf{x}) \log \hat{f}(\mathbf{x}) d\mathbf{x} \approx -\frac{1}{n''} \sum_{i=1}^{n''} \log \hat{f}(\mathbf{x}''_i),$$

where $\{\mathbf{x}''_i\}_{i=1}^{n''}$ is a separate nominal sample, obtained from the test data. Table 4 and 5 show that with this KL divergence as performance measure, VKDE performs best for almost all data sets and ϵ . When comparing RKDE and KDE based on Wilcoxon signed-rank test, T has a large value for every ϵ which makes it inconclusive whether one method is better than the other.

Data Set	ϵ						
	0.00	0.05	0.10	0.15	0.20	0.25	0.30
Banana	(3,1,2)	(3,1,2)	(3,1,2)	(3,1,2)	(3,1,2)	(3,1,2)	(3,1,2)
B. Cancer	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Diabetes	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(1,2,3)	(1,2,3)
F. Solar	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)
German	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Heart	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Image	(2,1,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Ringnorm	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Splice	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Thyroid	(2,1,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Twonorm	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Waveform	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)
Pima Indian	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(1,2,3)
Iris	(2,1,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
MNIST	(3,1,2)	(2,1,3)	(2,1,3)	(2,1,3)	(1,2,3)	(1,2,3)	(1,2,3)

Table 2: The ranks of (RKDE, KDE, VKDE) using $D_{KL}(\hat{f}||f_0)$ as a performance measure. For example, (2, 1, 3) means that KDE performs best, RKDE next, and VKDE worst.

7.2.3 ANOMALY DETECTION

In this experiment, we apply the density estimators in anomaly detection problems. If we had a pure sample from f_0 , we would estimate f_0 and use $\{\mathbf{x} : \hat{f}_0(\mathbf{x}) > \lambda\}$ as a detector. For each λ , we could get a false negative and false positive probability using test data. By varying λ , we would then obtain a receiver operating characteristic (ROC) and area under the curve (AUC). However, since we have a contaminated sample, we have to estimate f_0 robustly. Robustness can be checked by comparing the AUC of the anomaly detectors, where the density estimates are based on the contaminated training data (higher AUC means better performance).

Examples of the ROCs are shown in Figure 4. The RKDE provides better detection probabilities, especially at low false alarm rates. This results in higher AUC. For each pair of methods and each ϵ , R_1 , R_2 , T and p -values are shown in Table 7. The results indicate that RKDEs are significantly better than KDEs when $\epsilon \geq 0.20$ with significance level 0.05. RKDEs are also better than VKDEs when $\epsilon \geq 0.15$ but the difference is not significant. We also note that we have also evaluated the kernelized spatial depth (KSD) (Chen et al., 2009) in this setting. While this method does not yield a density estimate, it does aim to estimate density contours robustly. We found that the KSD performs worse in terms of AUC than either the RKDE or KDE, so those results are omitted (Kim and Scott, 2011).

method 1	method 2		ϵ						
			0.00	0.05	0.10	0.15	0.20	0.25	0.30
RKDE	KDE	R_1	26	67	78	83	94	101	103
		R_2	94	53	42	37	26	19	17
		T	26	53	42	37	26	19	17
		p -value	0.06	0.72	0.33	0.21	0.06	0.02	0.01
RKDE	VKDE	R_1	104	117	117	117	117	119	119
		R_2	16	3	3	3	3	1	1
		T	16	3	3	3	3	1	1
		p -value	0.01	0.00	0.00	0.00	0.00	0.00	0.00
VKDE	KDE	R_1	0	0	0	0	0	0	0
		R_2	120	120	120	120	120	120	120
		T	0	0	0	0	0	0	0
		p -value	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: The signed-rank statistics and p -values of the Wilcoxon signed-rank test using $D_{KL}(\hat{f}||f_0)$ as a performance measure. If R_1 is larger than R_2 , method 1 is better than method 2.

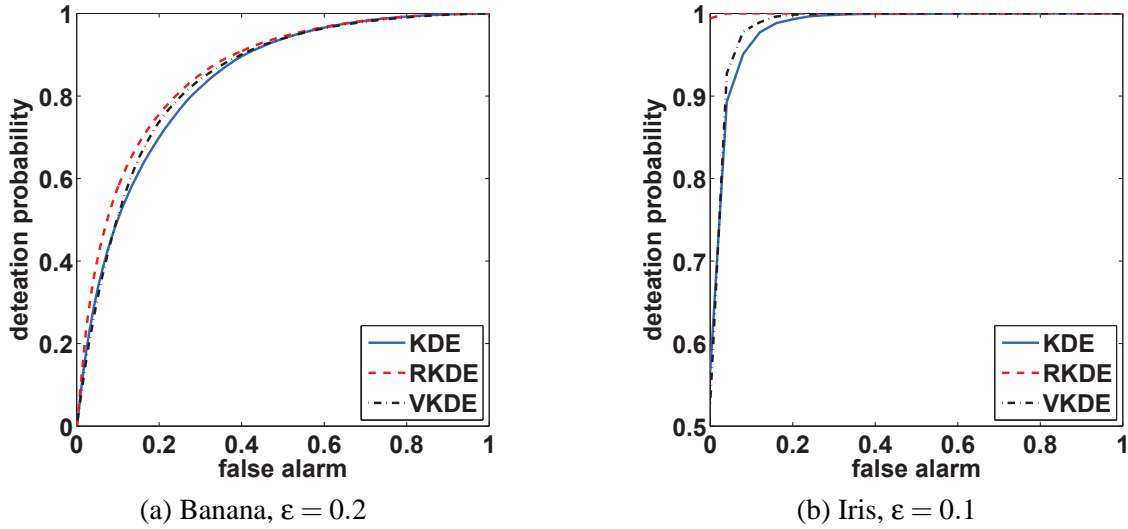


Figure 4: Examples of ROCs.

8. Conclusions

When kernel density estimators employ a smoothing kernel that is also a PSD kernel, they may be viewed as M -estimators in the RKHS associated with the kernel. While the traditional KDE corresponds to the quadratic loss, the RKDE employs a robust loss to achieve robustness to contamination of the training sample. **The RKDE is a weighted kernel density estimate, where smaller weights are given to more outlying data points.** These weights can be computed efficiently using

Data Set	ϵ						
	0.00	0.05	0.10	0.15	0.20	0.25	0.30
Banana	(3,1,2)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)
B. Cancer	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)
Diabetes	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)
F. Solar	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
German	(3,2,1)	(3,2,1)	(3,2,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
Heart	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
Image	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)
Ringnorm	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
Splice	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
Thyroid	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
Twonorm	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
Waveform	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(2,3,1)	(2,3,1)	(2,3,1)
Pima Indian	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)
Iris	(3,2,1)	(3,2,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
MNIST	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)

Table 4: The ranks of (RKDE, KDE, VKDE) using $D_{KL}(f_0 || \hat{f})$ as a performance measure. For example, (2, 1, 3) means that KDE performs best, RKDE next, and VKDE worst.

a kernelized iteratively re-weighted least squares algorithm. **The decreased sensitivity of RKDEs to contamination is further attested by the influence function, as well as experiments on anomaly detection and density estimation problems.**

Robust kernel density estimators are **nonparametric**, making no parametric assumptions on the data generating distributions. However, their success is still contingent on certain conditions being satisfied. Obviously, the percentage of contaminating data must be less than 50%; our experiments examine contamination up to around 25%. In addition, the contaminating distribution must be outlying with respect to the nominal distribution. Furthermore, the anomalous component should not be too concentrated, otherwise it may look like a mode of the nominal component. Such assumptions seem necessary given the unsupervised nature of the problem, and are implicit in our interpretation of the representer theorem and influence functions.

Although our focus has been on density estimation, in many applications the ultimate goal is not to estimate a density, but rather to estimate decision regions. Our methodology is immediately applicable to such situations, as evidenced by our experiments on anomaly detection. It is only necessary that the kernel be PSD here; the assumption that the kernel be nonnegative and integrate to one can clearly be dropped. This allows for the use of more general kernels, such as polynomial kernels, or kernels on non-Euclidean domains such as strings and trees. The learning problem here could be described as one-class classification with contaminated data.

In future work it would be interesting to investigate asymptotics, the bias-variance trade-off, and the efficiency-robustness trade-off of robust kernel density estimators, as well as the impact of different losses and kernels.

method 1	method 2		ϵ						
			0.00	0.05	0.10	0.15	0.20	0.25	0.30
RKDE	KDE	R_1	47	52	55	58	64	65	68
		R_2	73	68	65	62	56	55	52
		T	47	52	55	58	56	55	52
		p -value	0.49	0.68	0.80	0.93	0.85	0.80	0.68
RKDE	VKDE	R_1	0	0	0	0	0	0	0
		R_2	120	120	120	120	120	120	120
		T	0	0	0	0	0	0	0
		p -value	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VKDE	KDE	R_1	119	120	120	120	120	120	120
		R_2	1	0	0	0	0	0	0
		T	1	0	0	0	0	0	0
		p -value	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5: The signed-rank statistics and p -values of the Wilcoxon signed-rank test using $D_{KL}(f_0 \parallel \hat{f})$ divergence as a performance measure. If R_1 is larger than R_2 , method 1 is better than method 2.

9. Proofs

We begin with three lemmas and proofs. The first lemma will be used in the proofs of Lemma 11 and Theorem 9, the second one in the proof of Lemma 4, and the third one in the proof of Theorem 5.

Lemma 10 *Let $\mathbf{z}_1, \dots, \mathbf{z}_m$ be distinct points in \mathbb{R}^d . If $K = (k(\mathbf{z}_i, \mathbf{z}_j))_{i,j=1}^n$ is positive definite, then $\Phi(\mathbf{z}_i) = k(\cdot, \mathbf{z}_i)$'s are linearly independent.*

Proof $\sum_{i=1}^m \alpha_i \Phi(\mathbf{z}_i) = 0$ implies

$$\begin{aligned}
 0 &= \left\| \sum_{i=1}^m \alpha_i \Phi(\mathbf{z}_i) \right\|_{\mathcal{H}}^2 \\
 &= \left\langle \sum_{i=1}^m \alpha_i \Phi(\mathbf{z}_i), \sum_{j=1}^m \alpha_j \Phi(\mathbf{z}_j) \right\rangle_{\mathcal{H}} \\
 &= \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j k(\mathbf{z}_i, \mathbf{z}_j)
 \end{aligned}$$

and from positive definiteness of K , $\alpha_1 = \dots = \alpha_m = 0$. ■

Lemma 11 *Let \mathcal{H} be a RKHS associated with a kernel k , and $\mathbf{x}_1, \mathbf{x}_2$, and \mathbf{x}_3 be distinct points in \mathbb{R}^d . Assume that $K = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^3$ is positive definite. For any $g, h \in \mathcal{H}$ with $g \neq h$, $\Phi(\mathbf{x}_i) - g$ and $\Phi(\mathbf{x}_i) - h$ are linearly independent for some $i \in \{1, 2, 3\}$.*

Data Set	ϵ						
	0.00	0.05	0.10	0.15	0.20	0.25	0.30
Banana	(3,2,1)	(3,2,1)	(3,2,1)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)
B. Cancer	(2,1,3)	(2,1,3)	(2,1,3)	(1,3,2)	(1,3,2)	(1,3,2)	(2,3,1)
Diabetes	(3,1,2)	(3,2,1)	(2,3,1)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)
F. Solar	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(3,1,2)
German	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(1,2,3)	(1,2,3)	(1,2,3)
Heart	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
Image	(3,1,2)	(3,1,2)	(3,1,2)	(2,3,1)	(2,3,1)	(1,3,2)	(1,3,2)
Ringnorm	(2,1,3)	(2,1,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)	(1,2,3)
Splice	(1,2,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)	(2,1,3)
Thyroid	(3,1,2)	(3,2,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)	(2,3,1)
Twonorm	(3,2,1)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)
Waveform	(2,3,1)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)
Pima Indian	(3,1,2)	(3,2,1)	(2,3,1)	(2,3,1)	(2,3,1)	(1,3,2)	(1,3,2)
Iris	(3,1,2)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)	(1,3,2)
MNIST	(3,1,2)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)	(3,2,1)

Table 6: The ranks of (RKDE, KDE, VKDE) using $D_{KL}(f_0 || \hat{f})$ as a performance measure. For example, (2, 1, 3) means that KDE performs best, RKDE next, and VKDE worst.

method 1	method 2		ϵ						
			0.00	0.05	0.10	0.15	0.20	0.25	0.30
RKDE	KDE	R_1	26	46	67	90	95	96	99
		R_2	94	74	53	30	25	24	21
		T	26	46	53	30	25	24	21
		p -value	0.06	0.45	0.72	0.09	0.05	0.04	0.03
RKDE	VKDE	R_1	33	49	58	75	80	90	86
		R_2	87	71	62	45	40	30	34
		T	33	49	58	45	40	30	34
		p -value	0.14	0.56	0.93	0.42	0.28	0.09	0.15
VKDE	KDE	R_1	38	70	79	91	95	96	99
		R_2	82	50	41	29	25	24	21
		T	38	50	41	29	25	24	21
		p -value	0.23	0.60	0.30	0.08	0.05	0.04	0.03

Table 7: The signed-rank statistics of the Wilcoxon signed-rank test using AUC as a performance measure. If R_1 is larger than R_2 , method 1 is better than method 2.

Proof We will prove the lemma by contradiction. Suppose $\Phi(\mathbf{x}_i) - g$ and $\Phi(\mathbf{x}_i) - h$ are linearly dependent for all $i = 1, 2, 3$. Then, there exists $(\alpha_i, \beta_i) \neq (0, 0)$ for $i = 1, 2, 3$ such that

$$\alpha_1(\Phi(\mathbf{x}_1) - g) + \beta_1(\Phi(\mathbf{x}_1) - h) = \mathbf{0} \quad (12)$$

$$\alpha_2(\Phi(\mathbf{x}_2) - g) + \beta_2(\Phi(\mathbf{x}_2) - h) = \mathbf{0} \quad (13)$$

$$\alpha_3(\Phi(\mathbf{x}_3) - g) + \beta_3(\Phi(\mathbf{x}_3) - h) = \mathbf{0}. \quad (14)$$

Note that $\alpha_i + \beta_i \neq 0$ since $g \neq h$.

First consider the case $\alpha_2 = 0$. This gives $h = \Phi(\mathbf{x}_2)$, and $\alpha_1 \neq 0$ and $\alpha_3 \neq 0$. Then, (12) and (13) simplify to

$$\begin{aligned} g &= \frac{\alpha_1 + \beta_1}{\alpha_1} \Phi(\mathbf{x}_1) - \frac{\beta_1}{\alpha_1} \Phi(\mathbf{x}_2), \\ g &= \frac{\alpha_3 + \beta_3}{\alpha_3} \Phi(\mathbf{x}_3) - \frac{\beta_3}{\alpha_3} \Phi(\mathbf{x}_2), \end{aligned}$$

respectively. This is contradiction because $\Phi(\mathbf{x}_1)$, $\Phi(\mathbf{x}_2)$, and $\Phi(\mathbf{x}_3)$ are linearly independent by Lemma 10 and

$$\frac{\alpha_1 + \beta_1}{\alpha_1} \Phi(\mathbf{x}_1) + \left(\frac{\beta_3}{\alpha_3} - \frac{\beta_1}{\alpha_1} \right) \Phi(\mathbf{x}_2) - \frac{\alpha_3 + \beta_3}{\alpha_3} \Phi(\mathbf{x}_3) = \mathbf{0}$$

where $(\alpha_1 + \beta_1)/\alpha_1 \neq 0$.

Now consider the case where $\alpha_2 \neq 0$. Subtracting (13) multiplied by α_1 from (12) multiplied by α_2 gives

$$(\alpha_1\beta_2 - \alpha_2\beta_1)h = -\alpha_2(\alpha_1 + \beta_1)\Phi(\mathbf{x}_1) + \alpha_1(\alpha_2 + \beta_2)\Phi(\mathbf{x}_2).$$

In the above equation $\alpha_1\beta_2 - \alpha_2\beta_1 \neq 0$ because this implies $\alpha_2(\alpha_1 + \beta_1) = 0$ and $\alpha_1(\alpha_2 + \beta_2) = 0$, which, in turn, implies $\alpha_2 = 0$. Therefore, h can be expressed as $h = \lambda_1\Phi(\mathbf{x}_1) + \lambda_2\Phi(\mathbf{x}_2)$ where

$$\lambda_1 = -\frac{\alpha_2(\alpha_1 + \beta_1)}{\alpha_1\beta_2 - \alpha_2\beta_1}, \quad \lambda_2 = \frac{\alpha_1(\alpha_2 + \beta_2)}{\alpha_1\beta_2 - \alpha_2\beta_1}.$$

Similarly, from (13) and (14), $h = \lambda_3\Phi(\mathbf{x}_2) + \lambda_4\Phi(\mathbf{x}_3)$ where

$$\lambda_3 = -\frac{\alpha_3(\alpha_2 + \beta_2)}{\alpha_2\beta_3 - \alpha_3\beta_2}, \quad \lambda_4 = \frac{\alpha_2(\alpha_3 + \beta_3)}{\alpha_2\beta_3 - \alpha_3\beta_2}.$$

Therefore, we have $h = \lambda_1\Phi(\mathbf{x}_1) + \lambda_2\Phi(\mathbf{x}_2) = \lambda_3\Phi(\mathbf{x}_2) + \lambda_4\Phi(\mathbf{x}_3)$. Again, from the linear independence of $\Phi(\mathbf{x}_1)$, $\Phi(\mathbf{x}_2)$, and $\Phi(\mathbf{x}_3)$, we have $\lambda_1 = 0$, $\lambda_2 = \lambda_3$, $\lambda_4 = 0$. However, $\lambda_1 = 0$ leads to $\alpha_2 = 0$.

Therefore $\Phi(\mathbf{x}_i) - g$ and $\Phi(\mathbf{x}_i) - h$ are linearly independent for some $i \in \{1, 2, 3\}$. ■

Lemma 12 Given $\mathbf{X}_1, \dots, \mathbf{X}_n$, let $\mathcal{D}_n \subset \mathcal{H}$ be defined as

$$\mathcal{D}_n = \left\{ g \left| g = \sum_{i=1}^n w_i \cdot \Phi(\mathbf{X}_i), \quad w_i \geq 0, \quad \sum_{i=1}^n w_i = 1 \right. \right\}$$

Then \mathcal{D}_n is compact.

Proof Define

$$A = \left\{ (w_1, \dots, w_n) \in \mathbb{R}^n \mid w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\},$$

and a mapping W

$$W : (w_1, \dots, w_n) \in A \rightarrow \sum_{i=1}^n w_i \cdot \Phi(\mathbf{X}_i) \in \mathcal{H}.$$

Note that A is compact, W is continuous, and \mathcal{D}_n is the image of A under W . Since the continuous image of a compact space is also compact (Munkres, 2000), \mathcal{D}_n is compact. \blacksquare

9.1 Proof of Lemma 1

We begin by calculating the Gateaux differential of J . We consider the two cases: $\Phi(\mathbf{x}) - (g + \alpha h) = \mathbf{0}$ and $\Phi(\mathbf{x}) - (g + \alpha h) \neq \mathbf{0}$.

For $\Phi(\mathbf{x}) - (g + \alpha h) \neq \mathbf{0}$,

$$\begin{aligned} & \frac{\partial}{\partial \alpha} \rho(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \\ &= \psi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \cdot \frac{\partial}{\partial \alpha} \|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}} \\ &= \psi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \cdot \frac{\partial}{\partial \alpha} \sqrt{\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}^2} \\ &= \psi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \cdot \frac{\frac{\partial}{\partial \alpha} \|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}^2}{2\sqrt{\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}^2}} \\ &= \frac{\psi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}})}{2\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}} \cdot \frac{\partial}{\partial \alpha} \left(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}^2 - 2\langle \Phi(\mathbf{x}) - g, \alpha h \rangle_{\mathcal{H}} + \alpha^2 \|h\|_{\mathcal{H}}^2 \right) \\ &= \frac{\psi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}})}{\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}} \cdot \left(-\langle \Phi(\mathbf{x}) - g, h \rangle_{\mathcal{H}} + \alpha \|h\|_{\mathcal{H}}^2 \right) \\ &= \phi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \cdot (-\langle \Phi(\mathbf{x}) - (g + \alpha h), h \rangle_{\mathcal{H}}). \end{aligned} \tag{15}$$

For $\Phi(\mathbf{x}) - (g + \alpha h) = \mathbf{0}$,

$$\begin{aligned}
 & \frac{\partial}{\partial \alpha} \rho(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \\
 &= \lim_{\delta \rightarrow 0} \frac{\rho(\|\Phi(\mathbf{x}) - (g + (\alpha + \delta)h)\|_{\mathcal{H}}) - \rho(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}})}{\delta} \\
 &= \lim_{\delta \rightarrow 0} \frac{\rho(\|\delta h\|_{\mathcal{H}}) - \rho(0)}{\delta} \\
 &= \lim_{\delta \rightarrow 0} \frac{\rho(\delta \|h\|_{\mathcal{H}})}{\delta} \\
 &= \begin{cases} \lim_{\delta \rightarrow 0} \frac{\rho(0)}{\delta}, & h = \mathbf{0} \\ \lim_{\delta \rightarrow 0} \frac{\rho(\delta \|h\|_{\mathcal{H}})}{\delta \|h\|_{\mathcal{H}}} \cdot \|h\|_{\mathcal{H}}, & h \neq \mathbf{0} \end{cases} \\
 &= 0 \\
 &= \varphi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \cdot (-\langle \Phi(\mathbf{x}) - (g + \alpha h), h \rangle_{\mathcal{H}})
 \end{aligned} \tag{16}$$

where the second to the last equality comes from (A1) and the last equality comes from the facts that $\Phi(\mathbf{x}) - (g + \alpha h) = \mathbf{0}$ and $\varphi(0)$ is well-defined by (A2).

From (15) and (16), we can conclude that for any $g, h \in \mathcal{H}$, and $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}
 & \frac{\partial}{\partial \alpha} \rho(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \\
 &= \varphi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \cdot (-\langle \Phi(\mathbf{x}) - (g + \alpha h), h \rangle_{\mathcal{H}})
 \end{aligned} \tag{17}$$

Therefore,

$$\begin{aligned}
 \delta J(g; h) &= \frac{\partial}{\partial \alpha} J(g + \alpha h) \Big|_{\alpha=0} \\
 &= \frac{\partial}{\partial \alpha} \left(\frac{1}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - (g + \alpha h)\|_{\mathcal{H}}) \right) \Big|_{\alpha=0} \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \alpha} \rho(\|\Phi(\mathbf{X}_i) - (g + \alpha h)\|_{\mathcal{H}}) \Big|_{\alpha=0} \\
 &= \frac{1}{n} \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - (g + \alpha h)\|_{\mathcal{H}}) \cdot (-\langle \Phi(\mathbf{X}_i) - (g + \alpha h), h \rangle_{\mathcal{H}}) \Big|_{\alpha=0} \\
 &= -\frac{1}{n} \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) \cdot \langle \Phi(\mathbf{X}_i) - g, h \rangle_{\mathcal{H}} \\
 &= -\left\langle \frac{1}{n} \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{X}_i) - g), h \right\rangle_{\mathcal{H}} \\
 &= -\langle V(g), h \rangle_{\mathcal{H}}.
 \end{aligned}$$

The necessary condition for g to be a minimizer of J , that is, $g = \widehat{f}_{RKDE}$, is that $\delta J(g; h) = 0$, $\forall h \in \mathcal{H}$, which leads to $V(g) = \mathbf{0}$.

9.2 Proof of Theorem 2

From Lemma 1, $V(\hat{f}_{RKDE}) = \mathbf{0}$, that is,

$$\frac{1}{n} \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - \hat{f}_{RKDE}\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{X}_i) - \hat{f}_{RKDE}) = \mathbf{0}.$$

Solving for \hat{f}_{RKDE} , we have $\hat{f}_{RKDE} = \sum_{i=1}^n w_i \Phi(\mathbf{X}_i)$ where

$$w_i = \left(\sum_{j=1}^n \varphi(\|\Phi(\mathbf{X}_j) - \hat{f}_{RKDE}\|_{\mathcal{H}}) \right)^{-1} \cdot \varphi(\|\Phi(\mathbf{X}_i) - \hat{f}_{RKDE}\|_{\mathcal{H}}).$$

Since ρ is non-decreasing, $w_i \geq 0$. Clearly $\sum_{i=1}^n w_i = 1$

9.3 Proof of Lemma 4

J is strictly convex on \mathcal{H} if for any $0 < \lambda < 1$, and $g, h \in \mathcal{H}$ with $g \neq h$

$$J(\lambda g + (1 - \lambda)h) < \lambda J(g) + (1 - \lambda)J(h).$$

Note that

$$\begin{aligned} J(\lambda g + (1 - \lambda)h) &= \frac{1}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - \lambda g - (1 - \lambda)h\|_{\mathcal{H}}) \\ &= \frac{1}{n} \sum_{i=1}^n \rho(\|\lambda(\Phi(\mathbf{X}_i) - g) + (1 - \lambda)(\Phi(\mathbf{X}_i) - h)\|_{\mathcal{H}}) \\ &\leq \frac{1}{n} \sum_{i=1}^n \rho(\lambda\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}} + (1 - \lambda)\|\Phi(\mathbf{X}_i) - h\|_{\mathcal{H}}) \\ &\leq \frac{1}{n} \sum_{i=1}^n \lambda \rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) + (1 - \lambda) \rho(\|\Phi(\mathbf{X}_i) - h\|_{\mathcal{H}}) \\ &= \lambda J(g) + (1 - \lambda)J(h). \end{aligned}$$

The first inequality comes from the fact that ρ is non-decreasing and

$$\|\lambda(\Phi(\mathbf{X}_i) - g) + (1 - \lambda)(\Phi(\mathbf{X}_i) - h)\|_{\mathcal{H}} \leq \lambda\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}} + (1 - \lambda)\|\Phi(\mathbf{X}_i) - h\|_{\mathcal{H}},$$

and the second inequality comes from the convexity of ρ .

Under condition (i), ρ is strictly convex and thus the second inequality is strict, implying J is strictly convex. Under condition (ii), we will show that the first inequality is strict using proof by contradiction. Suppose the first inequality holds with equality. Since ρ is strictly increasing, this can happen only if

$$\|\lambda(\Phi(\mathbf{X}_i) - g) + (1 - \lambda)(\Phi(\mathbf{X}_i) - h)\|_{\mathcal{H}} = \lambda\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}} + (1 - \lambda)\|\Phi(\mathbf{X}_i) - h\|_{\mathcal{H}},$$

for $i = 1, \dots, n$. Equivalently, it can happen only if $(\Phi(\mathbf{X}_i) - g)$ and $(\Phi(\mathbf{X}_i) - h)$ are linearly dependent for all $i = 1, \dots, n$. However, from $n \geq 3$ and positive definiteness of K , there exist three distinct \mathbf{X}_i 's, say $\mathbf{Z}_1, \mathbf{Z}_2$, and \mathbf{Z}_3 with positive definite $K' = (k_{\sigma}(\mathbf{Z}_i, \mathbf{Z}_j))_{i,j=1}^3$. By Lemma 11, it must be the case that for some $i \in \{1, 2, 3\}$, $(\Phi(\mathbf{Z}_i) - g)$ and $(\Phi(\mathbf{Z}_i) - h)$ are linearly independent. Therefore, the inequality is strict, and thus J is strictly convex.

9.4 Proof of Theorem 5

First, we will prove the monotone decreasing property of $J(f^{(k)})$. Given $r \in \mathbb{R}$, define

$$u(x; r) = \rho(r) - \frac{1}{2}r\psi(r) + \frac{1}{2}\phi(r)x^2.$$

If ϕ is nonincreasing, then u is a surrogate function of ρ , having the following property (Huber, 1981):

$$u(r; r) = \rho(r) \tag{18}$$

$$u(x; r) \geq \rho(x), \quad \forall x. \tag{19}$$

Define

$$Q(g; f^{(k)}) = \frac{1}{n} \sum_{i=1}^n u(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}, \|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathcal{H}}).$$

Note that since ψ and ϕ are continuous, $Q(\cdot; \cdot)$ is continuous in both arguments.

From (18) and (19), we have

$$\begin{aligned} Q(f^{(k)}; f^{(k)}) &= \frac{1}{n} \sum_{i=1}^n u(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathcal{H}}, \|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathcal{H}}) \\ &= \frac{1}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathcal{H}}) \\ &= J(f^{(k)}) \end{aligned} \tag{20}$$

and

$$\begin{aligned} Q(g; f^{(k)}) &= \frac{1}{n} \sum_{i=1}^n u(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}, \|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathcal{H}}) \\ &\geq \frac{1}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) \\ &= J(g), \quad \forall g \in \mathcal{H} \end{aligned} \tag{21}$$

The next iterate $f^{(k+1)}$ is the minimizer of $Q(g; f^{(k)})$ since

$$\begin{aligned} f^{(k+1)} &= \sum_{i=1}^n w_i^{(k)} \Phi(\mathbf{X}_i) \\ &= \sum_{i=1}^n \frac{\phi(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathcal{H}})}{\sum_{j=1}^n \phi(\|\Phi(\mathbf{X}_j) - f^{(k)}\|_{\mathcal{H}})} \Phi(\mathbf{X}_i) \\ &= \arg \min_{g \in \mathcal{H}} \sum_{i=1}^n \phi(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathcal{H}}) \cdot \|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}^2 \\ &= \arg \min_{g \in \mathcal{H}} Q(g; f^{(k)}) \end{aligned} \tag{22}$$

From (20), (21), and (22),

$$J(f^{(k)}) = Q(f^{(k)}; f^{(k)}) \geq Q(f^{(k+1)}; f^{(k)}) \geq J(f^{(k+1)})$$

and thus $J(f^{(k)})$ monotonically decreases at every iteration. Since $\{J(f^{(k)})\}_{k=1}^\infty$ is bounded below by 0, it converges.

Next, we will prove that every limit point f^* of $\{f^{(k)}\}_{k=1}^\infty$ belongs to \mathcal{S} . Since the sequence $\{f^{(k)}\}_{k=1}^\infty$ lies in the compact set \mathcal{D}_n (see Theorem 2 and Lemma 12), it has a convergent subsequence $\{f^{(k_l)}\}_{l=1}^\infty$. Let f^* be the limit of $\{f^{(k_l)}\}_{l=1}^\infty$. Again, from (20), (21), and (22),

$$\begin{aligned} Q(f^{(k_{l+1})}; f^{(k_{l+1})}) &= J(f^{(k_{l+1})}) \\ &\leq J(f^{(k_l+1)}) \\ &\leq Q(f^{(k_l+1)}; f^{(k_l)}) \\ &\leq Q(g; f^{(k_l)}) \quad , \forall g \in \mathcal{H}, \end{aligned}$$

where the first inequality comes from the monotone decreasing property of $J(f^{(k)})$. By taking the limit on the both side of the above inequality, we have

$$Q(f^*; f^*) \leq Q(g; f^*) \quad , \forall g \in \mathcal{H}.$$

Therefore,

$$\begin{aligned} f^* &= \arg \min_{g \in \mathcal{H}} Q(g; f^*) \\ &= \sum_{i=1}^n \frac{\varphi(\|\Phi(\mathbf{X}_i) - f^*\|_{\mathcal{H}})}{\sum_{j=1}^n \varphi(\|\Phi(\mathbf{X}_j) - f^*\|_{\mathcal{H}})} \Phi(\mathbf{X}_i) \end{aligned}$$

and thus

$$\sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - f^*\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{X}_i) - f^*) = \mathbf{0}.$$

This implies $f^* \in \mathcal{S}$.

Now we will prove $\|f^{(k)} - \mathcal{S}\|_{\mathcal{H}} \rightarrow 0$ by contradiction. Suppose $\inf_{g \in \mathcal{S}} \|f^{(k)} - g\|_{\mathcal{H}} \not\rightarrow 0$. Then, there exists $\varepsilon > 0$ such that $\forall K \in \mathbb{N}, \exists k > K$ with $\inf_{g \in \mathcal{S}} \|f^{(k)} - g\|_{\mathcal{H}} \geq \varepsilon$. Thus, we can construct an increasing sequence of indices $\{k_l\}_{l=1}^\infty$ such that $\inf_{g \in \mathcal{S}} \|f^{(k_l)} - g\|_{\mathcal{H}} \geq \varepsilon$ for all $l = 1, 2, \dots$. Since $\{f^{(k_l)}\}_{l=1}^\infty$ lies in the compact set \mathcal{D}_n , it has a subsequence converging to some f^\dagger , and we can choose j such that $\|f^{(k_j)} - f^\dagger\|_{\mathcal{H}} < \varepsilon/2$. Since f^\dagger is also a limit point of $\{f^{(k)}\}_{k=1}^\infty$, $f^\dagger \in \mathcal{S}$. This is a contradiction because

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

9.5 Proof of Theorem 8

Since the RKDE is given as $\hat{f}_{RKDE}(\mathbf{x}; F) = f_F(\mathbf{x})$, the influence function for the RKDE is

$$\begin{aligned} IF(\mathbf{x}, \mathbf{x}'; \hat{f}_{RKDE}, F) &= \lim_{s \rightarrow 0} \frac{\hat{f}_{RKDE}(\mathbf{x}; F_s) - \hat{f}_{RKDE}(\mathbf{x}; F)}{s} \\ &= \lim_{s \rightarrow 0} \frac{f_{F_s}(\mathbf{x}) - f_F(\mathbf{x})}{s} \end{aligned}$$

and thus we need to find $\dot{f}_F \triangleq \lim_{s \rightarrow 0} \frac{f_{F_s} - f_F}{s}$.

As we generalize the definition of RKDE from \widehat{f}_{RKDE} to f_F , the necessary condition $V(\widehat{f}_{RKDE})$ also generalizes. However, a few things must be taken care of since we are dealing with integral instead of summation. Suppose ψ and ϕ are bounded by B' and B'' , respectively. Given a probability measure μ , define

$$J_\mu(g) = \int \rho(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) d\mu(\mathbf{x}). \quad (23)$$

From (17),

$$\begin{aligned} \delta J_\mu(g; h) &= \frac{\partial}{\partial \alpha} J_\mu(g + \alpha h) \Big|_{\alpha=0} \\ &= \frac{\partial}{\partial \alpha} \int \rho(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) d\mu(\mathbf{x}) \Big|_{\alpha=0} \\ &= \int \frac{\partial}{\partial \alpha} \rho(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) d\mu(\mathbf{x}) \Big|_{\alpha=0} \\ &= \int \phi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \cdot (-\langle \Phi(\mathbf{x}) - (g + \alpha h), h \rangle_{\mathcal{H}}) d\mu(\mathbf{x}) \Big|_{\alpha=0} \\ &= - \int \phi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \cdot \langle \Phi(\mathbf{x}) - g, h \rangle_{\mathcal{H}} d\mu(\mathbf{x}) \\ &= - \int \left\langle \phi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{x}) - g), h \right\rangle_{\mathcal{H}} d\mu(\mathbf{x}). \end{aligned}$$

The exchange of differential and integral is valid (Lang, 1993) since for any fixed $g, h \in \mathcal{H}$, and $\alpha \in (-1, 1)$

$$\begin{aligned} &\left| \frac{\partial}{\partial \alpha} \rho(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \right| \\ &= \phi(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}}) \cdot |-\langle \Phi(\mathbf{x}) - (g + \alpha h), h \rangle_{\mathcal{H}}| \\ &\leq B'' \cdot \|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathcal{H}} \cdot \|h\|_{\mathcal{H}} \\ &\leq B'' \cdot (\|\Phi(\mathbf{x})\|_{\mathcal{H}} + \|g\|_{\mathcal{H}} + \|h\|_{\mathcal{H}}) \cdot \|h\|_{\mathcal{H}} \\ &\leq B'' \cdot (\tau + \|g\|_{\mathcal{H}} + \|h\|_{\mathcal{H}}) \cdot \|h\|_{\mathcal{H}} < \infty. \end{aligned}$$

Since $\phi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{x}) - g)$ is strongly integrable, that is,

$$\int \|\phi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{x}) - g)\|_{\mathcal{H}} d\mu(\mathbf{x}) \leq B' < \infty,$$

its Bochner-integral (Berlinet and Thomas-Agnan, 2004)

$$V_\mu(g) \triangleq \int \phi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{x}) - g) d\mu(\mathbf{x})$$

is well-defined. Therefore, we have

$$\begin{aligned} \delta J_\mu(g; h) &= - \left\langle \int \phi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{x}) - g) d\mu(\mathbf{x}), h \right\rangle_{\mathcal{H}} \\ &= - \langle V_\mu(g), h \rangle_{\mathcal{H}}. \end{aligned}$$

and $V_\mu(f_\mu) = \mathbf{0}$.

From the above condition for f_{F_s} , we have

$$\begin{aligned}\mathbf{0} &= V_{F_s}(f_{F_s}) \\ &= (1-s) \cdot V_F(f_{F_s}) + sV_{\delta_{\mathbf{x}'}}(f_{F_s}), \quad \forall s \in [0, 1)\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{0} &= \lim_{s \rightarrow 0} (1-s) \cdot V_F(f_{F_s}) + \lim_{s \rightarrow 0} s \cdot V_{\delta_{\mathbf{x}'}}(f_{F_s}) \\ &= \lim_{s \rightarrow 0} V_F(f_{F_s}).\end{aligned}$$

Then,

$$\begin{aligned}\mathbf{0} &= \lim_{s \rightarrow 0} \frac{1}{s} \left(V_{F_s}(f_{F_s}) - V_F(f_F) \right) \\ &= \lim_{s \rightarrow 0} \frac{1}{s} \left((1-s)V_F(f_{F_s}) + sV_{\delta_{\mathbf{x}'}}(f_{F_s}) - V_F(f_F) \right) \\ &= \lim_{s \rightarrow 0} \frac{1}{s} \left(V_F(f_{F_s}) - V_F(f_F) \right) - \lim_{s \rightarrow 0} V_F(f_{F_s}) + \lim_{s \rightarrow 0} V_{\delta_{\mathbf{x}'}}(f_{F_s}) \\ &= \lim_{s \rightarrow 0} \frac{1}{s} \left(V_F(f_{F_s}) - V_F(f_F) \right) + \lim_{s \rightarrow 0} V_{\delta_{\mathbf{x}'}}(f_{F_s}) \\ &= \lim_{s \rightarrow 0} \frac{1}{s} \left(V_F(f_{F_s}) - V_F(f_F) \right) + \lim_{s \rightarrow 0} \varphi(\|\Phi(\mathbf{x}') - f_{F_s}\|) \cdot (\Phi(\mathbf{x}') - f_{F_s}) \\ &= \lim_{s \rightarrow 0} \frac{1}{s} \left(V_F(f_{F_s}) - V_F(f_F) \right) + \varphi(\|\Phi(\mathbf{x}') - f_F\|) \cdot (\Phi(\mathbf{x}') - f_F).\end{aligned}\tag{24}$$

where the last equality comes from the facts that $f_{F_s} \rightarrow f_F$ and continuity of φ .

Let U denote the mapping $\mu \mapsto f_\mu$. Then,

$$\begin{aligned}\dot{f}_F &\triangleq \lim_{s \rightarrow 0} \frac{f_{F_s} - f_F}{s} \\ &= \lim_{s \rightarrow 0} \frac{U(F_s) - U(F)}{s} \\ &= \lim_{s \rightarrow 0} \frac{U((1-s)F + s\delta_{\mathbf{x}'}) - U(F)}{s} \\ &= \lim_{s \rightarrow 0} \frac{U(F + s(\delta_{\mathbf{x}'} - F)) - U(F)}{s} \\ &= \delta U(F; \delta_{\mathbf{x}'} - F)\end{aligned}\tag{25}$$

where $\delta U(P; Q)$ is the Gateaux differential of U at P with increment Q . The first term in (24) is

$$\begin{aligned}
 & \lim_{s \rightarrow 0} \frac{1}{s} \left(V_F(f_{F_s}) - V_F(f_F) \right) \\
 &= \lim_{s \rightarrow 0} \frac{1}{s} \left(V_F(U(F_s)) - V_F(U(F)) \right) \\
 &= \lim_{s \rightarrow 0} \frac{1}{s} \left((V_F \circ U)(F_s) - (V_F \circ U)(F) \right) \\
 &= \lim_{s \rightarrow 0} \frac{1}{s} \left((V_F \circ U)(F + s(\delta_{\mathbf{x}'} - F)) - (V_F \circ U)(F) \right) \\
 &= \delta(V_F \circ U)(F; \delta_{\mathbf{x}'} - F) \\
 &= \delta V_F(U(F); \delta U(F; \delta_{\mathbf{x}'} - F)) \\
 &= \delta V_F(f_F; \dot{f}_F)
 \end{aligned} \tag{26}$$

where we apply the chain rule of Gateaux differential, $\delta(G \circ H)(u; x) = \delta G(H(u); \delta H(u; x))$, in the second to the last equality. Although \dot{f}_F is technically not a Gateaux differential since the space of probability distributions is not a vector space, the chain rule still applies.

Thus, we only need to find the Gateaux differential of V_F . For $g, h \in \mathcal{H}$

$$\begin{aligned}
 \delta V_F(g; h) &= \lim_{s \rightarrow 0} \frac{1}{s} \left(V_F(g + s \cdot h) - V_F(g) \right) \\
 &= \lim_{s \rightarrow 0} \frac{1}{s} \left(\int \varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{x}) - g - s \cdot h) dF(\mathbf{x}) \right. \\
 &\quad \left. - \int \varphi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \cdot (\Phi(\mathbf{x}) - g) dF(\mathbf{x}) \right) \\
 &= \lim_{s \rightarrow 0} \frac{1}{s} \int \left(\varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) - \varphi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \right) \cdot (\Phi(\mathbf{x}) - g) dF(\mathbf{x}) \\
 &\quad - \lim_{s \rightarrow 0} \frac{1}{s} \int \left(\varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) \cdot s \cdot h \right) dF(\mathbf{x}) \\
 &= \int \lim_{s \rightarrow 0} \frac{1}{s} \left(\varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) - \varphi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \right) \cdot (\Phi(\mathbf{x}) - g) dF(\mathbf{x}) \\
 &\quad - h \cdot \int \lim_{s \rightarrow 0} \varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) dF(\mathbf{x}) \\
 &= - \int \left(\frac{\psi'(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \cdot \|\Phi(\mathbf{x}) - g\|_{\mathcal{H}} - \psi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}})}{\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}^2} \cdot \frac{\langle h, \Phi(\mathbf{x}) - g \rangle_{\mathcal{H}}}{\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}} \right) \\
 &\quad \cdot (\Phi(\mathbf{x}) - g) dF(\mathbf{x}) \\
 &\quad - h \cdot \int \varphi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) dF(\mathbf{x})
 \end{aligned} \tag{27}$$

where in the last equality, we use the fact

$$\frac{\partial}{\partial s} \varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) = \varphi'(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) \cdot \frac{\langle \Phi(\mathbf{x}) - g - s \cdot h, h \rangle_{\mathcal{H}}}{\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}}$$

and

$$\varphi'(x) = \frac{d}{dx} \frac{\psi(x)}{x} = \frac{\psi'(x)x - \psi(x)}{x^2}.$$

The exchange of limit and integral is valid due to the dominated convergence theorem since under the assumption that φ is bounded and Lipschitz continuous with Lipschitz constant L ,

$$|\varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|)| < \infty, \quad \forall \mathbf{x}$$

and

$$\begin{aligned} & \left\| \frac{1}{s} \left(\varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) - \varphi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) \right) \cdot (\Phi(\mathbf{x}) - g) \right\|_{\mathcal{H}} \\ &= \frac{1}{s} |\varphi(\|\Phi(\mathbf{x}) - g - s \cdot h\|_{\mathcal{H}}) - \varphi(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}})| \cdot \|\Phi(\mathbf{x}) - g\|_{\mathcal{H}} \\ &\leq \frac{1}{s} L \cdot \|s \cdot h\|_{\mathcal{H}} \cdot (\|\Phi(\mathbf{x})\|_{\mathcal{H}} + \|g\|_{\mathcal{H}}) \\ &\leq L \cdot \|h\|_{\mathcal{H}} \cdot (\|\Phi(\mathbf{x})\|_{\mathcal{H}} + \|g\|_{\mathcal{H}}) < \infty, \quad \forall \mathbf{x}. \end{aligned}$$

By combining (24), (25), (26), and (27), we have

$$\begin{aligned} & \left(\int \varphi(\|\Phi(\mathbf{x}) - f_F\|) dF \right) \cdot \dot{f}_F \\ &+ \int \left(\frac{\langle \dot{f}_F, \Phi(\mathbf{x}) - f_F \rangle_{\mathcal{H}}}{\|\Phi(\mathbf{x}) - f_F\|^3} \cdot q(\|\Phi(\mathbf{x}) - f_F\|) \cdot (\Phi(\mathbf{x}) - f_F) \right) dF(\mathbf{x}) \\ &= (\Phi(\mathbf{x}') - f_F) \cdot \varphi(\|\Phi(\mathbf{x}') - f_F\|) \end{aligned}$$

where $q(x) = x\psi'(x) - \psi(x)$.

9.6 Proof of Theorem 9

With F_n instead of F , (11) becomes

$$\begin{aligned} & \left(\frac{1}{n} \sum_{i=1}^n \varphi(\|\Phi(\mathbf{X}_i) - f_{F_n}\|) \right) \cdot \dot{f}_{F_n} \\ &+ \frac{1}{n} \sum_{i=1}^n \left(\frac{\langle \dot{f}_{F_n}, \Phi(\mathbf{X}_i) - f_{F_n} \rangle_{\mathcal{H}}}{\|\Phi(\mathbf{X}_i) - f_{F_n}\|^3} \cdot q(\|\Phi(\mathbf{X}_i) - f_{F_n}\|) \cdot (\Phi(\mathbf{X}_i) - f_{F_n}) \right) \\ &= (\Phi(\mathbf{x}') - f_{F_n}) \cdot \varphi(\|\Phi(\mathbf{x}') - f_{F_n}\|). \end{aligned} \tag{28}$$

Let $r_i = \|\Phi(\mathbf{X}_i) - f_{F_n}\|$, $r' = \|\Phi(\mathbf{x}') - f_{F_n}\|$, $\gamma = \sum_{i=1}^n \varphi(r_i)$ and

$$d_i = \langle \dot{f}_{F_n}, \Phi(\mathbf{X}_i) - f_{F_n} \rangle_{\mathcal{H}} \cdot \frac{q(r_i)}{r_i^3}.$$

Then, (28) simplifies to

$$\gamma \cdot \dot{f}_{F_n} + \sum_{i=1}^n d_i \cdot (\Phi(\mathbf{X}_i) - f_{F_n}) = n \cdot (\Phi(\mathbf{x}') - f_{F_n}) \cdot \varphi(r')$$

Since $f_{F_n} = \sum_{i=1}^n w_i \Phi(\mathbf{X}_i)$, we can see that \dot{f}_{F_n} has a form of $\sum_{i=1}^n \alpha_i \Phi(\mathbf{X}_i) + \alpha' \Phi(\mathbf{x}')$. By substituting this, we have

$$\begin{aligned} & \gamma \sum_{j=1}^n \alpha_j \Phi(\mathbf{X}_j) + \gamma \cdot \alpha' \Phi(\mathbf{x}') + \sum_{i=1}^n d_i \left(\Phi(\mathbf{X}_i) - \sum_{k=1}^n w_k \Phi(\mathbf{X}_k) \right) \\ &= n \cdot \left(\Phi(\mathbf{x}') - \sum_{k=1}^n w_k \Phi(\mathbf{X}_k) \right) \cdot \varphi(r'). \end{aligned}$$

Since K' is positive definite, $\Phi(\mathbf{X}_i)$'s and $\Phi(\mathbf{x}')$ are linearly independent (see Lemma 10). Therefore, by comparing the coefficients of the $\Phi(\mathbf{X}_j)$'s and $\Phi(\mathbf{x}')$ in both sides, we have

$$\gamma \cdot \alpha_j + d_j - w_j \cdot \left(\sum_{i=1}^n d_i \right) = -w_j \frac{\Psi(r')}{r'} \cdot n \quad (29)$$

$$\gamma \alpha' = n \cdot \varphi(r'). \quad (30)$$

From (30), $\alpha' = n\varphi(r')/\gamma$. Let $q_i = q(r_i)/r_i^3$ and $\Phi(\mathbf{X}_i) - f_{F_n} = \sum_{k=1}^n w_{k,i} \Phi(\mathbf{X}_k)$ where

$$w_{k,i} = \begin{cases} -w_k & , \quad k \neq i \\ 1 - w_k & , \quad k = i. \end{cases}$$

Then,

$$\begin{aligned} d_i &= \frac{q(r_i)}{r_i^3} \left\langle \dot{f}_{F_n}, \Phi(\mathbf{X}_i) - f_{F_n} \right\rangle_{\mathcal{H}} \\ &= q_i \left\langle \sum_{j=1}^n \alpha_j \Phi(\mathbf{X}_j) + \alpha' \Phi(\mathbf{x}'), \sum_{k=1}^n w_{k,i} \Phi(\mathbf{X}_k) \right\rangle_{\mathcal{H}} \\ &= q_i \left(\sum_{j=1}^n \sum_{k=1}^n \alpha_j w_{k,i} k_{\sigma}(\mathbf{X}_j, \mathbf{X}_k) + \alpha' \sum_{k=1}^n w_{k,i} k_{\sigma}(\mathbf{x}', \mathbf{X}_k) \right) \\ &= q_i (\mathbf{e}_i - \mathbf{w})^T K \alpha + q_i \alpha' \cdot (\mathbf{e}_i - \mathbf{w})^T \mathbf{k}' \\ &= q_i (\mathbf{e}_i - \mathbf{w})^T (K \alpha + \alpha' \mathbf{k}') \end{aligned}$$

where $K := (k_{\sigma}(\mathbf{X}_i, \mathbf{X}_j))_{i,j=1}^n$ is a kernel matrix, \mathbf{e}_i denotes the i th standard basis vector, and $\mathbf{k}' = [k_{\sigma}(\mathbf{x}', \mathbf{X}_1), \dots, k_{\sigma}(\mathbf{x}', \mathbf{X}_n)]^T$. By letting $Q = \text{diag}([q_1, \dots, q_n])$,

$$\mathbf{d} = Q \cdot (I_n - \mathbf{1} \mathbf{w}^T) (K \alpha + \alpha' \mathbf{k}').$$

Thus, (29) can be expressed in matrix-vector form,

$$\begin{aligned} & \gamma \alpha + Q \cdot (I_n - \mathbf{1} \cdot \mathbf{w}^T) (K \alpha + \alpha' \mathbf{k}') - \mathbf{w} \cdot (\mathbf{1}^T Q \cdot (I_n - \mathbf{1} \cdot \mathbf{w}^T) (K \alpha + \alpha' \mathbf{k}')) \\ &= -n \cdot \mathbf{w} \varphi(r'). \end{aligned}$$

Thus, α can be found solving the following linear system of equations,

$$\begin{aligned} & \left\{ \gamma I_n + (I_n - \mathbf{1} \cdot \mathbf{w}^T)^T Q \cdot (I_n - \mathbf{1} \cdot \mathbf{w}^T) \cdot K \right\} \alpha \\ &= -n \cdot \varphi(r') \mathbf{w} - \alpha' (I_n - \mathbf{1} \cdot \mathbf{w}^T)^T Q \cdot (I_n - \mathbf{1} \cdot \mathbf{w}^T) \mathbf{k}'. \end{aligned}$$

Therefore,

$$\begin{aligned}
 IF(\mathbf{x}, \mathbf{x}'; \hat{f}_{RKDE}, F_n) &= \left\langle \Phi(\mathbf{x}), \dot{f}_{F_n} \right\rangle_{\mathcal{H}} \\
 &= \left\langle \Phi(\mathbf{x}), \sum_{i=1}^n \alpha_i \Phi(\mathbf{X}_i) + \alpha' \Phi(\mathbf{x}') \right\rangle_{\mathcal{H}} \\
 &= \sum_{i=1}^n \alpha_i k_{\sigma}(\mathbf{x}, \mathbf{X}_i) + \alpha' k_{\sigma}(\mathbf{x}, \mathbf{x}').
 \end{aligned}$$

The condition $\lim_{s \rightarrow 0} f_{F_{n,s}} = f_{F_n}$ is implied by the strict convexity of J . Given $\mathbf{X}_1, \dots, \mathbf{X}_n$ and \mathbf{x}' , define \mathcal{D}_{n+1} as in Lemma 12. From Theorem 2, $f_{F_{n,s}}$ and f_{F_n} are in \mathcal{D}_{n+1} . With the definition in (23),

$$\begin{aligned}
 J_{F_{n,s}}(g) &= \int \rho(\|\Phi(\mathbf{x}) - g\|_{\mathcal{H}}) dF_{n,s}(\mathbf{x}) \\
 &= \frac{(1-s)}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) + s \cdot \rho(\|\Phi(\mathbf{x}') - g\|_{\mathcal{H}}).
 \end{aligned}$$

Note that $J_{F_{n,s}}$ uniformly converges to J on \mathcal{D}_{n+1} , that is, $\sup_{g \in \mathcal{D}_{n+1}} |J_{F_{n,s}}(g) - J(g)| \rightarrow 0$ as $s \rightarrow 0$, since for any $g \in \mathcal{D}_{n+1}$

$$\begin{aligned}
 &|J_{F_{n,s}}(g) - J(g)| \\
 &= \left| \frac{(1-s)}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) + s \cdot \rho(\|\Phi(\mathbf{x}') - g\|_{\mathcal{H}}) - \frac{1}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) \right| \\
 &= \frac{s}{n} \sum_{i=1}^n \rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathcal{H}}) + s \cdot \rho(\|\Phi(\mathbf{x}') - g\|_{\mathcal{H}}) \\
 &\leq \frac{s}{n} \sum_{i=1}^n \rho(2\tau) + s \cdot \rho(2\tau) \\
 &= 2s \cdot \rho(2\tau)
 \end{aligned}$$

where in the inequality we use the fact that ρ is nondecreasing and

$$\begin{aligned}
 \|\Phi(\mathbf{x}) - g\|_{\mathcal{H}} &\leq \|\Phi(\mathbf{x})\| + \|g\|_{\mathcal{H}} \\
 &\leq 2\tau.
 \end{aligned}$$

since $g \in \mathcal{D}_{n+1}$, and by the triangle inequality.

Now, let $\varepsilon > 0$ and $B_{\varepsilon}(f_{F_n}) \subset \mathcal{H}$ be the open ball centered at f_{F_n} with radius ε . Since $\mathcal{D}_{n+1}^{\varepsilon} \triangleq \mathcal{D}_{n+1} \setminus B_{\varepsilon}(f_{F_n})$ is also compact, $\inf_{g \in \mathcal{D}_{n+1}^{\varepsilon}} J(g)$ is attained by some $g^* \in \mathcal{D}_{n+1}^{\varepsilon}$ by the extreme value theorem (Adams and Franzosa, 2008). Since f_{F_n} is unique, $M_{\varepsilon} = J(g^*) - J(f_{F_n}) > 0$. For sufficiently small s , $\sup_{g \in \mathcal{D}_{n+1}} |J_{F_{n,s}}(g) - J(g)| < M_{\varepsilon}/2$ and thus

$$J(g) - \frac{M_{\varepsilon}}{2} < J_{F_{n,s}}(g) < J(g) + \frac{M_{\varepsilon}}{2}, \quad \forall g \in \mathcal{D}_{n+1}.$$

Therefore,

$$\begin{aligned}
 \inf_{g \in \mathcal{D}_{n+1}^\varepsilon} J_{F_{n,s}}(g) &> \inf_{g \in \mathcal{D}_{n+1}^\varepsilon} J(g) - \frac{M_\varepsilon}{2} \\
 &= J(g^*) - \frac{M_\varepsilon}{2} \\
 &= J(f_{F_n}) + M_\varepsilon - \frac{M_\varepsilon}{2} \\
 &= J(f_{F_n}) + \frac{M_\varepsilon}{2} \\
 &> J_{F_{n,s}}(f_{F_n}).
 \end{aligned}$$

Since the minimum of $J_{F_{n,s}}$ is not attained on $\mathcal{D}_{n+1}^\varepsilon$, $f_{F_{n,s}} \in B_\varepsilon(f_{F_n})$. Since ε is arbitrary, $\lim_{s \rightarrow 0} f_{F_{n,s}} = f_{F_n}$.

Acknowledgments

This work was supported in part by NSF Award No. 0830490 and 1047871. The authors thank an anonymous reviewer for pointing out the generalization to non-reproducing kernels.

References

- I. S. Abramson. On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics*, 10(4):1217–1223, 1982.
- C. Adams and R. Franzosa. *Introduction to Topology Pure and Applied*. Pearson Prentice Hall, New Jersey, 2008.
- A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces In Probability And Statistics*. Kluwer Academic Publishers, Norwell, 2004.
- D. Berry, K. Chaloner, and J. Geweke. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*. Wiley, New York, 1996.
- K. D. Brabanter, K. Pelckmans, J. D. Brabanter, M. Debruyne, J.A.K. Suykens, M. Hubert, and B. D. Moor. Robustness of kernel based regression: A comparison of iterative weighting schemes. *Proceedings of the 19th International Conference on Artificial Neural Networks (ICANN)*, pages 100–110, 2009.
- L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.
- Y. Chen, X. Dang, H. Peng, and H. Bart. Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):288–305, 2009.
- P. Chhabra, C. Scott, E. D. Kolaczyk, and M. Crovella. Distributed spatial anomaly detection. *Proc. IEEE Conference on Computer Communications (INFOCOM)*, pages 1705–1713, 2008.

- A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.
- D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. *IEEE International Conference on Computer Vision*, 1:438–445, 2001.
- M. Debruyne, A. Christmann, M. Hubert, and J.A.K. Suykens. Robustness and stability of reweighted kernel based regression. *Technical Report 06-09, Department of Mathematics, K.U.Leuven, Leuven, Belgium*, 2008a.
- M. Debruyne, M. Hubert, and J.A.K. Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9:2377–2400, 2008b.
- M. Debruyne, M. Hubert, and J. V. Horebeek. Detecting influential observations in kernel PCA. *Computational Statistics & Data Analysis*, 54:3007–3019, 2010.
- M. Girolami and C. He. Probability density estimation from optimally condensed data samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1253–1264, OCT 2003.
- F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.
- P. Huber. Robust estimation of a location parameter. *Ann. Math. Statist*, 35:45, 1964.
- P. Huber. *Robust Statistics*. Wiley, New York, 1981.
- M. W. Jacobson and J. A. Fessler. An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms. *IEEE Transactions on Image Processing*, 16(10):2411–2422, October 2007.
- D. Kim. *Least Squares Mixture Decomposition Estimation*. Doctoral dissertation, Dept. of Statistics, Virginia Polytechnic Inst. and State Univ., 1995.
- J. Kim and C. Scott. Robust kernel density estimation. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3381–3384, 2008.
- J. Kim and C. Scott. L_2 kernel classification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(10):1822–1831, 2010.
- J. Kim and C. Scott. On the robustness of kernel density M-estimators. *to be published, Proceedings of the Twenty-Eighth International Conference on Machine Learning (ICML)*, 2011.
- S. Lang. *Real and Functional Analysis*. Springer, New York, 1993.
- K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *J. Computational and Graphical Stat.*, 9(1):1–20, March 2000.
- L. J. Latecki, A. Lazarevic, and D. Pokrajac. Outlier detection with kernel density functions. In *Proceedings of the 5th Int. Conf. on Machine Learning and Data Mining in Pattern Recognition*, pages 61–75, Berlin, Heidelberg, 2007. Springer-Verlag.

- D. G. Luenberger. *Optimization by Vector Space Methods*. Wiley-Interscience, New York, 1997.
- R. S. G. Mahapatruni and A. Gray. CAKE: Convex adaptive kernel density estimation. In D. Dunson G. Gordon and M. Dud, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*, volume 15, pages 498–506. JMLR: W&CP, 2011.
- J. R. Munkres. *Topology*. Prentice Hall, 2000.
- B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. *Proc. Annu. Conf. Comput. Learning Theory*, pages 416–426, 2001.
- D. W. Scott. *Multivariate Density Estimation*. Wiley, New York, 1992.
- C. Scovel, D. Hush, I. Steinwart, and J. Theiler. Radial kernels and their reproducing kernel Hilbert spaces. *Journal of Complexity*, 26:641–660, 2010.
- J. Shawe-Taylor and A. N. Dolia. A framework for probability density estimation. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*,, pages 468–475., 2007.
- B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CR, New York, 1986.
- L. Song, X. Zhang, A. J. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th Int. Conf. on Machine Learning, ICML '08*, pages 992–999, New York, NY, USA, 2008. ACM.
- I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.
- G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3): 1236–1265, 1992.
- V. N. Vapnik and S. Mukherjee. Support vector method for multivariate density estimation. In *Advances in Neural Information Processing Systems*, pages 659–665. MIT Press, 2000.
- A. Wibowo. Robust kernel ridge regression based on M-estimation. *Computational Mathematics and Modeling*, 20(4), 2009.
- F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.
- L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.
- J. Zhu, S. Hoi, and M. R.-T. Lyu. Robust regularized kernel regression. *IEEE Transaction on Systems, Man, and Cybernetics. Part B: Cybernetics*,, 38(6):1639–1644, December 2008.