Kernel Methods for Classification with Irregularly Sampled and Contaminated Data

by

Joo Seuk Kim

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering : Systems)
in The University of Michigan
2011

Doctoral Committee:

Assistant Professor Clayton D. Scott, Chair
Professor Jessy W. Grizzle
Associate Professor Ji Zhu
Assistant Professor Zeeshan Syed

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER 1

Introduction

In binary classification, we are given a training sample, consisting of exemplary objects from two classes, along with their class labels. We are then presented with new objects without class labels, called the test sample, and have to assign class labels to those objects such that some measure of performance is optimized. For example, in a spam email filtering, the training sample may consist of legitimate emails and spam emails that have been received so far in an email server. The goal in this application is to design a classifier that accurately predicts whether a newly received email is spam or not.

Design of a classifier consists of two main stages. The first stage is extracting features. Feature extraction is a process that transforms input data into a list of numerical features, or feature vectors. There exist several reasons for extracting features. One is that the input data at hand may not be in a mathematically suitable form. For example, in the spam mail filtering mentioned above, email messages are hard to use directly, and thus we may build feature vectors, each of whose elements counts the frequencies of a certain word that can be found in spam emails, e.g., "free", "viagra", "insurance", and "buy", etc. Another reason for feature extraction is dimensionality reduction. In applications like image processing, the input data are usually represented as vectors with dimension $\approx 10^6$. These high dimensional data may lead to high computational and storage complexity, and sometimes degraded performance. Feature extraction method such as principal component analysis generates vectors with fewer dimension $\ll 10^6$ while the reduced dimensional feature vectors still describe the data.

The next stage is learning a classifier. Suppose that the training sample as a result of feature extraction can be expressed as $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ where $\mathbf{X}_i \in \mathbb{R}^d$ is a feature vector and $Y_i \in \{-1, +1\}$ is a class label. A classifier is a function taking $\mathbf{x} \in \mathbb{R}^d$ as input and outputting a label,

$$g : \mathbb{R}^d \to \{\pm 1\}.$$

In general, we first select a family of functions that can approximate the relation between the features and the labels, and then choose a classifier from the family of functions according to a certain criteria. For example, in Fisher's linear discriminant rule [24], the classifier is chosen from a family of linear functions

$$\{\mathbf{x} \mapsto \operatorname{sgn}(f(\mathbf{x})) \mid f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}\}.$$

such that $\mathbf{w}$ maximizes the between-class scatter while minimizing the within-class scatter.

In this thesis, we focus on kernel methods for classification [56]. Kernel classifiers are an important family of classifiers that have drawn much attention recently for their ability to represent nonlinear decision boundaries and to scale well with increasing dimension d. A kernel classifier has the form

$$g(\mathbf{x}) = \operatorname{sign}\left\{ \sum_{i=1}^{n} \alpha_i Y_i k(\mathbf{x}, \mathbf{X}_i) \right\},$$

where α_i are parameters and k is a kernel function. For example, support vector machines (without offset) have this form, as does the standard kernel density estimate (KDE) plug-in classifier. These concepts are explained in detail below.

This thesis consists of three main chapters. Each chapter provides kernel methods for classification but with different focuses.

In Chapter 2, we focus on feature extraction for a particular medical application. In the application, we have to predict whether a post-operative patient needs to be admitted to an intensive care unit (ICU). The training data consists of irregularly sampled vital sign measurements from post-operative patients in telemetry. Due to irregular sampling it is undesirable to represent the measurements as vectors, which lead us to extract features from the measurements. The experimental results show that the proposed features, when paired with kernel methods, have more discriminating power than the features used by clinicians.

Chapter 3 focuses on a one-class classification problem with contaminated data. In one-class classification, we are given only one class of objects as training sample and the goal is to determine whether new objects belong to the same class of training sample or not. We further consider the case where the training sample is contaminated such that a small fraction of data do not belong to the class, which makes the problem harder. We deal with this problem by robustly estimating the density, or the level sets, of objects belonging to the given class, from the contaminated data. To achieve this, we combine a traditional kernel density estimator (KDE) with ideas from classical M-estimation. The robustness of our proposed density estimator is demonstrated with a representer theorem, the influence function, and experimental results.

Chapter 4 focuses on learning a classifier. We propose a kernel classifier that optimizes the L_2 or integrated squared error (ISE) of a “difference of densities”. The classifier is obtained as a solution of quadratic programming with an efficient algorithm, and has a sparse representation meaning that the classifier can be expressed in terms of small fraction of training samples. We provide statistical performance guarantees for the proposed L_2 kernel classifier in the form of a finite sample oracle inequality, and strong consistency in the sense of both ISE and probability of error.

Chapter 5 provides overall conclusion and suggestions on future research.

CHAPTER 2

Predicting ICU Admission in Postoperative Patients with Possible Sepsis

Sepsis represents a major factor in morbidity and mortality in postoperative patients. Therefore, patients with possible sepsis need to be admitted to an intensive care unit (ICU) for monitoring and treatment. The systemic inflammatory response syndrome (SIRS) criteria are binary statistics used to identify patients with sepsis, and are based on four physiological variables: body temperature, heart rate, breathing rate, and white blood cell count. However, the SIRS criteria have been criticized for having reduced specificity (high false positive rate), which diminishes their utility in clinical settings. This chapter presents a feature extraction method from the same four variables, and a methodology for predicting ICU admission in postoperative patients under moderate care.

Data are obtained from 1087 post-operative patients on telemetry who had thoracic surgery, 83 of which were admitted to an ICU. We propose to extract features that capture trends and variability in the SIRS variables, and apply existing kernel methods (two-class and one-class support vector machines) on those features to predict ICU admission. Since the physiological variables of patients in moderate care are sampled irregularly, the proposed features often have missing values. We adopt the zero-imputation method to account for these missing values.

We compare the predictive power of the proposed features to those of the SIRS criteria. Performance is evaluated not just when the patients are discharged or sent to the intensive care unit (ICU), but also some number of hours in advance. The experimental results show that using the proposed features leads to improvements over the SIRS criteria.

2.1 Introduction

Sepsis refers to a systemic response arising from infection [3]. In the United States, 0.8 to 2 million patients become septic every year, 30% of which are surgical patients, and hospital mortality for sepsis patients ranges from 18% to 60% [50, 67]. The number of sepsis-related deaths has tripled over the past 20 years due to the increase in the number of sepsis cases, even though the mortality rate has decreased [67]. Because of its high mortality, post-operative surgical patients with possible sepsis are frequently admitted to an intensive care unit (ICU) from moderate care/telemetry unit for monitoring and treatment. Delay in treatment is associated with mortality. Hence, timely prediction of ICU admission is critical.

The clinical definition of sepsis is the presence of systemic inflammatory response syndrome (SIRS), together with a known or suspected infection. The phrase systemic inflammatory response syndrome was proposed to describe an inflammatory state affecting the whole body, independent of its cause. SIRS is defined as the presence of two or more of the following “SIRS criteria”:

- a body temperature greater than 38° C or less than 36° C
- a heart rate greater than 90 beats per minute
- tachypnea, manifested by a respiratory rate greater than 20 breaths per minute, or hyperventilation, as indicated by a PaCO2 of less than 32 mm Hg

- an alteration in the white blood cell count, such as a count greater than 12,000/cu mm, a count less than 4,000/cu mm, or the presence of more than 10 percent immature neutrophils

within a certain time window, e.g., during the past 24 hours [3].

SIRS criteria are widely used by physicians as a way to identify patients with possible sepsis [7,22] and their ICU admission. However, it has been criticized for having reduced specificity (high false positive rate) at acceptable sensitivities, thus limiting its use in clinical settings [7,22,62]. One possible reason would be that the SIRS criteria depend only on the highest and/or lowest values of the four SIRS variables within the window.

One of the challenges in this prediction problem is that measurements obtained from patients in moderate care are sampled *irregularly*, meaning the time between consecutive samples is not constant. Unlike patients in an intensive care unit (ICU), whose vital signs are monitored constantly, patients in moderate care are monitored based on the severity of their condition and the availability of nursing staff. Due to irregular sampling, the number of measurements differs from patient to patient, and thus it is undesirable to represent the measurements as vectors. This motivates us to extract features from the SIRS variables with more discriminating power than the SIRS criteria. Irregular sampling also leads to situations where there are too few samples available in a given time window to compute some of the proposed features. In our data, about 20% of the features are missing due to irregular sampling.

After the feature extraction, we can apply the general framework of kernel methods, which have proven to be successful in a number of applications [56]. In particular, we employ particular kernel methods known as support vector machines. Based on the maximum-margin principle, SVMs employ kernels to generating nonlinear decision boundaries, and have empirically been shown to generalize well even in the presence of many irrelevant features.

The application of machine learning methods such as logistic regression, artificial neural networks, and support vector machines to sepsis-related problems have been explored in different patient monitoring environments [34, 69]. The authors were interested in prediction of death [34] or severe sepsis [69] from sepsis patients rather than predicting sepsis itself from surgical patients. In their settings, since the patients were monitored in an ICU, the patients' vital signs were observed regularly and frequently, and therefore more conventional methods could be applied. We note that the general methodology developed here is applicable to the case of regularly sampled data.

2.2 Problem Statement

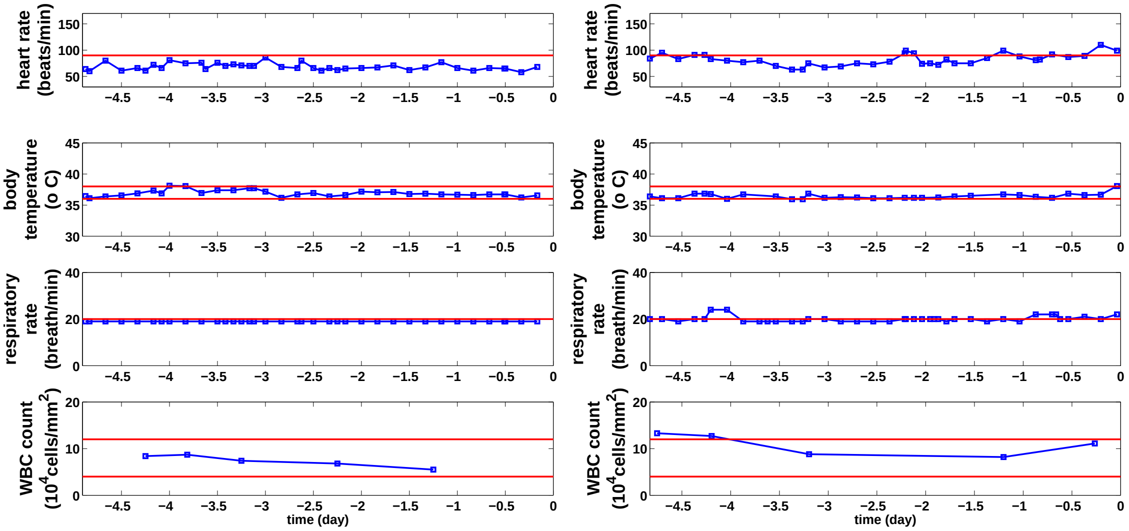
For concreteness, we first describe our motivating application and data before presenting the general methodology.

2.2.1 Concrete problem statement

Institutional review board approved this study at The University of Michigan Hospital, a large, tertiary care facility. The data used for this study are from patients who were admitted for thoracic surgery and post-operative care between 7/1/2007 and 1/30/2010. A perioperative electronic medical record (Centricity, General Electric Healthcare, Waukesha, WI) was used to identify patients who were subsequently admitted to a telemetry unit for post-operative care.

Following the SIRS criteria, the variables we used were heart rate, body temperature, respiratory rate, and white blood cell count. For convenience, we refer to the variables as vital signs, even though white blood cell count is technically not considered a vital sign. Hemodynamic and respiratory data were acquired either automatically by a validated electronic interface from the physiological monitors (General Electric Healthcare) or manually by nursing staff. All physiologic data were acquired for each telemetry unit following the admission order protocol and were validated by clinical nursing staff prior to the entry into the medical record. In step-down/telemetry units, patients are monitored only when necessary. Hence, the vital signs obtained are recorded at irregular time intervals. The vital signs were recorded up to the point where the patient was admitted to the intensive care unit (ICU) or discharged. There are 1087 post operative patients, 83 of which were admitted to the ICU.

We preprocessed the data in order to remove any obvious errors. For example, some body temperature readings had no indication of the temperature scale. If the recorded body temperature was greater than 60, we assumed that it was recorded in ° F and converted the value to ° C. We excluded patients whose recordings had none of the above mentioned vital signs, and any samples in the vital signs that did not make sense were dropped, e.g., heart rate samples equal to zero.



(a) a patient discharged

(b) a patient admitted to ICU

Fig. 2.1: Examples of vital signs recorded. Vertical lines corresponds to time (days). Horizontal lines correspond to the thresholds defining the SIRS criteria.

Figure 2.1 (a) and (b) show representative examples of vital signs of a patient discharged and a patient admitted to the ICU, respectively. In this figure, each patient's vital signs are time-shifted such that $t = 0$ corresponds to the time when he or she is discharged or admitted to the ICU. Note that the vital signs are sampled at irregular intervals and also differ in the number of samples. Our goal is to make accurate, early predictions of ICU admission.

2.2.2 Abstract problem statement

We denote the vital signs of a patient in a structured form $x = (d_1, \ldots, d_{N_d})$ and the set of all x as $\mathscr{X}$. In the cases of predicting ICU admission, the number of vital signs is $N_d = 4$. Each d_i corresponds to a vital sign of the patient and is an irregularly sampled time-series, i.e.,

$$d_i = (t_{i1}, v_{i1}, t_{i2}, v_{i2}, \ldots, t_{ip_i}, v_{ip_i})$$

for some $p_i \in \{0, 1, 2, \ldots\}$, where t_{ij} and v_{ij} represent the time and the value of the jth observed sample in d_i. Notice that within one pattern x, each d_i is obtained from irregular sampling, i.e., $(t_{i2} - t_{i1}), (t_{i3} - t_{i2}), \ldots, (t_{ip_i} - t_{i(p_i-1)})$ are typically distinct, and the number of observed samples for each d_i are different, i.e., $p_1, p_2, \ldots, p_{N_d}$ are typically distinct. Furthermore, for any two patterns x and $x' \in \mathscr{X}$, we typically have $p_i \neq p'_i$ for $i = 1, \ldots, N_d$, meaning different variables are recorded different numbers of times. The class label $y \in \{-1, 1\}$ of x is -1 if the patient corresponding to x is septic and 1 otherwise. The training data consists of labeled patients $(x_1, y_1), \ldots, (x_n, y_n)$, where in our application $n = 1004 + 83 = 1087$.

For each *training* patient, $t = 0$ corresponds to the time when he or she is admitted to the ICU or discharged. To assess the performance of early diagnosis, a test patient will be diagnosed not only when he or she is admitted to the ICU or discharged, but also

# of vital signs	N_d
the set of patterns	$\mathscr{X} = \{x : x = (d_1, \ldots, d_{N_d})\}$
vital signs	$d_i \in \mathbb{R}^{2p_i}, \quad \text{for some} \quad p_i, \quad \forall i = 1, \ldots, N_d$ $d_i = (t_{i1}, v_{i1}, t_{i2}, v_{i2}, \ldots, t_{ip_i}, v_{ip_i})$
class labels	$y \in \{-1, 1\}$

Table 2.1: Notations

some number of hours in advance. To do this, we will truncate the vital signs of a test patient beyond the time of prediction, and for this patient $t = 0$ corresponds to the time of prediction.

2.3 Proposed Feature Extraction Method

Our approach to this problem can be summarized as follows. First, we extract features from vital signs. Because of irregular sampling, some features could be missing. For example, features extracted from temperature are missing in about 3% of the patients and those from white blood cell count are missing in about 70% of the patients. In order to handle the missing data, we adopt the zero-imputation method where missing values are replaced with zero [43]. Once we obtain the imputed features, we are in a position where we can apply existing kernel-based machine learning algorithms, e.g., support vector machine.

We define our proposed features as follows. Let $\Phi_\Delta : \mathscr{X} \to \mathbb{R}^l$ denote such a feature map, which outputs a vector of length l whose elements consist of sample statistics from vital sign measurements, based on a time window of length Δ. The time window is defined as $[-\Delta, 0]$, and Φ_Δ only considers samples that are observed within the window. Samples observed outside the window are ignored. For each vital sign, the proposed features are composed of the mean, standard deviation, range, maximum positive change, maximum negative change, and slope of a line fit using least squares regression. Therefore, $l =$

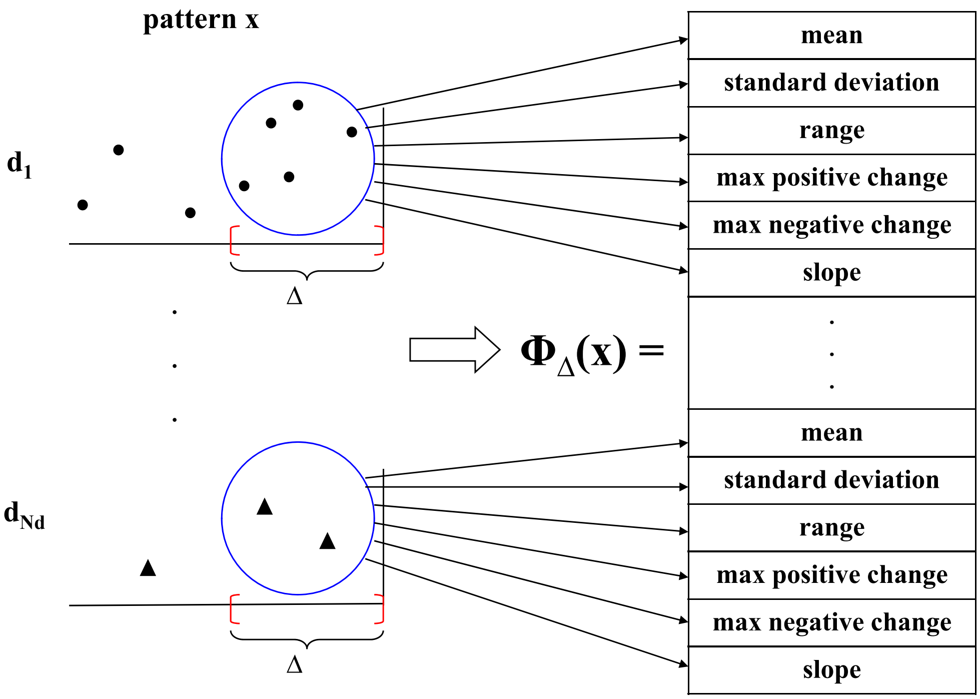


Fig. 2.2: The illustration of the proposed feature extraction process

$4 \times 6 = 24$. Figure 2.2 illustrates the procedure of the proposed feature extraction. After the extraction, we scale each feature to the range $[-1, 1]$ and impute zero for missing values.

2.4 Experiments

2.4.1 Experimental setting

Recall that there are 1087 post operative patients who had thoracic surgery, 83 of which were admitted to ICU. The data actually were obtained as two separate data: the first data set includes patients who were admitted to telemetry before 2008/10/27 and the second one includes patients after 2008/10/27. We refer to the first data set as Phase I data and the second data set as Phase II data. We also refer to patients admitted to the ICU as ICU

	# of non-ICU patients	# of ICU patients
Phase I data	486	37
Phase II data	518	46

Table 2.2: The number of ICU and non-ICU patients in the Phase I and Phase II data

patients and patient discharged as non-ICU patients. The number of ICU and non-ICU patients in the Phase I and II data is summarized in Table 2.2.

To show how the idea of introducing feature extraction works compared to SIRS criteria, we compare 3 feature sets, (a) the proposed features, (b) SIRS indicators, and (c) SIRS score. By following a clinical convention, all the features are based on vital sign observations during the last $\Delta = 24$ hour window. SIRS indicators are 4 binary variables, each of which indicates whether the corresponding condition in SIRS criteria is met. When there is no observation for a certain vital sign, the corresponding indicator is assumed to be 0. SIRS score is defined as the sum of the SIRS indicators, taking values $0-4$. Note that a diagnosis of SIRS is equivalent to SIRS score ≥ 2.

Among machine learning algorithms, we first apply the SVM with the linear kernel

$$k(x,x') = \langle \Phi_\Delta(x), \Phi_\Delta(x') \rangle$$

to the proposed features and SIRS indicators. Since SIRS score is just one variable, it is directly thresholded without any learning procedure. We also include experimental results when the OC-SVM (one class support vector machine) with Gaussian kernel

$$k(x,x') = \exp\bigl(-\|\Phi_\Delta(x) - \Phi_\Delta(x')\|^2/\sigma^2\bigr).$$

with $\sigma = 1$ is applied to the proposed features. The OC-SVM uses only the non-ICU admitted patients as training data, and is motivated by our findings in 2.4.2 below.

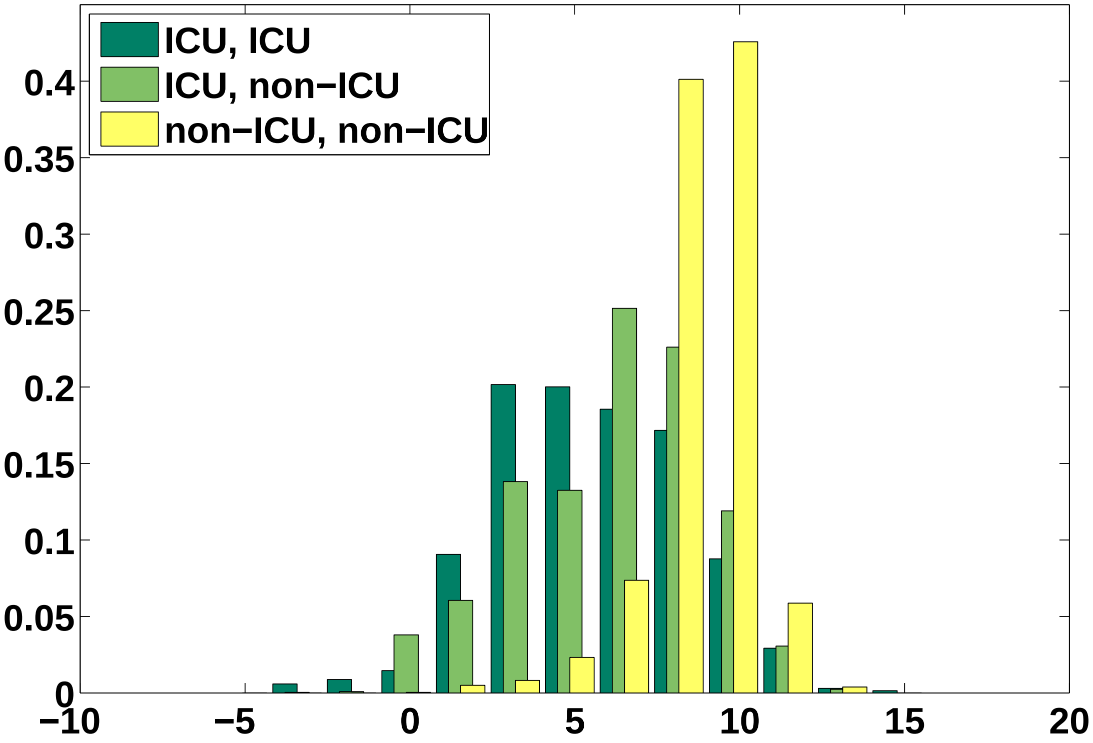


Fig. 2.3: Histogram of similarity $k(x, x')$

2.4.2 Exploratory results

A kernel can be considered as a measure of similarity between patterns expressed as an inner product in some feature space [56]. In Figure 2.3, we plot the histograms of kernel values between patients in the Phase I data (with the linear kernel) to show whether our proposed features capture the actual similarity between patients. As shown in the figure, similarities between non-ICU patients are overall lager than those between ICU patients and non-ICU patients. One interesting thing to note is the similarities between non-ICU patients. We expected that similarities between non-ICU patients would have larger values, but actually they don't. This observation can be explained as *healthy patients are alike but unhealthy patients are unhealthy in their own ways*. This motivated us to try the OC-SVM, since the ICU patients did not seem to form a homogeneous class.

2.4.3 Experimental results

We present experimental results where we train a classifier on the Phase I data and test on the Phase II data. We assess performance with the AUC (area under curve) of the ROC (receiver operating characteristic). We generate ROCs using different thresholds for the outputs of the decision function. We also assess the performances of the methods for early prediction. To do this, we truncate a certain amount of time (3, 6, and 12 hours) from vital signs of test patients.

The AUC plots are shown in Figure 2.4 (parameters are set as $C = 1.0$ for SVM and $\nu = 0.5$ for OC-SVM). We can easily see that OC-SVM with the proposed features is the best, and SVM with the proposed features is the second best. The investigation on similarity in the previous section explains why OC-SVM can provide better performance than SVM. The performances of SVM with SIRS indicator and SIRS score are similar and worse than the two other method by a significant amount. This suggests that the proposed features have more predictive power than SIRS indicators or SIRS score.

The results including other choices of C and ν are shown in Table 2.3. The results indicates that the performance does not depend much on the choices of these parameters. For all methods, AUCs decreases as we predict earlier. However, if we look at the relative performance, the AUC gain of OC-SVM with the proposed features over the methods using SIRS indicator or score is around 0.08 to 0.10 even for the early prediction.

2.5 Conclusion

In this chapter, we propose a method for predicting ICU admission for postoperative patients with possible sepsis. Our methodology is based on the feature extraction from the same physiological variables that define SIRS. Unlike the SIRS criteria, which only reflect the extreme values of vital signs in a given window, these proposed features also capture

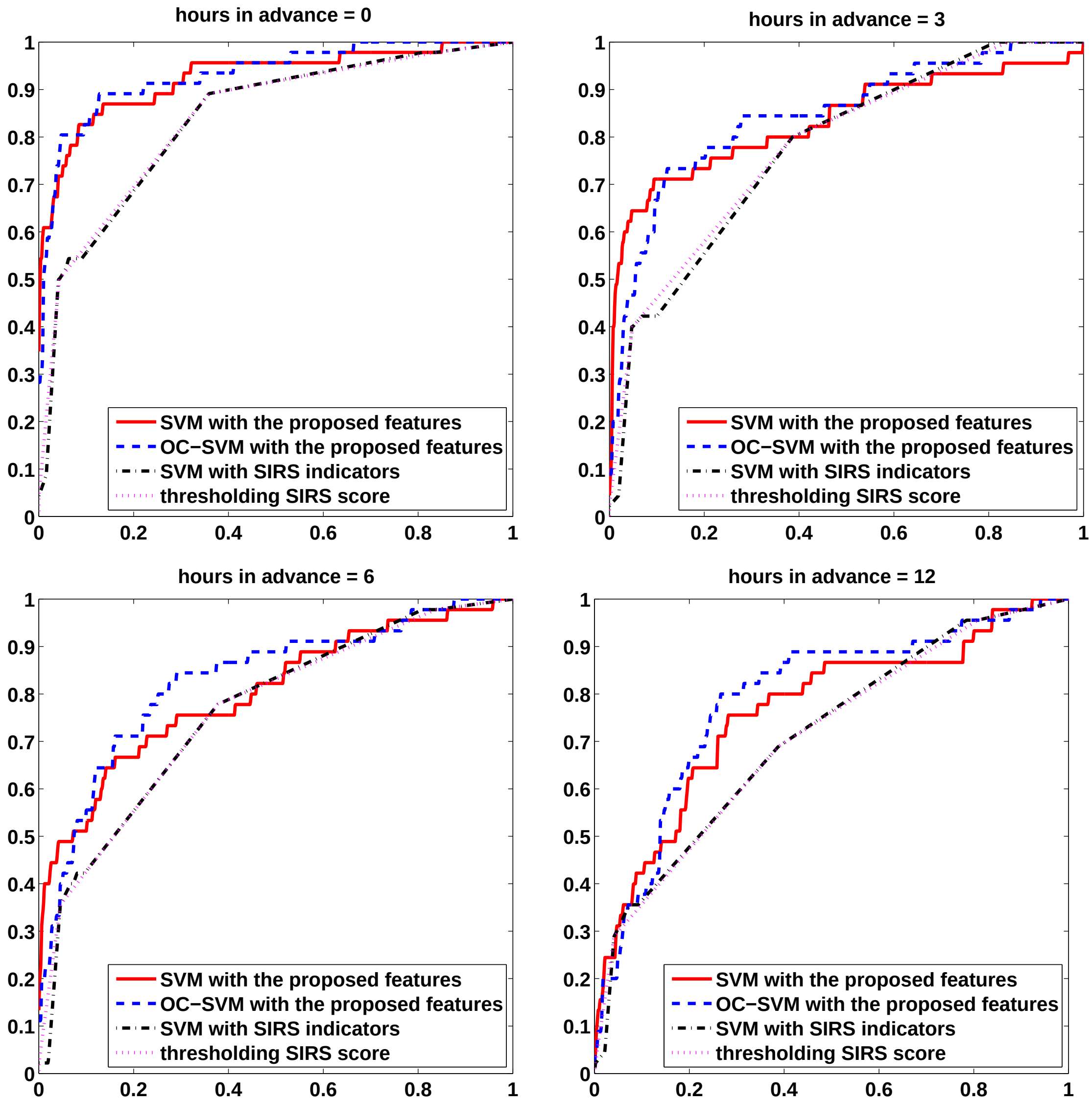


Fig. 2.4: ROC curves of the 4 methods. Parameters are set as $C = 1.0$ for SVM and $\nu = 0.5$ for OC-SVM.

the trends and variability of vital signs.

The experimental results show that the combination of the proposed features in kernel methods leads to significant improvements in predictive power compared to the more conventional SIRS score. We evaluated these methods based on their ability to predict the ICU admission several hours in advance of when the patients were actually transitioned to an ICU or discharged. For example, when making predictions for six hours in advance,

method	parameter	hours in advance			
		0	3	6	12
SVM with the proposed features	$C = 0.1$	0.93	0.81	0.81	0.74
	$C = 1.0$	0.93	0.83	0.83	0.76
	$C = 10.0$	0.90	0.82	0.81	0.74
OC-SVM with the proposed features	$\nu = 0.1$	0.93	0.86	0.84	0.79
	$\nu = 0.3$	0.93	0.85	0.84	0.79
	$\nu = 0.5$	0.93	0.84	0.83	0.79
SVM with SIRS indicators	$C = 0.1$	0.84	0.78	0.74	0.71
	$C = 1.0$	0.84	0.77	0.75	0.71
	$C = 10.0$	0.84	0.84	0.74	0.71
SIRS score	N/A	0.84	0.78	0.74	0.71

Table 2.3: AUC results on the different choices of C and ν

and assuming a specificity (false positive rate) of 20 percent, our methods achieve a true positive rate of 65 or 70 percent, while the basic SIRS criteria lead to a true positive rate of around 50 percent.

The proposed features were chosen a priori, and are intended to be simple yet general. Hand tuning of these features for particular variables may lead to further improvements.

CHAPTER 3

Robust Kernel Density Estimation

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 radial, 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.

3.1 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 [57, 60]. However, relatively little work has been done to understand or improve the KDE in situations where the training sample is contaminated. This chapter addresses a method of nonparametric density estimation that generalizes the KDE, and exhibits robustness to contamination of the training sample.

Consider training data following a contamination model

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{iid}{\sim} (1-p)f_0 + pf_1, \tag{3.1}$$

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.

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

(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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$ 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 [13]. 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 [28, 32]. The KDE based on a radial, 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 datasets 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 [1, 9, 64]. 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 [15, 42], 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 & Mukherjee [66] optimize a criterion based on the empirical cumulative distribution function over the class of weighted KDEs based on a PSD kernel. Shawe-Taylor & Dolia [59] provide a refined theoretical treatment of this approach. Song et al. [61] 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 [37]; Girolami & He [25]; Kim & Scott [38]; Mahapatruni & Gray [45] 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 [8, 14, 19, 20, 70, 75]. Robust surrogate losses for kernel-based classifiers have also been studied [74]. 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 [21]. The kernelized spatial depth was also proposed to estimate depth contours nonparametrically [12]. To our knowledge, the RKDE is the first application of M-estimation ideas in kernel density estimation.

In Section 3.2 we propose robust kernel density estimation. In Section 3.3 we present a representer theorem for the RKDE. In Section 3.4 we describe the KIRWLS algorithm and its convergence. The influence function is developed in Section 3.5, and experimental results are reported in Section 3.6. Conclusions are offered in Section 3.7. Section 3.8 contains proofs of theorems.

3.2 Robust Kernel Density Estimation

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

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

where k_σ is a kernel function with bandwidth σ. To ensure that $\widehat{f}_{KDE}(\mathbf{x})$ 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 *radial*, in that $k_\sigma(\mathbf{x}, \mathbf{x}') = g(\|\mathbf{x} - \mathbf{x}'\|^2)$ for some g.

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 $x_1, \ldots, x_m \in \mathbb{R}^d$. For radial kernels, this is equivalent to the condition that g is completely monotone, i.e.,

$$(-1)^k \frac{d^k}{dt^k} g(t) \geq 0, \quad \text{for all } k \geq 1, t > 0,$$

$$\lim_{t \to 0} g(t) = g(0),$$

and to the assumption that there exists a finite Borel measure μ on $\mathbb{R}^+ \triangleq [0,\infty)$ such that

$$k_\sigma(\mathbf{x},\mathbf{x}') = \int \exp\left(-t^2 \|\mathbf{x}-\mathbf{x}'\|^2\right) d\mu(t).$$

See [58]. Well-known examples of kernels satisfying all of the above properties are the Gaussian kernel

$$k_\sigma(\mathbf{x},\mathbf{x}') = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right), \tag{3.2}$$

the multivariate Student kernel

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

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 $\int k_\sigma(\mathbf{x},\cdot)\, d\mathbf{x} = 1$. The PSD assumption does, however, exclude several common kernels for density estimation, including those with finite support.

It is possible to associate every PSD kernel with a feature map and a Hilbert space. Although there are many ways to do this, we will consider the following canonical construction. Define $\Phi(\mathbf{x}) \triangleq k_\sigma(\cdot,\mathbf{x})$, which is called the *canonical feature map* associated with k_σ. Then define the Hilbert space of functions $\mathscr{H}$ to be the completion of the span of $\{\Phi(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d\}$. This space is known as the reproducing kernel Hilbert space (RKHS) associated with k_σ. See [63] for a thorough treatment of PSD kernels and RKHSs. For our purposes, the critical property of $\mathscr{H}$ is the so-called *reproducing property*. It states that for all $g \in \mathscr{H}$ and all $\mathbf{x} \in \mathbb{R}^d$, $g(\mathbf{x}) = \langle \Phi(\mathbf{x}), g \rangle_{\mathscr{H}}$. As a special case, taking $g = k_\sigma(\cdot, \mathbf{x}')$, we obtain

$$k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$$

for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. Therefore, the kernel evaluates the inner product of its arguments after they have been transformed by Φ.

For radial kernels, $\|\Phi(\mathbf{x})\|_{\mathscr{H}}$ is constant since

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

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

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

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

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

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

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 [32] to find a robust sample mean of the $\Phi(\mathbf{X}_i)$'s. For a robust loss function $\rho(x)$ on $x \geq 0$, the robust kernel density estimate is defined as

$$\widehat{f}_{RKDE} = \arg\min_{g \in \mathscr{H}} \sum_{i=1}^{n} \rho\left(\|\Phi(\mathbf{X}_i) - g\|_{\mathscr{H}}\right). \tag{3.3}$$

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

ρ, ψ is given by

$$\psi(x) = \begin{cases} x, & 0 \le x \le a \\ a, & a < x. \end{cases} \tag{3.4}$$

and for Hampel's ρ,

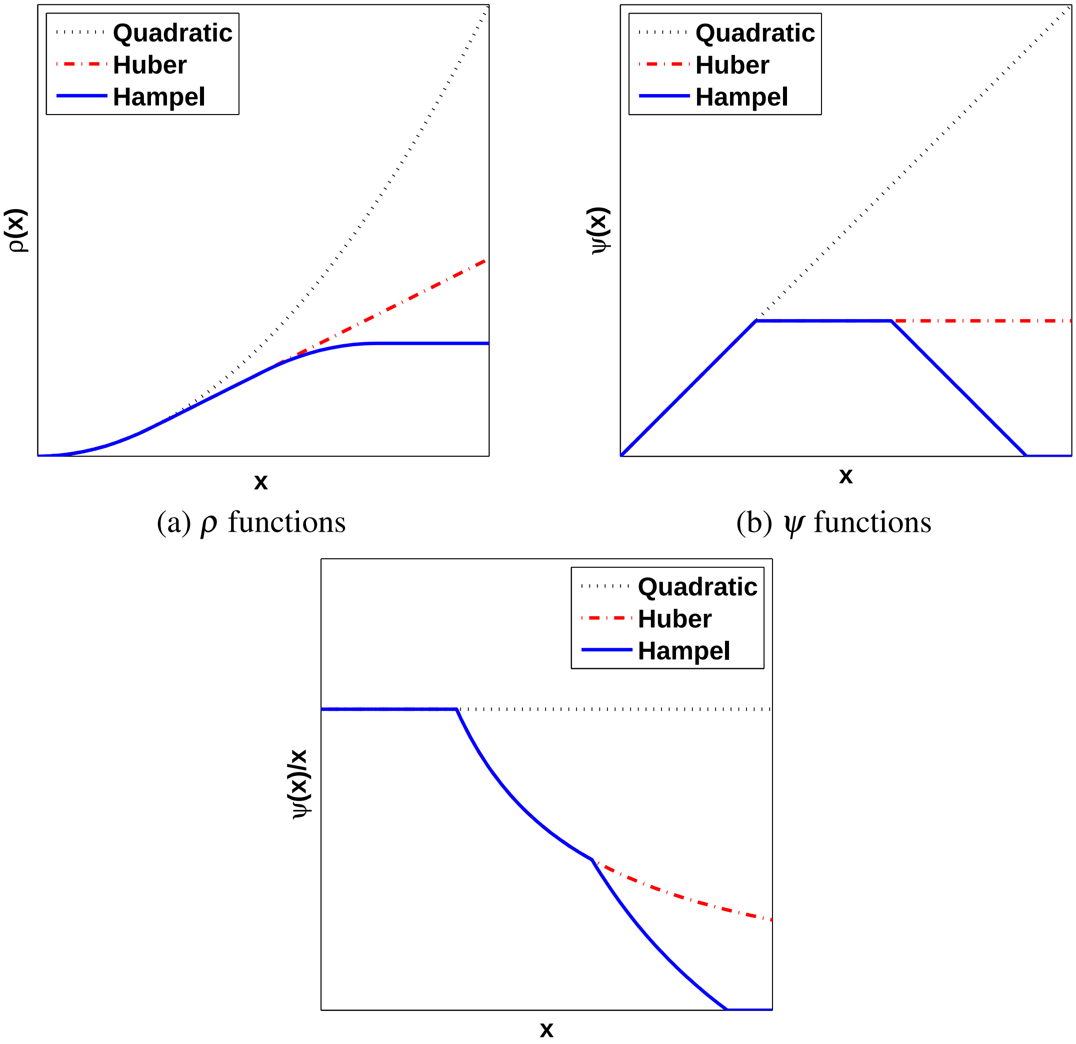
$$\psi(x) = \begin{cases} x, & 0 \le x < a \\ a, & a \le x < b \\ a \cdot (c-x)/(c-b), & b \le x < c \\ 0, & c \le x. \end{cases} \tag{3.5}$$

The functions $\rho(x), \psi(x)$, and $\psi(x)/x$ are plotted in Figure 3.1, 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 $\widehat{f}_{RKDE}$ 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 3.2 (a) shows a contour plot of a Gaussian mixture distribution on $\mathbb{R}^2$. Figure 3.2 (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 3.2 (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.

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

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



(a) ρ functions

(b) ψ functions

(c) $\psi(x)/x$

Fig. 3.1: The comparison between three different $\rho(x)$, $\psi(x)$, and $\psi(x)/x$: quadratic, Huber's, and Hampel's.

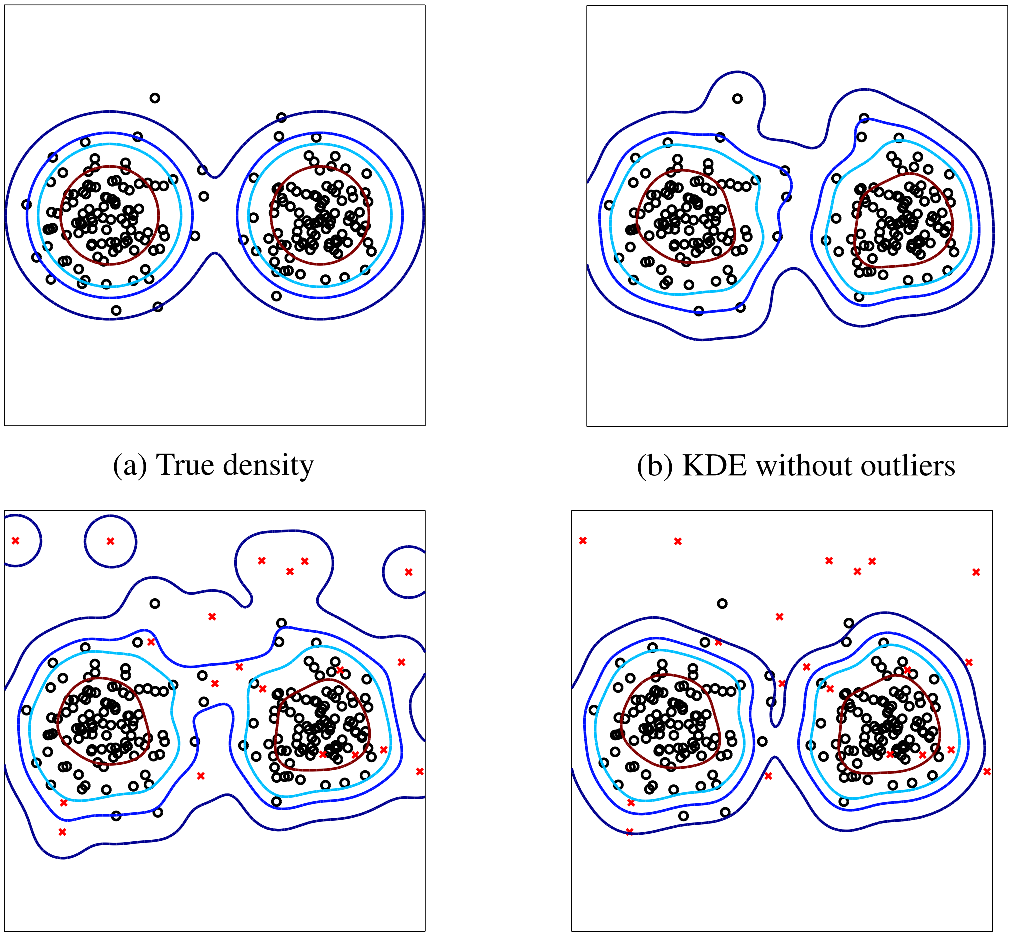
(A2) $\varphi(0) \triangleq \lim_{x\to 0} \frac{\psi(x)}{x}$ exists and is finite

(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.



(a) True density

(b) KDE without outliers

(c) KDE with outliers

(d) RKDE with outliers

Fig. 3.2: Contours of a nominal density and kernel density estimates 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.

3.3 Representer Theorem

In this section, we will describe how $\widehat{f}_{RKDE}(\mathbf{x})$ can be expressed as a weighted combination of the $k_\sigma(\mathbf{x}, \mathbf{X}_i)$'s. A formula for the weights explains how a robust sample mean in $\mathscr{H}$ translates to a robust nonparametric density estimate. We also present necessary and sufficient conditions for a function to be an RKDE. From (3.3), $\widehat{f}_{RKDE} = \arg\min_{g \in \mathscr{H}} J(g)$, where

$$J(g) = \frac{1}{n}\sum_{i=1}^{n} \rho(\|\Phi(\mathbf{X}_i) - g\|_{\mathscr{H}}). \tag{3.6}$$

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 $\mathscr{X}$ and a function $T : \mathscr{X} \to \mathbb{R}$, the Gateaux differential of T at $x \in \mathscr{X}$ with incremental $h \in \mathscr{X}$ is defined as

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

If $\delta T(x_0;h)$ is defined for all $h \in \mathscr{X}$, a necessary condition for T to have a minimum at x_0 is that $\delta T(x_0;h) = 0$ for all $h \in \mathscr{X}$ [44]. From this optimality principle, we have the following lemma.

Lemma 3.1. *Suppose assumptions (A1) and (A2) are satisfied. Then the Gateaux differential of J at $g \in \mathscr{H}$ with incremental $h \in \mathscr{H}$ is*

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

where $V : \mathscr{H} \to \mathscr{H}$ is given by

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

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

Lemma 3.1 is used to establish the following representer theorem, so named because $\widehat{f}_{RKDE}$ can be represented as a weighted combination of kernels centered at the data points. Similar results are known for supervised kernel methods [55].

Theorem 3.2. *Suppose assumptions (A1) and (A2) are satisfied. Then,*

$$\widehat{f}_{RKDE}(\mathbf{x}) = \sum_{i=1}^{n} w_i k_\sigma(\mathbf{x}, \mathbf{X}_i) \tag{3.7}$$

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

$$w_i \propto \varphi(\|\Phi(\mathbf{X}_i) - \widehat{f}_{RKDE}\|_{\mathscr{H}}). \tag{3.8}$$

It follows that $\widehat{f}_{RKDE}$ 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 $\|\Phi(\mathbf{X}_i) - \widehat{f}_{RKDE}\|_{\mathscr{H}}$ is large. Now for any $g \in \mathscr{H}$,

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

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

Theorem 3.2 provides a necessary condition for $\widehat{f}_{RKDE}$ to be the minimizer of (3.6). With an additional assumption on J, this condition is also sufficient.

Theorem 3.3. *Suppose that assumptions (A1) and (A2) are satisfied, and J is strictly convex. Then (3.7), (3.8), and $\sum_{i=1}^{n} w_i = 1$ are sufficient for $\widehat{f}_{RKDE}$ to be the minimizer of (3.6).*

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

Lemma 3.4. *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.

3.4 KIRWLS Algorithm and Its Convergence

In general, (3.3) does not have a closed form solution and $\widehat{f}_{RKDE}$ has to be found by an iterative algorithm. Fortunately, the iteratively re-weighted least squares (IRWLS) algorithm used in classical M-estimation [32] 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:

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

Intuitively, this procedure is seeking a fixed point of equations (3.7) and (3.8). The computation of $\|\Phi(\mathbf{X}_j) - f^{(k)}\|_{\mathscr{H}}$ can be done by observing

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

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

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

Recalling that $\Phi(\mathbf{x}) = k_\sigma(\cdot, \mathbf{x})$, after the kth 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 is discussed in the experimental study below.

KIRWLS can also be viewed as a kind of optimization transfer/majorize-minimize algorithm [33,41] with a quadratic surrogate for ρ. This perspective is used in our analysis in Section 3.8.4, where $f^{(k)}$ is seen to be the solution of a weighted least squares problem.

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

Theorem 3.5. *Suppose assumptions (A1) - (A3) are satisfied, and $\varphi(x)$ is nonincreasing. Let*

$$\mathscr{S} = \left\{ g \in \mathscr{H} \,\middle|\, V(g) = \mathbf{0} \right\}$$

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, $\mathscr{S} \neq \emptyset$ and

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

as $k \to \infty$.

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

Corollary 3.6. *Suppose that the assumptions in Theorem 3.5 hold and J is strictly convex. Then, $\{f^{(k)}\}_{k=1}^{\infty}$ converges to $\widehat{f}_{RKDE}$ in the $\mathscr{H}$-norm.*

This follows because under strict convexity of J, $|\mathscr{S}| = 1$.

3.5 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 [28]. The influence function (IF) for T at F is defined as

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

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 θ of Gaussian distribution is the sample mean $T(F)$,

$$T(F) = E_F[X] = \int x dF(x). \tag{3.9}$$

The influence function for $T(F)$ in (3.9) is

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

Since $|IF(x';T,F)|$ increases without bound as x' goes to $\pm\infty$, the estimator is considered 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 3.7 (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

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

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

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

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

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

$$\begin{aligned} IF(\mathbf{x},\mathbf{x}';\widehat{f}_{KDE},F) &= \lim_{s\to 0} \frac{\widehat{f}_{KDE}(\mathbf{x};F_s) - \widehat{f}_{KDE}(\mathbf{x};F)}{s} \\ &= \lim_{s\to 0} \frac{E_{F_s}[k_\sigma(\mathbf{x},\mathbf{X})] - E_F[k_\sigma(\mathbf{x},\mathbf{X})]}{s} \\ &= \lim_{s\to 0} \frac{-sE_F[k_\sigma(\mathbf{x},\mathbf{X})] + sE_{\delta_{\mathbf{x}'}}[k_\sigma(\mathbf{x},\mathbf{X})]}{s} \\ &= -E_F[k_\sigma(\mathbf{x},\mathbf{X})] + E_{\delta_{\mathbf{x}'}}[k_\sigma(\mathbf{x},\mathbf{X})] \\ &= -E_F[k_\sigma(\mathbf{x},\mathbf{X})] + k_\sigma(\mathbf{x},\mathbf{x}') \qquad (3.10) \end{aligned}$$

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

$$IF(\mathbf{x},\mathbf{x}';\widehat{f}_{KDE},F_n) = -\frac{1}{n}\sum_{i=1}^n k_\sigma(\mathbf{x},\mathbf{X}_i) + k_\sigma(\mathbf{x},\mathbf{x}'). \qquad (3.11)$$

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

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

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

Theorem 3.8. *Suppose assumptions (A1)-(A5) are satisfied. In addition, assume that $f_{F_s} \to f_F$ as $s \to 0$. If $\dot{f}_F \triangleq \lim_{s\to 0}\frac{f_{F_s}-f_F}{s}$ exists, then*

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

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

$$\begin{aligned}
&\left(\int \varphi(\|\Phi(\mathbf{x}) - f_F\|_{\mathscr{H}}) dF\right) \cdot \dot{f}_F \\
&+ \quad \int \left(\frac{\langle \dot{f}_F, \Phi(\mathbf{x}) - f_F \rangle_{\mathscr{H}}}{\|\Phi(\mathbf{x}) - f_F\|^3_{\mathscr{H}}} \cdot q(\|\Phi(\mathbf{x}) - f_F\|_{\mathscr{H}}) \cdot \big(\Phi(\mathbf{x}) - f_F\big) \right) dF(\mathbf{x}) \\
&= \quad (\Phi(\mathbf{x}') - f_F) \cdot \varphi(\|\Phi(\mathbf{x}') - f_F\|_{\mathscr{H}}). \qquad (3.12)
\end{aligned}$$

Unfortunately, for Huber or Hampel's ρ, there is no closed form solution for $\dot{f}_F$ of (3.12). However, if we work with F_n instead of F, we can find $\dot{f}_{F_n}$ explicitly. Let

$$\mathbf{1} = [1, \ldots, 1]^T,$$

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

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

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

and

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

where $\mathbf{w}$ gives the RKDE weights as in (3.7).

Theorem 3.9. *Suppose assumptions (A1)-(A5) are satisfied. In addition, assume that*

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

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

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

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

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

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

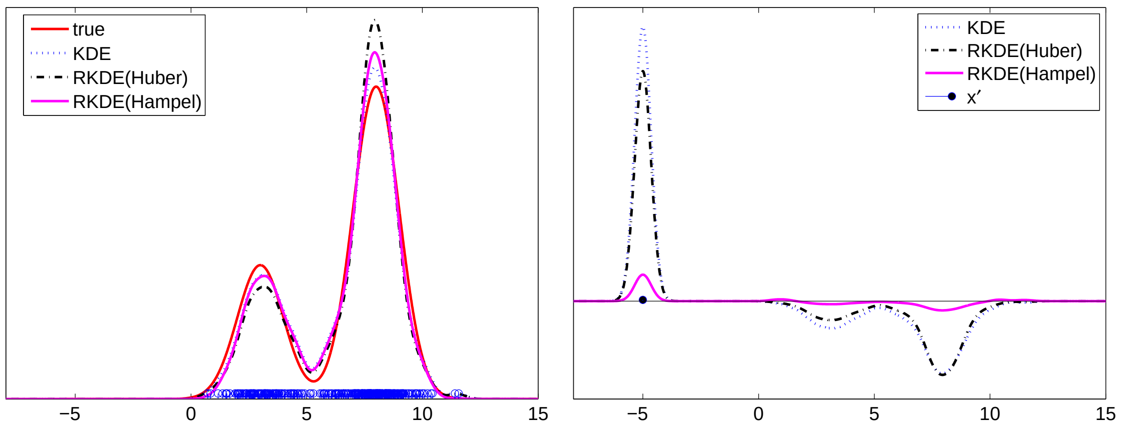
For a standard KDE, we have $\varphi \equiv 1$ and $\alpha' = 1$, in agreement with (3.11). For robust ρ, $\varphi(\|\Phi(\mathbf{x}') - f_{F_n}\|_{\mathscr{H}})$ can be viewed as a measure of "inlyingness", with more inlying points having larger values. This follows from the discussion just after Theorem 3.2. 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 $\|\mathbf{x}'\| \to \infty$. Therefore, when we compare the robustness of density estimates, we compare how close the influence functions are to the zero function.

Simulation results are shown in Figure 3.3 for a synthetic univariate distribution. Figure 3.3 (a) shows the density of the distribution, and three estimates. Figure 3.3 (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 3.6.2.

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

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



(a)

(b)

Fig. 3.3: (a) true density and density estimates. (b) IF as a function of $\mathbf{x}$ when $\mathbf{x}' = -5$

Thus $\alpha' = -\sum_{i=1}^{n} \alpha_i$ for a robust KDE. This suggests that since $\widehat{f}_{RKDE}$ 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 $IF(\mathbf{x}, \mathbf{x}'; \widehat{f}_{RKDE}, F_n)$ should be smaller overall when $\mathbf{x}'$ is an outlier. We confirm this in our experiments below.

3.6 Experiments

The experimental setup is described in 3.6.1, and results are presented in 3.6.2.

3.6.1 Experimental Setup

Data, methods, and evaluation are now discussed.

3.6.1.1 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 $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim f = (1-p) \cdot f_0 + p \cdot f_1$, 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.

3.6.1.2 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 (3.2) 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,

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

and the bandwidth σ_i is set as

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

where η is the mean of $\{\widehat{f}_{KDE}(\mathbf{X}_i)\}_{i=1}^{n}$ [1, 15]. There is another implementation of the

VKDE where σ_i is based on the distance to its k-th nearest neighbor [9]. However, this version did not perform as well and is therefore omitted.

For the RKDE, the parameters a, b, and c in (3.5) are set as follows. First, we compute $\widehat{f}_{med}$, the RKDE based on $\rho = |\cdot|$, and set $d_i = \|\Phi(\mathbf{X}_i) - \widehat{f}_{med}\|_{\mathscr{H}}$. 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 $f^{(1)} = \widehat{f}_{med}$ and terminate KIRWLS when

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

3.6.1.3 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 3.6.2. To compare a pair of methods across multiple data sets, we adopt the Wilcoxon signed-rank test [71]. 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 ith 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.

3.6.2 Experimental Results

We begin by studying influence functions.

3.6.2.1 Sensitivity using influence function

As the first measure of robustness, we compare the influence functions for KDEs and RKDEs, given in (3.11) and Theorem 3.9, 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

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

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, i.e, 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 3.1. The results clearly states that for all data sets, RKDEs are less affected by outliers than KDEs.

3.6.2.2 Kullback-Leibler (KL) divergence

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

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

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

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 3.1: 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.

The computation of D_{KL} is done as follows. Since we do not know the nominal f_0, it is estimated as $\widetilde{f}_0$, a KDE based on a separate nominal sample, obtained from the test data for each benchmark data set. Then, the integral is approximated by the sample mean, i.e.,

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

where $\{\mathbf{x}_i'\}_{i=1}^{n'}$ is an i.i.d sample from the estimated density $\widehat{f}$ with $n' = 2n = 2(n_0 + n_1)$. Note that the estimated KL divergence can have an infinite value when $\widetilde{f}_0(\mathbf{y}) = 0$ (to machine precision) and $\widehat{f}(\mathbf{y}) > 0$ 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). Table 3.2 summarizes the results.

When comparing RKDEs and KDEs, the results show that KDEs have smaller KL divergence than RKDEs with $\varepsilon = 0$. 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.

3.6.2.3 Anomaly detection

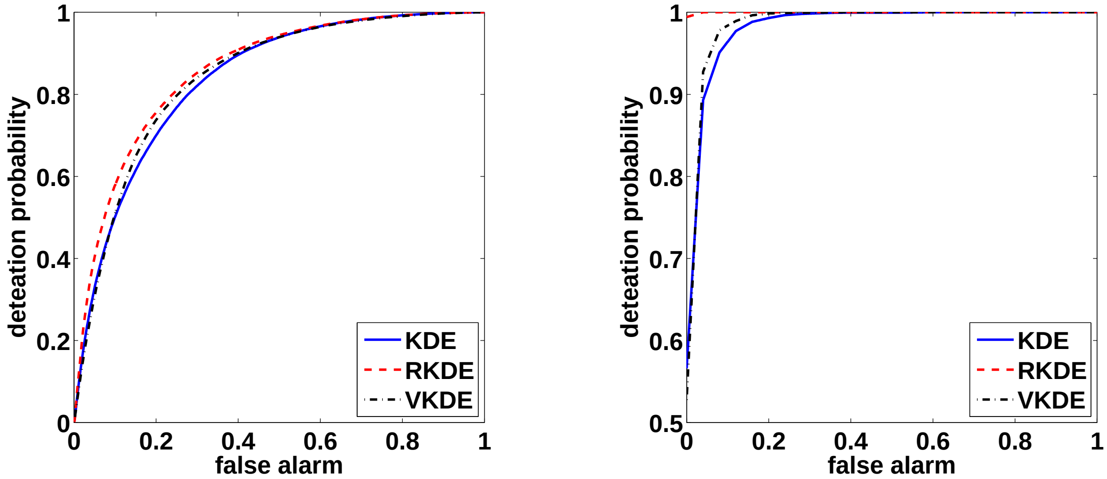
In this experiment, we apply the density estimators in anomaly detection problems. If we had a pure sample from f_0, we would estimate f_0 and use $\{\mathbf{x} : \widehat{f}_0(\mathbf{x}) > \lambda\}$ as a detector. For each λ, we could get a false negative and false positive probability using test

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 3.2: The signed-rank statistics and p-values of the Wilcoxon signed-rank test using KL divergence as a performance measure. If R_1 is larger than R_2, method 1 is better than method 2.

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 3.4. 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 3.3. The results indicate that RKDEs are significantly better than KDEs when $\varepsilon \geq 0.20$ with significance level 0.05. RKDEs are also better than VKDEs when $\varepsilon \geq 0.15$ but the difference is not significant. We also note that we have also evaluated the kernelized spatial depth (KSD) [12] 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 that either the RKDE or KDE, so those results are omitted [39].



(a) Banana, $\varepsilon = 0.2$

(b) Iris, $\varepsilon = 0.1$

Fig. 3.4: Examples of ROCs.

3.7 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 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.

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 3.3: 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.

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.

3.8 Proofs

We begin with three lemmas and proofs. The first lemma will be used in the proofs of Lemma 3.11 and Theorem 3.9, the second one in the proof of Lemma 3.4, and the third one in the proof of Theorem 3.5.

Lemma 3.10. *Let $\mathbf{z}_1,\ldots,\mathbf{z}_m$ 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

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

and from positive definiteness of K, $\alpha_1=\cdots=\alpha_m=0$. □

Lemma 3.11. *Let $\mathscr{H}$ 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 $g,h\in\mathscr{H}$ with $g\neq h$, $\Phi(\mathbf{x}_i)-g$ and $\Phi(\mathbf{x}_i)-h$ are linearly independent for some $i\in\{1,2,3\}$.*

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

$$\alpha_1(\Phi(\mathbf{x}_1)-g)+\beta_1(\Phi(\mathbf{x}_1)-h)=\mathbf{0} \tag{3.13}$$

$$\alpha_2(\Phi(\mathbf{x}_2)-g)+\beta_2(\Phi(\mathbf{x}_2)-h)=\mathbf{0} \tag{3.14}$$

$$\alpha_3(\Phi(\mathbf{x}_3)-g)+\beta_3(\Phi(\mathbf{x}_3)-h)=\mathbf{0}. \tag{3.15}$$

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, (3.13) and (3.14) simplify to

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

respectively. This is contradiction because $\Phi(\mathbf{x}_1)$, $\Phi(\mathbf{x}_2)$, and $\Phi(\mathbf{x}_3)$ are linearly independent by Lemma 3.10 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 (3.14) multiplied by α_1 from (3.13) 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 (3.14) and (3.15), $h = \lambda_3\Phi(\mathbf{x}_2) + \lambda_4\Phi(\mathbf{x}_3)$ 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 $i \in \{1, 2, 3\}$. $\square$

Lemma 3.12. *Given* $\mathbf{X}_1, \dots, \mathbf{X}_n$*, let* $\mathscr{D}_n \subset \mathscr{H}$ *be defined as*

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

Then, $\mathscr{D}_n$ *is compact.*

Proof. Define

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

and a mapping W

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

Note that A is compact, W is continuous, and $\mathscr{D}_n$ is the image of A under W. Since the continuous image of a compact space is also compact [49], $\mathscr{D}_n$ is compact. □

3.8.1 Proof of Lemma 3.1

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

$$\begin{aligned}
&\frac{\partial}{\partial \alpha} \rho\left(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\right) \\
&= \psi\big(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\big) \cdot \frac{\partial}{\partial \alpha} \|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}} \\
&= \psi\big(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\big) \cdot \frac{\partial}{\partial \alpha} \sqrt{\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}^2} \\
&= \psi\big(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\big) \cdot \frac{\frac{\partial}{\partial \alpha} \|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}^2}{2\sqrt{\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}^2}} \\
&= \frac{\psi\big(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\big)}{2\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}} \cdot \frac{\partial}{\partial \alpha} \left(\|\Phi(\mathbf{x}) - g\|_{\mathscr{H}}^2 - 2\big\langle \Phi(\mathbf{x}) - g, \alpha h \big\rangle_{\mathscr{H}} + \alpha^2 \|h\|_{\mathscr{H}}^2 \right) \\
&= \frac{\psi\big(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\big)}{\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}} \cdot \left(-\big\langle \Phi(\mathbf{x}) - g, h \big\rangle_{\mathscr{H}} + \alpha \|h\|_{\mathscr{H}}^2 \right) \\
&= \varphi\big(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\big) \cdot \big(-\big\langle \Phi(\mathbf{x}) - (g + \alpha h), h \big\rangle_{\mathscr{H}} \big). \qquad (3.16)
\end{aligned}$$

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

$$
\begin{aligned}
&\frac{\partial}{\partial \alpha} \rho\left(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\right) \\
= \quad & \lim_{\delta \to 0} \frac{\rho\left(\|\Phi(\mathbf{x}) - (g + (\alpha + \delta) h)\|_{\mathscr{H}}\right) - \rho\left(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\right)}{\delta} \\
= \quad & \lim_{\delta \to 0} \frac{\rho\left(\|\delta h\|_{\mathscr{H}}\right) - \rho\left(0\right)}{\delta} \\
= \quad & \lim_{\delta \to 0} \frac{\rho\left(\delta \|h\|_{\mathscr{H}}\right)}{\delta} \\
= \quad & \begin{cases} \lim_{\delta \to 0} \frac{\rho(0)}{\delta}, & h = \mathbf{0} \\ \lim_{\delta \to 0} \frac{\rho(\delta \|h\|_{\mathscr{H}})}{\delta \|h\|_{\mathscr{H}}} \cdot \|h\|_{\mathscr{H}}, & h \neq \mathbf{0} \end{cases} \\
= \quad & 0 \\
= \quad & \varphi\left(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\right) \cdot \left(-\left\langle \Phi(\mathbf{x}) - (g + \alpha h), h \right\rangle_{\mathscr{H}}\right) \qquad (3.17)
\end{aligned}
$$

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 (3.16) and (3.17), we can conclude that for any g, $h \in \mathscr{H}$, and $\mathbf{x} \in \mathbb{R}^d$,

$$
\begin{aligned}
&\frac{\partial}{\partial \alpha} \rho\left(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\right) \\
= \quad & \varphi\left(\|\Phi(\mathbf{x}) - (g + \alpha h)\|_{\mathscr{H}}\right) \cdot \left(-\left\langle \Phi(\mathbf{x}) - (g + \alpha h), h \right\rangle_{\mathscr{H}}\right) \qquad (3.18)
\end{aligned}
$$

Therefore,

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

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

3.8.2 Proof of Theorem 3.2

From Lemma 3.1, $V(\widehat{f}_{RKDE})=\mathbf{0}$, that is,

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

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

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

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

3.8.3 Proof of Lemma 3.4

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

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

Note that

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

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

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

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

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

for $i=1,\dots,n$. Equivalently, it can happen only if $(\Phi(\mathbf{X}_i)-g)$ and $(\Phi(\mathbf{X}_j)-h)$ 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 3.11, 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.

3.8.4 Proof of Theorem 3.5

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

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

If φ is nonincreasing, then u is a surrogate function of ρ, having the following property [31]:

$$u(r;r) = \rho(r) \tag{3.19}$$

$$u(x;r) \geq \rho(x), \quad \forall x. \tag{3.20}$$

Define

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

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

From (3.19) and (3.20), we have

$$\begin{aligned} Q(f^{(k)}; f^{(k)}) &= \frac{1}{n}\sum_{i=1}^{n} u\big(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathscr{H}}, \|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathscr{H}}\big) \\ &= \frac{1}{n}\sum_{i=1}^{n} \rho(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathscr{H}}) \\ &= J(f^{(k)}) \end{aligned} \tag{3.21}$$

and

$$\begin{aligned} Q(g; f^{(k)}) &= \frac{1}{n}\sum_{i=1}^{n} u\big(\|\Phi(\mathbf{X}_i) - g\|_{\mathscr{H}}, \|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathscr{H}}\big) \\ &\geq \frac{1}{n}\sum_{i=1}^{n} \rho\big(\|\Phi(\mathbf{X}_i) - g\|_{\mathscr{H}}) \\ &= J(g), \quad \forall g \in \mathscr{H} \end{aligned} \tag{3.22}$$

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

$$\begin{aligned} f^{(k+1)} &= \sum_{i=1}^{n} w_i^{(k)} \Phi(\mathbf{X}_i) \\ &= \sum_{i=1}^{n} \frac{\varphi(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathscr{H}})}{\sum_{j=1}^{n} \varphi(\|\Phi(\mathbf{X}_j) - f^{(k)}\|_{\mathscr{H}})} \Phi(\mathbf{X}_i) \\ &= \arg\min_{g\in\mathscr{H}} \sum_{i=1}^{n} \varphi(\|\Phi(\mathbf{X}_i) - f^{(k)}\|_{\mathscr{H}}) \cdot \|\Phi(\mathbf{X}_i) - g\|_{\mathscr{H}}^2 \\ &= \arg\min_{g\in\mathscr{H}} Q(g;f^{(k)}) \end{aligned} \tag{3.23}$$

From (3.21), (3.22), and (3.23),

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

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 $\mathscr{S}$. Since the sequence $\{f^{(k)}\}_{k=1}^{\infty}$ lies in the compact set $\mathscr{D}_n$ (see Theorem 3.2 and Lemma 3.12), 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 (3.21), (3.22), and (3.23),

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

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

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

Therefore,

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

and thus

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

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

Now we will prove $\|f^{(k)} - \mathscr{S}\|_{\mathscr{H}} \to 0$ by contradiction. Suppose $\inf_{g \in \mathscr{S}} \|f^{(k)} - g\|_{\mathscr{H}} \nrightarrow 0$. Then, there exists $\varepsilon > 0$ such that $\forall K \in \mathbb{N}$, $\exists k > K$ with $\inf_{g \in \mathscr{S}} \|f^{(k)} - g\|_{\mathscr{H}} \geq \varepsilon$. Thus, we can construct an increasing sequence of indices $\{k_l\}_{l=1}^{\infty}$ such that $\inf_{g \in \mathscr{S}} \|f^{(k_l)} - g\|_{\mathscr{H}} \geq \varepsilon$ for all $l = 1, 2, \ldots$. Since $\{f^{(k_l)}\}_{l=1}^{\infty}$ lies in the compact set $\mathscr{D}_n$, it has a subsequence converging to some $f^\dagger$, and we can choose j such that $\|f^{(k_j)} - f^\dagger\|_{\mathscr{H}} < \varepsilon/2$. Since $f^\dagger$ is also a limit point of $\{f^{(k)}\}_{k=1}^{\infty}$, $f^\dagger \in \mathscr{S}$. This is a contradiction because

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

3.8.5 Proof of Theorem 3.8

Since the RKDE is given as $\widehat{f}_{RKDE}(\mathbf{x}; F) = \langle \Phi(\mathbf{x}), f_F \rangle_{\mathscr{H}}$, the influence function for the RKDE is

$$\begin{aligned} IF(\mathbf{x}, \mathbf{x}'; \widehat{f}_{RKDE}, F) &= \lim_{s \to 0} \frac{\widehat{f}_{RKDE}(\mathbf{x}; F_s) - \widehat{f}_{RKDE}(\mathbf{x}; F)}{s} \\ &= \lim_{s \to 0} \frac{\langle \Phi(\mathbf{x}), f_{F_s} \rangle_{\mathscr{H}} - \langle \Phi(\mathbf{x}), f_F \rangle_{\mathscr{H}}}{s} \\ &= \left\langle \Phi(\mathbf{x}), \lim_{s \to 0} \frac{f_{F_s} - f_F}{s} \right\rangle_{\mathscr{H}} \end{aligned}$$

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

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

$$J_\mu(g) = \int \rho(\|\Phi(\mathbf{x}) - g\|_{\mathscr{H}})\, d\mu(\mathbf{x}). \tag{3.24}$$

From (3.18),

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

The exchange of differential and integral is valid [40] since for any fixed $g, h \in \mathscr{H}$, and $\alpha \in (-1,1)$

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

Since $\varphi(\|\Phi(\mathbf{x}) - g\|_{\mathscr{H}})\cdot\big(\Phi(\mathbf{x}) - g\big)$ is strongly integrable, i.e.,

$$\int \big\|\varphi\big(\|\Phi(\mathbf{x}) - g\|_{\mathscr{H}}\big)\cdot\big(\Phi(\mathbf{x}) - g\big)\big\|_{\mathscr{H}}\, d\mu(\mathbf{x}) \le B' < \infty,$$

its Bochner-integral [5]

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

is well-defined. Therefore, we have

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

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,

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

Then,

$$\begin{aligned}\mathbf{0} &= \lim_{s\to 0} \frac{1}{s}\Big(V_{F_s}(f_{F_s}) - V_F(f_F)\Big) \\ &= \lim_{s\to 0} \frac{1}{s}\Big((1-s)V_F(f_{F_s}) + sV_{\delta_{\mathbf{x}'}}(f_{F_s}) - V_F(f_F)\Big) \\ &= \lim_{s\to 0} \frac{1}{s}\Big(V_F(f_{F_s}) - V_F(f_F)\Big) - \lim_{s\to 0} V_F(f_{F_s}) + \lim_{s\to 0} V_{\delta_{\mathbf{x}'}}(f_{F_s}) \\ &= \lim_{s\to 0} \frac{1}{s}\Big(V_F(f_{F_s}) - V_F(f_F)\Big) + \lim_{s\to 0} V_{\delta_{\mathbf{x}'}}(f_{F_s}) \\ &= \lim_{s\to 0} \frac{1}{s}\Big(V_F(f_{F_s}) - V_F(f_F)\Big) + \lim_{s\to 0} \varphi(\|\Phi(\mathbf{x}') - f_{F_s}\|) \cdot (\Phi(\mathbf{x}') - f_{F_s}) \\ &= \lim_{s\to 0} \frac{1}{s}\Big(V_F(f_{F_s}) - V_F(f_F)\Big) + \varphi(\|\Phi(\mathbf{x}') - f_F\|) \cdot (\Phi(\mathbf{x}') - f_F). \qquad (3.25)\end{aligned}$$

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

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

$$\begin{aligned}
\dot{f}_F &\triangleq \lim_{s\to 0} \frac{f_{F_s} - f_F}{s} \\
&= \lim_{s\to 0} \frac{U(F_s) - U(F)}{s} \\
&= \lim_{s\to 0} \frac{U\big((1-s)F + s\delta_{\mathbf{x}'}\big) - U(F)}{s} \\
&= \lim_{s\to 0} \frac{U\big(F + s(\delta_{\mathbf{x}'} - F)\big) - U(F)}{s} \\
&= \delta U(F; \delta_{\mathbf{x}'} - F) \qquad (3.26)
\end{aligned}$$

where $\delta U(P;Q)$ is the Gateaux differential of U at P with increment Q. The first term in (3.25) is

$$\begin{aligned}
&\lim_{s\to 0} \frac{1}{s}\Big(V_F\big(f_{F_s}\big) - V_F\big(f_F\big)\Big) \\
&= \lim_{s\to 0} \frac{1}{s}\Big(V_F\big(U(F_s)\big) - V_F\big(U(F)\big)\Big) \\
&= \lim_{s\to 0} \frac{1}{s}\Big((V_F \circ U)\big(F_s\big) - (V_F \circ U)(F)\Big) \\
&= \lim_{s\to 0} \frac{1}{s}\Big((V_F \circ U)\big(F + s(\delta_{\mathbf{x}'} - F)\big) - (V_F \circ U)(F)\Big) \\
&= \delta(V_F \circ U)(F; \delta_{\mathbf{x}'} - F) \\
&= \delta V_F\big(U(F); \delta U(F; \delta_{\mathbf{x}'} - F)\big) \\
&= \delta V_F\big(f_F; \dot{f}_F\big) \qquad (3.27)
\end{aligned}$$

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 $g, h \in \mathscr{H}$

$$
\begin{aligned}
\delta V_F(g;h) = \lim_{s\to 0}\frac{1}{s}\Big(V_F(g+s\cdot h)-V_F(g)\Big) \\
= \;& \lim_{s\to 0}\frac{1}{s}\Big(\int \varphi(\|\Phi(\mathbf{x})-g-s\cdot h\|_{\mathscr{H}})\cdot(\Phi(\mathbf{x})-g-s\cdot h)dF(\mathbf{x}) \\
& \qquad\qquad -\int \varphi(\|\Phi(\mathbf{x})-g\|_{\mathscr{H}})\cdot(\Phi(\mathbf{x})-g)dF(\mathbf{x})\Big) \\
= \;& \lim_{s\to 0}\frac{1}{s}\int\Big(\varphi(\|\Phi(\mathbf{x})-g-s\cdot h\|_{\mathscr{H}})-\varphi(\|\Phi(\mathbf{x})-g\|_{\mathscr{H}})\Big)\cdot(\Phi(\mathbf{x})-g)dF(\mathbf{x}) \\
& -\lim_{s\to 0}\frac{1}{s}\int\Big(\varphi(\|\Phi(\mathbf{x})-g-s\cdot h\|_{\mathscr{H}})\cdot s\cdot h\Big)dF(\mathbf{x}) \\
= \;& \int\lim_{s\to 0}\frac{1}{s}\Big(\varphi(\|\Phi(\mathbf{x})-g-s\cdot h\|_{\mathscr{H}})-\varphi(\|\Phi(\mathbf{x})-g\|_{\mathscr{H}})\Big)\cdot(\Phi(\mathbf{x})-g)dF(\mathbf{x}) \\
& -h\cdot\int\lim_{s\to 0}\varphi(\|\Phi(\mathbf{x})-g-s\cdot h\|_{\mathscr{H}})dF(\mathbf{x}) \\
= \;& -\int\Bigg(\frac{\psi'(\|\Phi(\mathbf{x})-g\|_{\mathscr{H}})\cdot\|\Phi(\mathbf{x})-g\|_{\mathscr{H}}-\psi(\|\Phi(\mathbf{x})-g\|_{\mathscr{H}})}{\|\Phi(\mathbf{x})-g\|_{\mathscr{H}}^2}\cdot\frac{\langle h,\Phi(\mathbf{x})-g\rangle_{\mathscr{H}}}{\|\Phi(\mathbf{x})-g\|_{\mathscr{H}}}\Bigg) \\
& \qquad \cdot\big(\Phi(\mathbf{x})-g\big)\,dF(\mathbf{x}) \\
& -h\cdot\int\varphi(\|\Phi(\mathbf{x})-g\|_{\mathscr{H}})\,dF(\mathbf{x}) \qquad (3.28)
\end{aligned}
$$

where in the last equality, we use the fact

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

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,

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

and

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

By combining (3.25), (3.26), (3.27), and (3.28), we have

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

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

3.8.6 Proof of Theorem 3.9

With F_n instead of F, (3.12) becomes

$$\begin{aligned}
&\left(\frac{1}{n}\sum_{i=1}^{n} \varphi(\|\Phi(\mathbf{X}_i) - f_{F_n}\|) \right) \cdot \dot{f}_{F_n} \\
&+ \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\langle \dot{f}_{F_n}, \Phi(\mathbf{X}_i) - f_{F_n} \rangle_{\mathscr{H}}}{\|\Phi(\mathbf{X}_i) - f_{F_n}\|^3} \cdot q(\|\Phi(\mathbf{X}_i) - f_{F_n}\|) \cdot \big(\Phi(\mathbf{X}_i) - f_{F_n}\big) \right) \\
&= (\Phi(\mathbf{x}') - f_{F_n}) \cdot \varphi(\|\Phi(\mathbf{x}') - f_{F_n}\|). \qquad (3.29)
\end{aligned}$$

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

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

Then, (3.29) simplifies to

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

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 3.10). Therefore, by comparing the coefficients of the $\Phi(\mathbf{X}_j)$'s and $\Phi(\mathbf{x}')$ in both sides, we have

$$\gamma \cdot \alpha_j + d_j - w_j \cdot \left(\sum_{i=1}^{n} d_i \right) = -w_j \frac{\psi(r')}{r'} \cdot n \tag{3.30}$$

$$\gamma \alpha' = n \cdot \varphi(r'). \tag{3.31}$$

From (3.31), $\alpha' = n\varphi(r')/\gamma$. 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,

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

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

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

Thus, (3.30) can be expressed in matrix-vector form,

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

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

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

Therefore,

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

The condition $\lim_{s\to 0} f_{F_{n,s}} = f_{F_n}$ is implied by the strict convexity of J. Given $\mathbf{X}_1,\ldots,\mathbf{X}_n$ and $\mathbf{x}'$, define $\mathscr{D}_{n+1}$ as in Lemma 3.12. From Theorem 3.2, $f_{F_{n,s}}$ and f_{F_n} are in $\mathscr{D}_{n+1}$. With the definition in (3.24),

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

Note that $J_{F_{n,s}}$ uniformly converges to J on $\mathscr{D}_{n+1}$, i.e, $\sup_{g\in\mathscr{D}_{n+1}} |J_{F_{n,s}}(g) - J(g)| \to 0$ as

$s \to 0$, since for any $g \in \mathscr{D}_{n+1}$

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

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

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

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

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

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

Therefore,

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

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

CHAPTER 4

L_2 Kernel Classification

Nonparametric kernel methods are widely used and proven to be successful in many statistical learning problems. Well-known examples include the kernel density estimate (KDE) for density estimation and the support vector machine (SVM) for classification. We propose a kernel classifier that optimizes the L_2 or integrated squared error (ISE) of a "difference of densities". We focus on the Gaussian kernel, although the method applies to other kernels suitable for density estimation. Like a support vector machine (SVM), the classifier is sparse and results from solving a quadratic program. We provide statistical performance guarantees for the proposed L_2 kernel classifier in the form of a finite sample oracle inequality, and strong consistency in the sense of both ISE and probability of error. A special case of our analysis applies to a previously introduced ISE-based method for kernel density estimation. For dimensionality greater than 15, the basic L_2 kernel classifier performs poorly in practice. Thus, we extend the method through the introduction of a natural regularization parameter, which allows it to remain competitive with the SVM in high dimensions. Simulation results for both synthetic and real-world data are presented.

4.1 Introduction

In the binary classification problem we are given realizations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of a jointly distributed pair $(\mathbf{X}, Y)$, where $\mathbf{X} \in \mathbb{R}^d$ is a pattern and $Y \in \{-1, +1\}$ is a class label. The goal of classification is to build a classifier, i.e., a function taking $\mathbf{X}$ as input and outputting a label, such that some measure of performance is optimized. Kernel classifiers [56] are an important family of classifiers that have drawn much recent attention for their ability to represent nonlinear decision boundaries and to scale well with increasing dimension d. A kernel classifier (without offset) has the form

$$g(\mathbf{x}) = \operatorname{sgn}\left\{\sum_{i=1}^{n} \alpha_i Y_i k(\mathbf{x}, \mathbf{X}_i)\right\},$$

where α_i are parameters and k is a kernel function. For example, support vector machines (SVMs) without offset have this form [16], as does the standard kernel density estimate (KDE) plug-in rule.

In this chapter we employ an L_2 or integrated squared error (ISE) criterion to design the coefficients α_i of a kernel classifier. Like the SVM, L_2 kernel classifiers are the solutions of convex quadratic programs that can be solved efficiently using standard decomposition algorithms. In addition, the classifiers are sparse, meaning most of the coefficients $\alpha_i = 0$, which has advantages for representation and evaluation efficiency. The L_2 objective function also has appealing geometric interpretations in that it estimates a hyperplane in kernel feature space. Unlike the SVM, the most basic version of our method has no free parameters to be set by the user except perhaps the kernel bandwidth parameter. However, this basic L_2 kernel classifier is not competitive with the SVM for problems of dimensionality exceeding 15 to 20. Thus, we also extend the method to incorporate a regularization parameter, which allows it to remain competitive with the SVM in high dimensions.

We provide statistical performance guarantees for the proposed L_2 kernel classifier.

The linchpin of our analysis is a new concentration inequality bounding the deviation of a cross-validation based ISE estimate from the true ISE. This bound is then applied to prove an oracle inequality and consistency in both ISE and probability of error. In addition, as a special case of our analysis, we are able to deduce performance guarantees for the method of L_2 kernel density estimation described in [37] and [25], which has not previously been analyzed.

4.1.1 Related work

The ISE criterion has a long history in the literature on bandwidth selection for kernel density estimation [65] and more recently in parametric estimation [73]. The use of ISE for optimizing the weights of a KDE via quadratic programming was first described in [37] and later rediscovered in [25]. In [10], an ℓ_1 penalized ISE criterion was used to aggregate a finite number of pre-determined densities. Linear and convex aggregation of densities, based on an L_2 criterion, are studied in [53], where the densities are based on a finite dictionary or an independent sample. In contrast, our proposed method allows data-adaptive kernels, and does not require an independent (holdout) sample.

In classification, some connections relating SVMs and ISE are made in [36], although no new algorithms are proposed. The application of ISE based kernel method to classification problem is first studied in [29], where each class conditional density is estimated separately and plugged into the final classifier. However, our ISE criterion is a more natural choice for classification in that we directly estimate the difference of densities. It also leads to interesting geometric interpretations and relationships between our method and SVMs.

The "difference of densities" perspective has been applied to classification in other settings by several authors. In [27] and [46], a difference of densities is used to find smoothing parameters or kernel bandwidths. In [47], conditional densities are chosen among a parameterized set of densities to maximize the average (bounded) density differences. The relationship between consistency of ISE to the consistency of the probability of error is studied in [72]. Finally, Pelckmans et al. [52] considers a kernel classifier that maximizes the average (as opposed to worst-case) empirical margin. The resulting classifier amounts to an estimate of the difference of densities having uniform α_i's.

4.1.2 Organization

Section 4.2 introduces our L_2 criterion for classification and formulates the criterion as a quadratic program. Statistical performance guarantees are presented in Section 4.3. Geometric interpretations for the proposed method are provided in Section 4.4. Extension and variations of the basic method are presented in Section 4.5, including one extension that makes the method competitive in higher dimensions at the expense of an extra regularization parameter. We demonstrate experimental results in Section 4.6. Conclusions are offered in Section 4.7. Section 4.8 contains proofs of theorems.

4.2 L_2 Kernel Classification

Let $f_+(\mathbf{x})$ and $f_-(\mathbf{x})$ denote the class-conditional densities of the pattern given the label. From decision theory, the optimal classifier has the form

$$g^*(x) = \operatorname{sgn}\{f_+(\mathbf{x}) - \gamma f_-(\mathbf{x}).\}, \tag{4.1}$$

Denote the "difference of densities" (DOD) by $d_\gamma(\mathbf{x}) := f_+(\mathbf{x}) - \gamma f_-(\mathbf{x})$.

Here we view γ as a fixed parameter to be set by the user to reflect prior class probabilities and class-conditional error costs. For example, if we are interested in minimizing the probability of error, γ should be set to $\gamma^* = \frac{1-p}{p}$ where $0 < p < 1$ is the prior probabilities of the positive class. If p is unknown, we may set γ to be the natural empirical estimate for γ^*. We analyze this exact strategy in Section 4.3, and also employ in our experiments in Section 4.6.

Recall that we are given realizations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ where $\mathbf{X}_i \in \mathbb{R}^d$ is a pattern and $Y_i \in \{-1, +1\}$ is a class label. For convenience, we relabel Y so that it belongs to $\{1, -\gamma\}$ and denote $I_+ = \{i \mid Y_i = +1\}$ and $I_- = \{i \mid Y_i = -\gamma\}$. The class-conditional densities are modelled as KDEs with *variable weights* $\alpha = (\alpha_1, \ldots, \alpha_n)$,

$$\widehat{f}_+(\mathbf{x}; \alpha) = \sum_{i \in I_+} \alpha_i k_\sigma(\mathbf{x}, \mathbf{X}_i),$$

$$\widehat{f}_-(\mathbf{x}; \alpha) = \sum_{i \in I_-} \alpha_i k_\sigma(\mathbf{x}, \mathbf{X}_i)$$

with constraints $\alpha \in A$ where

$$A = \left\{ \alpha \Big| \sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i = 1, \quad \alpha_i \geq 0 \quad \forall i \right\}.$$

and

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

is the Gaussian kernel with bandwidth $\sigma > 0$. In general, σ is a tuning parameter that will need to set using standard model selection strategies, such as cross-validation. As explained in Section 4.5, other kernels besides the Gaussian also fit naturally into our framework.

We take as our goal to estimate $d_\gamma(\mathbf{x})$ directly with $\widehat{d_\gamma}(\mathbf{x}; \alpha) := \widehat{f}_+(\mathbf{x}; \alpha) - \gamma \widehat{f}_-(\mathbf{x}; \alpha)$, rather than to estimate $f_+(\mathbf{x})$ and $f_-(\mathbf{x})$ separately and "plug in" to (4.1) as in [29]. In particular, we propose to estimate α by minimizing the L_2 distance or ISE between the model $\widehat{d_\gamma}(\mathbf{x}; \alpha)$ and the truth $d_\gamma(\mathbf{x})$. The ISE associated with α is

$$\begin{aligned} ISE(\alpha) &= \| \widehat{d_\gamma}(\mathbf{x}; \alpha) - d_\gamma(\mathbf{x}) \|_{L_2}^2 \\ &= \int \left(\widehat{d_\gamma}(\mathbf{x}; \alpha) - d_\gamma(\mathbf{x}) \right)^2 d\mathbf{x} \\ &= \int \widehat{d_\gamma^2}(\mathbf{x}; \alpha)\, d\mathbf{x} - 2 \int \widehat{d_\gamma}(\mathbf{x}; \alpha)\, d_\gamma(\mathbf{x})\, d\mathbf{x} + \int d_\gamma^2(\mathbf{x})\, d\mathbf{x}. \end{aligned}$$

Since we do not know the true $d_\gamma(\mathbf{x})$, we need to estimate the second term in the above equation

$$H(\alpha) \triangleq \int \widehat{d_\gamma}(\mathbf{x};\alpha)\, d_\gamma(\mathbf{x})\, d\mathbf{x} \tag{4.2}$$

by $H_n(\alpha)$ which will be explained in detail in Section 4.2.1. Then, the empirical ISE becomes

$$\widehat{ISE}(\alpha) = \int \widehat{d_\gamma^2}(\mathbf{x};\alpha)\, d\mathbf{x} - 2H_n(\alpha) + \int d_\gamma^2(\mathbf{x})\, d\mathbf{x}. \tag{4.3}$$

Now, $\widehat{\alpha}$ is defined as

$$\widehat{\alpha} = \arg\min_{\alpha \in A} \widehat{ISE}(\alpha) \tag{4.4}$$

and the final classifier will be

$$g(\mathbf{x}) = \begin{cases} +1, & \widehat{d_\gamma}(\mathbf{x};\widehat{\alpha}) \geq 0 \\ -\gamma, & \widehat{d_\gamma}(\mathbf{x};\widehat{\alpha}) < 0. \end{cases} \tag{4.5}$$

4.2.1 Estimation of $H(\alpha)$

In this section, we propose a method of estimating $H(\alpha)$ in (4.2). The basic idea is to view $H(\alpha)$ as an expectation and estimate it using a sample average. We use a leave-one-out cross-validation (LOOCV) estimator, which is unbiased and facilitates our theoretical analysis. Note that the DOD can be expressed as

$$\widehat{d_\gamma}(\mathbf{x};\alpha) = \widehat{f_+}(\mathbf{x}) - \gamma \widehat{f_-}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i Y_i k_\sigma(\mathbf{x}, \mathbf{X}_i).$$

Then,

$$\begin{aligned} H(\alpha) &= \int \widehat{d_\gamma}(\mathbf{x};\alpha)\, d_\gamma(\mathbf{x})\, d\mathbf{x} \\ &= \int \widehat{d_\gamma}(\mathbf{x};\alpha)\, f_+(\mathbf{x})\, d\mathbf{x} - \gamma \int \widehat{d_\gamma}(\mathbf{x};\alpha)\, f_-(\mathbf{x})\, d\mathbf{x} \\ &= \int \sum_{i=1}^{n} \alpha_i Y_i k_\sigma(\mathbf{x},\mathbf{X}_i)\, f_+(\mathbf{x})\, d\mathbf{x} - \gamma \int \sum_{i=1}^{n} \alpha_i Y_i k_\sigma(\mathbf{x},\mathbf{X}_i)\, f_-(\mathbf{x})\, d\mathbf{x} \\ &= \sum_{i=1}^{n} \alpha_i Y_i h(\mathbf{X}_i) \end{aligned}$$

where

$$h(\mathbf{X}_i) \triangleq \int k_\sigma(\mathbf{x},\mathbf{X}_i)\, f_+(\mathbf{x})\, d\mathbf{x} - \gamma \int k_\sigma(\mathbf{x},\mathbf{X}_i)\, f_-(\mathbf{x})\, d\mathbf{x}. \tag{4.6}$$

We estimate each $h(\mathbf{X}_i)$ in (4.6) for $i = 1,\ldots,n$ using leave-one-out cross-validation

$$\widehat{h_i} \triangleq \begin{cases} \dfrac{1}{N_+ - 1} \displaystyle\sum_{j \in I_+, j \neq i} k_\sigma(\mathbf{X}_j,\mathbf{X}_i) - \dfrac{\gamma}{N_-} \displaystyle\sum_{j \in I_-} k_\sigma(\mathbf{X}_j,\mathbf{X}_i), & i \in I_+ \\ \dfrac{1}{N_+} \displaystyle\sum_{j \in I_+} k_\sigma(\mathbf{X}_j,\mathbf{X}_i) - \dfrac{\gamma}{N_- - 1} \displaystyle\sum_{j \in I_-, j \neq i} k_\sigma(\mathbf{X}_j,\mathbf{X}_i), & i \in I_- \end{cases}$$

where $N_+ = |I_+|, N_- = |I_-|$. Then, the estimate of $H(\alpha)$ is $H_n(\alpha) = \sum_{i=1}^{n} \alpha_i Y_i \widehat{h_i}$. We emphasize that here cross-validation is employed as a method of estimation, and is distinct from any procedure that may be used for tuning the bandwidth σ.

4.2.2 Optimization

The optimization problem (4.4) can be formulated as a quadratic program. The first term in (4.3) is

$$\begin{aligned} \int \widehat{d_\gamma^2}(\mathbf{x};\alpha)\, d\mathbf{x} &= \int \left(\sum_{i=1}^{n} \alpha_i Y_i k_\sigma(\mathbf{x},\mathbf{X}_i) \right)^2 d\mathbf{x} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j \int k_\sigma(\mathbf{x},\mathbf{X}_i)\, k_\sigma(\mathbf{x},\mathbf{X}_j)\, d\mathbf{x} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j k_{\sqrt{2}\sigma}(\mathbf{X}_i,\mathbf{X}_j) \end{aligned}$$

by the convolution theorem for Gaussian kernels [68]. As we have seen in Section 4.2.1, the second term $H_n(\alpha)$ in (4.3) is linear in α and can be expressed as $\sum_{i=1}^{n} \alpha_i c_i$ where $c_i = Y_i\widehat{h}_i$. Finally, since the third term does not depend on α, the optimization problem (4.4) becomes the following quadratic program (QP)

$$\widehat{\alpha} = \arg\min_{\alpha \in A} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j k_{\sqrt{2}\sigma}\left(\mathbf{X}_i, \mathbf{X}_j\right) - \sum_{i=1}^{n} c_i \alpha_i .$$

We refer to the resulting classifier as L2QP (L2 classification via Quadratic Programming). Since the Gaussian kernel is positive definite [56], the objective function in (4.2.3) is strictly convex if the $\mathbf{X}_i$'s are distinct, and thus has a unique solution. As discussed in [25], quadratic programs derived from ISE-based criteria induce sparse solutions, and the nonzero α_i's tend to be concentrated in regions of space with greater probability mass. Another explanation of this is presented in Section 4.4. The QP (4.2.3) is similar in some respects to the dual QP of the 2-norm SVM with hinge loss [16]. However, unlike the SVM, (4.2.3) does not include a regularization parameter, and therefore the computational cost required for training the L2QP classifier will typically be less than that of the SVM. The QP can be solved by a variant of the Sequential Minimal Optimization (SMO) algorithm [17].

4.2.3 SMO algorithm

Sequential Minimal Optimization (SMO) is a simple algorithm that can quickly solve the SVM QP problem without any extra matrix storage and without using time-consuming numerical QP optimization steps [17]. SMO decomposes the overall QP problem into the smallest possible optimization problem. This sub-problem can be solved analytically. An appropriate variant of SMO to solve (4.2.3) is detailed below following [25].

Given α, the algorithm optimizes two variables of α with other variables fixed. Two variables to be optimized should be chosen from $\{\alpha_i \mid i \in I_-\}$ or $\{\alpha_i \mid i \in I_+\}$. Otherwise, the variables which we are trying to optimize cannot change since the other variables are fixed and due to the constraints $\sum_{i\in I_-}\alpha_i = 1$ and $\sum_{i\in I_+}\alpha_i = 1$. Suppose that we choose two variables from $\{\alpha_i \mid i \in I_+\}$. For notational convenience, assume the two variables are α_1 and α_2 and $1,2 \in I_+$. Then, (4.2.3) reduces to

$$\begin{aligned} \min_{\alpha_1,\alpha_2} \quad & \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\alpha_i\alpha_j Q_{ij} + \sum_{i=1}^{2} d_i\alpha_i + D \\ \text{s.t} \quad & \alpha_1,\alpha_2 \geq 0, \quad \sum_{i=1}^{2}\alpha_i = \Delta \end{aligned}$$

where $D = \frac{1}{2}\sum_{i=3}^{n}\sum_{j=3}^{n}\alpha_i\alpha_j Q_{ij} - \sum_{i=3}^{n} c_i\alpha_i$ and

$$d_i = \sum_{j=3}^{n}\alpha_j Q_{ij} - c_i, \quad \Delta = 1 - \sum_{i\in I_+\setminus\{1,2\}}\alpha_i.$$

We discard D, which is independent of α_1 and α_2, and eliminate α_1 to obtain

$$\begin{aligned} \min_{\alpha_2} \quad & \frac{1}{2}(\Delta-\alpha_2)^2 Q_{11} + \alpha_2(\Delta-\alpha_2)Q_{12} + \frac{1}{2}\alpha_2^2 Q_{22} + (\Delta-\alpha_2)d_1 + \alpha_2 d_2 \qquad (4.7) \\ \text{s.t} \quad & 0 \leq \alpha_2 \leq \Delta. \end{aligned}$$

Since the objective function is quadratic and convex in one variable α_2, we can take the derivative of (4.7) and set it equal to zero. Then,

$$\alpha_2 = \frac{\Delta(Q_{11}-Q_{12}) + d_1 - d_2}{Q_{11} - 2Q_{12} + Q_{22}}. \qquad (4.8)$$

Let α^* denote the value before the optimization step. If we define $O_i := Q_{i1}\alpha_1^* + Q_{i2}\alpha_2^* + d_i = \sum_{j=1}^{n}\alpha_i^* Q_{ij} - c_i$, then (4.8) can be expressed as the update equation

$$\alpha_2 = \alpha_2^* + \frac{O_1 - O_2}{Q_{11} - 2Q_{12} + Q_{22}}. \qquad (4.9)$$

If α_2 is outside $[0,\Delta]$, we truncate it so that it is within $[0,\Delta]$. After finding α_2, α_1 can be recovered from $\alpha_1 = \Delta - \alpha_2$.

The optimality condition and the choice of α_i's can be found in the following way. There are three cases when choosing α_1 and α_2 : (a) Both are zero, (b) One is positive and the other is zero, (c) Both are positive.

Case (a): α_1 and α_2 are not updated because of nonnegativity constraints.

Case (b): Assume that α_2 is zero. From (4.9), α_2 is updated only when $O_1 - O_2 > 0$ and so is α_1

Case (c): α_1 and α_2 are updated only when $O_1 \neq O_2$.

The objective value will strictly decrease if and only if α_1 and α_2 are updated after optimization step. Therefore, the optimal solution should satisfy

$$O_i \geq O_j \quad \text{for} \quad \alpha_i = 0, \alpha_j > 0 \tag{4.10}$$

$$O_i = O_j \quad \text{for} \quad \alpha_i, \alpha_j > 0. \tag{4.11}$$

The convergence to the global minimum is thus guaranteed by choosing two α_i's which do not satisfy (4.10) or (4.11) for each optimization step. The optimization procedure for two variables from $\{\alpha_i \in I_-\}$ is similar.

4.3 Statistical Performance Analysis

We give theoretical performance guarantees for our proposed method. We assume that $\{\mathbf{X}_i\}_{i \in I_+}$ and $\{\mathbf{X}_i\}_{i \in I_-}$ are i.i.d samples from $f_+(\mathbf{x})$ and $f_-(\mathbf{x})$, respectively, and treat N_+ and N_- as deterministic variables n_+ and n_- such that $n_+ \to \infty$ and $n_- \to \infty$ as $n \to \infty$. Proofs are found in Section 4.8.

4.3.1 Concentration inequality for $H_n(\alpha)$

Lemma 4.1. *Conditioned on $\mathbf{X}_i$, $\widehat{h}_i$ is an unbiased estimator of $h(\mathbf{X}_i)$, i.e,*

$$\mathbf{E}\left[\widehat{h}_i \middle| \mathbf{X}_i\right] = h(\mathbf{X}_i).$$

Furthermore, for any $\varepsilon > 0$

$$\mathbf{P}\left\{\sup_{\alpha \in A} \left|H_n(\alpha) - H(\alpha)\right| > \varepsilon\right\} \leq 2n\left(e^{-c(n_+ - 1)\varepsilon^2} + e^{-c(n_- - 1)\varepsilon^2}\right)$$

where $c = 2\left(\sqrt{2\pi}\sigma\right)^{2d} / (1+\gamma)^4$.

Lemma 4.1 implies that $H_n(\alpha) \to H(\alpha)$ almost surely for all $\alpha \in A$ simultaneously, provided that σ, n_+ and n_- evolve as functions of n such that $n_+\sigma^{2d}/\ln n \to \infty$ and $n_-\sigma^{2d}/\ln n \to \infty$.

4.3.2 Oracle inequality

Next, we establish on oracle inequality, which relates the performance of our estimator to that of the best possible kernel classifier.

Theorem 4.2. *Let* $\varepsilon > 0$ *and set* $\delta = \delta(\varepsilon) = 2n\left(e^{-c(n_+-1)\varepsilon^2} + e^{-c(n_--1)\varepsilon^2}\right)$ *where* $c = 2\left(\sqrt{2\pi}\sigma\right)^{2d} / (1+\gamma)^4$. *Then, with probability at least* $1-\delta$

$$ISE(\widehat{\alpha}) \leq \inf_{\alpha \in A} ISE(\alpha) + 4\varepsilon.$$

Proof. From Lemma 4.1, with probability at least $1-\delta$

$$\left| ISE(\alpha) - \widehat{ISE}(\alpha) \right| \leq 2\varepsilon, \quad \forall \alpha \in A$$

by using the fact $ISE(\alpha) - \widehat{ISE}(\alpha) = 2(H_n(\alpha) - H(\alpha))$. Then, with probability at least $1-\delta$, for all $\alpha \in A$, we have

$$ISE(\widehat{\alpha}) \leq \widehat{ISE}(\widehat{\alpha}) + 2\varepsilon \leq \widehat{ISE}(\alpha) + 2\varepsilon \leq ISE(\alpha) + 4\varepsilon$$

where the second inequality holds from the definition of $\widehat{\alpha}$. This proves the theorem. □

4.3.3 ISE consistency

Next, we have a theorem stating that $ISE(\widehat{\alpha})$ converges to zero in probability.

Theorem 4.3. *Suppose that for* $f = f_+$ *and* f_-, *the Hessian* $\mathscr{H}_f(\mathbf{x})$ *exists and each entry of* $\mathscr{H}_f(\mathbf{x})$ *is piecewise continuous and square integrable. If* σ, n_+ *and* n_- *evolve as functions of* n *such that* $\sigma \to 0$, $n_+\sigma^{2d}/\ln n \to \infty$ *and* $n_+\sigma^{2d}/\ln n \to \infty$, *then* $ISE(\widehat{\alpha}) \to 0$ *in probability as* $n \to \infty$

This result intuitively follows from the oracle inequality since the standard Parzen window density estimate is consistent and uniform weights are among the simplex A. The rigorous proof is presented in Section 4.8.2.

4.3.4 Bayes error consistency

In classification, we are ultimately interested in minimizing the probability of error. The consistency with respect to the probability of error could be easily shown if we set γ to $\gamma^* = \frac{1-p}{p}$ and apply Theorem 3 in [72], where $0 < p < 1$ is the prior probability of the positive class. However, since p is unknown, we must estimate γ^*. Let us now assume $\{\mathbf{X}_i\}_{i=1}^n$ is an i.i.d sample from $f(\mathbf{x}) = pf_+(\mathbf{x}) + (1-p)f_-(\mathbf{x})$. Then N_+ and N_- are binomial random variables, and we may estimate γ^* as $\gamma = \frac{N_-}{N_+}$. The next theorem says the L_2 kernel classifier is consistent with respect to the probability of error.

Theorem 4.4. *Suppose that the assumptions in Theorem 4.3 are satisfied. In addition, suppose that $f_- \in L_2(\mathbb{R})$, i.e. $\|f_-\|_2 < \infty$. Let $\gamma = N_-/N_+$ be an estimate of $\gamma^* = \frac{1-p}{p}$. If σ evolves as a function of n such that $\sigma \to 0$ and $n\sigma^{2d}/\ln n \to \infty$ as $n \to \infty$, then the L_2 kernel classifier is consistent. In other words, given training data $\mathbf{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$, the classification error*

$$L_n = \mathbf{P}\left\{ sgn\left\{ \widehat{d}_\gamma(\mathbf{X}; \widehat{\alpha}) \right\} \neq Y \mid \mathbf{D}_n \right\}$$

converges to the Bayes error L^* in probability as $n \to \infty$.

The proof is given in Section 4.8.3.

4.3.5 Application to density estimation

By setting $\gamma = 0$, our goal becomes estimating f_+ and we recover the L_2 kernel density estimate of [37] and [25] using leave-one-out cross-validation. Given an i.i.d sample

$\mathbf{X}_1, \ldots, \mathbf{X}_n$ from $f(\mathbf{x})$, the L_2 kernel density estimate of $f(\mathbf{x})$ is defined as

$$\widehat{f}(\mathbf{x}; \widehat{\alpha}) = \sum_{i=1}^{n} \widehat{\alpha}_i k_\sigma(\mathbf{x}, \mathbf{X}_i)$$

with $\widehat{\alpha}_i$'s optimized such that

$$\widehat{\alpha} = \underset{\substack{\sum \alpha_i = 1 \\ \alpha_i \geq 0}}{\arg\min} \quad \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k_{\sqrt{2}\sigma}(\mathbf{X}_i, \mathbf{X}_j) - \sum_{i=1}^{n} \alpha_i \left(\frac{1}{n-1} \sum_{j \neq i} k_\sigma(\mathbf{X}_i, \mathbf{X}_j) \right).$$

Our concentration inequality, oracle inequality, and L_2 consistency result immediately extend to provide the same performance guarantees for this method. In particular, we state the following corollaries.

Corollary 4.5. *Let $\varepsilon > 0$ and set $\delta = \delta(\varepsilon) = 2ne^{-c(n-1)\varepsilon^2}$ where $c = 2\left(\sqrt{2\pi}\sigma\right)^{2d}$. Then, with probability at least $1 - \delta$*

$$\int \left(\widehat{f}(\mathbf{x}; \widehat{\alpha}) - f(\mathbf{x}) \right)^2 d\mathbf{x} \leq \inf_{\substack{\sum \alpha_i = 1 \\ \alpha_i \geq 0}} \int \left(\widehat{f}(\mathbf{x}; \alpha) - f(\mathbf{x}) \right)^2 d\mathbf{x} + 4\varepsilon.$$

Corollary 4.6. *Suppose that the Hessian $\mathscr{H}_f(\mathbf{x})$ of a density function $f(\mathbf{x})$ exists and each entry of $\mathscr{H}_f(\mathbf{x})$ is piecewise continuous and square integrable. If $\sigma \to 0$ and $n\sigma^{2d}/\ln n \to \infty$ as $n \to \infty$, then*

$$\int \left(\widehat{f}(\mathbf{x}; \widehat{\alpha}) - f(\mathbf{x}) \right)^2 d\mathbf{x} \to 0$$

in probability.

4.4 Geometric Interpretations

In this section, we present two geometric interpretations of the L2QP classifier.

4.4.1 Separating hyperplane in kernel feature space

The first interpretation views the QP (4.2.3) as the dual of a primal problem defined in a kernel feature space. The corresponding primal problem is

$$\begin{aligned} \min_{\mathbf{w},\xi_+,\xi_-} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \xi_+ + \xi_- \qquad (4.12) \\ \text{s.t} \quad & Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i)\rangle \geq c_i - \xi_+, \quad \text{for } i \in I_+ \\ & Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i)\rangle \geq c_i - \xi_-, \quad \text{for } i \in I_- \end{aligned}$$

where $\Phi_\sigma(\mathbf{x})$ is the implicit kernel mapping into the feature space associated with the Gaussian kernel Hilbert space [56].

This primal formulation differs from that of the standard 2-norm SVM with hinge loss (and without offset) in the following aspects. First, in the right hand sides of the constraints, c_i's appear instead of 1. This means if c_i is larger (i.e. $\mathbf{X}_i$ is accurately classified by the Parzen window plug-in formula), this modified SVM places more emphasis on correctly classifying $\mathbf{X}_i$. Second, there exist only two slack variables ξ_+ and ξ_-, one per class, and these are not required to be nonnegative. Finally, after finding the optimal solution $\widehat{\mathbf{w}} = \sum_{i=1}^n \widehat{\alpha}_i Y_i \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i)$, the final classifier takes the sign of the inner product between $\widetilde{\mathbf{w}} = \sum_{i=1}^n \widehat{\alpha}_i Y_i \Phi_\sigma(\mathbf{X}_i)$ and $\Phi_\sigma(\mathbf{x})$, not between $\widehat{\mathbf{w}}$ and $\Phi_{\sqrt{2}\sigma}(\mathbf{x})$, i.e.,

$$g(\mathbf{x}) = \operatorname{sgn}\left\{ \langle \widetilde{\mathbf{w}}, \Phi_\sigma(\mathbf{x}) \rangle \right\} = \operatorname{sgn}\left\{ \sum_{i=1}^n \widehat{\alpha}_i Y_i k_\sigma(\mathbf{x}, \mathbf{X}_i) \right\}.$$

The primal offers another explanation of why the points with nonzero α_i's are concentrated in regions of space with greater probability mass. First note that since we are minimizing ξ_+ and ξ_-, they satisfy

$$\begin{aligned} \xi_+ &= \max_{i \in I_+} \left\{ c_i - Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i)\rangle \right\}, \\ \xi_- &= \max_{i \in I_+} \left\{ c_i - Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i)\rangle \right\}. \end{aligned}$$

As α_i is the Lagrangian multiplier associated with each constraint, the optimal α_i should satisfy the Karush-Kuhn-Tucker(KKT) conditions, in particular the complimentary slackness condition. Thus, for nonzero $\alpha_i > 0$ the associated constraint should be met with equality, i.e.,

$$\begin{aligned}
& c_i - Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i) \rangle = \xi_+ \\
&= \max_{j \in I_+} \left\{ c_j - Y_j \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_j) \rangle \right\} \text{ for } \alpha_i > 0, i \in I_+ \\
& c_i - Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i) \rangle = \xi_- \\
&= \max_{j \in I_+} \left\{ c_j - Y_j \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_j) \rangle \right\} \text{ for } \alpha_i > 0, i \in I_- .
\end{aligned}$$

Therefore, we can see that if c_i is larger, $c_i - Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i) \rangle$ is more likely to be a maximum value and thus the corresponding α_i is nonzero. Since $c_i = Y_i \widehat{h}(X_i)$, where $\widehat{h}$ is the Parzen window plug-in estimate of d_γ, it tends to be largest in regions of space with high probability mass.

In Section 4.5, we introduce an extension of the L_2 kernel classification that amounts to augmenting the primal with an additional parameter multiplying the slack variables.

4.4.2 Weighted centroids in kernel feature space

Another interpretation can be obtained by expressing the L_2 criterion itself in the kernel feature space, not considering it as a dual problem. Define

$$\mathbf{m}_\sigma^+ = \frac{1}{N_+} \sum_{i \in I_+} \Phi_\sigma(\mathbf{X}_i), \quad \mathbf{m}_\sigma^- = \frac{1}{N_-} \sum_{i \in I_-} \Phi_\sigma(\mathbf{X}_i)$$

$$\mathbf{m}_\sigma^+(\alpha) = \sum_{i \in I_+} \alpha_i \Phi_\sigma(\mathbf{X}_i), \quad \mathbf{m}_\sigma^-(\alpha) = \sum_{i \in I_-} \alpha_i \Phi_\sigma(\mathbf{X}_i).$$

With this notation, by adding the constant term $\left(\frac{1}{N_+-1}+\frac{\gamma^2}{N_--1}\right)k_\sigma(\mathbf{0},\mathbf{0})$, the L_2 objective function may be expressed as

$$\frac{1}{2}\|\mathbf{m}^+_{\sqrt{2}\sigma}(\alpha)-\gamma\mathbf{m}^-_{\sqrt{2}\sigma}(\alpha)\|^2-\langle\mathbf{m}^+_\sigma(\alpha),\frac{N_+}{N_+-1}\mathbf{m}^+_\sigma-\gamma\mathbf{m}^-_\sigma\rangle$$
$$-\gamma\langle\mathbf{m}^-_\sigma(\alpha),\mathbf{m}^+_\sigma-\gamma\frac{N_-}{N_--1}\mathbf{m}^-_\sigma\rangle.$$

Since $\frac{N_+-1}{N_+}$ and $\frac{N_--1}{N_-}$ approach to 1 as n_+ and n_- go to ∞, for large n, (4.2.3) is equivalent to

$$\widehat{\alpha}=\arg\min_{\alpha\in A}\quad\frac{1}{2}\|\mathbf{m}^+_{\sqrt{2}\sigma}(\alpha)-\gamma\mathbf{m}^-_{\sqrt{2}\sigma}(\alpha)\|^2-\langle\mathbf{m}^+_\sigma(\alpha)-\gamma\mathbf{m}^-_\sigma(\alpha),\mathbf{m}^+_\sigma-\gamma\mathbf{m}^-_\sigma\rangle.$$

This has an appealing geometric interpretation. The first term, by itself, gives rise to the max-margin hyperplane in feature space in the case of separable data [4, 18]. In particular, because of the constraints $\sum_{i\in I_+}\alpha_i=\sum_{i\in I_-}\alpha_i=1$ and $\alpha_i\geq 0,\ \forall i$, the first term is minimized when $\mathbf{m}^+_{\sqrt{2}\sigma}(\alpha)$ and $\mathbf{m}^-_{\sqrt{2}\sigma}(\alpha)$ are on the boundaries of their respective convex hulls, giving rise to the maximum margin separating hyperplane. The second term tries to align $\mathbf{m}^+_\sigma(\alpha)-\gamma\mathbf{m}^-_\sigma(\alpha)$ with $\mathbf{m}^+_\sigma-\gamma\mathbf{m}^-_\sigma$, which is the normal vector defining the nearest centroid classifier. Interestingly, with $\gamma=1$, the nearest centroid classifier in feature space is identical to the Parzen window plug-in classifier [56] up to an offset term. Thus we may say that the second term regularizes the SVM (an alternative to the SVM's soft-margin-based regularization), or the first term sparsifies the Parzen window. Note, however, that the first and second terms involve different kernel bandwidths, so that the two terms correspond to different Hilbert spaces.

4.5 Variations and Extensions

4.5.1 Weighted L_2 distance in Fourier domain

One variation of the L2QP classifier is obtained by minimizing the weighted L_2 distance in the Fourier domain. For density estimation, weighted ISE applied to characteristic functions was previously considered in a parametric setting in [30, 51]. We denote the Fourier transforms of $\widehat{d_\gamma}(\mathbf{x};\alpha)$ and $d_\gamma(\mathbf{x})$ by $\widehat{D}_\gamma(\omega;\alpha)$ and $D_\gamma(\omega)$ respectively, each of which is a difference of characteristic functions. Define the weighted L_2 distance associated with α

$$ISE_\lambda(\alpha) := \int \left| \widehat{D}_\gamma(\omega;\alpha) - D_\gamma(\omega) \right|^2 e^{-\lambda^2\omega^2} d\omega,$$

where $\lambda \geq 0$ is a fixed parameter. The effect of the weighting term $e^{-\lambda^2\omega^2}$ and the choice of λ will be discussed below. We may write

$$\begin{aligned} ISE_\lambda(\alpha) =& \int \left| \widehat{D}_\gamma(\omega;\alpha) e^{-\lambda^2\omega^2/2} - D_\gamma(\omega) e^{-\lambda^2\omega^2/2} \right|^2 d\omega \\ =& 2\pi \int \left(\widehat{d_\gamma}(\mathbf{x};\alpha) * k_\lambda(\mathbf{x},0) - d_\gamma(\mathbf{x}) * k_\lambda(\mathbf{x},0) \right)^2 d\mathbf{x} \\ =& 2\pi \int \left(\sum_{i=1}^n \alpha_i Y_i k_{\sqrt{\sigma^2+\lambda^2}}(\mathbf{x},\mathbf{X}_i) - \int d_\gamma(\mathbf{x}') k_\lambda(\mathbf{x},\mathbf{x}') d\mathbf{x}' \right)^2 d\mathbf{x} \end{aligned} \quad (4.13)$$

where the second equality holds by Parseval's theorem and $*$ denotes convolution.

After expanding the square in (4.13), the first term becomes

$$\int \left(\sum_{i=1}^n \alpha_i Y_i k_{\sqrt{\sigma^2+\lambda^2}}(\mathbf{x},\mathbf{X}_i) \right)^2 d\mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j Y_i Y_j k_\rho(\mathbf{X}_i,\mathbf{X}_j) \quad (4.14)$$

where $\rho = \sqrt{2\sigma^2 + 2\lambda^2}$ by the convolution theorem for Gaussian kernels [68]. The second term can be written

$$\begin{aligned} &\int \left(\sum_{i=1}^n \alpha_i Y_i k_{\sqrt{\sigma^2+\lambda^2}}(\mathbf{x},\mathbf{X}_i) \right) \cdot \left(\int d_\gamma(\mathbf{x}') k_\lambda(\mathbf{x},\mathbf{x}') d\mathbf{x}' \right) d\mathbf{x} \\ =& \sum_{i=1}^n \alpha_i Y_i \cdot \int d_\gamma(\mathbf{x}') \left(\int k_{\sqrt{\sigma^2+\lambda^2}}(\mathbf{x},\mathbf{X}_i) k_\lambda(\mathbf{x},\mathbf{x}') d\mathbf{x} \right) d\mathbf{x}' \\ =& \sum_{i=1}^n \alpha_i Y_i \int d_\gamma(\mathbf{x}') k_{\sqrt{\sigma^2+2\lambda^2}}(\mathbf{x}',\mathbf{X}_i) d\mathbf{x}' \approx \sum_{i=1}^n \alpha_i \tilde{c}_i \end{aligned}$$

where we used leave-one-out cross-validation estimate in the last step and

$$\tilde{c}_i \triangleq \begin{cases} Y_i \Big(\frac{1}{N_+ - 1} \sum_{j \in I_+, j \neq i} k_{\sqrt{\sigma^2+2\lambda^2}}(\mathbf{X}_j,\mathbf{X}_i) - \frac{\gamma}{N_-} \sum_{j \in I_-} k_{\sqrt{\sigma^2+2\lambda^2}}(\mathbf{X}_j,\mathbf{X}_i) \Big), & i \in I_+ \\ Y_i \Big(\frac{1}{N_+} \sum_{j \in I_+} k_{\sqrt{\sigma^2+2\lambda^2}}(\mathbf{X}_j,\mathbf{X}_i) - \frac{\gamma}{N_- - 1} \sum_{j \in I_-, j \neq i} k_{\sqrt{\sigma^2+2\lambda^2}}(\mathbf{X}_j,\mathbf{X}_i) \Big), & i \in I_-. \end{cases}$$

Therefore, an empirical minimizer of the weighted L^2 distance $ISE_\lambda(\alpha)$ is obtained by solving

$$\widehat{\alpha} = \underset{\alpha \in A}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j k_\rho \left(\mathbf{X}_i, \mathbf{X}_j \right) - \sum_{i=1}^{n} \tilde{c}_i \alpha_i.$$

From the Fourier domain definition of $ISE_\lambda(\alpha)$, we may interpret the Gaussian weight function $e^{-\lambda^2 \omega^2}$ as a low-pass filter that de-emphasizes high-frequency content in the unknown densities. Thus larger values of λ place more emphasis on the slowly varying features of $d_\gamma(\mathbf{x})$. A similar interpretation results if we consider the effect of λ in the $\mathbf{x}$ domain. In (4.13), we see that α is chosen to optimize the L^2 distance between an α-weighted DOD with kernel bandwidth $\sqrt{\sigma^2 + \lambda^2}$, and a uniformly weighted DOD with kernel bandwidth λ. That is, the "target" DOD is increasingly smooth as λ increases.

We refer to this method as L2QP-*k* where *k* determines λ through $\lambda = k \cdot \sigma$. Since L2QP-0 corresponds to the previous L2QP, L2QP-*k* is a generalization of L2QP method. Our experiments have primarily focused on $\lambda = 0$ and $\lambda = \sigma$, the latter being motivated by the belief that the "target" DOD and final classifier should be accurately represented by the same kernel bandwidth. Our evidence thus far suggest that both of these choices of λ, as well as others much larger, lead to comparable classifiers. We have observed, however, that smaller values of λ tend to yield sparser classifiers.

4.5.2 L_2 criterion with inequality constraints

Our theoretical analysis carries through if we replace the constraint set $A = \{\alpha : \alpha_i \geq 0, \sum_{i \in I_+} \alpha_i = \sum_{i \in I_-} \alpha_i = 1\}$ with the set

$$A' = \left\{ \alpha : \alpha_i \geq 0, \, \big(1 - \sum_{i \in I_+} \alpha_i\big) = \gamma\big(1 - \sum_{i \in I_-} \alpha_i\big) \geq 0 \right\}.$$

By requiring $\big(1 - \sum_{i \in I_+} \alpha_i\big) = \gamma\big(1 - \sum_{i \in I_-} \alpha_i\big)$, we still enforce that d_γ integrate to the true value of $1 - \gamma$. However, by allowing that the coefficients in each class sum to less than one, we allow for the possibility that some positive and negative coefficients might "cancel out" in regions of space where f_+ and f_- overlap. This could potentially lead to even sparser solutions.

4.5.3 L_2 criterion without constraints

Since our goal is classification and not density estimation, it is not necessary that $\widehat{f_+}(\mathbf{x};\alpha)$ and $\widehat{f_-}(\mathbf{x};\alpha)$ be proper density estimates, and hence the constraints $\alpha \in A$ may be dropped. In this case, the unconstrained quadratic objective function, in the matrix/vector form,

$$\frac{1}{2}\alpha^T Q\alpha - \widetilde{\mathbf{c}}^T\alpha$$

is minimized by the solution of

$$Q\alpha = \widetilde{\mathbf{c}} \tag{4.15}$$

where $\widetilde{\mathbf{c}} = [\tilde{c}_1, \tilde{c}_2, \cdots, \tilde{c}_n]^T$, $Q := \left(Y_i Y_j K_{ij}\right)_{i,j=1}^n$, $K := \left(k_\rho\left(\mathbf{X}_i, \mathbf{X}_j\right)\right)_{i,j=1}^n$. If K is positive definite and $\gamma \neq 0$, then Q is also positive definite, and thus the objective is strictly convex.

The optimization problem now becomes the problem of solving a linear system of equations (4.15). It is similar in that respect to the 2-norm SVM with squared error loss, or least-squares SVM (LS-SVM) [35], but again does not include a regularization parameter. The resulting L2LE-k (L2 classification via Linear system of Equations) classifier is not sparse, again like the LS-SVM. Since Q is positive definite, (4.15) can be solved efficiently by the conjugate gradient descent (CGD) algorithm [54].

4.5.4 Other kernels

Our methodology allows for any kernel $k(\mathbf{x}, \mathbf{x}')$ such that $k(\mathbf{x}, \mathbf{x}') \geq 0$ and, for any fixed $\mathbf{x}'$, $\int k(\mathbf{x}, \mathbf{x}')d\mathbf{x} = 1$, e.g., the multivariate Cauchy kernel,

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

or the multivariate Laplacian (product) kernel,

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

The L_2 kernel classifier is still the solution of

$$\widehat{\alpha} = \underset{\alpha\in A}{\arg\min} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j Y_i Y_j K_{ij} - \sum_{i=1}^{n} c_i\alpha_i,$$

where

$$K_{ij} = \int k(\mathbf{x},\mathbf{X}_i)k(\mathbf{x},\mathbf{X}_j)d\mathbf{x}$$

and c_i is as before.

We make two important observations regarding this QP. First, from the identity

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j Y_i Y_j K_{ij} = \int\left(\sum_{i=1}^{n}\alpha_i Y_i k(\mathbf{x},\mathbf{X}_i)\right)^2 d\mathbf{x},$$

we see that the matrix $(Y_iY_jK_{ij})_{i,j=1}^n$ is always positive definite, and therefore the QP is strictly convex, provided the $\mathbf{X}_i$ are distinct. Second, it is desirable that K_{ij} be easily computable. For some kernels, like the Gaussian, the integral has a closed form expression. For example, the multivariate Cauchy kernel satisfies

$$k_{2\sigma}(\mathbf{X}_i,\mathbf{X}_j) = \int k_\sigma(\mathbf{x},\mathbf{X}_i)\cdot k_\sigma(\mathbf{x},\mathbf{X}_j)d\mathbf{x},$$

[6] and the multivariate Laplacian (product) kernel satisfies

$$\begin{aligned}&\int k_\sigma(\mathbf{x},\mathbf{X}_i)\cdot k_\sigma(\mathbf{x},\mathbf{X}_j)d\mathbf{x}\\ &= \frac{1}{(4\sigma)^d}\prod_{l=1}^{d}\left(1+\frac{|\mathbf{X}_{i,l}-\mathbf{X}_{j,l}|}{\sigma}\right)\exp\left\{-\frac{\|\mathbf{X}_i-\mathbf{X}_j\|_1}{\sigma}\right\}.\end{aligned}$$

For kernels without such a formula, values of the integral may still be pre-computed and stored. For radially symmetric kernels, such as an alternative multivariate Laplacian kernel [26], $k_\sigma(\mathbf{x},\mathbf{X}_i) = C\cdot\exp\bigl(-\|\mathbf{x}-\mathbf{X}_i\|/\sigma\bigr)$, where C is a normalizing constant, this entails a simple one-dimensional table, as K_{ij} will depend only on $\|\mathbf{X}_i-\mathbf{X}_j\|$. We experimented briefly with multivariate Cauchy kernels, but did not see significant differences compared to the Gaussian.

4.5.5 Regularization for high dimensional data

Our experimental results show that our L2QP thus far discussed perform poorly on most high dimensional data. Similarly, in [29], where the class conditional densities are estimated separately based on the L_2 criterion, the authors only consider low dimensional data (the 20 dimensional German dataset was reduced to 7 dimensional). In this section, we offer an explanation for this phenomenon. We also present a variation that significantly improves the performance in high dimensions at the expense of introducing a new regularization parameter that must be tuned.

To understand the impact of dimension on L2QP, it is important to realize that the method involves Gaussian kernels of bandwidth $\sqrt{2}\sigma$ (quadratic term) and σ (linear term). The normalizing constants for these kernels are $\left(4\pi\sigma^2\right)^{-d/2}$ and $\left(2\pi\sigma^2\right)^{-d/2}$, respectively. The ratio of the second normalizing constant to the first one is $\sqrt{2}^d$. In other words, the ratio exponentially increases as a function of dimension and thus in high dimensional data the linear term in (4.2.3) dominates the quadratic term. In this case, minimizing (4.2.3) causes a few data points associated with larger c_i's to monopolize the weights and yields a too sparse solution.

To address this problem, we introduce a new parameter $\eta > 0$ that balances the linear term and the quadratic term

$$\min_{\alpha \in A} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j k_{\sqrt{2}\sigma}\left(\mathbf{X}_i, \mathbf{X}_j\right) - \frac{1}{\eta}\sum_{i=1}^{n} c_i \alpha_i.$$

The corresponding primal is

$$\begin{aligned} \min_{\mathbf{w},\xi_+,\xi_-} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \eta\left(\xi_+ + \xi_-\right) & (4.16) \\ \text{s.t} \quad & Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i)\rangle \geq c_i - \xi_+, \quad \text{for } i \in I_+ \\ & Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i)\rangle \geq c_i - \xi_-, \quad \text{for } i \in I_-. \end{aligned}$$

Therefore, (4.12) can be thought as a special case of (4.16) where the regularization parameter η is set to 1. In the primal point of view, η controls the trade off between the complexity of the classifier, $||\mathbf{w}||^2$ and how much the classifier fits to Parzen window plug-in classifier $c_i - Y_i \cdot \langle \mathbf{w}, \Phi_{\sqrt{2}\sigma}(\mathbf{X}_i) \rangle$. This new algorithm may also be viewed as minimizing an estimated of a modified ISE, given by

$$\begin{aligned} ISE^{\eta}(\alpha) &= ||\widehat{d_\gamma}(\mathbf{x};\alpha) - \frac{1}{\eta} d_\gamma(\mathbf{x})\,||^2_{L_2} \\ &= \int \left(\widehat{d_\gamma}(\mathbf{x};\alpha) - \frac{1}{\eta} d_\gamma(\mathbf{x}) \right)^2 d\mathbf{x}. \end{aligned}$$

This new method may also be combined with the Fourier domain extension discussed previously, and we refer to resulting classifier as L2QP$_\eta$-k.

4.6 Experiments

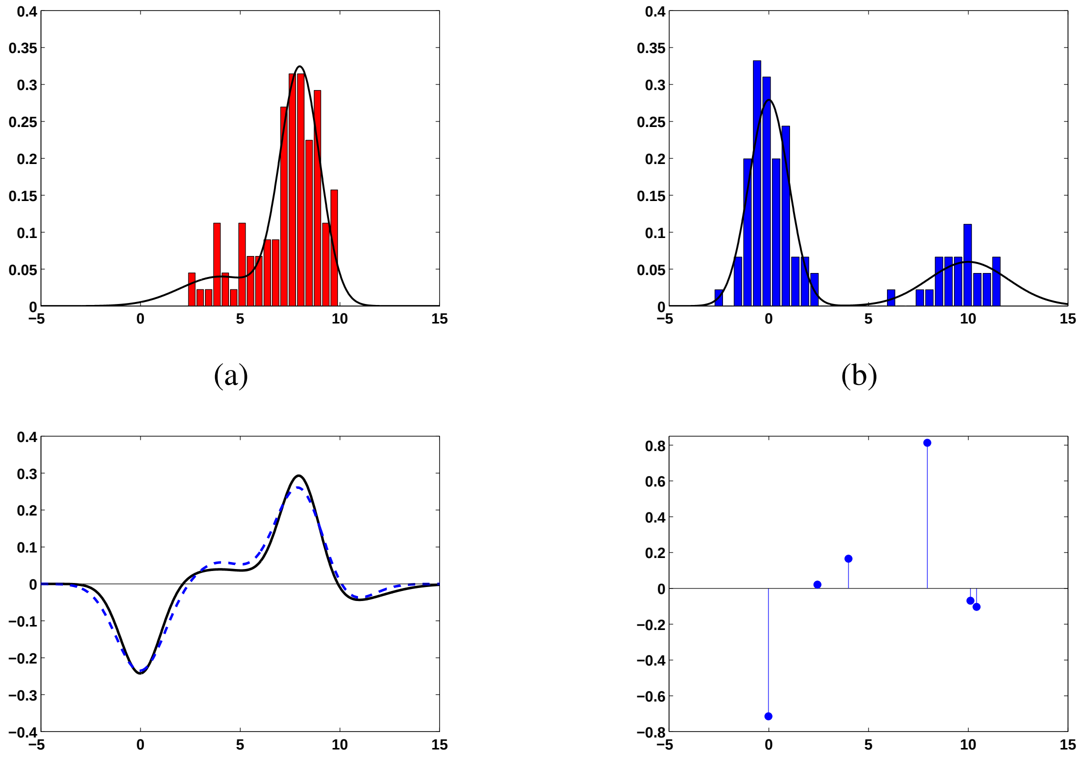
We implement our methods(L2QP-0, L2QP-1, L2QP$_\eta$-0, L2QP$_\eta$-1) based on LIBSVM [11] by modifying an SMO subroutine (see Section 4.2.3). For comparison, we also experiment with the 2-norm SVM with hinge loss (S-SVM, S for “soft margin”), the 2-norm SVM with hinge loss and $C \to \infty$ (H-SVM, H for “hard margin”), and a plug-in classifier based on Parzen window density estimates (Parzen).

To illustrate some of the basic properties of L_2 kernel classifiers, we first experiment with 1 dimensional data. Both classes are equally likely and

$$\begin{aligned} f_+(\mathbf{x}) &= 0.2\phi(\mathbf{x};4,\sqrt{2}) + 0.8\phi(\mathbf{x};8,1) \\ f_-(\mathbf{x}) &= 0.7\phi(\mathbf{x};0,1) + 0.3\phi(\mathbf{x};10,\sqrt{2}) \end{aligned}$$

where $\phi(\mathbf{x};\mu,\sigma)$ is a univariate Gaussian pdf with mean μ and variance σ^2. We build a L2QP-0 classifier from 200 training samples. To find a classifier with the smallest probability error, we set $\gamma = N_-/N_+$ and use 5-fold-cross validation to estimate the bandwidth σ from a logarithmically spaced grid of 50 points from 10^{-2} to 10^1.

The results are shown in Figure 4.1. The estimate $\widehat{d_\gamma}(\mathbf{x};\widehat{\alpha})$ is fairly close to the true $d_\gamma(\mathbf{x})$. For $\widehat{\alpha}_i > 0$, $\widehat{\alpha}_i Y_i$ are shown at the corresponding $\mathbf{X}_i$ in Figure 4.1 (d) and the number of nonzero weights is 9.



(a) (b)

(c) (d)

Fig. 4.1: (a) $f_+(\mathbf{x})$ and histogram of its samples (b) $f_-(\mathbf{x})$ and histogram of its samples (c) $d_\gamma(\mathbf{x})$ (solid line) and $\widehat{d_\gamma}(\mathbf{x};\widehat{\alpha})$ (dashed line) (d) Sparsity of the proposed method

Next, we demonstrate our algorithms on 18 artificial and real-world benchmark datasets, available online[1] [11, 48]. There are 100 randomly permuted partitions of each dataset into training and test sets (20 for Image, Splice, Adult, Mushrooms and Web). The dimension and sample sizes[2] of each dataset are summarized in Table 4.1. We set

[1] http://www.fml.tuebingen.mpg.de/Members/ for the first 13 datasets and http://www.csie.ntu.edu.tw/˜cjlin/libsvmtools/datasets/ for the last 5 datasets.

[2] The Adult and Web datasets were subsampled owing to their large size.

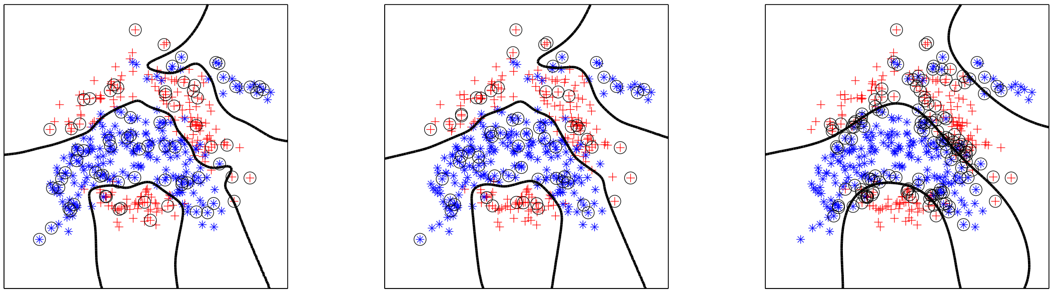
Dataset	# of training data	# of test data	input dimension
Banana	400	4900	2
B. Cancer	200	77	9
Diabetes	468	300	8
F. Solar	666	400	9
German	700	300	20
Heart	170	100	13
Image	1300	1010	18
Ringnorm	400	7000	20
Splice	1000	2175	60
Thyroid	140	75	5
Titanic	150	2051	3
Twonorm	400	7000	20
Waveform	400	4600	21
Adult	3000	3000	123
Ionosphere	251	100	34
mushrooms	4124	4000	112
Sonar	108	100	60
Web	3000	3000	300

Table 4.1: General information about benchmark datasets

$\gamma = N_-/N_+$ to minimize the probability of error. The parameters to be tuned are σ (all methods), C (S-SVM), and η (L2QP$_\eta$-k, $k = 0,1$). The following grids were used. For L2QP-0, L2QP-1, and Parzen, we search a logarithmically spaced grid of 50 points from 10^{-2} to 10^1 for σ. For the SVMs, we search the grid $2^{-2}, 2^{-1}, ..., 2^7$ for σ and for S-SVM we searched $2^{-5}, 2^{-3}, ..., 2^{15}$ for C. For L2QP$_\eta$-0 and L2QP$_\eta$-1, we searched a logarithmically spaced grid of 11 points from 10^{-2} to 10^1 for σ, and a logarithmically spaced grid of 10 points from 1 to $\sqrt{2}^d$ for η. The grids were chosen to ensure that the two-parameter methods searched grids of the same size. The parameters were taken to be the same for all partitions. Each parameter was determined by taking the median estimate based on the first five training sets. On each of these training sets, we use 5-fold-cross validation to determine the best parameters.

For the ‘banana’ data set, we plot the decision boundary of the L2QP-0, L2QP-1, and

S-SVM in Figure 4.2 along with training samples. The number of training samples is 400 and the first partition of the dataset is used. The number of non-zero weights of each method are 77, 66, and 142, respectively. The decision boundaries of L2QP-0 and L2QP-1 slightly differ in that L2QP-1 shows smoother boundary than L2QP-0.



(a) L2QP-0 (b) L2QP-1 (b) S-SVM

Fig. 4.2: Decision boundary along with positive samples (+) and negative samples ($*$) for banana dataset. Points whose corresponding α_i are nonzero are enclosed by ◯ (a) L2QP-0 (b) L2QP-1 (c) S-SVM

The results for all the datasets are presented in Table 4.2, 4.3, and 4.4. They show the average probability of error, the average percentage of nonzero coefficients (reflecting the sparsity), and training time over all permutations, respectively. Time indicates the total time required to build a classifier, including the cross-validation search for free parameters.

From these results, we can see that the L2QP-0 and L2QP-1 methods shows comparable performance to SVMs except on some high dimensional datasets, e.g., German, Image, Splice, Waveform, Adult and Ionosphere. For low dimensional datasets, the default value $\eta = 1$ works well, but for dimensionality exceeding 15, this default method tends to be too sparse, as explained in Section 4.5.5. Significantly improved performance on high-dimensional data results from optimizing η. The L2QP$_\eta$-0 and L2QP$_\eta$-1 are comparable to SVMs for almost all datasets; their prediction accuracy is $2-3\%$ worse on average. The primary exception is the Splice data. A likely explanation for this is that the dataset

	H-SVM	S-SVM	L2QP-0	L2QP-1	L2QP$_\eta$-0	L2QP$_\eta$-1	Parzen
Banana	10.9±0.5	10.8±0.5	11.1±0.6	*10.7±0.4*	10.7±0.5	**10.6±0.4**	11.3±0.6
B.Cancer	27.4±4.8	26.9±4.6	26.5±4.6	26.4±4.4	27.6±4.7	*25.2±4.2*	**24.7±4.2**
Diabetes	*23.8±1.9*	**23.2±1.8**	26.5±2.4	26.8±2.4	26.2±2.3	26.6±1.9	26.0±2.1
F.Solar	37.8±4.6	**32.3±1.8**	35.7±3.4	35.5±1.8	37.3±1.8	*34.0±2.0*	36.2±1.9
German	*24.2±2.2*	**24.2±2.1**	29.4±2.1	28.4±2.7	26.1±2.8	25.5±2.6	25.3±2.5
Heart	19.4±4.0	**15.6±3.4**	17.5±4.2	16.8±3.6	17.4±4.0	*16.7±3.8*	18.0±3.5
Image	*3.3±0.7*	**3.0±0.7**	28.7±7.8	9.3±4.5	3.5±0.6	3.7±0.5	3.4±0.5
Ringnorm	**1.6±0.1**	*2.0±0.2*	2.4±0.2	2.4±0.2	2.4±0.2	2.4±0.2	2.4±0.2
Splice	**10.8±0.7**	*11.3±0.6*	38.9±4.3	36.0±5.2	19.2±1.5	29.1±10.1	26.0±1.9
Thyroid	5.3±2.3	5.0±2.2	5.2±2.2	4.9±2.3	4.6±2.4	*4.3±2.5*	**4.2±2.1**
Titanic	*22.4±1.0*	22.8±0.7	23.0±0.4	22.9±0.6	23.2±1.5	23.2±1.7	**22.2±1.1**
Twonorm	3.6±0.6	*2.9±0.2*	6.9±3.6	3.9±0.4	3.3±0.7	5.4±4.0	**2.5±0.2**
Waveform	13.2±1.2	**10.0±0.4**	14.2±0.8	13.5±1.2	11.6±0.7	11.2±1.1	*10.7±0.8*
Adult	*15.9±0.8*	**15.7±0.8**	19.5±1.5	21.6±1.2	18.3±0.8	20.0±0.8	18.4±0.6
Ionosphere	*5.7±2.4*	**5.5±2.1**	29.4±4.1	29.6±4.2	12.9±3.8	9.5±2.8	13.2±3.3
Mushrooms	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Sonar	**15.5±3.7**	*15.6±3.6*	16.9±3.7	16.8±3.6	16.3±3.8	16.2±3.6	*15.6±3.6*
Web	2.3±0.3	**1.9±0.3**	2.9±0.3	2.9±0.3	3.6±1.3	2.9±0.3	*2.1±0.3*

Table 4.2: Probability of Error. Best method in bold face, second best emphasized

consists of only categorical features and thus density based methods may not be suitable.

Training time shows L2QP-0 and L2QP-1 are significantly faster than the SVM, a reflection of not having to search for an additional regularization parameter. Regarding sparsity, the L_2 methods are often much sparser than the SVMs. One noticeable exception is the Ringnorm data. We have discovered that allowing the two classes to have separate bandwidths (which easily fits within our framework) leads to greatly improved performance here, as well as sparse L_2 classifiers. To maintain a uniform presentation, however, we do not present detailed results for this extension.

Finally, we remark that the hard margin SVM was considered as alternative method having only one tuning parameter, like L2QP-0 and L2QP-1. In reality, however, we were only able to implement H-SVM by taking C very large in the S-SVM. Since the problem

	H-SVM	S-SVM	L2QP-0	L2QP-1	L2QP$_\eta$-0	L2QP$_\eta$-1	Parzen
Banana	26.2±2.5	38.5±2.2	19.9±1.5	16.5±1.1	13.5±1.0	12.7±1.0	100±0.0
B.Cancer	46.9±2.5	55.2±2.9	3.1±0.9	6.7±1.0	17.2±1.9	23.3±2.3	100±0.0
Diabetes	43.1±1.6	52.3±1.7	3.2±0.6	16.3±1.2	6.5±1.1	9.1±1.0	100±0.0
F.Solar	60.6±3.9	76.5±1.8	30.0±2.5	46.7±2.9	46.7±2.4	53.9±3.2	100±0.0
German	52.7±1.6	57.7±1.5	0.5±0.1	6.9±0.5	63.7±2.7	40.7±1.8	100±0.0
Heart	27.8±3.1	50.7±3.0	1.8±0.5	10.4±1.7	20.0±3.5	22.9±3.5	100±0.0
Image	30.2±0.9	7.9±0.8	7.8±2.0	75.8±23.2	79.8±1.0	73.0±1.0	100±0.0
Ringnorm	53.9±3.3	21.5±1.4	100±0.0	100±0.0	100±0.0	100±0.0	100±0.0
Splice	70.8±1.4	17.8±1.9	0.4±0.1	0.3±0.1	40.0±4.7	17.4±1.1	100±0.0
Thyroid	26.7±2.2	32.7±2.4	31.0±2.7	30.8±2.5	60.5±3.1	57.9±3.0	100±0.0
Titanic	45.3±6.5	48.0±6.8	36.5±11.5	44.0±8.2	70.2±4.9	70.4±4.1	100±0.0
Twonorm	19.7±2.1	13.9±1.1	0.7±0.2	5.2±1.2	3.9±0.9	4.5±0.7	100±0.0
Waveform	23.9±2.8	33.9±2.4	96.0±7.0	39.5±4.4	88.6±3.8	42.9±2.6	100±0.0
Adult	32.0±1.4	37.2±1.4	0.13±0.03	0.11±0.02	0.81±0.14	4.1±0.3	100±0.0
Ionosphere	56.7±2.1	34.3±1.7	1.2±0.4	3.1±3.7	32.6±7.2	53.9±2.0	100±0.0
Mushrooms	4.5±0.2	97.4±0.2	100±0.0	100±0.0	51.0±0.7	18.4±0.7	100±0.0
Sonar	89.2±2.5	89.6±2.3	100±0.0	100±0.0	100±0.0	99.2±0.7	100±0.0
Web	7.4±0.5	7.4±0.6	8.2±0.5	8.2±0.5	90.9±1.1	8.2±0.5	100±0.0

Table 4.3: Percentage of non zero weights

is not feasible for C too large, depending on σ, it was actually necessary to search for C after all (not reflected in reported run times). In addition, the running time for large C was far greater than that of any other approach.

4.7 Conclusion

In this chapter, the L_2 kernel classification method is proposed which minimizes the L_2 distance between the true unknown difference of densities $d_\gamma(\mathbf{x})$ and an estimator $\widehat{d_\gamma}(\mathbf{x};\alpha)$. Like the SVM, it is the solution of a convex quadratic program and has a sparse representation.

Through the development of a novel concentration inequality, we have established statistical performance guarantees on the L_2 kernel classifier. The results also specialize to

	H-SVM	S-SVM	L2QP-0	L2QP-1	L2QP$_\eta$-0	L2QP$_\eta$-1	Parzen
Banana	4140.14	82.01	37.29	51.22	89.17	124.84	-
B.Cancer	2181.51	24.48	10.52	10.80	29.22	47.07	-
Diabetes	30222.68	165.03	52.10	50.42	139.38	135.02	-
F.Solar	15058.31	335.95	79.93	88.83	363.91	849.55	-
German	7310.32	237.69	137.66	142.33	431.41	423.13	-
Heart	173.17	6.60	8.54	8.27	22.72	23.25	-
Image	2047.28	1056.40	397.61	418.93	1353.30	1367.40	-
Ringnorm	105.68	127.04	46.36	48.44	144.95	145.96	-
Splice	2583.97	1323.10	387.71	371.54	1329.09	1386.79	-
Thyroid	8.65	12.12	5.22	4.93	12.78	12.54	-
Titanic	173.12	239.78	4.53	4.55	14.63	15.58	-
Twonorm	12.46	95.32	46.38	44.97	141.88	139.34	-
Waveform	425.29	127.46	47.15	45.21	143.47	140.28	-
Adult	11705.50	10924.42	1016.06	989.54	9929.85	16359.32	-
Ionosphere	15.87	64.89	9.41	9.48	67.01	68.53	-
Mushrooms	861.73	13337.38	3027.31	2944.65	28368.16	29076.56	-
Sonar	4.42	24.16	2.95	3.16	19.25	18.70	-
Web	499.02	5086.96	871.05	858.64	8703.72	11241.11	-

Table 4.4: Time (s): run time, including cross-validation search for a regularization parameter where appropriate and training time for all permutations.

give performance guarantees for an existing method of L_2 kernel density estimation. The oracle inequality here has been applied to deduce consistency of the procedure (in both ISE and probability of error), but we suspect it may also yield adaptive rates of convergence.

Although formulated in terms of the L_2 distance on the difference of densities, the L_2 kernel classifier has geometric interpretations that more clearly reveal similarities and differences to the SVM. One of these interpretations motivates the incorporation of a regularization parameter into the approach, which allows the method to remain competitive with the SVM for dimensionality $d > 15$.

4.8 Proofs

4.8.1 Proof of Lemma 4.1

Note that for any given i, $\left(k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right)\right)_{j\neq i}$ are independent and bounded by $M = 1/\left(\sqrt{2\pi}\sigma\right)^d$. For random vectors $\mathbf{Z}\sim f_+(\mathbf{x})$ and $\mathbf{W}\sim f_-(\mathbf{x})$, $h(\mathbf{X}_i)$ in (6) can be expressed as

$$h(\mathbf{X}_i) = \mathbf{E}\left[k_\sigma(\mathbf{Z},\mathbf{X}_i)\mid \mathbf{X}_i\right] - \gamma\mathbf{E}\left[k_\sigma(\mathbf{W},\mathbf{X}_i)\mid \mathbf{X}_i\right].$$

Since $\mathbf{X}_i\sim f_+(\mathbf{x})$ for $i\in I_+$ and $\mathbf{X}_i\sim f_-(\mathbf{x})$ for $i\in I_-$, it can be easily shown that

$$\mathbf{E}\left[\widehat{h}_i\mid \mathbf{X}_i\right] = h(\mathbf{X}_i).$$

For $i\in I_+$,

$$\begin{aligned}
&\mathbf{P}\left\{\left|\widehat{h}_i - h(\mathbf{X}_i)\right| > \varepsilon \,\middle|\, \mathbf{X}_i = \mathbf{x}, E\right\}\\
\leq\ & \mathbf{P}\left\{\left|\frac{1}{n_+-1}\sum_{j\in I_+, j\neq i} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right) - \mathbf{E}\left[k_\sigma(\mathbf{Z},\mathbf{X}_i)\mid\mathbf{X}_i\right]\right| > \frac{\varepsilon}{1+\gamma}\,\middle|\,\mathbf{X}_i=\mathbf{x}\right\}\\
&+\mathbf{P}\left\{\left|\frac{\gamma}{n_-}\sum_{j\in I_-} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right) - \gamma\mathbf{E}\left[k_\sigma(\mathbf{W},\mathbf{X}_i)\mid\mathbf{X}_i\right]\right| > \frac{\gamma\varepsilon}{1+\gamma}\,\middle|\,\mathbf{X}_i=\mathbf{x}\right\} \qquad (4.17)
\end{aligned}$$

Since we are conditioning on E, the first term in (4.17) is

$$\begin{aligned}
&\mathbf{P}\left\{\left|\sum_{j\in I_+, j\neq i} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right) - (n_+-1)\mathbf{E}\left[k_\sigma(\mathbf{Z},\mathbf{X}_i)\mid\mathbf{X}_i\right]\right| > \frac{(n_+-1)\,\varepsilon}{1+\gamma}\,\middle|\,\mathbf{X}_i=\mathbf{x}\right\}\\
=\ & \mathbf{P}\left\{\left|\sum_{j\in I_+, j\neq i} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right) - \mathbf{E}\left[\sum_{j\in I_+, j\neq i} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right)\mid\mathbf{X}_i\right]\right| > \frac{(n_+-1)\,\varepsilon}{(1+\gamma)}\,\middle|\,\mathbf{X}_i=\mathbf{x}\right\}\\
=\ & \mathbf{P}\left\{\left|\sum_{j\in I_+, j\neq i} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right) - \mathbf{E}\left[\sum_{j\in I_+, j\neq i} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right)\mid\mathbf{X}_i\right]\right| > \frac{(n_+-1)\,\varepsilon}{(1+\gamma)}\,\middle|\,\mathbf{X}_i=\mathbf{x}\right\}\\
\leq\ & 2e^{-2(n_+-1)\varepsilon^2/(1+\gamma)^2M^2}.
\end{aligned}$$

where the last inequality holds by Hoeffding's inequality [23]. The second term in (4.17) is

$$\begin{aligned}
&\mathbf{P}\left\{\left|\sum_{j\in I_-} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right)-n_-\mathbf{E}\left[k_\sigma\left(\mathbf{W},\mathbf{X}_i\right)\mid\mathbf{X}_i\right]\right|>\frac{n_-\varepsilon}{1+\gamma}\,\middle|\,\mathbf{X}_i=\mathbf{x}\right\}\\
\leq\ &\mathbf{P}\left\{\left|\sum_{j\in I_-} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right)-\mathbf{E}\left[\sum_{j\in I_-} k_\sigma\left(\mathbf{X}_j,\mathbf{X}_i\right)\mid\mathbf{X}_i\right]\right|>\frac{n_-\varepsilon}{1+\gamma}\,\middle|\,\mathbf{X}_i=\mathbf{x}\right\}\\
\leq\ &2e^{-2n_-\varepsilon^2/(1+\gamma)^2M^2}\leq 2e^{-2(n_--1)\varepsilon^2/(1+\gamma)^2M^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbf{P}\left\{\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\varepsilon\right\}=\sum_{\mathbf{x}}\mathbf{P}\left\{\mathbf{X}_i=\mathbf{x}\right\}\cdot\mathbf{P}\left\{\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\varepsilon\,\middle|\,\mathbf{X}_i=\mathbf{x}\right\}\\
\leq\ &\sum_{\mathbf{x}}\mathbf{P}\left\{\mathbf{X}_i=\mathbf{x}\right\}\left(2e^{-2(n_+-1)\varepsilon^2/(1+\gamma)^2M^2}+2e^{-2(n_--1)\varepsilon^2/(1+\gamma)^2M^2}\right)\\
=\ &2e^{-2(n_+-1)\varepsilon^2/(1+\gamma)^2M^2}+2e^{-2(n_--1)\varepsilon^2/(1+\gamma)^2M^2}.
\end{aligned}$$

In a similar way, it can be shown that for $i\in I_-$,

$$\mathbf{P}\left\{\left|\widehat{h}_i-h\left(\mathbf{X}_i\right)\right|>\varepsilon\right\}\leq 2e^{-2(n_+-1)\varepsilon^2/(1+\gamma)^2M^2}+2e^{-2(n_--1)\varepsilon^2/(1+\gamma)^2M^2}.$$

Then,

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{\alpha\in A}\left|H_n(\alpha)-H(\alpha)\right|>\varepsilon\right\}=\mathbf{P}\left\{\sup_{\alpha\in A}\left|\sum_{i=1}^{n}\alpha_i Y_i\left(\widehat{h}_i-h(\mathbf{X}_i)\right)\right|>\varepsilon\right\}\\
\leq\ &\mathbf{P}\left\{\sup_{\alpha\in A}\sum_{i=1}^{n}\alpha_i\left|Y_i\right|\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\varepsilon\right\}\\
=\ &\mathbf{P}\left\{\sup_{\alpha\in A}\sum_{i\in I_+}^{n}\alpha_i\left|\widehat{h}_i-h(\mathbf{X}_i)\right|+\sum_{i\in I_-}^{n}\alpha_i\gamma\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\varepsilon\right\}\\
\leq\ &\mathbf{P}\left\{\sup_{\alpha\in A}\sum_{i\in I_+}^{n}\alpha_i\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\frac{\varepsilon}{1+\gamma}\right\}+\mathbf{P}\left\{\sup_{\alpha\in A}\sum_{i\in I_-}^{n}\alpha_i\gamma\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\frac{\gamma\varepsilon}{1+\gamma}\right\}\\
=\ &\mathbf{P}\left\{\max_{i\in I_+}\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\frac{\varepsilon}{1+\gamma}\right\}+\mathbf{P}\left\{\max_{i\in I_-}\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\frac{\varepsilon}{1+\gamma}\right\}\\
=\ &\mathbf{P}\left\{\bigcup_{i\in I_+}\left\{\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\frac{\varepsilon}{1+\gamma}\right\}\right\}+\mathbf{P}\left\{\bigcup_{i\in I_-}\left\{\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\frac{\varepsilon}{1+\gamma}\right\}\right\}\\
\leq\ &\sum_{i\in I_+}\mathbf{P}\left\{\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\frac{\varepsilon}{1+\gamma}\right\}+\sum_{i\in I_-}\mathbf{P}\left\{\left|\widehat{h}_i-h(\mathbf{X}_i)\right|>\frac{\varepsilon}{1+\gamma}\right\}\\
\leq\ &n_+\left(2e^{-2(n_+-1)\varepsilon^2/(1+\gamma)^4M^2}+2e^{-2(n_--1)\varepsilon^2/(1+\gamma)^4M^2}\right)\\
&+n_-\left(2e^{-2(n_+-1)\varepsilon^2/(1+\gamma)^4M^2}+2e^{-2(n_--1)\varepsilon^2/(1+\gamma)^4M^2}\right)\\
=\ &n\left(2e^{-2(n_+-1)\varepsilon^2/(1+\gamma)^4M^2}+2e^{-2(n_--1)\varepsilon^2/(1+\gamma)^4M^2}\right).
\end{aligned}$$

4.8.2 Proof of Theorem 4.3

Define $\mathbf{u}=(u_1,\ldots,u_n)$ such that $u_i=1/n_+$ for $i\in I_+$ and $u_i=1/n_-$ for $i\in I_-$. By the similar argument for the convergence of MISE of kernel density estimate [57], it can be shown, using a multivariate Taylor series, that

$$\begin{aligned}
&MISE(\mathbf{u};n_+,n_-)=\mathbf{E}\left[ISE(\mathbf{u})\right]\\
=\ &\int Var\left(\widehat{d_\gamma}(\mathbf{x};\mathbf{u})\right)+bias^2\left(\widehat{d_\gamma}(\mathbf{x};\mathbf{u})\right)d\mathbf{x}\\
=\ &\left\{\frac{1}{n_+\sigma^d}+\frac{\gamma^2}{n_-\sigma^d}\right\}R(k)+\frac{1}{4}\sigma^4R\left(tr\left\{\mathscr{H}_{d_\gamma}\right\}\right)+o\left(n_+^{-1}\sigma^{-d}+n_-^{-1}\sigma^{-d}+\sigma^4\right)
\end{aligned}$$

where $R(f)=\int f^2(\mathbf{x})\,d\mathbf{x}$ and $\mathscr{H}_f$ represent the Hessian matrix of f. Therefore, $ISE(\mathbf{u})$ converges to 0 in probability since $\sigma\to 0$, $n_+\sigma^d\to\infty$ and $n_+\sigma^d\to\infty$ as $n\to\infty$. Furthermore,

$$\begin{aligned}\mathbf{P}\{ISE(\widehat{\alpha})>\varepsilon\} &= \mathbf{P}\left\{ISE(\widehat{\alpha})>\varepsilon, ISE(\mathbf{u})>\frac{\varepsilon}{2}\right\}+\mathbf{P}\left\{ISE(\widehat{\alpha})>\varepsilon, ISE(\mathbf{u})\leq\frac{\varepsilon}{2}\right\}\\ &\leq \mathbf{P}\left\{ISE(\mathbf{u})>\frac{\varepsilon}{2}\right\}+\mathbf{P}\left\{ISE(\widehat{\alpha})>ISE(\mathbf{u})+\frac{\varepsilon}{2}\right\}.\end{aligned}$$

From the consistency of $ISE(\mathbf{u})$ and the oracle inequality stated in Theorem 4.2, $ISE(\widehat{\alpha})$ converges to 0 in probability.

4.8.3 Proof of Theorem 4.4

First note that in the previous analyses we treat N_+, N_- and γ as deterministic variables but now we turn to the case where these variables are random. Thus, some of the previous results should be restated considering this.

Lemma 4.7. *γ converges to γ^* with probability 1.*

Proof. Note that N_+ and N_- are binomial random variables with (n,p) and (n,q) where $q=1-p$. From the Hoeffding's inequality, we know that for $\forall\varepsilon>0$

$$\begin{aligned}&\mathbf{P}\left\{\frac{N_+}{n}-p>\varepsilon\right\}\leq e^{-2n\varepsilon^2}, \quad \mathbf{P}\left\{\frac{N_+}{n}-p<-\varepsilon\right\}\leq e^{-2n\varepsilon^2}\\ &\mathbf{P}\left\{\frac{N_-}{n}-q>\varepsilon\right\}\leq e^{-2n\varepsilon^2}, \quad \mathbf{P}\left\{\frac{N_-}{n}-q<-\varepsilon\right\}\leq e^{-2n\varepsilon^2}.\end{aligned}$$

Then, for any $\varepsilon > 0$

$$\begin{aligned}
\mathbf{P}_n(\varepsilon) &\triangleq \mathbf{P}\left\{\left|\frac{N_-}{N_+} - \frac{q}{p}\right| > \varepsilon\right\} = \mathbf{P}\{|pN_- - qN_+| > \varepsilon pN_+\} \\
&= \mathbf{P}\left\{|pN_- - qN_+| > \varepsilon pN_+, N_+ \geq \frac{np}{2}\right\} + \mathbf{P}\left\{|pN_- - qN_+| > \varepsilon pN_+, N_+ < \frac{np}{2}\right\} \\
&\leq \mathbf{P}\left\{|pN_- - qN_+| > \varepsilon p \cdot \frac{np}{2}\right\} + \mathbf{P}\left\{N_+ < \frac{np}{2}\right\} \\
&\leq \mathbf{P}\left\{|pN_- - pqn + pqn - qN_+| > \frac{n\varepsilon p^2}{2}\right\} + \mathbf{P}\left\{N_+ - pn < -\frac{np}{2}\right\} \\
&\leq \mathbf{P}\left\{|pN_- - pqn| > \frac{n\varepsilon p^3}{2}\right\} + \mathbf{P}\left\{|qN_+ - pqn| > \frac{n\varepsilon p^2 q}{2}\right\} + \mathbf{P}\left\{N_+ - pn < -\frac{np}{2}\right\} \\
&= \mathbf{P}\left\{\left|\frac{N_-}{n} - q\right| > \frac{\varepsilon p^2}{2}\right\} + \mathbf{P}\left\{\left|\frac{N_+}{n} - p\right| > \frac{\varepsilon p^2}{2}\right\} + \mathbf{P}\left\{\frac{N_+}{n} - p < -\frac{p}{2}\right\} \\
&\leq 4\exp\left(-\frac{n\varepsilon^2 p^4}{2}\right) + \exp\left(-\frac{np^2}{2}\right).
\end{aligned}$$

Since $\sum_{n=1}^{\infty} \mathbf{P}_n(\varepsilon) < \infty$ for all $\varepsilon > 0$, γ converges to γ^* with probability 1. □

Lemma 4.8. *Suppose the assumptions in Theorem 4.4 are satisfied. For any* $\varepsilon' > 0$, $\mathbf{P}\{ISE(\widehat{\alpha}) > \inf_{\alpha \in A} ISE(\alpha) + \varepsilon'\}$ *converges to 0.*

Proof. We need to restate Theorem 4.2 as follows. For any $\delta > 0$,

$$\mathbf{P}\left\{ISE(\widehat{\alpha}) > \inf_{\alpha \in A} ISE(\alpha) + 4\sqrt{\frac{\ln(2n/\delta)}{c[\min(N_+, N_-) - 1]}}\,\middle|\, N_+ = n_+, N_- = n_-\right\} \leq \delta$$

since

$$\sqrt{\frac{\ln(2n/\delta)}{c[\min(n_+, n_-) - 1]}} \leq \varepsilon \leq \sqrt{\frac{\ln(2n/\delta)}{c[\max(n_+, n_-) - 1]}}.$$

Let us define $c' = 2\left(\sqrt{2\pi}\sigma\right)^{2d} / (1 + 2\gamma^*)^4$ and an event $D = \left\{N_+ \geq \frac{np}{2}, N_- \geq \frac{n(1-p)}{2}, \gamma \leq 2\gamma^*\right\}$.

Then,

$$\begin{aligned}
&\mathbf{P}\left\{ISE(\widehat{\alpha}) > \inf_{\alpha \in A} ISE(\alpha) + 4\sqrt{\frac{2\ln(2n/\delta)}{c'[\min(np, n(1-p)) - 1]}}\right\} \\
&\leq \mathbf{P}\{D^c\} + \mathbf{P}\{D\} \cdot \mathbf{P}\left\{ISE(\widehat{\alpha}) > \inf_{\alpha \in A} ISE(\alpha) + 4\sqrt{\frac{2\ln(2n/\delta)}{c'[\min(np, n(1-p)) - 1]}}\,\middle|\, D\right\}.
\end{aligned}$$

The first term converges to 0 from the strong law of large numbers and Lemma 4.7. The second term becomes

$$
\begin{aligned}
&\mathbf{P}\left\{ ISE\left(\widehat{\alpha}\right) > \inf_{\alpha \in A} ISE\left(\alpha\right) + 4\sqrt{\frac{2\ln\left(2n/\delta\right)}{c'[\min\left(np, n\left(1-p\right)\right)-1]}} \middle| D \right\} \\
\leq\ & \mathbf{P}\left\{ ISE\left(\widehat{\alpha}\right) > \inf_{\alpha \in A} ISE\left(\alpha\right) + 4\sqrt{\frac{\ln\left(2n/\delta\right)}{c[\min\left(N_+, N_-\right)-1]}} \middle| D \right\} \\
=\ & \sum \mathbf{P}\left\{ ISE\left(\widehat{\alpha}\right) > \inf_{\alpha \in A} ISE\left(\alpha\right) + 4\sqrt{\frac{\ln\left(2n/\delta\right)}{c[\min\left(N_+, N_-\right)-1]}} \middle| D, N_+ = n_+, N_- = n_- \right\} \\
& \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
\leq\ & \sum \delta \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} = \delta.
\end{aligned}
$$

For any $\delta > 0$, we can make $4\sqrt{\frac{2\ln(2n/\delta)}{c'[\min(np, n(1-p))-1]}}$ smaller than ε' as $n \to \infty$, provided that $\ln n / n\sigma^d \to 0$ as $n \to 0$. Therefore, $\mathbf{P}\left\{ISE\left(\widehat{\alpha}\right) > \inf_{\alpha \in A} ISE\left(\alpha\right) + \varepsilon'\right\}$ converges to 0. □

Lemma 4.9. *Suppose the assumptions in Theorem 4.4 are satisfied. Then, $ISE\left(\mathbf{u}\right)$ converges to 0 in probability.*

Proof. Define an event $D = \left\{N_+ \geq \frac{np}{2}, N_- \geq \frac{n(1-p)}{2}, \gamma \leq 2\gamma^*\right\}$. For any $\varepsilon > 0$,

$$
\mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \varepsilon\right\} \leq \mathbf{P}\left\{D^c\right\} + \mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \varepsilon, D\right\}.
$$

The first term converges to 0 from the strong law of large numbers and Lemma 4.7. Let define a set $S = \left\{(n_+, n_-) \,\middle|\, n_+ \geq \frac{np}{2}, n_- \geq \frac{n(1-p)}{2}, \frac{n_-}{n_+} \leq 2\gamma^*\right\}$. Then,

$$\begin{aligned}
&\mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \varepsilon, D\right\} \\
&= \sum \mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \varepsilon, D \,\middle|\, N_+ = n_+, N_- = n_-\right\} \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
&= \sum_{(n_+,n_-)\in S} \mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \varepsilon \,\middle|\, N_+ = n_+, N_- = n_-\right\} \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
&\leq \sum_{(n_+,n_-)\in S} \frac{\mathbf{E}\left[ISE\left(\mathbf{u}\right) \,\middle|\, N_+ = n_+, N_- = n_-\right]}{\varepsilon} \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
&\leq \frac{1}{\varepsilon} \sum_{(n_+,n_-)\in S} \left[\frac{1}{n\sigma^d}\left(\frac{2}{p} + \frac{8\gamma^{*2}}{1-p}\right) R\left(k\right) + \frac{1}{4}\sigma^4 R\left(tr\left\{\mathscr{H}_{d_\gamma}\right\}\right) + o\left(n^{-1}\sigma^{-d} + \sigma^4\right)\right] \\
&\quad \cdot \mathbf{P}\left\{N_+ = n_+, N_- = n_-\right\} \\
&\leq \frac{1}{\varepsilon}\left(\frac{1}{n\sigma^d}\left\{\frac{2}{p} + \frac{2\gamma^{*2}}{1-p}\right\} R\left(k\right) + \frac{1}{4}\sigma^4 R\left(tr\left\{\mathscr{H}_{d_\gamma}\right\}\right) + o\left(n^{-1}\sigma^{-d} + \sigma^4\right)\right)
\end{aligned}$$

where the second to the last step, we used $MISE\left(\mathbf{u}; n_+, n_-\right)$ formula in explained in Section 4.8.2 and the fact that for $(n_+, n_-) \in S$,

$$\frac{1}{n_+\sigma^d} + \frac{1}{n_-\sigma^d} \leq \frac{2}{np\sigma^d} + \frac{2}{n(1-p)\sigma^d} = \frac{1}{n\sigma^d}\left(\frac{2}{p} + \frac{2}{1-p}\right)$$

Therefore, $ISE\left(\mathbf{u}\right)$ converges to 0 since $\sigma \to 0$ and $n\sigma^d \to \infty$ as $n \to \infty$. □

Now let's prove Theorem 4.4. From Theorem 3 in [72], it suffices to show that

$$\int \left(\widehat{d_\gamma}\left(\mathbf{x}; \widehat{\alpha}\right) - d_{\gamma^*}\left(\mathbf{x}\right)\right)^2 d\mathbf{x} \to 0$$

in probability. Note that

$$\begin{aligned}
\|\widehat{d_\gamma}\left(\mathbf{x}; \widehat{\alpha}\right) - d_{\gamma^*}\left(\mathbf{x}\right)\|_{L^2} &= \|\widehat{d_\gamma}\left(\mathbf{x}; \widehat{\alpha}\right) - d_\gamma\left(\mathbf{x}\right) + \left(\gamma - \gamma^*\right) f_-\left(\mathbf{x}\right)\|_{L^2} \\
&\leq \|\widehat{d_\gamma}\left(\mathbf{x}; \widehat{\alpha}\right) - d_\gamma\left(\mathbf{x}\right)\|_{L_2} + \|\left(\gamma - \gamma^*\right) f_-\left(\mathbf{x}\right)\|_{L^2} \\
&= \sqrt{ISE\left(\widehat{\alpha}\right)} + |\gamma - \gamma^*| \cdot \|f_-\left(\mathbf{x}\right)\|_{L^2}. \qquad (4.18)
\end{aligned}$$

For the first term in (4.18), $\mathbf{P}\left\{ISE\left(\widehat{\alpha}\right) > \varepsilon\right\}$ converges to 0 in probability since

$$\mathbf{P}\left\{SE\left(\widehat{\alpha}\right) > \varepsilon\right\} \leq \mathbf{P}\left\{ISE\left(\widehat{\alpha}\right) > ISE\left(\mathbf{u}\right) + \frac{\varepsilon}{2}\right\} + \mathbf{P}\left\{ISE\left(\mathbf{u}\right) > \frac{\varepsilon}{2}\right\}$$

and from Lemma 4.8 and 4.9. The second term in (4.18) also converges to 0 in probability from Lemma 4.7. This proves the theorem.

CHAPTER 5

Conclusion

In this thesis, we present classification algorithms for both an application-driven problem (Chapter 2) and more theoretically driven problems (Chapter 3 and Chapter 4). When we dealing with these problems, we focus on kernel methods that have been quite successful and effective in various applications.

In Chapter 2, we apply a kernel method to a specific medical application, where we have to predict ICU admission for post-operative patients with possible sepsis. We have shown that feature extraction based on heuristics, paired with kernel methods, can lead to significant performance gain over those used by clinicians. As a future work, more experiments could be done if well-cleaned data is available. It would be also interesting to find more theoretically-driven features and see how they work compared to the proposed features. Another future research direction would be to design a suitable kernel for irregularly sampled vital signs that captures the similarity between patients.

In Chapter 3, we develop the novel, nonparametric density estimation for contaminated data. The RKDE is obtained as a robust sample mean in the RKHS associated to the kernel through M-estimation. The RKDE is expressed as a weighted kernel density estimate with a robust property that smaller weights are given to more outlying data points. The investigation of the influence function further confirms that RKDEs are less sensitive to contamination than traditional KDEs. 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.

In Chapter 4, the L_2 kernel classification method is proposed that minimizes the L_2 distance between "difference of densities". Like the SVM, it is the solution of a convex quadratic program, and has a sparse representation as well as the geometric interpretation in the associated RKHS. Statistical performance guarantees on the L_2 kernel classifier are established, which also specialize to give performance guarantees for an existing method of L_2 kernel density estimation.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] I. S. Abramson. On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics*, 10(4):1217–1223, 1982.

[2] C. Adams and R. Franzosa. *Introduction to Topology Pure and Applied*. Pearson Prentice Hall, New Jersey, 2008.

[3] American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Critical Care Medicine*, 20:864–874, 1992.

[4] K. Bennett and E. Bredensteiner. Duality and geometry in SVM classifiers. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 57–64, San Francisco, 2000. Morgan Kaufmann.

[5] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces In Probability And Statistics*. Kluwer Academic Publishers, Norwell, 2004.

[6] D. Berry, K. Chaloner, and J. Geweke. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*. Wiley, New York, 1996.

[7] Grant V. Bochicchio and et al. Persistent systemic inflammatory response syndrome is predictive of nosocomial infection in trauma. *The Journal of Trauma*, 53(2):245–250, 2002.

[8] 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.

[9] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

[10] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparse density estimation with l_1 penalties. *Proceedings of 20th Annual Conference on Learning Theory, COLT 2007, Lecture Notes in Artificial Intelligence, v4539*, pages 530– 543, 2007.

[11] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[12] Yixin Chen, Xin Dang, Hanxiang Peng, and Henry Bart. Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):288–305, 2009.

[13] Parminder Chhabra, Clayton D. Scott, Eric D. Kolaczyk, and Mark Crovella. Distributed spatial anomaly detection. *Proc. IEEE Conference on Computer Communications (INFOCOM)*, pages 1705–1713, 2008.

[14] A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

[15] 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.

[16] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[17] John C.Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Technical Report MSR-TR-98-14*, April 2001.

[18] D. Crisp and C. Burges. A geometric interpretation of ν-SVM classifiers. In *Neural Information Processing Systems 12*, 1999.

[19] 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*, 2008.

[20] 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, 2008.

[21] Michiel Debruyne, Mia Hubert, and Johan Van Horebeek. Detecting influential observations in kernel pca. *Computational Statistics & Data Analysis*, 54:3007–3019, 2010.

[22] R. Phillip Dellinger and et al. Surviving sepsis campaign guidelines for management of severe sepsis and septic shock. *Critical Care Medicine*, 36(1):296–327, 2008.

[23] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001.

[24] Ronald Aylmer Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[25] Mark Girolami and Chao He. Probability density estimation from optimally condensed data samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1253–1264, OCT 2003.

[26] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *J. Machine Learning Res.*, 6:2075–2129, 2005.

[27] Peter Hall and Matthew P.Wand. On nonparametric discrimination using density differeces. *Biometrika*, 75(3):541–547, Sept 1988.

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

[29] Chao He and Mark Girolami. Novelty detection employing an L_2 optimal nonparametric density estimator. *Pattern Recognition Letters 25*, 25:1389–1397, 2004.

[30] C.R. Heathcote. The integrated squared error estimation of parameters. *Biometrika*, 64:255–264, 1977.

[31] P. Huber. *Robust Statistics*. Wiley, New York, 1981.

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

[33] 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.

[34] F. Jaimes, J. Farbiarz, Diego Alvarez, and Carlos Martínez. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Critical Care*, 9:R150–R156, 2005.

[35] Suykens J.A.K. and Vandewalle J. Least squares support vector machine classifiers. *Neural Processing Letters*, 44(8):293–300, Jun 1999.

[36] Robert Jenssen, Deniz Erdogmus, Jose C.Principe, and Torbjørn Eltoft. Towards a unification of information theoretic learning and kernel method. In *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP2004),* Sao Luis, Brazil, pages 93–102, Sept 2004.

[37] D. Kim. *Least Squares Mixture Decomposition Estimation*. unpublished doctoral dissertation, Dept. of Statistics, Virginia Polytechnic Inst. and State Univ., 1995.

[38] J. Kim and C. Scott. L_2 kernel classification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(10):1822–1831, 2010.

[39] 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.

[40] Serge Lang. *Real and Functional Analysis*. Spinger, New York, 1993.

[41] 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.

[42] 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.

[43] Roderick Little and Donald Rubin. *Statistical analysis with missing data*. Wiley, 2nd edition, 2002.

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

[45] R. S. G. Mahapatruni and A. Gray. CAKE: Convex adaptive kernel density estimation. In G. Gordon, D. Dunson, 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.

[46] M. Di Marzio and C.C. Taylor. Kernel density classification and boosting: an L_2 analysis. *Statistics and Computing*, 15:113–123(11), April 2005.

[47] P. Meinicke, T. Twellmann, and H. Ritter. Discriminative densities from maximum contrast estimation. In *Advances in Neural Information Proceeding Systems 15,* Vancouver, Canada, pages 985–992, 2002.

[48] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12:181–201, Mar 2001.

[49] J. R. Munkres. *Topology*. Prentice Hall, 2000.

[50] James M. O'Brien and et al. Sepsis. *The American Journal of Medicine*, 120:1012–1022, 2007.

[51] A.S. Paulson, E.W. Holcomb, and R.A. Leitch. The estimation of the parameters of the stable laws. *Biometrika*, 62:163–170, 1975.

[52] Kristiaan Pelckmans, Johan A.K. Suykens, and Bart De Moor. A risk minimization principle for a class of parzen estimators. *Advances in Neural Information Processing Systems 20*, December 2007.

[53] Ph. Rigollet and A.B. Tsybakov. Linear and convex aggregation of density estimators. `https://hal.ccsd.cnrs.fr/ccsd-00068216`, 2004.

[54] Jonathan R.Schechuk. An introduction to the conjugate gradient method without the agonizing pain. *Technical Report MSR-TR-98-14*, August 1994.

[55] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. *Proc. Annu. Conf. Comput. Learning Theory*, pages 416–426, 2001.

[56] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[57] D. W. Scott. *Multivariate Density Estimation*. Wiley, New York, 1992.

[58] Clint Scovel, Don Hush, Ingo Steinwart, and James Theiler. Radial kernels and their reproducing kernel hilbert spaces. *Journal of Complexity*, 26:641–660, 2010.

[59] 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.

[60] B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall/CR, New York, 1986.

[61] L. Song, X. Zhang, A. 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.

[62] Charles L. Sprung and et al. An evaluation of systemic inflammatory response syndrome signs in the Sepsis Occurrence in Acutely ill Patients (SOAP) study. *Intensive Care Medicine*, 32(3):421–427, 2006.

[63] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

[64] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.

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

[66] 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.

[67] Todd R. Vogel, Viktor T. Dombrovskiy, and Stephen F. Lowry. Trends in postoperative sepsis: Are we improving outcomes? *Surgical infections*, 10(1):71–78, 2009.

[68] M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

[69] B. Wang. *The prediction of severe sepsis using SVM model*. PhD thesis, National Chung Chen University, 2006.

[70] A. Wibowo. Robust kernel ridge regression based on M-estimation. *Computational Mathematics and Modeling*, 20(4), 2009.

[71] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[72] Charles T. Wolverton and Terry J. Wagner. Asymptotically optimal discriminant fucntions for pattern classification. *IEEE Trans. Info. Theory*, 15(2):258–265, Mar 1969.

[73] David W.Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics 43*, pages 274–285, 2001.

[74] 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.

[75] 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.