Functional Analytic Perspectives on Nonparametric Density Estimation

by

Robert A. Vandermeulen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2016

Doctoral Committee:

Associate Professor Clayton Scott, Chair
Professor Alfred Hero
Assistant Professor Rajesh Nadakuditi
Assistant Professor Ambuj Tewari

For all the people.

ACKNOWLEDGEMENTS

I would like to thank my parents, my family, and my friends, especially my graduate school friends Eric, Madison, Matt, Mitch, Nick, Pat, and Paul for the moral support.

Thank you to my committee for taking the time to review my thesis.

Finally I would like to thank Professor Scott for his guidance as well as for being the sole professor who accepted my graduate school application.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Functional Analytic Perspectives on Nonparametric Density Estimation

by

Robert A. Vandermeulen

Chair: Clayton Scott

Nonparametric density estimation is a classic problem in statistics. In the standard estimation setting, when one has access to iid samples from an unknown distribution, there exist several established and well-studied nonparametric density estimators. Yet there remains interesting alternative settings which are less well-studied. This work considers two such settings. First we consider the case where the data contains some contamination, i.e. a portion of the data is not distributed according to the density we would like to estimate. In this setting one would like an estimator which is robust to the contaminating data. An approach to this was suggested in *Kim and Scott* (2012). The estimator in that paper was analytically and experimentally shown to be robust, but no consistency result was presented. In Chapter II it is demonstrated that this estimator is indeed consistent for a class of convex losses. Chapter III introduces a new robust kernel density estimator based on scaling and projection in Hilbert space. This estimator is proven to be consistent and will converge to the true density provided certain assumptions on the contaminating distribution. Its efficacy is demonstrated experimentally by applying it to several datasets. Chapter IV considers a different setting which can be thought of as nonparametric mixture modelling. Here one would

like to estimate multiple densities with access to groups of samples where each sample in a group is known to be distributed according the same unknown density. Tight identifiability bounds and a highly general algorithm for recovery of the densities are presented for this setting.

Functional analysis is a unifying theme of these problems. Hilbert spaces in particular are used extensively for the construction of estimators and mathematical analysis.

CHAPTER I

Introduction

There are two major thrusts of research presented in this thesis and they will be introduced separately.

1.1 Kernel Density Estimation

Density estimation is one of the oldest problems in statistics. Given data one would like to estimate its underlying distribution. Oftentimes the data is known to come from a parametric class of distributions, such as the class Gaussian distributions. In this case one simply needs to estimate the parameters of the distribution. When the data come form a class which is too complicated to effectively model, or perhaps wholly unknown distribution, one resorts to using a nonparametric density estimator. There are several examples of such estimators, but arguably the most commonly used estimator is the kernel density estimator. The estimator is as follows. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pdf and $X_1, \ldots, X_n$ be iid samples from f. Let $k_\sigma(x, x')$ be a radial smoothing kernel of the form $k_\sigma(x, x') = \sigma^{-d} q(\|x - x'\|_2 / \sigma)$ for some function $q \geq 0$ such that $q(\|\cdot\|_2)$ is a pdf on $\mathbb{R}^d$. Then

$$\bar{f}_\sigma^n := \frac{1}{n} \sum_{i=1}^{n} k_\sigma(\cdot, X_i)$$

is the well-known kernel density estimator (KDE) (*Silverman* (1986), *Scott* (1992), *Devroye and Lugosi* (2001)).

This estimator has many desirable properties. Foremost it is universally consistent. If we allow $n \to \infty$ and $\sigma \to 0$ with a rate satisfying $n\sigma^d \to \infty$ then we have that $\left\| f - \bar{f}_\sigma^n \right\| \xrightarrow{p} 0$ in both the 1 and 2 norms *Devroye and Lugosi* (2001). With more restrictive assumptions on the kernel, density f, and rate on σ the consistency extends further to the ∞ norm *Giné and Guillou* (2002). The KDE also avoids boundary issues associated with another popular nonparametric density estimator, the histogram *Silverman* (1986).

One issue with kernel density estimators is a lack of robustness. This work contains two major contributions to the problem of robust kernel density estimation. The first contribution is proving the consistency of a previously proposed robust kernel density estimator.

1.1.1 Consistency of The Robust Kernel Density Estimator

In *Kim and Scott* (2012) the authors suggest a modification of the KDE to induce robustness. In order to construct this estimator we additionally assume that k_σ is positive-semidefinite. Thus $k_\sigma(x, x') = \langle \Phi_\sigma(x), \Phi_\sigma(x') \rangle_{\mathcal{H}_\sigma}$, where $\mathcal{H}_\sigma$ is the reproducing kernel Hilbert space (RKHS) associated with k_σ (*Aronszajn*, 1950), and $\Phi_\sigma(x) := k_\sigma(\cdot, x)$ is the canonical feature map (*Steinwart and Christmann*, 2008). Some kernels satisfying these properties include the multivariate Gaussian, Laplacian, and Student kernels.

With this notation, the KDE may be written as

$$\bar{f}_\sigma^n = \frac{1}{n} \sum_{i=1}^{n} \Phi_\sigma(X_i),$$

the mean of the mapped data. The sample mean is easily shown to be the unique

solution of a least squares problem

$$\bar{f}_\sigma^n = \underset{g \in \mathcal{H}_\sigma}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left\| g - \Phi_\sigma \left(X_i \right) \right\|_{\mathcal{H}_\sigma}^2 .$$

Replacing the squared loss with a robust loss ρ, yields a *robust kernel density estimator*:

$$f_\sigma^n = \underset{g \in \mathcal{H}_\sigma}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho \left(\left\| g - \Phi_\sigma \left(X_i \right) \right\|_{\mathcal{H}_\sigma} \right) . \tag{1.1}$$

This construction was first introduced by *Kim and Scott* (2012) where they established several properties including a representer theorem, a convergent iterative algorithm, and the influence function. The representer theorem states that

$$f_\sigma^n = \sum_{i=1}^{n} \alpha_i \Phi_\sigma \left(X_i \right) ,$$

where $\alpha_i \geq 0$ and $\sum_1^n \alpha_i = 1$.

In this work we will establish consistency of the RKDE in the L^1 norm for a class of convex losses.

1.1.2 Related Work

The consistency of kernel density estimators has been established under the L^1 norm with very weak assumptions on distribution and kernel (*Devroye and Lugosi*, 2001). Necessary conditions on n and σ for L^1 consistency of the KDE are $n \to \infty$ with $\sigma \to 0$ and rate on bandwidth $n\sigma^d \to \infty$. Sup-norm consistency has also been established for a less general class of kernels and densities requiring more restrictive regularity conditions (*Silverman* (1978), *Stute* (1982), *Einmahl and Mason* (2000), *Deheuvels* (2000), *Giné and Guillou* (2002), *Gine et al.* (2004), *Wied and Weissbach* (2012)).

Consistency proofs tend to proceed by decomposing the error into a stochastic estimation error and a non-stochastic approximation error, namely

$$\left\|\bar{f}_\sigma^n - f\right\| \leq \left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\| + \left\|\bar{f}_\sigma - f\right\|,$$

Where $\bar{f}_\sigma = \int k_\sigma(\cdot, x) f(x)\, dx = \int \Phi_\sigma(x) f(x)\, dx$. The right summand is shown to go to zero analytically and the left summand is shown to go to zero with techniques from empirical process theory. We will show a simple proof of the consistency of the KDE using this decomposition and Bennett's inequality for Hilbert space to control the stochastic term. However, this decomposition is less fruitful for the RKDE, for which f_σ does not have a closed form expression (see Section 5.1). Instead, we use a completely different technique by investigating the convergent iterative algorithm used to compute the RKDE in *Kim and Scott* (2012).

1.1.3 Scale and Project Kernel Density Estimator

In Chapter III we introduce a new robust kernel density estimator. We consider the situation where most observations come from a target density f_{tar} but some observations are drawn from a contaminating density f_{con}, so our observed samples come from the density $f_{obs} = (1-\varepsilon) f_{tar} + \varepsilon f_{con}$. It is not known which component a given observation comes from. When considering this scenario in the infinite sample setting we would like to construct some transform that, when applied to f_{obs}, yields f_{tar}. We introduce a new formalism to describe transformations that "decontaminate" f_{obs} under sufficient conditions on f_{tar} and f_{con}. We focus on a specific nonparametric condition on f_{tar} and f_{con} that reflects the intuition that the contamination manifests in low density regions of f_{tar}. In the finite sample setting, we seek a nonparametric density estimator that converges to f_{tar} asymptotically. Thus, we construct a weighted KDE where the kernel weights are lower in low density regions and higher in high

density regions. To do this we multiply the standard KDE by a real value greater than one (scale) and then find the closest pdf to the scaled KDE in the L^2 norm (project), resulting in a scaled and projected kernel density estimator (SPKDE). Because the squared L^2 norm penalizes point-wise differences between functions quadratically, this causes the SPKDE to draw weight from the low density areas of the KDE and move it to high density areas to get a more uniform difference to the scaled KDE. The asymptotic limit of the SPKDE is a scaled and shifted version of f_{obs}. Given our proposed sufficient conditions on f_{tar} and f_{con}, the SPKDE can asymptotically recover f_{tar}.

In this work we present a new formalism for nonparametric density estimation, necessary and sufficient conditions for decontamination, the construction of the SPKDE, and a proof of consistency. We also include experimental results applying the algorithm to benchmark datasets with comparisons to the RKDE, traditional KDE, and an alternative robust KDE implementation. Many of our results and proof techniques are novel in KDE literature.

1.2 Nonparametric Mixture Models

Chapter IV addresses a different sort of problem which is related to mixture modelling. A finite mixture model $\mathscr{P}$ is a probability measure over a space of probability measures where $\mathscr{P}(\{\mu_i\}) = w_i > 0$ for some finite collection of probability measures $\mu_1, \ldots, \mu_m$ and $\sum_{i=1}^{m} w_i = 1$. A realization from this mixture model first randomly selects some mixture component $\mu \sim \mathscr{P}$ and then draws from μ. Mixture models have seen extensive use in statistics and machine learning.

A central theoretical question concerning mixture models is that of identifiability. A mixture model is said to be *identifiable* if there is no other mixture model that defines the same distribution over the observed data. Classically mixture models were concerned with the case where the observed data $X_1, X_2, \ldots$ are iid with X_i

distributed according to some unobserved random measure μ_i with $\mu_i \overset{iid}{\sim} \mathscr{P}$. This situation is equivalent to $X_i \overset{iid}{\sim} \sum_{j=1}^{m} w_j \mu_j$. If we impose no restrictions on the mixture components $\mu_1, \ldots, \mu_m$ one could easily concoct many choices of μ_j and w_j which yield an identical distribution on X_i. Because of this, most previous work on identifiability assumes some sort of structure on $\mu_1, \ldots, \mu_m$, such as Gaussianity *Anderson et al.* (2014); *Bruni and Koch* (1985); *Yakowitz and Spragins* (1968). In this work we consider an alternative scenario where we make no assumptions on $\mu_1, \ldots, \mu_m$ and instead have access to groups of samples that are known to come from the same component. We will call these groups of samples "random groups." Mathematically a random group is a random element X_i where $X_i = (X_{i,1}, \ldots, X_{i,n})$ with $X_{i,1}, \ldots, X_{i,n} \overset{iid}{\sim} \mu_i$ and $\mu_i \overset{iid}{\sim} \mathscr{P}$.

In this setting identifiability is now concerned with the distribution over X_i and the value of n, the number of samples in each random group. We call a mixture of measures $\mathscr{P}$ *n-identifiable* if it is the *simplest* mixture model (in terms of number of mixture components) that yields the observed distribution on X_i. We also introduce a concept which is stronger than identifiability. We call $\mathscr{P}$ *n-determined* if it is the *only* mixture model that yields the observed distribution on X_i. In this work we show that every mixture model with m components is $(2m-1)$-identifiable and $2m$-determined. Furthermore we show that any mixture model with linearly independent components is 3-identifiable and 4-determined. We also show that a mixture model with jointly irreducible components is 2-determined. These results hold for any mixture model over any space and cannot be improved. Finally, using these results, we demonstrate some new and old results on the identifiability of multinomial mixture models.

We also include algorithms for the recovery of the mixture components culminating in a algorithm for the recovery of mixtures of categorical distributions with m arbitrary mixture components provided $2m-1$ samples per group. We include experimental results showing that this algorithm does indeed recover the mixture

components from data.

1.2.1 Previous Work

In classical mixture model theory identifiability is achieved by making assumptions about the mixture components. Some assumptions which yield identifiability are Gaussian or binomial mixture components *Bruni and Koch* (1985); *Teicher* (1963). If one makes no assumptions on the mixture components then one must leverage some other type of structure in order to achieve identifiability. An example of such structure exists in the context of multiview models. In a multiview model samples have the form $X_i = (X_{i,1}, \ldots, X_{i,n})$ and the distribution of X_i is defined by $\sum_{i=1}^{m} w_i \prod_{j=1}^{n} \mu_i^j$. In *Allman et al.* (2009) it was shown that if μ_i^j are probability distributions on $\mathbb{R}$ with $\mu_1^j, \ldots, \mu_m^j$ linearly independent for all j and $n \geq 3$, then the model is identifiable.

The setting which we investigate is a special case of the multiview model where $\mu_i^j = \mu_i^{j'}$ for all i, j, j'. If the sample space of the μ_i is finite then this problem is exactly the topic modelling problem with a finite number of topics and one topic for each document. In topic modelling each μ_i is a "topic" and the sample space is a finite collection of words. This setting is well studied and it has been shown that one can recover the true topics provided certain assumptions on the topics *Allman et al.* (2009); *Anandkumar et al.* (2014); *Arora et al.* (2012). This problem was studied for arbitrary topics in *Rabani et al.* (2014). In this paper the authors introduce an algorithm that recovers any mixture of m topics provided $2m-1$ words per document. They also show, in a result analogous to our own, that this $2m - 1$ value cannot be improved. Our proof techniques are quite different than those used in *Rabani et al.* (2014), hold for arbitrary sample spaces, and are less complex. Additional connections to previous work are given in Chapter IV.

CHAPTER II

Consistency of Robust Kernel Density Estimators

In this chapter we present a proof of the consistency of the robust kernel density estimator (RKDE) described in *Kim and Scott* (2012). First we will introduce the statistical setting for the estimator and quickly review the classic kernel density estimator (KDE). Next we demonstrate a new proof of the consistency of the KDE. Components of this proof will be useful for proving the consistency of the RKDE. After this we introduce the RKDE and a few results from *Kim and Scott* (2012). Then we will prove the consistency of the RKDE. For readability many of the lemmas will only include proof sketches and full proofs can be found at the end of the chapter.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pdf and $X_1, \ldots, X_n$ be iid samples from f. Let $k_\sigma(x, x')$ be a radial smoothing kernel of the form $k_\sigma(x, x') = \sigma^{-d} q(\|x - x'\|_2 / \sigma)$ for some function $q \geq 0$ such that $q(\|\cdot\|_2)$ is a pdf on $\mathbb{R}^d$. Then

$$\bar{f}_\sigma^n := \frac{1}{n} \sum_{i=1}^{n} k_\sigma(\cdot, X_i)$$

is the well-known KDE (*Silverman* (1986), *Scott* (1992), *Devroye and Lugosi* (2001)). We will additionally assume that k_σ is positive semi-definite. Thus $k_\sigma(x, x') = \langle \Phi_\sigma(x), \Phi_\sigma(x') \rangle_{\mathcal{H}_\sigma}$, where $\mathcal{H}_\sigma$ is the reproducing kernel Hilbert space (RKHS) associated with k_σ (*Aronszajn*, 1950), and $\Phi_\sigma(x) := k_\sigma(\cdot, x)$ is the canonical feature map (*Steinwart and Christmann*, 2008). With this notation we have that the KDE

can be represented as

$$\bar{f}_\sigma^n = \frac{1}{n}\sum_{i=1}^{n} \Phi_\sigma(X_i).$$

Using basic techniques from calculus of variations it is straightforward to show that the KDE is equal to the minimizer of a least squares problem

$$\bar{f}_\sigma^n = \arg\min_{g\in\mathcal{H}_\sigma} \frac{1}{n}\sum_{i=1}^{n} \left\|g - \Phi_\sigma\left(X_i\right)\right\|_{\mathcal{H}_\sigma}^2 .$$

By replacing the squared loss with a robust loss ρ we arrive at the RKDE from *Kim and Scott* (2012)

$$\arg\min_{g\in\mathcal{H}_\sigma} \frac{1}{n}\sum_{i=1}^{n} \rho\left(\left\|g - \Phi_\sigma\left(X_i\right)\right\|_{\mathcal{H}_\sigma}\right).$$

Note that for radial kernels we have

$$\begin{aligned}\left\|\Phi_\sigma\left(x\right)\right\|_{\mathcal{H}_\sigma} =& \sqrt{\sigma^{-d} q\left(\left\|x - x\right\|_2 / \sigma\right)} \\ =& \sqrt{q(0)}\sigma^{-d/2}\end{aligned}$$

which does not depend on x. Because of this, we will abuse notation slightly and let $\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma} \triangleq \left\|\Phi_\sigma\left(x\right)\right\|_{\mathcal{H}_\sigma}$. Note that as $\sigma \to 0$, $\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}$ grows without bound, a fact we will use frequently. Throughout this chapter σ will implicitly be a function of n, such that $\sigma \to 0$ as $n \to \infty$. We will use f_σ^n to denote the RKDE for a general loss ρ and $\bar{f}_\sigma^n$ to denote the special case corresponding to $\rho\left(\cdot\right) = \left(\cdot\right)^2$, i.e. the classic KDE.

2.1 Novel KDE Consistency Proof

First we will introduce a construction that will be used frequently throughout the chapter:

$$\mathcal{D}_\sigma = \left\{ \int \Phi_\sigma(x) d\nu(x) \middle| \nu \text{ is a probability measure} \right\}.$$

Note that this and all Hilbert space valued integrals are Bochner integrals; see *Steinwart and Christmann* (2008) for a basic introduction to Bochner integrals. For this chapter these integrals can be thought of as the convolution of the kernel with a measure. This in turn implies that all elements of $\mathcal{D}_\sigma$ are pdfs. In fact all of the density estimators in this chapter will be an element of some $\mathcal{D}_\sigma$.

We will now present a novel proof of L^1 consistency of the kernel density estimator.

Theorem II.1. *If $n \to \infty$ and $\sigma \to 0$ with $n\sigma^d \to \infty$ then $\left\| \bar{f}_\sigma^n - f \right\|_1 \xrightarrow{p} 0$.*

Proof. Let $\bar{f}_\sigma = \mathbb{E}_{X \sim f} \left[\Phi_\sigma (X) \right]$. By the triangle inequality we have

$$\left\| f - \bar{f}_\sigma^n \right\|_1 \leq \left\| f - \bar{f}_\sigma \right\|_1 + \left\| \bar{f}_\sigma^n - \bar{f}_\sigma \right\|_1.$$

The left term in the sum goes to zero by elementary analysis (*Devroye and Lugosi*, 2001). We only need to show that $\left\| \bar{f}_\sigma^n - \bar{f}_\sigma \right\|_1 \xrightarrow{p} 0$. First we show convergence in the RKHS.

Lemma II.2. *Let $\varepsilon > 0$. For sufficiently small σ,*

$$\mathbb{P}\left(\left\| \bar{f}_\sigma^n - \bar{f}_\sigma \right\|_{\mathcal{H}_\sigma} \geq \varepsilon \right) \leq \exp\left\{ -\frac{n\varepsilon^2}{4 \left\| \Phi_\sigma \right\|_{\mathcal{H}_\sigma}^2} \right\}.$$

Therefore if $n \to \infty$ and $\sigma \to 0$ with $n\sigma^d \to \infty$, then $\left\| \bar{f}_\sigma^n - \bar{f}_\sigma \right\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$.

Proof Sketch. Observe that

$$\mathbb{E}\left[\bar{f}_\sigma^n\right] = \mathbb{E}\left[\frac{1}{n}\sum_1^n \Phi_\sigma\left(X_i\right)\right] = \mathbb{E}_{X\sim f}\left[\Phi_\sigma\left(X\right)\right] = \bar{f}_\sigma.$$

This fact combined with Bennett's inequality for Hilbert space yields the inequality in the lemma, after some trivial manipulations. The second part of the lemma is a simple consequence of the inequality. □

The previous lemma follows from Bennett's inequality for Hilbert space, but Hoeffding's or Bernstein's inequality for Hilbert space would also suffice (*Pinelis*, 1994). For other examples of simple proofs using concentration inequalities see *Caponnetto and Vito* (2007) and *Bauer et al.* (2007). The next lemma allows us to bound L^1 norms over sets of finite Lebesgue measure. Let λ denote Lebesgue measure.

Lemma II.3. *Let $S \in \mathbb{R}^d$ be a set with finite Lebesgue measure and $g \in \mathcal{H}_\sigma$. Then*

$$\int_S |g(x)|\, dx \leq 2\sqrt{\lambda(S)}\, \|g\|_{\mathcal{H}_\sigma}.$$

Proof Sketch. We will present a proof for the situation where $g > 0$. For the general case we can split the following integral into two parts corresponding to the subsets of

S where g is positive and g is negative. We have,

$$\begin{aligned}
\left(\int_S g(x)\,dx\right)^2 &= \left(\int_S \langle \Phi_\sigma(x), g\rangle_{\mathcal{H}_\sigma}\,dx\right)^2 \\
&= \left(\left\langle \int_S \Phi_\sigma(x)\,dx, g\right\rangle_{\mathcal{H}_\sigma}\right)^2 \\
&\leq \left\|\int_S \Phi_\sigma(x)\,dx\right\|^2_{\mathcal{H}_\sigma} \|g\|^2_{\mathcal{H}_\sigma} \\
&= \int_S\int_S \langle \Phi_\sigma(x), \Phi_\sigma(x')\rangle_{\mathcal{H}_\sigma}\,dxdx'\,\|g\|^2_{\mathcal{H}_\sigma} \\
&= \int_S\int_S k_\sigma(x,x')\,dxdx'\,\|g\|^2_{\mathcal{H}_\sigma} \\
&\leq \int_S 1dx\,\|g\|^2_{\mathcal{H}_\sigma} \\
&= \lambda(S)\,\|g\|^2_{\mathcal{H}_\sigma}.
\end{aligned}$$

□

For pdfs embedded in RKHSs, Lemma II.3 allows us to show that $\mathcal{H}_\sigma$ convergence implies L^1 convergence.

Lemma II.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pdf and g^n_σ and h^n_σ be sequences of (possibly random) densities in a sequence of spaces $\mathcal{D}_\sigma$ (again σ is implicitly a function of n). If $\|g^n_\sigma - f\|_1 \xrightarrow{p} 0$ and $\|g^n_\sigma - h^n_\sigma\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$ then $\|g^n_\sigma - h^n_\sigma\|_1 \xrightarrow{p} 0$.*

Proof Sketch. Define $B(y,r)$ to be the open ball centered at y with radius r and χ_S to be the indicator function on the set S. Let $\varepsilon > 0$. Choose r large enough that $\int_{B(0,r)^C} f(x)\,dx < \varepsilon/3$ (this is possible by Lemma II.11 in Section 2.3). Since $B(0,r)$

and $B(0,r)^C$ partition $\mathbb{R}^d$ we have

$$\begin{aligned}\left\|g_\sigma^n - h_\sigma^n\right\|_1 &= \left\|\left(g_\sigma^n - h_\sigma^n\right)\left(\chi_{B(0,r)} + \chi_{B(0,r)^C}\right)\right\|_1 \\ &= \left\|\left(g_\sigma^n - h_\sigma^n\right)\chi_{B(0,r)}\right\|_1 + \left\|\left(g_\sigma^n - h_\sigma^n\right)\chi_{B(0,r)^C}\right\|_1 . \end{aligned} \tag{2.1}$$

The left summand goes to zero in probability by Lemma II.3 so it becomes bounded by $\varepsilon/3$ with probability going to one. Since $\left\|\left(f - g_\sigma^n\right)\chi_{B(0,r)^C}\right\|_1 \xrightarrow{p} 0$ we have $\left\|g_\sigma^n \chi_{B(0,r)^C}\right\|_1 \xrightarrow{p} \left\|f\chi_{B(0,r)^C}\right\|_1 < \varepsilon/3$. Since g_σ^n and h_σ^n are densities and both of them are converging to have the same amount of mass in $B(0,r)$, their mass in $B(0,r)^C$ must also be converging. This means $\left|\left\|h_\sigma^n \chi_{B(0,r)^C}\right\|_1 - \left\|g_\sigma^n \chi_{B(0,r)^C}\right\|_1\right| \xrightarrow{p} 0$ so $\left\|h_\sigma^n \chi_{B(0,r)^C}\right\|_1$ becomes bounded by $\varepsilon/3$ with probability going to one. Thus the right summand of (2.1) becomes bounded by $2\varepsilon/3$ with high probability. Putting these results together we have $\left\|g_\sigma^n - h_\sigma^n\right\|_1 < \varepsilon$ with probability going to one.

□

The previous lemma is a bit more general than is necessary for the current theorem, but it will be handy later. In this case g_σ^n in the last lemma is replaced by $\bar{f}_\sigma$ and h_σ^n is replaced with $\bar{f}_\sigma^n$, thus completing our proof of Theorem II.1. □

It is worth noting that Lemma II.2 also implies consistency with respect to L^2 and L^∞ norms, assuming suitable conditions ensuring that the approximation error goes to zero. L^2 consistency is implied as long as $k_\sigma(\cdot, x) \in L^2\left(\mathbb{R}^d\right)$ for all $x \in \mathbb{R}^d$, (in particular, k_σ need not be a reproducing kernel) because Lemma II.2 holds for general Hilbert spaces. L^∞ consistency follows from the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\bar{f}_\sigma^n(x) - \bar{f}_\sigma(x)\right| &= \left|\left\langle \Phi_\sigma(x), \bar{f}_\sigma^n - \bar{f}_\sigma\right\rangle_{\mathcal{H}_\sigma}\right| \\ &\leq \|\Phi_\sigma\|_{\mathcal{H}_\sigma} \left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} .\end{aligned}$$

Unfortunately the $\|\Phi_\sigma\|_{\mathcal{H}_\sigma}$ term in the last line yields a suboptimal rate on the band-

width, $n\sigma^{2d} \to \infty$.

2.2 RKDE Consistency

We begin by reviewing some results about the RKDE.

2.2.1 Previous Results

Before we prove consistency of the RKDE, we will introduce some additional technical background on the RKDE from *Kim and Scott* (2012). First we will define some properties ρ may have. Let $\rho : [0, \infty) \to [0, \infty)$, $\psi \triangleq \rho'$, and $\varphi(x) \triangleq \psi(x)/x$. Consider the following properties:

(B1) ρ is strictly convex

(B2) ρ is strictly increasing, $\rho(0) = 0$ and $\rho(x)/x \to 0$ as $x \to 0$

(B3) $\varphi(0) := \lim_{x \to 0} \frac{\psi(x)}{x}$ exists and is finite

(B4) ψ is bounded

(B5) ρ'' exists and is nonincreasing on $(0, \infty)$

(B6) φ is nonincreasing.

Some examples of losses satisfying all of these properties are $\rho(x) = \sqrt{x^2 + 1} - 1$, $\rho(x) = x \arctan(x)$, and $\rho(x) = x - \log(1 + x)$. It is easy to show that property (B1) guarantees the existence and uniqueness of f_σ^n (*Kim and Scott*, 2012). Let f be a pdf and $X_1, \cdots, X_n$ be iid samples from f. Let $J_\sigma^n(\cdot)$ be the empirical risk introduced in (1.1). Taking the Gateaux derivative of the risk gives us

$$\delta J_\sigma^n(g; h) = -\left\langle \frac{1}{n} \sum_1^n \varphi\left(\|\Phi_\sigma(X_i) - g\|_{\mathcal{H}_\sigma}\right)(\Phi_\sigma(X_i) - g), h \right\rangle_{\mathcal{H}_\sigma}.$$

If (B2) and (B3) are satisfied then a necessary condition for $g = f_\sigma^n$ is that the Gateaux derivative at g is 0 for all directions h, which is equivalent to left term in the inner product being 0 (Lemma 1 *Kim and Scott* (2012)). A straightforward algebraic

manipulation of the last condition gives us

$$\frac{\sum_1^n \varphi\left(\left\|\Phi_\sigma\left(X_i\right)-g\right\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma\left(X_i\right)}{\sum_1^n \varphi\left(\left\|\Phi_\sigma\left(X_j\right)-g\right\|_{\mathcal{H}_\sigma}\right)}=g.$$

With this in mind we introduce the following functional,

$$\begin{aligned} R_\sigma^n:\mathcal{H}_\sigma\to\mathcal{H}_\sigma:g\mapsto R_\sigma^n(g) &= \frac{\int \varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x)d\mu_n(x)}{\int \varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)d\mu_n(y)} \\ &= \sum_1^n \alpha_i\left(g\right)k_\sigma\left(\cdot,X_i\right) \end{aligned}$$

where

$$\alpha_i\left(g\right)=\frac{\varphi\left(\left\|\Phi_\sigma(X_i)-g\right\|_{\mathcal{H}_\sigma}\right)}{\sum_1^n \varphi\left(\left\|\Phi_\sigma(X_j)-g\right\|_{\mathcal{H}_\sigma}\right)}$$

and μ_n is the empirical measure corresponding to the sample. This function is the Iterated Reweighted Least Squares algorithm (IRWLS) from *Kim and Scott* (2012), which is used to compute the RKDE in practice. From Corollary 6 in *Kim and Scott* (2012) it is easy to show that if (B1), (B2), (B3), (B5), and (B6) are satisfied (note that (B4) is used later), the sequence $\left\{R_\sigma^n\left(0\right),R_\sigma^n\left(R_\sigma^n\left(0\right)\right),\ldots\right\}$ converges in $\mathcal{H}_\sigma$ to f_σ^n, which is the unique fixed point of R_σ^n.

2.2.2 Consistency Theorem and Proof

Theorem II.5. *Let $f\in L^2\left(\mathbb{R}^d\right)$ and let ρ satisfy (B1)-(B6). If $n\sigma^d\to\infty$ and $\sigma\to 0$ as $n\to\infty$ then $\left\|f_\sigma^n-f\right\|_1\xrightarrow{p}0$.*

We know that ψ is bounded by (B4). In the proofs that follow it will be assumed, for simplicity, that $\sup_x\psi\left(x\right)=1$. Note that any loss with bounded ψ can be adapted such that $\sup_x\psi\left(x\right)=1$. This is done by dividing ρ by $\sup_x\psi\left(x\right)$ and does not affect the RKDE. The longer and more technical proof sketches are contained in a subsection

after this one.

The following lemma helps us establish the behavior of elements in $\mathcal{D}_\sigma$ with large norms.

Lemma II.6. *For all* $g \in \mathcal{D}_\sigma$, $\|g\|_{\mathcal{H}_\sigma}^2 \leq \|g\|_\infty$.

Proof. By the definition of $\mathcal{D}_\sigma$, let $g = \int \Phi_\sigma(x) d\nu(x)$, where ν is a probability measure.

$$\begin{aligned}\|g\|_{\mathcal{H}_\sigma}^2 &= \langle g, g \rangle_{\mathcal{H}_\sigma} = \left\langle \int \Phi_\sigma(x) d\nu(x), g \right\rangle_{\mathcal{H}_\sigma} = \int \langle \Phi_\sigma(x), g \rangle_{\mathcal{H}_\sigma} \, d\nu(x) \\ &= \int g(x) d\nu(x) \leq \int \|g\|_\infty \, d\nu(x) = \|g\|_\infty .\end{aligned}$$

□

This lemma allows us to show that an element in $\mathcal{D}_\sigma$ with large norm will have most of its mass concentrated around one point. An element of $\mathcal{D}_\sigma$ having most of the mass around one point causes its general risk to be large. The Vapnik-Chervonenkis inequality allows us to show that all such elements will, with high probability, have high empirical risk.

Lemma II.7. *If* $\sigma \to 0$ *and* $n \to \infty$ *then* $\mathbb{P}\left(\|f_\sigma^n\|_{\mathcal{H}_\sigma}^2 \geq \frac{9}{10} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\right) \to 0$.

The constant $\frac{9}{10}$ was chosen simply for convenience, it could be replaced with any positive value less than one.

The following result will be used to prove Lemma II.9 and Theorem II.5.

Lemma II.8. $\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \leq \|f\|_2$.

Proof. Using the Cauchy-Schwarz inequality and Young's inequality (*Devroye and*

Lugosi, 2001) we have

$$\begin{aligned}
\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma}^2 &= \left\langle \int f(x)\Phi_\sigma(x)dx, \int f(y)\Phi_\sigma(y)dy \right\rangle_{\mathcal{H}_\sigma} \\
&= \int f(x) \left\langle \Phi_\sigma(x), \int f(y)\Phi_\sigma(y)dy \right\rangle_{\mathcal{H}_\sigma} dx \\
&= \int f(x) \left(f * k_\sigma\right)(x)dx \\
&= \langle f, f * k_\sigma \rangle_2 \\
&\leq \|f\|_2 \|f * k_\sigma\|_2 \\
&\leq \|f\|_2 \|f\|_2 \|k_\sigma\|_1 \\
&= \|f\|_2^2 .
\end{aligned}$$

□

Lemma II.7 shows that f_σ^n is, with high probability, in a ball of radius $\sqrt{\frac{9}{10}} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}$. Lemma II.9 shows that, on that ball, R_σ^n is a contraction mapping.

Lemma II.9. *Let $n \to \infty$, $\sigma \to 0$, and $n\sigma^d \to \infty$. There exists C_R such that, with probability going to one, the restriction of R_σ^n to $B_{\mathcal{H}_\sigma}\left(0, \sqrt{\frac{9}{10}} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}\right)$ is Lipschitz continuous with Lipschitz constant $C_R \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1}$.*

This lemma is the final key to proving Theorem II.5.

Proof of Theorem II.5. Using the triangle inequality we get

$$\|f - f_\sigma^n\|_1 \leq \left\|f - \bar{f}_\sigma^n\right\|_1 + \left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_1 .$$

We know the left term of the summand goes to zero in probability by Theorem II.1, so it is sufficient to show that the right summand goes to zero in probability. By Lemma II.4 it is sufficient to show that $\left\|f_\sigma^n - \bar{f}_\sigma^n\right\|_{\mathcal{H}_\sigma}$ goes to zero in probability.

Notice that $R_\sigma^n(0) = \bar{f}_\sigma^n$ and recall $R_\sigma^n\left(f_\sigma^n\right) = f_\sigma^n$. Using Lemma II.7 and II.9, with probability going to 1, the following holds

$$\begin{aligned}
\left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_{\mathcal{H}_\sigma} &= \|R_\sigma^n(0) - R_\sigma^n\left(f_\sigma^n\right)\|_{\mathcal{H}_\sigma} \\
&\leq \|f_\sigma^n - 0\|_{\mathcal{H}_\sigma} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} C_R \\
&< \sqrt{\frac{9}{10}} \|\Phi_\sigma\|_{\mathcal{H}_\sigma} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} C_R \\
&= \sqrt{\frac{9}{10}} C_R.
\end{aligned}$$

Since $\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$ and $\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \leq \|f\|_2 < \infty$ (by Lemma II.8), for arbitrary $s > 0$ we have $\left\|\bar{f}_\sigma^n\right\|_{\mathcal{H}_\sigma} < \|f\|_2 + s$ with probability going to one. Applying the contraction mapping steps again we get, with probability going to 1, that

$$\begin{aligned}
\left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_{\mathcal{H}_\sigma} &= \|R_\sigma^n(0) - R_\sigma^n\left(f_\sigma^n\right)\|_{\mathcal{H}_\sigma} \\
&\leq \|f_\sigma^n - 0\|_{\mathcal{H}_\sigma} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} C_R \\
&\leq \left(\left\|f_\sigma^n - \bar{f}_\sigma^n\right\|_{\mathcal{H}_\sigma} + \left\|\bar{f}_\sigma^n\right\|_{\mathcal{H}_\sigma}\right) \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} C_R \\
&\leq \left(\sqrt{\frac{9}{10}} C_R + \|f\|_2 + s\right) \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} C_R.
\end{aligned}$$

The last line goes to zero as $\sigma \to 0$, completing our proof. □

2.2.3 Proof Sketches

Proof Sketch of Lemma II.7. We know that $f_\sigma^n \in \mathcal{D}_\sigma$, so to prove this lemma we will show that as $n \to \infty$ and $\sigma \to 0$, all vectors in $\mathcal{D}_\sigma$ with $\mathcal{H}_\sigma$-norm greater than or equal to $\sqrt{\frac{9}{10}} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}$ will have empirical risk greater than the zero vector. Define $J_\sigma^n : \mathcal{H}_\sigma \to \mathbb{R}$ as the empirical risk function

$$J_\sigma^n(g) = \frac{1}{n} \sum_1^n \rho\left(\|\Phi_\sigma(X_i) - g\|_{\mathcal{H}_\sigma}\right).$$

Let g_σ^n be the minimizer of J_σ^n when restricted to vectors in $\mathcal{D}_\sigma$ with $\mathcal{H}_\sigma$-norm greater than or equal to $\sqrt{\frac{9}{10}}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}$. By Lemma II.6 there must exist x^* such that $g_\sigma^n\left(x^*\right) \geq \frac{9}{10}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2$, this causes most of of the mass of g_σ^n to reside near x^*. It is possible to show that, given any $r>0$ and $\varepsilon>0$, for sufficiently small σ, that $\sup_{x \in B\left(x^*, r\right)^C} g_\sigma^n(x)<\frac{3}{20}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2+\varepsilon$. As n gets large, J_σ^n becomes well approximated by J_σ where

$$J_\sigma(g)=\int \rho\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right) f(x)\, dx. \tag{2.2}$$

We will substitute J_σ for J_σ^n (in the formal proof we work with J_σ^n and invoke the VC inequality to relate it to the population risk). Since ρ is increasing, the following holds for sufficiently small σ,

$$\begin{aligned}
J_\sigma\left(g_\sigma^n\right) & \geq \int_{B\left(x^*, r\right)^C} \rho\left(\left\|\Phi_\sigma(x)-g_\sigma^n\right\|_{\mathcal{H}_\sigma}\right) f(x)\, dx \\
& \geq \int_{B\left(x^*, r\right)^C} \rho\left(\sqrt{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2-2\left\langle g_\sigma^n, \Phi_\sigma(x)\right\rangle_{\mathcal{H}_\sigma}+\left\|g_\sigma^n\right\|_{\mathcal{H}_\sigma}^2}\right) f(x)\, dx \\
& \geq \int_{B\left(x^*, r\right)^C} \rho\left(\sqrt{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2-2\left(\frac{3}{20}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2+\varepsilon\right)+\left\|g_\sigma^n\right\|_{\mathcal{H}_\sigma}^2}\right) f(x)\, dx.
\end{aligned}$$

Since ε can be set to be arbitrarily small and $\left\|g_\sigma^n\right\|_{\mathcal{H}_\sigma}^2 \geq \frac{9}{10}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2$ the last term has an approximate lower bound of

$$\begin{aligned}
& \gtrsim \int_{B\left(x^*, r\right)^C} \rho\left(\sqrt{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2-\frac{6}{20}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2+\frac{9}{10}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2}\right) f(x)\, dx \\
& \geq \rho\left(\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma} \sqrt{\frac{32}{20}}\right) \inf_y \int_{B(y, r)^C} f(x)\, dx.
\end{aligned}$$

Finally r can be chosen to be sufficiently small so that $\inf_y \int_{B(y,r)^C} f(x)\,dx$ is arbitrarily close to one. Thus as $n \to \infty$ and $\sigma \to 0$, with probability going to one

$$J_\sigma^n\left(g_\sigma^n\right) \gtrsim \rho\left(\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\sqrt{\frac{32}{20}}\right).$$

Now, notice that

$$\begin{aligned} J_\sigma^n(0) &= \frac{1}{n}\sum_1^n \rho\left(\left\|\Phi_\sigma(X_i) - 0\right\|_{\mathcal{H}_\sigma}\right) \\ &= \rho\left(\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\right). \end{aligned}$$

It then follows that, with probability going to one, $J_\sigma^n\left(g_\sigma^n\right) > J_\sigma^n(0)$. □

Proof Sketch of Lemma II.9. Let $g, h \in B_{\mathcal{H}_\sigma}\left(0, \sqrt{\frac{9}{10}}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\right)$. We have

$$\begin{aligned} &\left\|R_\sigma^n(g) - R_\sigma^n(h)\right\|_{\mathcal{H}_\sigma} \\ &= \left\| \frac{\int \varphi\left(\left\|\Phi_\sigma(x) - g\right\|_{\mathcal{H}_\sigma}\right)\Phi(x)\,d\mu_n(x)}{\int \varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)d\mu_n(y)} - \frac{\int \varphi\left(\left\|\Phi_\sigma(x') - h\right\|_{\mathcal{H}_\sigma}\right)\Phi(x)\,d\mu_n(x')}{\int \varphi\left(\left\|\Phi_\sigma(y') - h\right\|_{\mathcal{H}_\sigma}\right)d\mu_n(y')} \right\|_{\mathcal{H}_\sigma}. \end{aligned} \tag{2.3}$$

Note that all integrals are over the same measure. Consider the situation if the integrals were evaluated at one point, we have that

$$\begin{aligned} &\left| \frac{\varphi\left(\left\|\Phi_\sigma(x) - g\right\|_{\mathcal{H}_\sigma}\right)}{\varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)} - \frac{\varphi\left(\left\|\Phi_\sigma(x) - h\right\|_{\mathcal{H}_\sigma}\right)}{\varphi\left(\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}\right)} \right| \\ &= \left| \frac{N}{\varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}\right)} \right| \end{aligned} \tag{2.4}$$

where

$$N = \varphi\left(\left\|\Phi_\sigma(x) - g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}\right)\ldots$$
$$-\varphi\left(\left\|\Phi_\sigma(x) - h\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right).$$

We will now find a lower bound on the denominator. Note that since g and h live in $B_{\mathcal{H}_\sigma}\left(0, \|\Phi_\sigma\|_{\mathcal{H}_\sigma}\sqrt{\frac{9}{10}}\right)$, that $\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}$ and $\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}$ grow without bound as $\sigma \to 0$. Since ρ is convex ψ must be increasing and since ψ has a supremum of 1, $\psi(z)$ is well approximated by 1 for large z. Thus we have, for small σ that the denominator is well approximated as follows

$$\begin{aligned}
\varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}\right) =& \frac{\psi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)}{\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}}\frac{\psi\left(\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}\right)}{\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}} \\
\approx& \frac{1}{\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}} \\
\geq& \frac{1}{\|\Phi_\sigma\|^2_{\mathcal{H}_\sigma}\left(1 + \sqrt{9/10}\right)^2} \\
=& C_D\|\Phi_\sigma\|^{-2}_{\mathcal{H}_\sigma}
\end{aligned}$$

where $C_D = \left(1 + \sqrt{9/10}\right)^{-2}$. We will now find an upper bound on the numerator. By the triangle inequality

$$\begin{aligned}
&\left|\varphi\left(\left\|\Phi_\sigma(x) - g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}\right)\ldots\right. \\
&\qquad\left.-\varphi\left(\left\|\Phi_\sigma(x) - h\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)\right| \\
&\leq \left|\varphi\left(\left\|\Phi_\sigma(x) - g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - h\right\|_{\mathcal{H}_\sigma}\right)\ldots\right. \\
&\qquad\left.-\varphi\left(\left\|\Phi_\sigma(x) - g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)\right|\ldots \\
&\qquad+\left|\varphi\left(\left\|\Phi_\sigma(x) - g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)\ldots\right. \\
&\qquad\left.-\varphi\left(\left\|\Phi_\sigma(x) - h\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y) - g\right\|_{\mathcal{H}_\sigma}\right)\right|.
\end{aligned}$$

Consider the second summand,

$$\begin{aligned}
&\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\dots\right. \\
&\qquad\left.-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\right| \\
&=\varphi\left(\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}\right)\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(y)-h\right\|_{\mathcal{H}_\sigma}\right)\right| \\
&\leq\frac{1}{\left\|\Phi_\sigma(y)-g\right\|_{\mathcal{H}_\sigma}}\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\right| \\
&\leq\frac{1}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\left(1-\sqrt{\frac{9}{10}}\right)}\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\right|.
\end{aligned} \tag{2.5}$$

Just as $\varphi(z)$ becomes well approximated by $\frac{1}{z}$ for large z, $\varphi'(z)$ becomes well approximated by $\frac{-1}{z^2}$. Using this it can be shown that there exists $C_L>0$ such that, for sufficiently small σ, $\varphi\left(\left\|\Phi_\sigma(y)-\cdot\right\|_{\mathcal{H}_\sigma}\right)$ is Lipschitz continuous on $B_{\mathcal{H}_\sigma}\left(0,\sqrt{\frac{9}{10}}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\right)$ with Lipschitz constant $\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-2}C_L$. Now we have

$$\left|\varphi\left(\left\|\Phi_\sigma(x)-g\right\|_{\mathcal{H}_\sigma}\right)-\varphi\left(\left\|\Phi_\sigma(x)-h\right\|_{\mathcal{H}_\sigma}\right)\right|\leq\left\|g-h\right\|_{\mathcal{H}_\sigma}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-2}C_L.$$

It now follows that (2.5) is less than or equal to $\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-3}C_N$ for some $C_N>0$. Returning to (2.4), we can now show that it has an upper bound of $\frac{2\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-3}C_N}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-2}C_D}=\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^{-1}\frac{2C_N}{C_D}$. This generally describes the behavior of the values found in (2.3). To take care of the $\int\Phi_\sigma(x)\,d\mu_n(x)$ terms, note that by Theorem II.1 $\left\|\int\Phi_\sigma(x)\,d\mu_n(x)-\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma}\xrightarrow{p}0$ if $n\sigma^d\to\infty$. By Lemma II.8, $\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma}\leq\left\|f\right\|_2$ so $\left\|\int\Phi_\sigma(x)\,d\mu_n(x)\right\|_{\mathcal{H}_\sigma}$ becomes bounded with high probability, thus completing our proof sketch. □

2.3 Proofs of Lemmas

For convenience the proofs have been split up into two subsections, one for proofs from the KDE section and the other for proofs from the RKDE section.

2.3.1 KDE Consistency Proofs

The following lemma is a Hilbert space version of Bennett's inequality (*Smale and Zhou*, 2007) and will be used in the proof of Lemma II.2.

Lemma II.10. *Let $\mathcal{H}$ be a Hilbert space and $\{\xi_i\}_{i=1}^m$ be m ($m < \infty$) independent random variables with values in $\mathcal{H}$. Also, assume that for each i, $\|\xi_i\|_{\mathcal{H}} \leq B < \infty$ almost surely. Let $\delta^2 = \sum_{i=1}^m E\left[\|\xi_i\|_{\mathcal{H}}^2\right]$. Then*

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\xi_i - \mathbb{E}\left[\xi_i\right]\right)\right\|_{\mathcal{H}} \geq \varepsilon\right) \leq \exp\left\{-\frac{m\varepsilon}{2B}\log\left(1 + \frac{mB\varepsilon}{\delta^2}\right)\right\}, \forall \varepsilon > 0.$$

Proof of Lemma II.2. We will apply Lemma II.10. From the lemma statement let $\xi_i = \Phi_\sigma(X_i)$ and $m = n$ yielding, for all $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}\left(\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \geq \varepsilon\right) &\leq \exp\left\{-\frac{n\varepsilon}{2\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}\log\left(1 + \frac{n\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\varepsilon}{n\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2}\right)\right\} \\ &= \exp\left\{-\frac{n\varepsilon}{2\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}\log\left(1 + \frac{\varepsilon}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}\right)\right\}.\end{aligned}$$

As $\sigma \to 0$ then $1 + \frac{\varepsilon}{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}} \to 1$ so for sufficiently small σ

$$\log\left(1 + \frac{\varepsilon}{\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}\right) \geq \frac{\varepsilon}{2\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}}$$

and

$$\mathbb{P}\left(\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \geq \varepsilon\right) \leq \exp\left\{-\frac{n\varepsilon^2}{4\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2}\right\}$$

which goes to zero as $\frac{n}{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2} \to \infty$, or equivalently $n\sigma^d \to \infty$. So $\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$.

□

Proof of Lemma II.3. Let $S^+ = \{s | s \in S, g(s) \geq 0\}$ and $S^- = S \setminus S^+$. We have

$$\begin{aligned}
\int_S |g(x)|\, dx &= \int_{S^+} g(x)dx + \int_{S^-} -g(x')dx' \\
&= \int_{S^+} \langle g, \Phi_\sigma(x)\rangle_{\mathcal{H}_\sigma}\, dx + \int_{S^-} \langle -g, \Phi_\sigma(x')\rangle_{\mathcal{H}_\sigma}\, dx' \\
&= \left\langle g, \int_{S^+} \Phi_\sigma(x)dx \right\rangle_{\mathcal{H}_\sigma} + \left\langle -g, \int_{S^-} \Phi_\sigma(x')dx' \right\rangle_{\mathcal{H}_\sigma} \\
&\leq \|g\|_{\mathcal{H}_\sigma} \left(\left\| \int_{S^+} \Phi_\sigma(x)dx \right\|_{\mathcal{H}_\sigma} + \left\| \int_{S^-} \Phi_\sigma(x')dx' \right\|_{\mathcal{H}_\sigma} \right). \qquad (2.6)
\end{aligned}$$

Now consider

$$\begin{aligned}
\left\| \int_{S^+} \Phi_\sigma(x)dx \right\|^2_{\mathcal{H}_\sigma} &= \left\langle \int_{S^+} \Phi_\sigma(x)dx, \int_{S^+} \Phi_\sigma(x')dx' \right\rangle_{\mathcal{H}_\sigma} \\
&= \int_{S^+}\int_{S^+} \langle \Phi_\sigma(x), \Phi_\sigma(x')\rangle_{\mathcal{H}_\sigma}\, dxdx' \\
&= \int_{S^+}\int_{S^+} k_\sigma(x, x')\, dxdx' \\
&\leq \int_{S^+} 1dx' \\
&= \lambda(S^+)
\end{aligned}$$

and a similar result can be shown for S^-. Plugging back into (2.6) we get

$$\begin{aligned}
\int_S |g(x)|\, dx &\leq \|g\|_{\mathcal{H}_\sigma} \left(\sqrt{\lambda(S^+)} + \sqrt{\lambda(S^-)} \right) \\
&\leq \|g\|_{\mathcal{H}_\sigma} 2\sqrt{\lambda(S)}.
\end{aligned}$$

□

Lemma II.11. *Let f be a pdf, $\varepsilon > 0$, and $y \in \mathbb{R}^d$. There exists $r > 0$ such that*

$$\int_{B(y,r)} f(x)dx \geq 1 - \varepsilon.$$

or equivalently

$$\int_{B(y,r)^C} f(x)dx < \varepsilon.$$

Proof. We will prove the second statement. Consider the following, where $i \in \mathbb{N}$,

$$\int_{B(y,i)^C} f(x)\,dx = \int \chi_{B(y,i)^C}(x)\, f(x)\,dx.$$

Clearly as $i \to \infty$, $\chi_{B(y,i)^C} f \to 0$ pointwise. Since $\chi_{B(y,i)^C} f$ is dominated by f, $\int \chi_{B(y,i)^C}(x)\, f(x)\,dx \to \int 0dx = 0$ by the dominated convergence theorem. Thus there exists $n \in \mathbb{N}$ where $\int_{B(y,n)^C} f(x)\,dx < \varepsilon$. □

Proof of Lemma II.4. Let $\varepsilon > 0$; by Lemma II.11 let $r > 0$ such that $\left\| f\chi_{B(0,r)^C} \right\|_1 < \varepsilon/3$. From Lemma II.3 we have

$$\left\| (g_\sigma^n - h_\sigma^n)\chi_{B(0,r)} \right\|_1 \xrightarrow{p} 0.$$

Since $\|g_\sigma^n - f\|_1 \xrightarrow{p} 0$, we have $\left\|g_\sigma^n \chi_{B(0,r)}\right\|_1 \xrightarrow{p} \left\|f\chi_{B(0,r)}\right\|_1$, and therefore

$$\begin{aligned}
\left|\left\|h_\sigma^n \chi_{B(0,r)^C}\right\|_1 - \left\|f\chi_{B(0,r)^C}\right\|_1\right| &= \left|\left(1 - \left\|h_\sigma^n \chi_{B(0,r)}\right\|_1\right) - \left(1 - \left\|f\chi_{B(0,r)}\right\|_1\right)\right| \\
&= \left|\left\|h_\sigma^n \chi_{B(0,r)}\right\|_1 - \left\|f\chi_{B(0,r)}\right\|_1\right| \\
&\leq \left\|\left(h_\sigma^n - f\right)\chi_{B(0,r)}\right\|_1 \\
&\leq \left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)}\right\|_1 + \left\|\left(g_\sigma^n - f\right)\chi_{B(0,r)}\right\|_1 \\
&\xrightarrow{p} 0.
\end{aligned}$$

Thus, $\left\|h_\sigma^n \chi_{B(0,r)^C}\right\|_1 \xrightarrow{p} \left\|f\chi_{B(0,r)^C}\right\|_1$. Since $\left\|f\chi_{B(0,r)^C}\right\|_1 < \varepsilon/3$, we have

$$\mathbb{P}\left(\left\|h_\sigma^n \chi_{B(0,r)^C}\right\|_1 \geq \varepsilon 5/12\right) \to 0. \tag{2.7}$$

Now to finish the proof,

$$\begin{aligned}
&\mathbb{P}\left(\|h_\sigma^n - g_\sigma^n\|_1 > \varepsilon\right) \\
&\quad = \mathbb{P}\left(\left\|(h_\sigma^n - g_\sigma^n)\chi_{B(0,r)}\right\|_1 + \left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)^C}\right\|_1 > \varepsilon\right) \\
&\quad \leq \mathbb{P}\left(\left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)}\right\|_1 \geq \varepsilon/4\right) + \mathbb{P}\left(\left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)^C}\right\|_1 > 3\varepsilon/4\right)
\end{aligned}$$

We've already shown the left summand goes to zero, now we take care of the right term

$$\begin{aligned}
&\mathbb{P}\left(\left\|\left(h_\sigma^n - g_\sigma^n\right)\chi_{B(0,r)^C}\right\|_1 > 3\varepsilon/4\right) \\
&\quad \leq \mathbb{P}\left(\left\|h_\sigma^n \chi_{B(0,r)^C}\right\|_1 + \left\|g_\sigma^n \chi_{B(0,r)^C}\right\|_1 > 3\varepsilon/4\right) \\
&\quad \leq \mathbb{P}\left(\left\|h_\sigma^n \chi_{B(0,r)^C}\right\|_1 \geq 5\varepsilon/12\right) + \mathbb{P}\left(\left\|g_\sigma^n \chi_{B(0,r)^C}\right\|_1 > \varepsilon/3\right)
\end{aligned}$$

The left summand goes to zero by (2.7). Since $\left\|g_\sigma^n \chi_{B(0,r)^C} - f\chi_{B(0,r)^C}\right\|_1 \to 0$ and $\left\|f\chi_{B(0,r)^C}\right\|_1 < \frac{\varepsilon}{3}$, with probability going to one, we have $\left\|g_\sigma^n \chi_{B(0,r)^C}\right\|_1 \leq \varepsilon/3$ and the

right summand goes to zero. This completes our proof. □

2.3.2 RKDE Consistency Proofs

Lemma II.12. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pdf. For all $\varepsilon > 0$, there exists $s > 0$ such that $\int_{B(z,s)} f(x)\,dx \leq \varepsilon$ for all $z \in \mathbb{R}^d$.*

Proof. We will proceed by contradiction. Let $\{x_i\}_1^\infty$ be a sequence in $\mathbb{R}^d$ such that $\int_{B(x_i,1/i)} f(x)\,dx > \varepsilon$. Clearly the sequence must be bounded or else f would not be a pdf. Let x_{i_j} be a convergent subsequence and let x' be its limit. Let $\{r_j\}_1^\infty$ be a sequence in $\mathbb{R}^+$ converging to zero with $B\left(x_{i_j}, 1/i_j\right) \subset B(x', r_j)$. So we have $\int_{B(x',r_j)} f(x)\,dx > \varepsilon$, for all j. We know

$$\int\limits_{B(x',r_j)} f(x)\,dx = \int \chi_{B(x',r_j)}(x)\, f(x)\,dx$$

and $f\chi_{B(x',r_j)} \to 0$ pointwise. Since $f\chi_{B(x',r_j)}$ is dominated by f, the dominated convergence theorem yields

$$\begin{aligned}
\lim_{j\to\infty} \int\limits_{B(x',r_j)} f(x)\,dx &= \lim_{j\to\infty} \int f(x)\,\chi_{B(x',r_j)}(x)\,dx \\
&= \int \lim_{j\to\infty} f(x)\,\chi_{B(x',r_j)}(x)\,dx \\
&= \int 0 dx \\
&= 0
\end{aligned}$$

but $\int_{B(x',r_j)} f(x)\,dx > \varepsilon$, a contradiction. □

Corollary II.13. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pdf with associated measure μ, $\varepsilon > 0$ and $r > 0$. There exists $s > 0$ such that for all $x \in \mathbb{R}^d$, $\mu\left(B(x, r+s) \setminus B(x,r)\right) < \varepsilon$.*

Proof. We will omit a full proof; the general strategy is the same as the previous proof. Find a series of annuli with width decreasing to zero that have probability

greater than ε. Next find a convergent subsequence of annuli centers, let its limit be x'. Finally construct a series of annuli centered at x' with probability measure greater than ε and width going to zero and arrive at the same contradiction. □

Lemma II.14. *Let $s > 0$. If $\sigma \to 0$ then $\sigma^{-d} q(s/\sigma) \to 0$.*

Proof. We will proceed by contradiction. Suppose $\sigma^{-d} q(s/\sigma)$ does not converge to zero, then there exists $C > 0$ such that we can find arbitrarily small σ satisfying

$$\sigma^{-d} q(s/\sigma) > C. \tag{2.8}$$

It is well known that there exists C_d such that the Lebesgue measure of a ball in $\mathbb{R}^d$ of radius r is $C_d r^d$. Since q is nonincreasing (*Scovel et al.*, 2010) this along with (2.8) implies that there exists arbitrarily small σ satisfying

$$\begin{aligned}\int_{B(0,s)} \sigma^{-d} q(\|x\|_2/\sigma)\, dx &\geq \int_{B(0,s)} \sigma^{-d} q(s/\sigma)\, dx \\ &> C_d s^d C\end{aligned}$$

where the last term must be less than or equal to 1. Now, by Lemma II.11, there exists $r > 0$ such that

$$\int_{B(0,r)} q(\|x\|_2)\, dx = \int_{B(0,r\sigma)} \sigma^{-d} q(\|x\|_2/\sigma)\, dx \geq 1 - \frac{C_d s^d C}{2}.$$

For sufficiently small σ we have

$$\begin{aligned}
1 \geq & \int\limits_{B(0,r\sigma)} \sigma^{-d} q\left(\|x\|_2/\sigma\right) dx \\
\geq & \int\limits_{B(0,r\sigma)} \sigma^{-d} q\left(\|x'\|_2/\sigma\right) dx' + \int\limits_{B(0,s)\setminus B(0,r\sigma)} \sigma^{-d} q\left(\|x\|_2/\sigma\right) dx \\
\geq & 1 - \frac{C_d s^d C}{2} + \int\limits_{B(0,s)\setminus B(0,r\sigma)} \sigma^{-d} q\left(\|x\|_2/\sigma\right) dx.
\end{aligned}$$

Because q is nonincreasing this is greater than or equal to

$$1 - \frac{C_d s^d C}{2} + C_d \left(s^d - (r\sigma)^d\right) \sigma^{-d} q\left(s/\sigma\right).$$

As $\sigma \to 0$, $C_d \left(s^d - (r\sigma)^d\right) \to C_d s^d$, so by (2.8) we can find some σ where the last term is greater than or equal to

$$1 - \frac{C_d s^d C}{2} + C_d s^d C \frac{2}{3}.$$

The last line is greater than 1, a contradiction.

□

Proof of Lemma II.7. Let conv be the convex hull operator. Define

$$Q_\sigma^n = \operatorname{conv}\left(\Phi_\sigma(X_1), \ldots, \Phi_\sigma(X_n)\right) \bigcap B_{\mathcal{H}_\sigma}\left(0, \sqrt{\frac{9}{10}}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}\right)^C.$$

Clearly $Q_\sigma^n \subset \mathcal{D}_\sigma$ since $\Phi_\sigma\left(X_i\right)$ is a density for all i. By the representer theorem in *Kim and Scott* (2012), $f_\sigma^n \in \operatorname{conv}\left(\Phi_\sigma\left(X_1\right), \ldots, \Phi_\sigma\left(X_n\right)\right)$. We also know that f_σ^n is the minimizer of J_σ^n, where $J_\sigma^n : \mathcal{H}_\sigma \to \mathbb{R}$ is the empirical risk function

$$J_\sigma^n\left(g\right) = \frac{1}{n}\sum_{i=1}^n \rho\left(\left\|\Phi_\sigma\left(X_i\right) - g\right\|_{\mathcal{H}_\sigma}\right).$$

From these facts if we can show

$$\mathbb{P}\left(J_\sigma^n(0) < J_\sigma^n(g), \forall g \in Q_\sigma^n\right) \to 0$$

then we have proven the lemma.

Since Q_σ^n is compact and J_σ^n is continuous (*Kim and Scott*, 2012) the set

$$\arg\min_{g \in Q_\sigma^n} J_\sigma^n(g)$$

contains at least one element. Let g_σ^n be an arbitrary minimizer of J_σ^n restricted to Q_σ^n. Let μ be the measure associated with f. From Lemma II.12 we can choose $r > 0$ such that $\mu\left(B\left(x, r\right)\right) \leq \frac{1}{10}$, for all $x \in \mathbb{R}^d$. Choose $s > 0$ such that $\mu\left(B\left(x, r+s\right)^C\right) \geq \frac{4}{5}$, for all $x \in \mathbb{R}^d$. The previous statement is satisfied by finding s such that, for all x, $\mu\left(B\left(x, r+s\right) \setminus B\left(x, r\right)\right) < \frac{1}{10}$, which is possible by Corollary II.13. By Lemma II.6 we know there exists x^* such that $g_\sigma^n(x^*) \geq \frac{9}{10}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2$ (x^* is implicitly a function of n). By the definition of Q_σ^n, let $g_\sigma^n = \sum_{i=1}^n \beta_i \Phi_\sigma(X_i)$ with $\beta_i \geq 0$ and $\sum_1^n \beta_i = 1$. Since $g_\sigma^n(x^*) \geq \frac{9}{10}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2$ and q is nonincreasing we have

$$\begin{aligned}
\frac{9}{10}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2 &\leq \sum_{i=1}^n \beta_i k_\sigma(X_i, x^*) \\
&= \sum_{i:X_i \in B(x^*,r)} \beta_i k_\sigma(X_i, x^*) + \sum_{j:X_j \in B(x^*,r)^C} \beta_j k_\sigma(X_j, x^*) \\
&= \sum_{i:X_i \in B(x^*,r)} \beta_i k_\sigma(X_i, x^*) + \sum_{j:X_j \in B(x^*,r)^C} \beta_j \sigma^{-d} q\left(\left\|X_j - x^*\right\|_2 / \sigma\right) \\
&\leq \sum_{i:X_i \in B(x^*,r)} \beta_i \left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2 + \sigma^{-d} q\left(r/\sigma\right)
\end{aligned}$$

The last line is due to the fact q must be nonincreasing (*Scovel et al.*, 2010). From Lemma II.14 we know that $\sigma^{-d} q(r/\sigma) \to 0$ as $\sigma \to 0$, so for sufficiently small σ we

have

$$\frac{17}{20}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2 < \sum_{i:X_i\in B(x^*,r)} \beta_i \left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2$$

and thus

$$\frac{17}{20} < \sum_{i:X_i\in B(x^*,r)} \beta_i. \tag{2.9}$$

Again, since q nonincreasing, for sufficiently small σ

$$\begin{aligned}
\sup_{y\in B(x^*,r+s)^C} g_\sigma^n(y) &= \sup_{y\in B(x^*,r+s)^C} \sum_{i=1}^n \beta_i k_\sigma\left(X_i,y\right) \\
&= \sup_{y\in B(x^*,r+s)^C} \sum_{i:X_i\in B(x^*,r)} \beta_i k_\sigma\left(X_i,y\right) \\
&+ \sum_{j:X_j\in B(x^*,r)^C} \beta_j \left\langle \Phi_\sigma\left(y\right), \Phi_\sigma\left(X_j\right)\right\rangle_{\mathcal{H}_\sigma} \\
&\leq \sigma^{-d} q\left(s/\sigma\right) + \sum_{j:X_j\in B(x^*,r)^C} \beta_j \left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2.
\end{aligned}$$

From this, (2.9) and because $\sigma^{-d}q(s/\sigma) \to 0$ as $\sigma \to 0$, for arbitrary $\varepsilon > 0$ we have, for sufficiently small σ,

$$\sup_{y\in B(x^*,r+s)^C} g_\sigma^n(y) < \varepsilon + \frac{3}{20}\left\|\Phi_\sigma\right\|_{\mathcal{H}_\sigma}^2.$$

Recall that we assumed that $\sup_x \psi(x) = \sup_x \rho'\left(x\right) = 1$ and $\rho\left(0\right) = 0$. Because ρ is

strictly increasing, for sufficiently small σ,

$$\begin{aligned}
J_\sigma^n(g_\sigma^n) &= \frac{1}{n}\sum_{i=1}^{n}\rho\left(\|\Phi_\sigma(X_i)-g_\sigma^n\|_{\mathcal{H}_\sigma}\right)\\
&= \frac{1}{n}\sum_{i:X_i\in B(x^*,r+s)}\rho\left(\|\Phi_\sigma(X_i)-g_\sigma^n\|_{\mathcal{H}_\sigma}\right)\dots\\
&\quad+\frac{1}{n}\sum_{j:X_j\in B(x^*,r+s)^C}\rho\left(\|\Phi_\sigma(X_j)-g_\sigma^n\|_{\mathcal{H}_\sigma}\right)\\
&\geq \frac{1}{n}\sum_{j:X_j\in B(x^*,r+s)^C}\rho\left(\|\Phi_\sigma(X_j)-g_\sigma^n\|_{\mathcal{H}_\sigma}\right)\\
&= \frac{1}{n}\sum_{j:X_j\in B(x^*,r+s)^C}\rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2-2g_\sigma^n(X_j)+\|g_\sigma^n\|_{\mathcal{H}_\sigma}^2}\right)\\
&\geq \frac{1}{n}\sum_{j:X_j\in B(x^*,r+s)^C}\rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2-2\left(\frac{3}{20}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2+\varepsilon\right)+\frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2}\right)\\
&= \mu_n\left(B(x^*,r+s)^C\right)\rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20}-2\varepsilon}\right)\\
&\geq \inf_x \mu_n\left(B\left(x,r+s\right)^C\right)\rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20}-2\varepsilon}\right).
\end{aligned}$$

Since ρ is strictly convex we know that ψ is strictly increasing. Because ψ has a supremum of 1 and is strictly increasing we know that for any $1>\varepsilon_\psi>0$ there exists b_ψ such that for all $x>b_\psi$, $\psi(x)>1-\varepsilon_\psi$. Then, for sufficiently small σ,

$$\begin{aligned}
\rho\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20}-2\varepsilon}\right) &= \int_0^{\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20}-2\varepsilon}}\psi\left(x\right)dx\\
&\geq \int_{b_\psi}^{\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20}-2\varepsilon}}\psi\left(x\right)dx\\
&\geq (1-\varepsilon_\psi)\left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2\frac{32}{20}-2\varepsilon}-b_\psi\right) \qquad (2.10)
\end{aligned}$$

For sufficiently small σ we have

$$\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 \frac{32}{20} - 2\varepsilon} \geq \|\Phi_\sigma\|_{\mathcal{H}_\sigma} \sqrt{\frac{32}{20}} - 2\varepsilon.$$

Since the complements of all open balls, in this case, all balls with radius $r+s$, have a finite shattering dimension (*Devroye and Lugosi*, 2001), and by our choice of r and s we know, with probability going to one, that $\inf_x \mu_n \left(B\left(x, r+s\right)^C \right) \to \inf_x \mu \left(B\left(x, r+s\right)^C \right) \geq 0.8$. Because of this for any $\varepsilon_B > 0$ we have, with probability going to one, that $\inf_x \mu_n \left(B\left(x, r+s\right)^C \right) \geq 0.8 - \varepsilon_B$. Since $\frac{4}{5}\sqrt{\frac{32}{20}} > 1$, we can choose ε_ψ and ε_B such that $\left(\frac{4}{5} - \varepsilon_B\right)(1-\varepsilon_\psi)\sqrt{\frac{32}{20}} > 1$. Using these facts with (2.10) we have, for sufficiently small σ, with probability going to one

$$\begin{aligned} J_\sigma^n\left(g_\sigma^n\right) &\geq \inf_x \mu_n \left(B\left(x, r+s\right)^C \right) (1-\varepsilon_\psi) \left(\sqrt{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2 \frac{32}{20} - 2\varepsilon} - b_\psi \right) \\ &\geq \left(\frac{4}{5} - \varepsilon_B \right) (1-\varepsilon_\psi) \left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma} \sqrt{\frac{32}{20}} - 2\varepsilon - b_\psi \right) \\ &> \|\Phi_\sigma\|_{\mathcal{H}_\sigma}. \end{aligned}$$

Now consider

$$\begin{aligned} J_\sigma^n(0) &= \frac{1}{n} \sum_{i=1}^n \rho\left(\|\Phi_\sigma(X_i) - 0\|_{\mathcal{H}_\sigma} \right) \\ &= \rho\left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma} \right) \\ &= \int_0^{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}} \psi(x)\, dx + \rho(0) \\ &\leq \int_0^{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}} 1 dx \\ &= \|\Phi_\sigma\|_{\mathcal{H}_\sigma}. \end{aligned}$$

So as $n \to \infty$ and $\sigma \to 0$ we have

$$\mathbb{P}\left(J_\sigma^n(g_\sigma^n) \leq J_\sigma^n(0)\right) \to 0,$$

thus finishing the proof. □

Proof of Lemma II.9. Let $g, h \in \mathcal{H}_\sigma$ such that $\|g\|_{\mathcal{H}_\sigma}^2 \leq \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$ and $\|h\|_{\mathcal{H}_\sigma}^2 \leq \frac{9}{10}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^2$. Cross multiplication gives us

$$\begin{aligned}
&\|R_\sigma^n(g) - R_\sigma^n(h)\|_{\mathcal{H}_\sigma} \\
&= \left\| \frac{\int \varphi\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x) d\mu_n(x)}{\int \varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right) d\mu_n(y)} - \frac{\int \varphi\left(\|\Phi_\sigma(x') - h\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x') d\mu_n(x')}{\int \varphi\left(\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}\right) d\mu_n(y')} \right\|_{\mathcal{H}_\sigma} \\
&= \left\| \frac{A}{B} \right\|_{\mathcal{H}_\sigma}
\end{aligned}$$

where

$$\begin{aligned}
A = &\left[\int \varphi\left(\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x)\, d\mu_n(x)\right]\left[\int \varphi\left(\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}\right) d\mu_n(y)\right] \\
&- \left[\int \varphi\left(\|\Phi_\sigma(x') - h\|_{\mathcal{H}_\sigma}\right)\Phi_\sigma(x')\, d\mu_n(x')\right]\left[\int \varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right) d\mu_n(y)\right]
\end{aligned}$$

and

$$B = \left[\int \varphi\left(\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}\right) d\mu_n(y')\right]\left[\int \varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right) d\mu_n(y)\right].$$

Note that $A \in \mathcal{H}_\sigma$ and $B \in \mathbb{R}^+$. We will now find a lower bound on B. As shown in the proof for Lemma II.7 there exists $b > 0$ such that $\psi(x) > 1/2$ for all $x \geq b$. By

the reverse triangle inequality

$$\begin{aligned}\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma} &\geq \left|\|\Phi_\sigma\|_{\mathcal{H}_\sigma} - \|h\|_{\mathcal{H}_\sigma}\right| \\ &\geq \|\Phi_\sigma\|_{\mathcal{H}_\sigma}\left(1 - \sqrt{\frac{9}{10}}\right)\end{aligned}$$

which grows without bound as $\sigma \to 0$. So for sufficiently small σ

$$\begin{aligned}\varphi\left(\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}\right) &= \frac{\psi\left(\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}\right)}{\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}} \\ &\geq \frac{1}{2\|\Phi_\sigma(y') - h\|_{\mathcal{H}_\sigma}} \\ &\geq \frac{1}{2\left(\|\Phi_\sigma\|_{\mathcal{H}_\sigma} + \|h\|_{\mathcal{H}_\sigma}\right)} \\ &\geq \frac{1}{\|\Phi_\sigma\|_{\mathcal{H}_\sigma} 2\left(1 + \sqrt{\frac{9}{10}}\right)}.\end{aligned}$$

A similar result can be shown for $\varphi\left(\|\Phi_\sigma(y) - g\|_{\mathcal{H}_\sigma}\right)$, so there exists $C_B > 0$ such that, for sufficiently small σ,

$$B \geq \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2} C_B.$$

Now we will focus on A. To make the following manipulations simpler we will let

$$\varphi\left(\|\Phi_\sigma(z) - k\|_{\mathcal{H}_\sigma}\right) = T_\sigma(z, k).$$

A is equal to

$$\begin{aligned}
&\left[\int T_\sigma\left(x,g\right)\Phi_\sigma(x)d\mu_n(x)\right]\left[\int T_\sigma\left(y',h\right)d\mu_n(y')\right]\ldots\\
&\quad-\left[\int T_\sigma\left(x',h\right)\Phi_\sigma(x')d\mu_n(x')\right]\left[\int T_\sigma\left(y,g\right)d\mu_n(y)\right]\\
&=\int\left\{T_\sigma\left(x,g\right)\Phi_\sigma(x)\left[\int T_\sigma\left(y',h\right)d\mu_n(y')\right]\ldots\right.\\
&\quad\left.-T_\sigma\left(x,h\right)\Phi_\sigma(x)\left[\int T_\sigma\left(y,g\right)d\mu_n(y)\right]\right\}d\mu_n(x)\\
&=\int\Phi_\sigma(x)\left[T_\sigma\left(x,g\right)\left[\int T_\sigma\left(y',h\right)d\mu_n(y)\right]\ldots\right.\\
&\quad\left.-T_\sigma\left(x,h\right)\left[\int T_\sigma\left(y,g\right)d\mu_n(y)\right]\right]d\mu_n(x)\\
&=\int\int\Phi_\sigma(x)\big[T_\sigma\left(x,g\right)T_\sigma\left(y,h\right)-T_\sigma\left(x,h\right)T_\sigma\left(y,g\right)\big]d\mu_n(y)d\mu_n(x).
\end{aligned}$$

We will now bound the inner term. Using the triangle inequality we have

$$\begin{aligned}
&\big|T_\sigma\left(x,g\right)T_\sigma\left(y,h\right)-T_\sigma\left(x,h\right)T_\sigma\left(y,g\right)\big| \qquad (2.11)\\
&<\big|T_\sigma\left(x,g\right)T_\sigma\left(y,h\right)-T_\sigma\left(x,g\right)T_\sigma\left(y,g\right)\big|+\big|T_\sigma\left(x,g\right)T_\sigma\left(y,g\right)-T_\sigma\left(x,h\right)T_\sigma\left(y,g\right)\big|\\
&=T_\sigma\left(x,g\right)\big|T_\sigma\left(y,h\right)-T_\sigma\left(y,g\right)\big|+T_\sigma\left(y,g\right)\big|T_\sigma\left(x,g\right)-T_\sigma\left(x,h\right)\big|.
\end{aligned}$$

We will bound the second summand in the last equality; a similar technique can

bound the first summand.

$$\begin{aligned}
\varphi\left(\|\Phi_\sigma(y)-g\|_{\mathcal{H}_\sigma}\right) &= \frac{\psi\left(\|\Phi_\sigma(y)-g\|_{\mathcal{H}_\sigma}\right)}{\|\Phi_\sigma(y)-g\|_{\mathcal{H}_\sigma}} \\
&\leq \frac{1}{\|\Phi_\sigma(y)-g\|_{\mathcal{H}_\sigma}} \\
&\leq \frac{1}{\left|\|\Phi_\sigma\|_{\mathcal{H}_\sigma}-\|g\|_{\mathcal{H}_\sigma}\right|} \\
&\leq \frac{1}{\|\Phi_\sigma\|_{\mathcal{H}_\sigma}\left(1-\sqrt{\frac{9}{10}}\right)}. \qquad (2.12)
\end{aligned}$$

A similar result can be shown for $\varphi\left(\|\Phi_\sigma(x)-g\|_{\mathcal{H}_\sigma}\right)$.

Consider $z \geq \|\Phi_\sigma\|_{\mathcal{H}_\sigma}\left(1-\sqrt{\frac{9}{10}}\right)$, then

$$\begin{aligned}
\left|\varphi'(z)\right| &= \left|\left(\frac{\psi(z)}{z}\right)'\right| \\
&= \left|\frac{z\psi'(z)-\psi(z)}{z^2}\right| \\
&\leq \frac{|z\psi'(z)|+|\psi(z)|}{z^2}.
\end{aligned}$$

We will now analyze the behaviour of ψ', specifically, there exists sufficiently large r such that $\psi'(x) \leq \frac{1}{x}$ for all $x \geq r$. We will proceed by contradiction. Suppose this is not the case. Then there exist positive numbers t_1, t_2 and t_3 such that $\psi'(t_i) > \frac{1}{t_i}$ and $\frac{t_i}{t_{i+1}} < \frac{1}{3}$. We know ψ' is nonincreasing by (B5) and nonnegative; we also know ψ

is bounded above by 1 so

$$\begin{aligned}
1 &\geq \int_0^{\infty} \psi'(x)\,dx \\
&\geq \int_{t_1}^{t_2} \psi'(x)\,dx + \int_{t_2}^{t_3} \psi'(y)\,dy \\
&\geq \frac{t_2 - t_1}{t_2} + \frac{t_3 - t_2}{t_3} \\
&\geq 2 - \frac{2}{3},
\end{aligned}$$

a contradiction. From this we have that for sufficiently large z,

$$\begin{aligned}
\frac{|z\psi'(z)| + |\psi(z)|}{z^2} &\leq \frac{z^{\frac{1}{z}} + 1}{z^2} \\
&= \frac{2}{z^2}.
\end{aligned}$$

Thus, for sufficiently small σ, on the space $\left[\left(1 - \sqrt{\frac{9}{10}}\right)\|\Phi_\sigma\|_{\mathcal{H}_\sigma}, \infty\right)$, φ is Lipschitz continuous with Lipschitz constant $2\left(1 - \sqrt{\frac{9}{10}}\right)^{-2}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2}$. Therefore we have

$$\begin{aligned}
&|\varphi(\|\Phi_\sigma(x) - g\|) - \varphi(\|\Phi_\sigma(x) - h\|)| \\
&\quad \leq \left|\|\Phi_\sigma(x) - g\|_{\mathcal{H}_\sigma} - \|\Phi_\sigma(x) - h\|_{\mathcal{H}_\sigma}\right| 2\left(1 - \sqrt{\frac{9}{10}}\right)^{-2}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2} \\
&\quad \leq \|g - h\|_{\mathcal{H}_\sigma} 2\left(1 - \sqrt{\frac{9}{10}}\right)^{-2}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2}.
\end{aligned}$$

Combining the last inequality with (2.12) we have that for sufficiently small σ, (2.11) is less than or equal to

$$4\|g - h\|_{\mathcal{H}_\sigma}\left(1 - \sqrt{\frac{9}{10}}\right)^{-3}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3}.$$

Using this bound we can do the following. Let

$$\tau \triangleq \left[T_\sigma(x,g)\,T_\sigma(y,h) - T_\sigma(x,h)\,T_\sigma(y,g)\right],$$

and

$$\tau' \triangleq \left[T_\sigma(x',g)\,T_\sigma(y',h) - T_\sigma(x',h)\,T_\sigma(y',g)\right],$$

and

$$\kappa \triangleq 4\,\|g-h\|_{\mathcal{H}_\sigma}\left(1-\sqrt{\frac{9}{10}}\right)^{-3}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3},$$

we have

$$\begin{aligned}
\|A\|_{\mathcal{H}_\sigma}^2 &= \left\| \int\int \Phi_\sigma(x)\tau d\mu_n(x)d\mu_n(y)\right\|_{\mathcal{H}_\sigma}^2 \\
&= \left\langle \int\int \Phi_\sigma(x)\tau d\mu_n(x)d\mu_n(y), \int\int \Phi_\sigma(x')\tau' d\mu_n(x')d\mu_n(y')\right\rangle_{\mathcal{H}_\sigma} \\
&= \int\int\int\int \tau\tau' \left\langle \Phi_\sigma(x),\Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma} d\mu_n(y)d\mu_n(y')d\mu_n(x)d\mu_n(x').
\end{aligned}$$

Since $\langle \Phi_\sigma(x),\Phi_\sigma(x')\rangle_{\mathcal{H}_\sigma} \geq 0$ for all x, x', for sufficiently small σ, the last line is less than or equal to

$$\begin{aligned}
&\int\int\int\int \kappa^2 \left\langle \Phi_\sigma(x),\Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma} d\mu_n(y)d\mu_n(y')d\mu_n(x)d\mu_n(x') \\
&= \int\int \kappa^2 \left\langle \Phi_\sigma(x),\Phi_\sigma(x')\right\rangle_{\mathcal{H}_\sigma} d\mu_n(x)d\mu_n(x') \\
&= \kappa^2 \left\| \int \Phi_\sigma(x)\,d\mu_n(x)\right\|_{\mathcal{H}_\sigma}^2.
\end{aligned}$$

Returning to the original notation, this means, for sufficiently small σ

$$\|A\|_{\mathcal{H}_\sigma} \leq \left\| \int \Phi_\sigma(x)\,d\mu_n(x)\right\|_{\mathcal{H}_\sigma} 4\,\|g-h\|_{\mathcal{H}_\sigma}\left(1-\sqrt{\frac{9}{10}}\right)^{-3}\|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3}.$$

From our proof of the consistency of the KDE we know that

$$\left\| \int \Phi_\sigma(x) d\mu_n(x) - \bar{f}_\sigma \right\|_{\mathcal{H}_\sigma} \xrightarrow{p} 0$$

and from Lemma II.8 $\left\|\bar{f}_\sigma\right\|_{\mathcal{H}_\sigma} \leq \|f\|_2$ so $\left\|\int \Phi_\sigma(x) d\mu_n(x)\right\|_{\mathcal{H}_\sigma}$ is bounded by some constant with probability going to one. Note that this is the only probabilistic step, which does not depend on g or h, so the result holds over the whole ball in $\mathcal{H}_\sigma$. So there exists $C_A > 0$ such that

$$\|A\|_{\mathcal{H}_\sigma} \leq \|g - h\|_{\mathcal{H}_\sigma} \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3} C_A$$

with probability going to one (we can omit "for sufficiently small σ" since $\sigma \to 0$ as $n \to \infty$). Finally we get with probability going to one as $n\sigma^d \to \infty$

$$\begin{aligned}
\left\| \frac{A}{B} \right\|_{\mathcal{H}_\sigma} &= \frac{\|A\|_{\mathcal{H}_\sigma}}{B} \\
&\leq \|g - h\|_{\mathcal{H}_\sigma} \frac{C_A \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-3}}{C_B \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-2}} \\
&= \|g - h\|_{\mathcal{H}_\sigma} C_R \|\Phi_\sigma\|_{\mathcal{H}_\sigma}^{-1} .
\end{aligned}$$

□

CHAPTER III

Robust Kernel Density Estimation by Scaling and Projection in Hilbert Space

In this chapter we introduce a new type of robust kernel density estimator we call the *scale and project kernel density estimator*. To do this we first introduce and analyze general contamination models for nonparametric density estimation and propose a contamination model for our estimator. Next we construct an estimator and show that it will asymptotically approach the desired decontaminated density if the assumptions of the contamination model are satisfied. Finally we demonstrate that the estimator is effective, even when the contamination model is not satisfied, by applying the algorithm to several datasets with varying amounts of contamination and comparing its performance to other estimators.

3.1 Nonparametric Contamination Models and Decontamination Procedures for Density Estimation

What assumptions are necessary and sufficient on a target and contaminating density in order to theoretically recover the target density is a question that, to the best of our knowledge, is completely unexplored in a nonparametric setting. We will approach this problem in the infinite sample setting, where we know $f_{obs} =$

$(1-\varepsilon)f_{tar}+\varepsilon f_{con}$ and ε, but do not know f_{tar} or f_{con}. To this end we introduce a new formalism. Let $\mathcal{D}$ be the set of all pdfs on $\mathbb{R}^d$. We use the term *contamination model* to refer to any subset $\mathcal{V} \subset \mathcal{D} \times \mathcal{D}$, i.e. a set of pairs (f_{tar}, f_{con}). Let $R_\varepsilon : \mathcal{D} \to \mathcal{D}$ be a set of transformations on $\mathcal{D}$ indexed by $\varepsilon \in [0, 1)$. We say that R_ε *decontaminates* $\mathcal{V}$ if for all $(f_{tar}, f_{con}) \in \mathcal{V}$ and $\varepsilon \in [0, 1)$ we have $R_\varepsilon((1-\varepsilon)f_{tar} + \varepsilon f_{con}) = f_{tar}$.

One may wonder whether there exists some set of contaminating densities, $\mathcal{D}_{con}$, and a transformation, R_ε, such that R_ε decontaminates $\mathcal{D} \times \mathcal{D}_{con}$. In other words, does there exist some set of contaminating densities for which we can recover any target density? It turns out this is impossible if $\mathcal{D}_{con}$ contains at least two elements.

Proposition III.1. *Let $\mathcal{D}_{con} \subset \mathcal{D}$ contain at least two elements. There does not exist any transformation R_ε which decontaminates $\mathcal{D} \times \mathcal{D}_{con}$.*

Proof. Let $f \in \mathcal{D}$ and $g, g' \in \mathcal{D}_{con}$ such that $g \neq g'$. Let $\varepsilon \in (0, \frac{1}{2})$. Clearly $f_{tar} \triangleq \frac{f(1-2\varepsilon)+g\varepsilon}{1-\varepsilon}$ and $f'_{tar} \triangleq \frac{f(1-2\varepsilon)+\varepsilon g'}{1-\varepsilon}$ are both elements of $\mathcal{D}$. Note that

$$(1-\varepsilon)f_{tar} + \varepsilon g' = (1-\varepsilon)f'_{tar} + \varepsilon g.$$

In order for R_ε to decontaminate $\mathcal{D}$ with respect to $\mathcal{D}_{con}$, we need

$$R_\varepsilon\left((1-\varepsilon)f_{tar} + \varepsilon g'\right) = f_{tar}$$

and

$$R_\varepsilon\left((1-\varepsilon)f'_{tar} + \varepsilon g\right) = f'_{tar},$$

which is impossible since $f_{tar} \neq f'_{tar}$. □

This proposition imposes significant limitations on what contamination models can be decontaminated. For example, suppose we know that f_{con} is Gaussian with known covariance matrix and unknown mean. Proposition III.1 says we cannot design R_ε so that it can decontaminate $(1-\varepsilon)f_{tar} + \varepsilon f_{con}$ for all $f_{tar} \in \mathcal{D}$. In other words,

it is impossible to design an algorithm capable of removing Gaussian contamination (for example) from arbitrary target densities. Furthermore, if R_ε decontaminates $\mathcal{V}$ and $\mathcal{V}$ is fully nonparametric (i.e. for all $f \in \mathcal{D}$ there exists some $f' \in \mathcal{D}$ such that $(f, f') \in \mathcal{V}$) then for each (f_{tar}, f_{con}) pair, f_{con} must satisfy some properties which depend on f_{tar}.

3.1.1 Proposed Contamination Model

For a function $f : \mathbb{R}^d \to \mathbb{R}$ let $\text{supp}(f)$ denote the support of f. We introduce the following contamination assumption:

Assumption (A). For the pair (f_{tar}, f_{con}), there exists u such that $f_{con}(x) = u$ for almost all (in the Lebesgue sense) $x \in \text{supp}(f_{tar})$ and $f_{con}(x') \leq u$ for almost all $x' \notin \text{supp}(f_{tar})$.

See Figure 3.1 for an example of a density satisfying this assumption. Because f_{con} must be uniform over the support of f_{tar} a consequence of Assumption A is that $\text{supp}(f_{tar})$ has finite Lebesgue measure. Let $\mathcal{V}_A$ be the contamination model containing all pairs of densities which satisfy Assumption A. Note that $\bigcup_{(f_{tar}, f_{con}) \in \mathcal{V}_A} f_{tar}$ is exactly all densities whose support has finite Lebesgue measure, which includes all densities with compact support.

The uniformity assumption on f_{con} is a common "noninformative" assumption on the contamination. Furthermore, this assumption is supported by connections to one-class classification. In that problem, only one class (corresponding to our f_{tar}) is observed for training, but the testing data is drawn from f_{obs} and must be classified. The dominant paradigm for nonparametric one-class classification is to estimate a level set of f_{tar} from the one observed training class *Theiler and Cai* (2003); *Lanckriet et al.* (2003); *Steinwart et al.* (2005); *Vert and Vert* (2006); *Sricharan and Hero* (2011); *Schölkopf et al.* (2001), and classify test data according to that level set. Yet level sets only yield optimal classifiers (i.e. likelihood ratio tests) under the uniformity

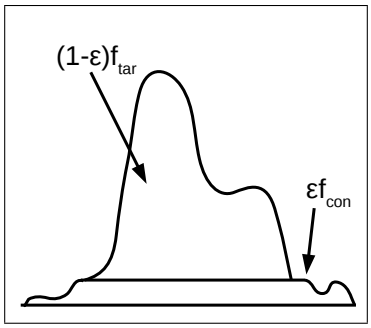

Figure 3.1: Density with contamination satisfying Assumption A

assumption on f_{con}, so that these methods are implicitly adopting this assumption. Furthermore, a uniform contamination prior has been shown to optimize the worst-case detection rate among all choices for the unknown contamination density *El-Yaniv and Nisenson* (2007). Finally, our experiments demonstrate that the SPKDE works well in practice, even when Assumption A is significantly violated.

3.1.2 Decontamination Procedure

Under Assumption A f_{tar} is present in f_{obs} and its shape is left unmodified (up to a multiplicative factor) by f_{con}. To recover f_{tar} it is necessary to first scale f_{obs} by $\beta = \frac{1}{1-\varepsilon}$ yielding

$$\frac{1}{1-\varepsilon}\left((1-\varepsilon)f_{tar} + \varepsilon f_{con}\right) = f_{tar} + \frac{\varepsilon}{1-\varepsilon}f_{con}.$$

After scaling we would like to slice off $\frac{\varepsilon}{1-\varepsilon}f_{con}$ from the bottom of $f_{tar} + \frac{\varepsilon}{1-\varepsilon}f_{con}$. This transform is achieved by

$$\max\left\{0, f_{tar} + \frac{\varepsilon}{1-\varepsilon}f_{con} - \alpha\right\}, \tag{3.1}$$

where α is set such that (3.1) is a pdf (which in this case is achieved with $\alpha = r\frac{\varepsilon}{1-\varepsilon}$). We will now show that this transform is well defined in a general sense. Let f be a

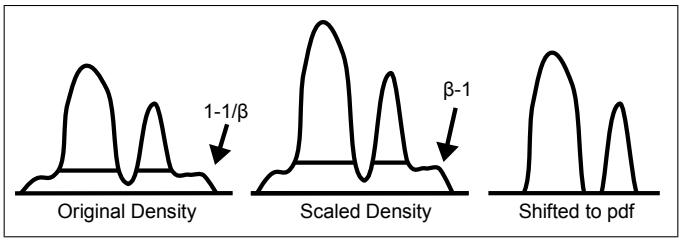

Figure 3.2: Infinite sample SPKDE transform. Arrows indicate the area under the line.

pdf and let

$$g_{\beta,\alpha} = \max\{0, \beta f(\cdot) - \alpha\}$$

where the max is defined pointwise. The following lemma shows that it is possible to slice off the bottom of any scaled pdf to get a transformed pdf and that the transformed pdf is unique.

Lemma III.2. *For fixed $\beta > 1$ there exists a unique $\alpha' > 0$ such that $\|g_{\beta,\alpha'}\|_{L^1} = 1$.*

Figure 3.2 demonstrates this transformation applied to a pdf. We define the following transform $R_\varepsilon^A : \mathcal{D} \to \mathcal{D}$ where $R_\varepsilon^A(f) = \max\left\{\frac{1}{1-\varepsilon} f(\cdot) - \alpha, 0\right\}$ where α is such that $R_\varepsilon^A(f)$ is a pdf. The remaining mathematical proofs for this chapter are deferred to Appendix A.

Proposition III.3. *R_ε^A decontaminates $\mathcal{V}_A$.*

The proof of this proposition is an intermediate step for the proof for Theorem III.8. For any two subsets of $\mathcal{V}, \mathcal{V}' \subset \mathcal{D} \times \mathcal{D}$, R_ε decontaminates $\mathcal{V}$ and $\mathcal{V}'$ iff R_ε decontaminates $\mathcal{V} \bigcup \mathcal{V}'$. Because of this, every decontaminating transform has a maximal set which it can decontaminate. Assumption A is both sufficient and necessary for decontamination by R_ε^A, i.e. the set $\mathcal{V}_A$ is maximal.

Proposition III.4. *Let $\{(q, q')\} \in \mathcal{D} \times \mathcal{D}$ and $(q, q') \notin \mathcal{V}_A$. R_ε^A cannot decontaminate $\{(q, q')\}$.*

3.1.3 Other Possible Contamination Models

The model described previously is just one of many possible models. An obvious approach to robust kernel density estimation is to use an anomaly detection algorithm and construct the KDE using only non-anomalous samples. We will investigate this model under a couple of anomaly detection schemes and describe their properties.

One of the most common methods for anomaly detection is the level set method. For a probability measure μ this method attempts to find the set S with smallest Lebesgue measure such that $\mu(S)$ is above some threshold, t, and declares samples outside of that set as being anomalous. For a density f this is equivalent to finding λ such that $\int_{\{x|f(x)\geq\lambda\}} f(y)dy = t$ and declaring samples were $f(X) < \lambda$ as being anomalous. Let $X_1, \ldots, X_n$ be iid samples from f_{obs}. Using the level set method for a robust KDE, we would construct a density $\widehat{f}_{obs}$ which is an estimate of f_{obs}. Next we would select some threshold $\lambda > 0$ and declare a sample, X_i, as being anomalous if $\widehat{f}_{obs}(X_i) < \lambda$. Finally we would construct a KDE using the non-anomalous samples. Let $\chi_{\{\cdot\}}$ be the indicator function. Applying this method in the infinite sample situation, i.e. $\widehat{f}_{obs} = f_{obs}$, would cause our non-anomalous samples to come from the density $p(x) = \frac{f_{obs}(x)\chi_{\{f_{obs}(x)>\lambda\}}}{\tau}$ where $\tau = \int \chi_{\{f(y)>\lambda\}} f(y)dy$. See Figure 3.3. Perfect recovery of f_{tar} using this method requires $\varepsilon f_{con}(x) \leq f_{tar}(x)\,(1-\varepsilon)$ for all x and that f_{con} and f_{tar} have disjoint supports. The first assumption means that this density estimator can only recover f_{tar} if it has a drop off on the boundary of its support, whereas Assumption A only requires that f_{tar} have finite support. See the last diagram in Figure 3.3. Although these assumptions may be reasonable in certain situations, we find them less palatable than Assumption A. We also evaluate this approach experimentally later and find that it performs poorly.

Another approach based on anomaly detection would be to find the connected components of f_{obs} and declare those that are, in some sense, small as being anomalous. A "small" connected component may be one that integrates to a small value, or

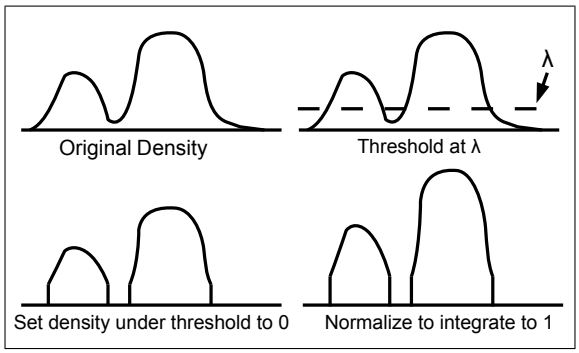

Figure 3.3: Infinite sample version of the level set rejection KDE

which has a small mode. Unfortunately this approach also assumes that f_{tar} and f_{con} have disjoint supports. There are also computational issues with this anomaly detection scheme; finding connected components, finding modes, and numerical integration are computationally difficult.

To some degree, R_ε^A actually achieves the objectives of the previous two robust KDEs. For the first model, the R_ε^A does indeed set those regions of the pdf that are below some threshold to zero. For the second, if the magnitude of the level at which we choose to slice off the bottom of the contaminated density is larger than the mode of the anomalous component then the anomalous component will be eliminated.

3.2 Scaled Projection Kernel Density Estimator

Here we consider approximating R_ε^A in a finite sample situation. Let $f \in L^2\left(\mathbb{R}^d\right)$ be a pdf and $X_1, \ldots, X_n$ be iid samples from f. Let $k_\sigma\left(x, x'\right)$ be a radial smoothing kernel with bandwidth σ such that $k_\sigma\left(x, x'\right) = \sigma^{-d} q\left(\left\|x - x'\right\|_2 / \sigma\right)$, where $q\left(\left\|\cdot\right\|_2\right) \in L^2\left(\mathbb{R}^d\right)$ and is a pdf. The classic kernel density estimator is:

$$\bar{f}_\sigma^n := \frac{1}{n} \sum_1^n k_\sigma\left(\cdot, X_i\right).$$

In practice ε is usually not known and Assumption A is violated. Because of this

we will scale our density by $\beta > 1$ rather than $\frac{1}{1-\varepsilon}$. For a density f define

$$Q_\beta(f) \triangleq \max\left\{\beta f\left(\cdot\right) - \alpha, 0\right\},$$

where $\alpha = \alpha(\beta)$ is set such that the RHS is a pdf. β can be used to tune robustness with larger β corresponding to more robustness (setting β to all the following transforms simply yields the KDE). Given a KDE we would ideally like to apply Q_β directly and search over α until $\max\left\{\beta \bar{f}_\sigma^n\left(\cdot\right) - \alpha, 0\right\}$ integrates to 1. Such an estimate requires multidimensional numerical integration and is not computationally tractable. The SPKDE is an alternative approach that always yields a density and manifests the transformed density in its asymptotic limit.

We now introduce the construction of the SPKDE. Let $\mathcal{D}_\sigma^n$ be the convex hull of $k_\sigma\left(\cdot, X_1\right), \ldots, k_\sigma\left(\cdot, X_n\right)$ (the space of weighted kernel density estimators). The SPKDE is defined as

$$f_{\sigma,\beta}^n := \arg\min_{g \in \mathcal{D}_\sigma^n} \left\|\beta \bar{f}_\sigma^n - g\right\|_{L^2},$$

which is guaranteed to have a unique minimizer since $\mathcal{D}_\sigma^n$ is closed and convex and we are projecting in a Hilbert space (*Bauschke and Combettes* (2011) Theorem 3.14). If we represent $f_{\sigma,\beta}^n$ in the form

$$f_{\sigma,\beta}^n = \sum_1^n a_i k_\sigma\left(\cdot, X_i\right),$$

then the minimization problem is a quadratic program over the vector $a = [a_1, \ldots, a_n]^T$, with a restricted to the probabilistic simplex, Δ^n. Let G be the Gram

matrix of $k_\sigma(\cdot, X_1), \ldots, k_\sigma(\cdot, X_n)$, that is

$$\begin{aligned} G_{ij} &= \langle k_\sigma(\cdot, X_i), k_\sigma(\cdot, X_j) \rangle_{L^2} \\ &= \int k_\sigma(x, X_i) k_\sigma(x, X_j) \, dx. \end{aligned}$$

Let $\mathbf{1}$ be the ones vector and $b = G\mathbf{1}\frac{\beta}{n}$, then the quadratic program is

$$\min_{a \in \Delta^n} a^T G a - 2b^T a.$$

Since G is a Gram matrix, and therefore positive-semidefinite, this quadratic program is convex. Furthermore, the integral defining G_{ij} can be computed in closed form for many kernels of interest. For example for the Gaussian kernel

$$k_\sigma(x, x') = \left(2\pi\sigma^2\right)^{-\frac{d}{2}} \exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right) \implies G_{ij} = k_{\sqrt{2}\sigma}(X_i, X_j),$$

and for the Cauchy kernel *Berry et al.* (1996)

$$k_\sigma(x, x') = \frac{\Gamma\left(\frac{1+d}{2}\right)}{\pi^{(d+1)/2} \cdot \sigma^d} \left(1 + \frac{\|x - x'\|^2}{\sigma^2}\right)^{-\frac{1+d}{2}} \implies G_{ij} = k_{2\sigma}(X_i, X_j).$$

We now present some results on the asymptotic behavior of the SPKDE. Let $\mathcal{D}$ be the set of all pdfs in $L^2\left(\mathbb{R}^d\right)$. The infinite sample version of the SPKDE is

$$f'_\beta = \arg\min_{h \in \mathcal{D}} \|\beta f - h\|_{L^2}^2 .$$

It is worth noting that projection operators in Hilbert space, like the one above, are known to be well defined if the convex set we are projecting onto is closed and convex. $\mathcal{D}$ is not closed in $L^2\left(\mathbb{R}^d\right)$, but this turns out not to be an issue because of the form of βf. For details see the proof of Lemma III.5 in the supplemental material.

Lemma III.5. *$f'_\beta = \max\{\beta f(\cdot) - \alpha, 0\}$ where α is set such that $\max\{\beta f(\cdot) - \alpha, 0\}$ is a pdf.*

Given the same rate on bandwidth necessary for consistency of the traditional KDE, the SPKDE converges to its infinite sample version in its asymptotic limit.

Theorem III.6. *Let $f \in L^2(\mathbb{R}^d)$. If $n \to \infty$ and $\sigma \to 0$ with $n\sigma^d \to \infty$ then $\left\| f^n_{\sigma,\beta} - f'_\beta \right\|_{L^2} \xrightarrow{p} 0$.*

Because $f^n_{\sigma,\beta}$ is a sequence of pdfs and $f'_\beta \in L^2(\mathbb{R}^d)$, it is possible to show L^2 convergence implies L^1 convergence.

Corollary III.7. *Given the conditions in the previous theorem statement,*

$$\left\| f^n_{\sigma,\beta} - f'_\beta \right\|_{L^1} \xrightarrow{p} 0.$$

To summarize, the SPKDE converges to a transformed version of f. In the next section we will show that under Assumption A and with $\beta = \frac{1}{1-\varepsilon}$, the SPKDE converges to f_{tar}.

3.2.1 SPKDE Decontamination

Let $f_{tar} \in L^2(\mathbb{R}^d)$ be a pdf having support with finite Lebesgue measure and let f_{tar} and f_{con} satisfy Assumption A. Let $X_1, X_2, \ldots, X_n$ be iid samples from $f_{obs} = (1-\varepsilon) f_{tar} + \varepsilon f_{con}$ with $\varepsilon \in [0, 1)$. Finally let $f^n_{\sigma,\beta}$ be the SPKDE constructed from $X_1, \ldots, X_n$, having bandwidth σ and robustness parameter β. We have

Theorem III.8. *Let $\beta = \frac{1}{1-\varepsilon}$. If $n \to \infty$ and $\sigma \to 0$ with $n\sigma^d \to \infty$ then $\left\| f^n_{\sigma,\beta} - f_{tar} \right\|_{L_1} \xrightarrow{p} 0$.*

To our knowledge this result is the first of its kind, wherein a nonparametric density estimator is able to asymptotically recover the underlying density in the presence of contaminated data.

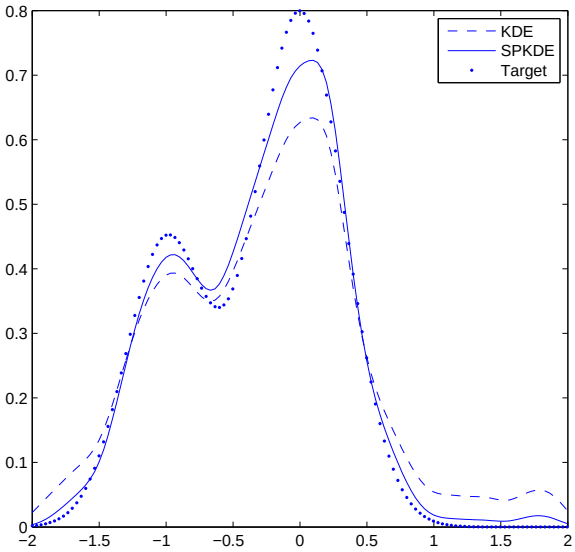

Figure 3.4: KDE and SPKDE in the presence of uniform noise

3.3 Experiments

For all of the experiments optimization was performed using projected gradient descent. The projection onto the probabilistic simplex was done using the algorithm developed in *Duchi et al.* (2008) (which was actually originally discovered a few decades ago *Brucker* (1984); *Pardalos and Kovoor* (1990)).

3.3.1 Synthetic Data

To show that the SPKDE's theoretical properties are manifested in practice we conducted an idealized experiment where the contamination is uniform and the contamination proportion is known. Figure 3.4 exhibits the ability of the SPKDE to compensate for uniform noise. Samples for the density estimator came from a mixture of the "Target" density with a uniform contamination on $[-2, 2]$, sampling from the contamination with probability $\varepsilon = 0.2$. This experiment used 500 samples and the robustness parameter β was set to $\frac{1}{1-\varepsilon} = \frac{5}{4}$ (the value for perfect asymptotic decontamination).

The SPKDE performs well in this situation and yields a scaled and shifted version of the standard KDE. This scale and shift is especially evident in the preservation of the bump on the right hand side of Figure 3.4.

3.3.2 Datasets

In our remaining experiments we investigate two performance metrics for different amounts of contamination. We perform our experiments on 12 classification datasets (names given in the supplemental material) where the 0 label is used as the target density and the 1 label is the anomalous contamination. This experimental setup **does not** satisfy Assumption A. The training datasets are constructed with n_0 samples from label 0 and $\frac{\varepsilon}{1-\varepsilon}n_0$ samples from label 1, thus making an ε proportion of our samples come from the contaminating density. For our experiments we use the values $\varepsilon = 0, 0.05, 0.1, 0.15, 0.20, 0.25, 0.30$. Given some dataset we are interested in how well our density estimators $\widehat{f}$ estimate the density of the 0 class of our dataset, f_{tar}. Each test is performed on 15 permutations of the dataset. The experimental setup here is similar to the setup in Kim & Scott *Kim and Scott* (2012), the most significant difference being that σ is set differently.

3.3.3 Performance Criteria

First we investigate the Kullback-Leibler (KL) divergence

$$D_{KL}\left(\widehat{f}||f_0\right) = \int \widehat{f}\,(x) \log\left(\frac{\widehat{f}\,(x)}{f_0\,(x)}\right) dx.$$

This KL divergence is large when $\widehat{f}$ estimates f_0 to have mass where it does not. For example, in our context, $\widehat{f}$ makes mistakes because of outlying contamination. We estimate this KL divergence as follows. Since we do not have access to f_0, it is estimated from the testing sample using a KDE, $\widetilde{f}_0$. The bandwidth for $\widetilde{f}_0$ is set

using the testing data with a LOOCV line search minimizing $D_{KL}\left(f_0||\widetilde{f}_0\right)$, which is described in more detail below. We then approximate the integral using a sample mean by generating samples from $\widehat{f}$, $\{x_i'\}_1^{n'}$ and using the estimate

$$D_{KL}\left(\widehat{f}||f_0\right) \approx \frac{1}{n'}\sum_1^{n'} \log\left(\frac{\widehat{f}\left(x_i'\right)}{\widetilde{f}_0\left(x_i'\right)}\right).$$

The number of generated samples n' is set to double the number of training samples.

Since KL divergence isn't symmetric we also investigate

$$D_{KL}\left(f_0||\widehat{f}\right) = \int f_0\left(x\right)\log\left(\frac{f_0\left(x\right)}{\widehat{f}\left(x\right)}\right)dx = C - \int f_0\left(y\right)\log\left(\widehat{f}\left(y\right)\right)dy,$$

where C is a constant not depending on $\widehat{f}$. This KL divergence is large when f_0 has mass where $\widehat{f}$ does not. The final term is easy to estimate using expectation. Let $\{x_i''\}_1^{n''}$ be testing samples from f_0 (not used for training). The following is a reasonable approximation

$$-\int f_0\left(y\right)\log\left(\widehat{f}\left(y\right)\right)dy \approx -\frac{1}{n''}\sum_1^{n''}\log\left(\widehat{f}\left(x_i''\right)\right).$$

For a given performance metric and contamination amount, we compare the mean performance of two density estimators across datasets using the Wilcoxon signed rank test *Wilcoxon* (1945). Given N datasets we first rank the datasets according to the absolute difference between performance criterion, with h_i being the rank of the ith dataset. For example if the jth dataset has the largest absolute difference we set $h_j = N$ and if the kth dataset has the smallest absolute difference we set $h_k = 1$. We let R_1 be the sum of the h_is where method one's metric is greater than metric two's and R_2 be the sum of the h_is where method two's metric is larger. The test statistic is $\min(R_1, R_2)$, which we do not report. Instead we report R_1 and R_2 and the p-value that the two methods do not perform the same on average. $R_i < R_j$ is indicative of

method i performing better than method j.

3.3.4 Methods

The data were preprocessed by scaling to fit in the unit cube. This scaling technique was chosen over whitening because of issues with singular covariance matrices. The Gaussian kernel was used for all density estimates. For each permutation of each dataset, the bandwidth parameter is set using the training data with a LOOCV line search minimizing $D_{KL}\left(f_{obs}||\widehat{f}\right)$, where $\widehat{f}$ is the KDE based on the contaminated data and f_{obs} is the observed density. This metric was used in order to maximize the performance of the traditional KDE in KL divergence metrics. For the SPKDE the parameter β was chosen to be 2 for all experiments. This choice of β is based on a few preliminary experiments for which it yielded good results over various sample contamination amounts. The construction of the RKDE follows exactly the methods outlined in the "Experiments" section of Kim & Scott *Kim and Scott* (2012). It is worth noting that the RKDE depends on the loss function used and that the Hampel loss used in these experiments very aggressively suppresses the kernel weights on the tails. Because of this we expect that RKDE performs well on the $D_{KL}\left(\widehat{f}||f_0\right)$ metric. We also compare the SPKDE to a kernel density estimator constructed from samples declared non-anomalous by a level set anomaly detection as described in Section 3.1.3. To do this we first construct the classic KDE, $\bar{f}_\sigma^n$ and then reject those samples in the lower 10th percentile of $\bar{f}_\sigma^n(X_i)$. Those samples not rejected are used in a new KDE, the "rejKDE" using the same σ parameter.

3.3.5 Results

We present the results of the Wilcoxon signed rank tests in Table 3.1. Experimental results for each dataset can be found in the supplemental material. From the results it is clear that the SPKDE is effective at compensating for contamination

Table 3.1: Wilcoxon signed rank test results

	Wilcoxon Test Applied to $D_{KL}\left(\widehat{f}\|\|f_0\right)$						
ε	0	0.05	0.1	0.15	0.2	0.25	0.3
SPKDE	5	0	1	2	0	0	0
KDE	73	78	77	76	78	78	78
p-value	.0049	5e-4	1e-3	.0015	5e-4	5e-4	5e-4
SPKDE	53	59	58	67	63	61	63
RKDE	25	19	20	11	15	17	15
p-value	0.31	0.13	0.15	.027	.064	.092	.064
SPKDE	0	0	1	1	0	2	0
rejKDE	78	78	77	77	78	76	78
p-value	5e-4	5e-4	1e-3	1e-3	5e-4	.0015	5e-4
	Wilcoxon Test Applied to $D_{KL}\left(f_0\|\|\widehat{f}\right)$						
ε	0	0.05	0.1	0.15	0.2	0.25	0.3
SPKDE	37	30	27	21	17	16	17
KDE	41	48	51	57	61	62	61
p-value	.91	.52	.38	.18	.092	.078	.092
SPKDE	14	14	14	10	10	12	12
RKDE	64	64	64	68	68	66	66
p-value	.052	.052	.052	.021	.021	.034	.034
SPKDE	29	21	19	15	13	9	11
rejKDE	49	57	59	63	65	69	67
p-value	.47	.18	.13	.064	.043	.016	.027

in the $D_{KL}\left(\widehat{f}||f_0\right)$ metric, albeit not quite as well as the RKDE. The main advantage of the SPKDE over the RKDE is that it significantly outperforms the RKDE in the $D_{KL}\left(f_0||\widehat{f}\right)$ metric. The rejKDE performs significantly worse than the SPKDE on almost every experiment. Remarkably the SPKDE outperforms the KDE in the situation with no contamination ($\varepsilon = 0$) for both performance metrics.

CHAPTER IV

An Operator Theoretic Approach to Nonparametric Mixture Models

We begin this chapter by formally developing a notion of identifiability for nonparametric mixture models. Next we state several tight bounds for the identifiability of finite mixture models. After a quick introduction to tensor products of Hilbert spaces we prove these bounds. Next we introduce a highly general method for recovering nonparametric mixture components and demonstrate that that this method will asymptotically recover the mixture components in a finite discrete sample space. Finally we include experimental results demonstrating that the recovery method does indeed work in practice.

4.1 Problem Setup

We treat this problem in a general setting. For any measurable space we define δ_x as the Dirac measure at x. For Υ a set, σ-algebra, or measure, we denote $\Upsilon^{\times a}$ to be the standard a-fold product associated with that object. Let $\mathbb{N}$ be the set of integers greater than or equal to zero and $\mathbb{N}_+$ be the integers strictly greater than 0. For $k \in \mathbb{N}_+$, we define $[k] \triangleq \mathbb{N}_+ \cap [1, k]$. Let Ω be a set containing more than one element. This set is the sample space of our data. Let $\mathcal{F}$ be a σ-algebra over

Ω. Assume $\mathcal{F} \neq \{\emptyset, \Omega\}$, i.e. $\mathcal{F}$ contains nontrivial events. We denote the space of probability measures over this space as $\mathcal{D}(\Omega, \mathcal{F})$, which we will shorten to $\mathcal{D}$. We will equip $\mathcal{D}$ with the σ-algebra $2^{\mathcal{D}}$ so that each Dirac measure over $\mathcal{D}$ is unique. Define $\Delta(\mathcal{D}) \triangleq \text{span}(\{\delta_x : x \in \mathcal{D}\})$. This is the ambient space where our mixtures of probability measures live. Let $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ be a probability measure in $\Delta(\mathcal{D})$. Let $\mu \sim \mathscr{P}$ and $X_1, \ldots, X_n \overset{iid}{\sim} \mu$. Here X is a random group sample, which was described in the introduction. We will denote $X = (X_1, \ldots, X_n)$.

We now derive the probability law of X. Let $A \in \mathcal{F}^{\times n}$. Letting $\mathbb{P}$ reflect both the draw of $\mu \sim \mathscr{P}$ and $X_1, \ldots, X_n \overset{iid}{\sim} \mu$, we have

$$\begin{aligned} \mathbb{P}(X \in A) &= \sum_{i=1}^{m} \mathbb{P}(X \in A | \, \mu = \mu_i) \mathbb{P}(\mu = \mu_i) \\ &= \sum_{i=1}^{m} w_i \mu_i^{\times n}(A). \end{aligned}$$

The second equality follows from Lemma 3.10 in *Kallenberg* (2002). So the probability law of X is

$$\sum_{i=1}^{m} w_i \mu_i^{\times n}. \tag{4.1}$$

We want to view the probability law of X as a function of $\mathscr{P}$ in a mathematically rigorous way, which requires a bit of technical buildup. Let $\mathscr{Q} \in \Delta(\mathcal{D})$. From the definition of $\Delta(\mathcal{D})$ it follows that $\mathscr{Q}$ admits the representation

$$\mathscr{Q} = \sum_{i=1}^{r} \alpha_i \delta_{\nu_i}.$$

From the well-ordering principle there must exist some representation with minimal r and we define this r as the *order* of $\mathscr{Q}$. We can show that the minimal representation of any $\mathscr{Q} \in \Delta(\mathcal{D})$ is unique up to permutation of its indices.

Lemma IV.1. *Let $\mathscr{Q} \in \Delta(\mathcal{D})$ and admit minimal representations $\mathscr{Q} = \sum_{i=1}^{r} \alpha_i \delta_{\nu_i} = \sum_{j=1}^{r} \alpha'_j \delta_{\nu'_j}$. There exists some permutation $\psi : [r] \to [r]$ such that $\nu_{\psi(i)} = \nu'_i$ and $\alpha_{\psi(i)} = \alpha'_i$ for all i.*

Henceforth when we define an element of $\Delta(\mathcal{D})$ with a summation we will assume that the summation is a minimal representation.

Definition IV.2. We call $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ a *mixture of measures* if it is a probability measure in $\Delta(\mathcal{D})$. The elements $\mu_1, \ldots, \mu_m$, are called *mixture components*.

Any minimal representation of a mixture of measures $\mathscr{P}$ with m components satisfies $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ with $w_i > 0$ for all i and $\sum_{i=1}^{m} w_i = 1$. Hence any mixture of measures is a convex combination of Dirac measures at elements in $\mathcal{D}$.

For a measurable space $(\Psi, \mathcal{G})$ we define $\mathscr{M}(\Psi, \mathcal{G})$ as the space of all finite signed measures over $(\Psi, \mathcal{G})$. We can now introduce the operator $V_n : \Delta(\mathcal{D}) \to \mathscr{M}(\Omega^{\times n}, \mathcal{F}^{\times n})$. For a minimal representation $\mathscr{Q} = \sum_{i=1}^{r} \alpha_i \delta_{\nu_i}$, we define V_n, with $n \in \mathbb{N}_+$, as

$$V_n(\mathscr{Q}) = \sum_{i=1}^{r} \alpha_i \nu_i^{\times n}.$$

This mapping is well defined as a consequence of Lemma IV.1. From this definition we have that $V_n(\mathscr{P})$ is simply the law of X which we derived earlier. In the following definitions, two mixtures of measures are considered equal if they define the same measure.

Definition IV.3. We call a mixture of measures, $\mathscr{P}$, *n-identifiable* if there does not exist a different mixture of measures $\mathscr{P}'$, with order no greater than the order of $\mathscr{P}$, such that $V_n(\mathscr{P}) = V_n(\mathscr{P}')$.

Definition IV.4. We call a mixture of measures, $\mathscr{P}$, *n-determined* if there exists no other mixture of measures $\mathscr{P}'$ such that $V_n(\mathscr{P}) = V_n(\mathscr{P}')$.

Definition IV.3 and IV.4 are central objects of interest in this work. Given a mixture of measures, $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ then $V_n(\mathscr{P})$ is equal to $\sum_{i=1}^{m} w_i \mu_i^{\times n}$, the measure from which X is drawn. If $\mathscr{P}$ is not n-identifiable then we know that there exists a different mixture of measures that is no more complex (in terms of number of mixture components) than $\mathscr{P}$ which induces the same distribution on X. Practically speaking this means we need more samples in each random group X in order for the full richness of $\mathscr{P}$ to be manifested in X. A stronger version of n-identifiability is n-determinedness where we enforce the requirement that our mixture of measures be the *only* mixture of measures (of any order) that admits the distribution on X.

A quick note on terminology. We use the term "mixture of measures" rather than "mixture model" to emphasize that a mixture of measures should be interpreted a bit differently than a typical mixture model. A "mixture model" connotes a probability measure on the sample space of observed data Ω, whereas a "mixture of measures" connotes a probability measure on the sample space of the unobserved latent measures $\mathcal{D}$.

4.2 Main Results

The first result is a bound on the n-identifiability of all mixtures of measures with m or fewer components. This bound cannot be uniformly improved.

Theorem IV.5. *Let $(\Omega, \mathcal{F})$ be a measurable space. Mixtures of measures with m components are $(2m-1)$-identifiable.*

Theorem IV.6. *Let $(\Omega, \mathcal{F})$ be a measurable space with $\mathcal{F} \neq \{\emptyset, \Omega\}$. For all m, there exists a mixture of measures with $m \geq 2$ components that is not $(2m-2)$-identifiable.*

The following lemmas convey the unsurprising fact that n-identifiability is, in some sense, monotonic.

Lemma IV.7. *If a mixture of measures is n-identifiable then it is q-identifiable for all $q > n$.*

Lemma IV.8. *If a mixture of measures is not n-identifiable then it is not q-identifiable for any $q < n$.*

Viewed alternatively these results say that $n = 2m - 1$ is the smallest value for which V_n is injective over the set of mixtures of measures with m or fewer components.

We also present an analogous bound for n-determinedness. This bound also cannot be improved.

Theorem IV.9. *Let $(\Omega, \mathcal{F})$ be a measurable space. Mixtures of measures with m components are $2m$-determined.*

Theorem IV.10. *Let $(\Omega, \mathcal{F})$ be a measurable space with $\mathcal{F} \neq \{\emptyset, \Omega\}$. For all m, there exists a mixture of measures with m components that is not $(2m-1)$-determined.*

Again n-determinedness is monotonic in the number of samples per group.

Lemma IV.11. *If a mixture of measures is n-determined then it is q-determined for all $q > n$.*

Lemma IV.12. *If a mixture of measures is not n-determined then it is not q-determined for any $q < n$.*

This collection of results can be interpreted in an alternative way. Consider some pair of mixtures of measures $\mathscr{P}, \mathscr{P}'$. If $n \geq 2m$ and either mixture of measures is of order m or less, then $V_n(\mathscr{P}) = V_n(\mathscr{P}')$ implies $\mathscr{P} = \mathscr{P}'$. Furthermore $n = 2m$ is the smallest value of n for which the previous statement is true for all pairs of mixtures of measures.

Our definitions of n-identifiability, n-determinedness, and their relation to previous works on identifiability deserve a bit of discussion. Some previous works on

identifiability contain results related to what we call "identifiability" and others contain results related what we call "determinedness." Both of these are simply called "identifiability" in these works. For example in *Yakowitz and Spragins* (1968) it is shown that different finite mixtures of multivariate Gaussian distributions will always yield different distributions, a result which we could call "determinedness." Alternatively, in *Teicher* (1963) it is demonstrated that mixtures of binomial distributions, with a fixed number of trials n for every mixture component, are identifiable provided we only consider mixtures with m mixture components and $n \geq 2m - 1$. In this result allowing for more mixture components may destroy identifiability and thus this is what *we* would call an "identifiability" result. The fact that the value $2m - 1$ occurs in both the previous binomial mixture model result and Theorem IV.5 is not a coincidence. We will demonstrate a new determinedness result for multinomial mixtures models later in the chapter, under the assumption that $n \geq 2m$. We will prove these results using Theorems IV.5 and IV.9. To our knowledge our work is the first to consider both identifiability and determinedness.

Finally we also include results that are analogous to previously shown results for the discrete setting. We note that our proof techniques are markedly different than the previous proofs for the discrete case.

Theorem IV.13. *If $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ is a mixture of measures where $\mu_1, \ldots, \mu_m$ are linearly independent then $\mathscr{P}$ is* 3*-identifiable.*

This bound is tight as a consequence of Theorem IV.6 with $m = 2$ since any pair of distinct measures must be linearly independent.

A version of this theorem was first proven in *Allman et al.* (2009) by making use of Kruskal's Theorem *Kruskal* (1977). Kruskal's Theorem demonstrates that order 3 tensors over $\mathbb{R}^d$ admit unique decompositions (up to scaling and permutation) given certain linear independence assumptions. Our proof makes no use of Kruskal's Theorem and demonstrates that n-identifiability for linearly independent mixture

components need not be attached to the discrete version in any way. An efficient algorithm for recovering linearly independent mixture components for discrete sample spaces with 3 samples per random group is described in *Anandkumar et al.* (2014). Interestingly, with one more sample per group, these mixtures of measures become determined.

Theorem IV.14. *If $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ is a mixture of measures where $\mu_1, \ldots, \mu_m$ are linearly independent then $\mathscr{P}$ is* 4*-determined.*

This bound is tight as a result of Theorem IV.10 with $m = 2$.

Our final result is related to the "separability condition" found in *Donoho and Stodden* (2003). The separability condition in the discrete case requires that, for each mixture component μ_i, there exists $B_i \in \mathcal{F}$ such that $\mu_i(B_i) > 0$ and $\mu_j(B_i) = 0$ for all $i \neq j$. There exists a generalization of the separability condition, known as *joint irreducibility*.

Definition IV.15. A collection of probability measures $\mu_1, \ldots, \mu_m$ are said to be *jointly irreducible* if $\sum_{i=1}^{m} w_i \mu_i$ being a probability measure implies $w_i \geq 0$.

In other words, any probability measure in the span of $\mu_1, \ldots, \mu_m$ must be a convex combination of those measures. It was shown in *Blanchard and Scott* (2014) that separability implies joint irreducibility, but not visa-versa. In that paper it was also shown that joint irreducibility implies linear independence, but the converse does not hold.

Theorem IV.16. *If $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ is a mixture of measures where $\mu_1, \ldots, \mu_m$ are jointly irreducible then $\mathscr{P}$ is* 2*-determined.*

A straightforward consequence of the corollary of Theorem 1 in *Donoho and Stodden* (2003) is that any mixture of measures on a finite discrete space with jointly irreducible components is 2-identifiable. The result in *Donoho and Stodden* (2003)

is concerned with the uniqueness of nonnegative matrix factorizations and Theorem IV.16, when applied to a finite discrete space, can be posed as a special case of the result in *Donoho and Stodden* (2003). In the context of nonnegative matrix factorization the result in *Donoho and Stodden* (2003) is significantly more general than our result. In another sense our result is more general since it applies to spaces where joint irreducibility and the separability condition are not equivalent. Furthermore *Donoho and Stodden* (2003) only implies that the mixture of measures in Theorem IV.16 are identifiable. The determinedness result is, as far as we know, totally new.

4.3 Tensor Products of Hilbert Spaces

Our proofs will rely heavily on the geometry of tensor products of Hilbert spaces which we will introduce in this section.

4.3.1 Overview of Tensor Products

First we introduce tensor products of Hilbert spaces. To our knowledge there does not exist a rigorous construction of the tensor product Hilbert space which is both succinct and intuitive. Because of this we will simply state some basic facts about tensor products of Hilbert spaces and hopefully instill some intuition for the uninitiated by way of example. A thorough treatment of tensor products of Hilbert spaces can be found in *Kadison and Ringrose* (1983).

Let H and H' be Hilbert spaces. From these two Hilbert spaces the "simple tensors" are elements of the form $h \otimes h'$ with $h \in H$ and $h' \in H'$. We can treat the simple tensors as being the basis for some inner product space H_0, with the inner product of simple tensors satisfying

$$\langle h_1 \otimes h_1', h_2 \otimes h_2' \rangle = \langle h_1, h_2 \rangle \langle h_1', h_2' \rangle .$$

The tensor product of H and H' is the completion of H_0 and is denoted $H \otimes H'$. To avoid potential confusion we note that the notation just described is standard in operator theory literature. In some literature our definition of H_0 is denoted as $H \otimes H'$ and our definition of $H \otimes H'$ is denoted $H \widehat{\otimes} H'$.

As an illustrative example we consider the tensor product $L^2(\mathbb{R}) \otimes L^2(\mathbb{R})$. It can be shown that there exists an isomorphism between $L^2(\mathbb{R}) \otimes L^2(\mathbb{R})$ and $L^2(\mathbb{R}^2)$ that maps the simple tensors to separable functions *Kadison and Ringrose* (1983), $f \otimes f' \mapsto f(\cdot)f'(\cdot)$. We can demonstrate this isomorphism with a simple example. Let $f, g, f', g' \in L^2(\mathbb{R})$. Taking the $L^2(\mathbb{R}^2)$ inner product of $f(\cdot)f'(\cdot)$ and $g(\cdot)g'(\cdot)$ gives us

$$\begin{aligned} \int\int (f(x)f'(y))\,(g(x)g'(y))dxdy &= \int f(x)g(x)dx \int f'(y)g'(y)dy \\ &= \langle f, g\rangle \langle f', g'\rangle \\ &= \langle f \otimes f', g \otimes g'\rangle . \end{aligned}$$

Beyond tensor product we will need to define tensor power. To begin we will first show that tensor products are, in a certain sense, associative. Let H_1, H_2, H_3 be Hilbert spaces. Proposition 2.6.5 in *Kadison and Ringrose* (1983) states that there is a unique unitary operator, $U : (H_1 \otimes H_2) \otimes H_3 \to H_1 \otimes (H_2 \otimes H_3)$, that satisfies the following for all $h_1 \in H_1, h_2 \in H_2, h_3 \in H_3$,

$$U((h_1 \otimes h_2) \otimes h_3) = h_1 \otimes (h_2 \otimes h_3).$$

This implies that for any collection of Hilbert spaces, $H_1, \ldots, H_n$, the Hilbert space $H_1 \otimes \cdots \otimes H_n$ is defined unambiguously regardless of how we decide to associate the products. In the space $H_1 \otimes \cdots \otimes H_n$ we define a simple tensor as a vector of the form $h_1 \otimes \cdots \otimes h_n$ with $h_i \in H_i$. In *Kadison and Ringrose* (1983) it is shown that

$H_1 \otimes \cdots \otimes H_n$ is the closure of the span of these simple tensors. To conclude this primer on tensor products we introduce the following notation. For a Hilbert space H we denote $H^{\otimes n} = \underbrace{H \otimes H \otimes \cdots \otimes H}_{\text{n times}}$ and for $h \in H$, $h^{\otimes n} = \underbrace{h \otimes h \otimes \cdots \otimes h}_{\text{n times}}$.

4.3.2 Tensor Rank

A tool we will use frequently in our proofs is *tensor rank*, which behaves similarly to matrix rank.

Definition IV.17. Let $h \in H^{\otimes n}$ where H is a Hilbert space. The *rank* of h is the smallest natural number r such that $h = \sum_{i=1}^{r} h_i$ where h_i are simple tensors.

In an infinite dimensional Hilbert space it is possible for a tensor to have infinite rank. We will only be concerned with finite rank tensors.

4.3.3 Some Results for Tensor Product Spaces

We derive some technical results concerning tensor product spaces that will be useful for the rest of the chapter. These lemmas are similar to or are straightforward extensions of previous results which we needed to modify for our particular purposes. Let $(\Psi, \mathcal{G}, \gamma)$ be a σ-finite measure space. We have the following lemma that connects tensor power of a L^2 space to the L^2 space of the product measure. Proofs of many of the lemmas in this chapter are deferred to the Appendix B.1.

Lemma IV.18. *There exists a unitary transform*

$$U : L^2(\Psi, \mathcal{G}, \gamma)^{\otimes n} \to L^2\left(\Psi^{\times n}, \mathcal{G}^{\times n}, \gamma^{\times n}\right)$$

such that, for all $f_1, \ldots, f_n \in L^2(\Psi, \mathcal{G}, \gamma)$,

$$U(f_1 \otimes \cdots \otimes f_n) = f_1(\cdot) \cdots f_n(\cdot).$$

The following lemma is used in the proof of Lemma IV.18 as well as the proof of Theorem IV.6.

Lemma IV.19. *Let $H_1, \ldots, H_n, H_1', \ldots, H_n'$ be a collection of Hilbert spaces and $U_1, \ldots, U_n$ a collection of unitary operators with $U_i : H_i \to H_i'$ for all i. There exists a unitary operator $U : H_1 \otimes \cdots \otimes H_n \to H_1' \otimes \cdots \otimes H_n'$ satisfying $U(h_1 \otimes \cdots \otimes h_n) = U_1(h_1) \otimes \cdots \otimes U_n(h_n)$ for all $h_1 \in H_1, \ldots, h_n \in H_n$.*

A statement of the following lemma for $\mathbb{R}^d$ can be found in *Comon et al.* (2008). We present our own proof for the Hilbert space setting.

Lemma IV.20. *Let $n > 1$ and let $h_1, \ldots, h_n$ be elements of a Hilbert space such that no elements are zero and no pairs of elements are collinear. Then $h_1^{\otimes n-1}, \ldots, h_n^{\otimes n-1}$ are linearly independent.*

The following lemma is a Hilbert space version of a well known property for positive semi-definite matrices.

Lemma IV.21. *Let $h_1, \ldots, h_m$ be elements of a Hilbert space. The rank of $\sum_{i=1}^m h_i^{\otimes 2}$ is the dimension of* $\operatorname{span}(\{h_1, \ldots, h_m\})$.

4.4 Proofs of Theorems

With the tools developed in the previous sections we can now prove our theorems. First we introduce one additional piece of notation. For a function f on a domain $\mathcal{X}$ we define $f^{\times k}$ as simply the product of the function k times on the domain $\mathcal{X}^{\times k}$, $\underbrace{f(\cdot) \cdots f(\cdot)}_{\text{k times}}$. For a set, σ-algebra, or measure the notation continues to denote the standard k-fold product.

In these proofs we will be making extensive use of various L^2 spaces. These spaces will be equivalence classes of functions which are equal almost everywhere with respect to the measure associated with that space. When considering elements of these spaces,

equality will always mean almost everywhere equality with respect to the measure associated with that space. When performing integrals or other manipulations of elements in L^2 spaces, we will be performing operations that do not depend on the representative of the equivalence class. The following lemma will be quite useful.

Lemma IV.22. *Let $\gamma_1 \ldots, \gamma_m, \pi_1 \ldots, \pi_l$ be probability measures on a measurable space $(\Psi, \mathcal{G})$, $a_1 \ldots, a_m, b_1, \ldots b_l \in \mathbb{R}$, and $n \in \mathbb{N}_+$. If*

$$\sum_{i=1}^{m} a_i \gamma_i^{\times n} = \sum_{j=1}^{l} b_j \pi_j^{\times n}$$

then for all $n' \in \mathbb{N}_+$ with $n' \leq n$ we have that

$$\sum_{i=1}^{m} a_i \gamma_i^{\times n'} = \sum_{j=1}^{l} b_j \pi_j^{\times n'}.$$

Proof of Theorem IV.5. We proceed by contradiction. Suppose there exist $m, l \in \mathbb{N}_+$ with $l \leq m$ such that there two different mixtures of measures $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i} \neq \mathscr{P}' = \sum_{j=1}^{l} b_j \delta_{\nu_j}$, and

$$\sum_{i=1}^{m} a_i {\mu_i}^{\times 2m-1} = \sum_{j=1}^{l} b_j \nu_j^{\times 2m-1}. \tag{4.2}$$

By the well-ordering principle there exists a minimal m such that the previous statement holds. For that minimal m there exists a minimal l such that the previous statement holds. We will assume that the m and l are both minimal in this way. This assumption implies that $\mu_i \neq \nu_j$ for all i, j. To prove this we will assume that there exists i, j such that $\mu_i = \nu_j$, and show that this assumption leads to a contradiction. Without loss of generality we will assume that $\mu_m = \nu_l$. We will consider the three cases where $a_m = b_l$, $a_m > b_l$, and $a_m < b_l$.

Case 1. If $a_m = b_l$ then we have that

$$\sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \mu_i^{\times 2m-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-b_l} \nu^{\times 2m-1}$$

and from Lemma IV.22 we have

$$\sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \mu_i^{\times 2(m-1)-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-b_l} \nu^{\times 2(m-1)-1}.$$

Setting $\mathscr{P} = \sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \delta_{\mu_i}$ and $\mathscr{P}' = \sum_{j=1}^{l-1} \frac{b_j}{1-b_l} \delta_{\nu_j}$, we have that

$$V_{2(m-1)-1}(\mathscr{P}) = V_{2(m-1)-1}(\mathscr{P}')$$

which contradicts the minimality of m.

Case 2. If $a_m > b_l$ then we have

$$\sum_{i=1}^{m-1} \frac{a_i}{1-b_l} \mu_i^{\times 2m-1} + \frac{a_m - b_l}{1-b_l} \mu_m^{\times 2m-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-b_l} \nu_j^{\times 2m-1}$$

which contradicts the minimality of l by an argument similar to that in Case 1.

Case 3 If $a_m < b_l$ we have that

$$\sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \mu_i^{\times 2m-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-a_m} \nu_j^{\times 2m-1} + \frac{b_l - a_m}{1-a_m} \nu_l^{\times 2m-1}.$$

Again we will use arguments similar to the one used in Case 1. If $l = m$ then swapping the mixtures associated with m and l gives us a pair of mixtures of measures which violates the minimality of l. If $l < m$ then from Lemma IV.22

we have that

$$\sum_{i=1}^{m-1} \frac{a_i}{1-a_m} \mu_i^{\times 2(m-1)-1} = \sum_{j=1}^{l-1} \frac{b_j}{1-a_m} \nu_j^{\times 2(m-1)-1} + \frac{b_l - a_m}{1-a_m} \nu_l^{\times 2(m-1)-1},$$

which violates the minimality of m.

We have now established that $\mu_i \neq \nu_j$, for all i, j. We will use the following lemma to embed the mixture components in a Hilbert space.

Lemma IV.23. *Let $\gamma_1, \ldots, \gamma_n$ be finite measures on a measurable space $(\Psi, \mathcal{G})$. There exists a finite measure π and non-negative functions $f_1, \ldots, f_n \in L^1(\Psi, \mathcal{G}, \pi) \cap L^2(\Psi, \mathcal{G}, \pi)$ such that, for all i and all $B \in \mathcal{G}$*

$$\gamma_i(B) = \int_B f_i d\pi.$$

From Lemma IV.23 there exists a finite measure ξ and non-negative functions $p_1, \ldots, p_m, q_1, \ldots, q_l \in L^1(\Omega, \mathcal{F}, \xi) \cap L^2(\Omega, \mathcal{F}, \xi)$ such that, for all $B \in \mathcal{F}$, $\mu_i(B) = \int_B p_i d\xi$ and $\nu_j(B) = \int_B q_j d\xi$ for all i, j. Clearly no two of these functions are equal (in the ξ-almost everywhere sense). If one of the functions were a scalar multiple of another, for example $p_1 = \alpha p_2$ for some $\alpha \neq 1$, it would imply

$$\mu_1(\Omega) = \int p_1 d\xi = \int \alpha p_2 d\xi = \alpha.$$

This is not true so no pair of these functions are collinear.

We can use the following lemma to extend this new representation to a product measure.

Lemma IV.24. *Let $(\Psi, \mathcal{G})$ be a measurable space, γ and π a pair of finite measures on that space, and f a nonnegative function in $L^1(\Psi, \mathcal{G}, \pi)$ such that, for all $A \in \mathcal{G}$,*

$\gamma(A) = \int_A f d\pi$. *Then for all* n, *for all* $B \in \mathcal{G}^{\times n}$ *we have*

$$\gamma^{\times n}(B) = \int_B f^{\times n} d\pi^{\times n}.$$

Thus for any $R \in \mathcal{F}^{\times 2m-1}$ we have

$$\begin{aligned}
\int_R \sum_{i=1}^{m} a_i p_i^{\times 2m-1} d\xi^{\times 2m-1} &= \sum_{i=1}^{m} a_i \mu_i^{\times 2m-1}(R) \\
&= \sum_{j=1}^{l} b_j \nu_j^{\times 2m-1}(R) \\
&= \int_R \sum_{j=1}^{l} b_j q_j^{\times 2m-1} d\xi^{\times 2m-1}.
\end{aligned}$$

The following lemma is a well known result in real analysis (Proposition 2.23 in *Folland* (1999)), but it is worth mentioning explicitly.

Lemma IV.25. *Let* $(\Psi, \mathcal{G}, \gamma)$ *be a measure space and* $f, g \in L^1(\Psi, \mathcal{G}, \gamma)$. *Then* $f = g$ γ*-almost everywhere iff, for all* $A \in \mathcal{G}$, $\int_A f d\gamma = \int_A g d\gamma$.

From this lemma it follows that

$$\sum_{i=1}^{m} a_i p_i^{\times 2m-1} = \sum_{j=1}^{l} b_j q_j^{\times 2m-1}.$$

Applying the U^{-1} operator from Lemma IV.18 to the previous equation yields

$$\sum_{i=1}^{m} a_i p_i^{\otimes 2m-1} = \sum_{j=1}^{l} b_j q_j^{\otimes 2m-1}.$$

Since $l + m \leq 2m$ Lemma IV.20 states that

$$p_1^{\otimes 2m-1}, \ldots, p_m^{\otimes 2m-1}, q_1^{\otimes 2m-1}, \ldots, q_l^{\otimes 2m-1}$$

are all linearly independent and thus $a_i = 0$ and $b_j = 0$ for all i, j, a contradiction. $\square$

Proof of Theorem IV.6. To prove this theorem we will construct a pair of mixture of measures, $\mathscr{P} \neq \mathscr{P}'$ which both contain m components and satisfy $V_{2m-2}(\mathscr{P}) = V_{2m-2}(\mathscr{P}')$. From our definition of $(\Omega, \mathcal{F})$ we know there exists $F \in \mathcal{F}$ such that F and F^C are nonempty. Let $x \in F$ and $x' \in F^C$. It follows that δ_x and $\delta_{x'}$ are different probability measures on $(\Omega, \mathcal{F})$. The theorem follows from the next lemma. We will prove the lemma after the theorem proof.

Lemma IV.26. *Let $(\Psi, \mathcal{G})$ be a measurable space and γ, γ' be distinct probability measures on that space. Let $\varepsilon_1, \ldots, \varepsilon_t$ be $t \geq 3$ distinct values in $[0, 1]$. Then there exist $\beta_1, \ldots, \beta_t$, a permutation $\sigma : [t] \to [t]$, and $l \in \mathbb{N}_+$ such that*

$$\sum_{i=1}^{l} \beta_i \left(\varepsilon_{\sigma(i)}\gamma + \left(1 - \varepsilon_{\sigma(i)}\right)\gamma'\right)^{\times t-2} = \sum_{j=l+1}^{t} \beta_j \left(\varepsilon_{\sigma(j)}\gamma + \left(1 - \varepsilon_{\sigma(j)}\right)\gamma'\right)^{\times t-2}$$

where $\beta_i > 0$ for all i, $\sum_{i=1}^{l} \beta_i = \sum_{j=l+1}^{t} \beta_j = 1$, and $l, t - l \geq \left\lfloor \frac{t}{2} \right\rfloor$.

Let $\varepsilon_1, \ldots, \varepsilon_{2m} \in [0, 1]$ be distinct and let $\mu_i = \varepsilon_i \delta_x + (1 - \varepsilon_i)\delta_{x'}$ for $i \in [2m]$. From Lemma IV.26 with $t = 2m$ there exists a permutation $\sigma : [2m] \to [2m]$ and $\beta_1, \ldots, \beta_{2m}$ such that

$$\sum_{i=1}^{m} \beta_i \mu_{\sigma(i)}^{\times 2m-2} = \sum_{j=m+1}^{2m} \beta_j \mu_{\sigma(j)}^{\times 2m-2},$$

with $\sum_{i=1}^{m} \beta_i = \sum_{j=m+1}^{2m} \beta_j = 1$ and $\beta_i > 0$ for all i.

If we let $\mathscr{P} = \sum_{i=1}^{m} \beta_i \delta_{\mu_{\sigma(i)}}$ and $\mathscr{P}' = \sum_{j=m+1}^{2m} \beta_j \delta_{\mu_{\sigma(j)}}$, we have that $V_{2m-2}(\mathscr{P}) = V_{2m-2}(\mathscr{P}')$ and $\mathscr{P} \neq \mathscr{P}'$ since $\mu_1, \ldots, \mu_{2m}$ are distinct. $\square$

For the next proof we will introduce some notation. For a tensor $U \in \mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_l}$ we define $U_{i_1,\ldots,i_l}$ to be the entry in the $[i_1, \ldots, i_l]$ location of U.

Proof of Lemma IV.26. From Lemma IV.23, there exists a finite measure π and non-negative functions $f, f' \in L^1(\Psi, \mathcal{G}, \pi) \cap L^2(\Psi, \mathcal{G}, \pi)$ such that, for all $A \in \mathcal{G}$, $\gamma(A) = \int_A f d\pi$ and $\gamma'(A) = \int_A f' d\pi$.

Let H_2 be the Hilbert space associated with the subspace in $L^2(\Psi, \mathcal{G}, \pi)$ spanned by f and f'. Let $(f_i)_{i=1}^t$ be non-negative functions in $L^1(\Psi, \mathcal{G}, \pi) \cap L^2(\Psi, \mathcal{G}, \pi)$ with $f_i = \varepsilon_i f + (1 - \varepsilon_i) f'$. Clearly f_i is a pdf over π for all i and there are no pair in this collection which are collinear. Since H_2 is isomorphic to $\mathbb{R}^2$ there exists a unitary operator $U : H_2 \to \mathbb{R}^2$. From Lemma IV.19 there exists a unitary operator $U_{t-2} : H_2^{\otimes t-2} \to \mathbb{R}^{2^{\otimes t-2}}$, with $U_{t-2}(h_1 \otimes \cdots \otimes h_{t-2}) = U(h_1) \otimes \cdots \otimes U(h_{t-2})$. Because U is unitary it follows that

$$U_{t-2}\left(\text{span}\left(\left\{h^{\otimes t-2} : h \in H_2\right\}\right)\right) = \text{span}\left(\left\{x^{\otimes t-2} : x \in \mathbb{R}^2\right\}\right).$$

An order r tensor, $A_{i_1,\ldots,i_r}$, is *symmetric* if $A_{i_1,\ldots,i_r} = A_{i_{\psi(1)},\ldots,i_{\psi(r)}}$ for any $i_1, \ldots, i_r$ and permutation $\psi : [r] \to [r]$. A consequence of Lemma 4.2 in *Comon et al.* (2008) is that $\text{span}\left(\{x^{\otimes t-2} : x \in \mathbb{R}^2\}\right) \subset S^{t-2}(\mathbb{C}^2)$, the space of all symmetric order $t-2$ tensors over $\mathbb{C}^2$. Complex symmetric tensor spaces will always be viewed as a vector space over the complex numbers and real symmetric tensor spaces will be always be viewed as a vector space over the real numbers.

From Proposition 3.4 in *Comon et al.* (2008) it follows that the dimension of $S^{t-2}(\mathbb{C}^2)$ is $\begin{pmatrix} 2+t-2-1 \\ t-2 \end{pmatrix} = t-1$. From this it follows that $\dim S^{t-2}(\mathbb{R}^2) \leq t-1$. To see this consider some set of linearly dependent tensors $x_1, \ldots, x_r \in S^{t-2}(\mathbb{C}^2)$ each containing only real valued entries, i.e. the tensors are in $S^{t-2}(\mathbb{R}^2)$. Then it follows that there exists $c_1, \ldots, c_r \in \mathbb{C}$ such that

$$\sum_{i=1}^{r} c_i x_i = 0.$$

Let $\Re$ denote the real component when applied to an element of $\mathbb{C}$, and the real component applied entrywise when applied to a tensor. We have that

$$0 = \Re\left(\sum_{i=1}^{r} c_i x_i\right) = \sum_{i=1}^{r} \Re\left(c_i x_i\right) = \sum_{i=1}^{r} \Re\left(c_i\right) x_i.$$

Thus it follows that $x_1, \ldots x_r$ are linearly dependent in $S^{t-2}\left(\mathbb{R}^2\right)$ and thus the dimensionality bound holds. From this we get that

$$\dim\left(\operatorname{span}\left(\left\{h^{\otimes t-2} : h \in H_2\right\}\right)\right) \leq t-1.$$

The bound on the dimension of $\operatorname{span}\left(\{h^{\otimes t-2} : h \in H_2\}\right)$ implies that $\left(f_i^{\otimes t-2}\right)_{i=1}^{t}$ are linearly dependent. Conversely Lemma IV.20 implies that removing a single vector from $\left(f_i^{\otimes t-2}\right)_{i=1}^{t}$ yields a set of vectors which are linearly independent. It follows that there exists $(\alpha_i)_{i=1}^{t}$ with $\alpha_i \neq 0$ for all i and

$$\sum_{i=1}^{t} \alpha_i f_i^{\otimes t-2} = 0. \tag{4.3}$$

There exists a permutation $\sigma : [t] \to [t]$ such that $\alpha_{\sigma(i)} < 0$ for all $i \in [l]$ and $\alpha_{\sigma(j)} > 0$ for all $j > l$ with $l \leq \left\lfloor \frac{t}{2} \right\rfloor$ (ensuring that $l \leq \left\lfloor \frac{t}{2} \right\rfloor$ may also require multiplying (4.3) by -1). This σ appears in the lemma statement, but for the remainder of the proof we will simply assume without loss of generality that $\alpha_i < 0$ for $i \in [l]$ with $l \leq \left\lfloor \frac{t}{2} \right\rfloor$.

From this we have

$$\sum_{i=1}^{l} -\alpha_i f_i^{\otimes t-2} = \sum_{j=l+1}^{t} \alpha_j f_j^{\otimes t-2}. \tag{4.4}$$

From Lemma IV.18 we have

$$\sum_{i=1}^{l} -\alpha_i f_i^{\times t-2} = \sum_{j=l+1}^{t} \alpha_j f_j^{\times t-2}$$

and thus

$$\begin{aligned}\int \sum_{i=1}^{l} -\alpha_i f_i^{\times t-2} d\pi^{\times t-2} &= \int \sum_{j=l+1}^{t} \alpha_j f_j^{\times t-2} d\pi^{\times t-2} \\ \Rightarrow \sum_{i=1}^{l} -\alpha_i &= \sum_{j=l+1}^{t} \alpha_j.\end{aligned}$$

Let $r = \sum_{i=1}^{l} -\alpha_i$. We know $r > 0$ so dividing both sides of (4.4) by r gives us

$$\sum_{i=1}^{l} -\frac{\alpha_i}{r} f_i^{\otimes t-2} = \sum_{j=l+1}^{t} \frac{\alpha_j}{r} f_j^{\otimes t-2}$$

where the left and the right side are convex combinations. Let $(\beta_i)_{i=1}^{t}$ be positive numbers with $\beta_i = \frac{-\alpha_i}{r}$ for $i \in [l]$ and $\beta_j = \frac{\alpha_j}{r}$ for $j \in [t] \setminus [l]$. This gives us

$$\sum_{i=1}^{l} \beta_i f_i^{\otimes t-2} = \sum_{j=l+1}^{t} \beta_j f_j^{\otimes t-2}. \tag{4.5}$$

We will now consider 3 cases for the value of t.

If $t = 3$ then $l = 1$ and $l, t - l \geq \lfloor \frac{t}{2} \rfloor$ is satisfied.

If t is divisible by two then we can do the following,

$$\sum_{i=1}^{l} \beta_i f_i^{\otimes \frac{t}{2}-1} \otimes f_i^{\otimes \frac{t}{2}-1} = \sum_{j=l+1}^{t} \beta_j f_j^{\otimes \frac{t}{2}-1} \otimes f_j^{\otimes \frac{t}{2}-1}.$$

Consider the elements in the last inequality as order two tensors in $L^2(\Psi, \mathcal{G}, \pi)^{\otimes \frac{t}{2}-1} \otimes L^2(\Psi, \mathcal{G}, \pi)^{\otimes \frac{t}{2}-1}$. From Lemma IV.20 and Lemma IV.21 we have that the RHS of the previous equation has rank at least $\frac{t}{2}$ and since $l \leq \frac{t}{2}$ it follows that $l = \frac{t}{2}$. Again we have that $l, t - l \geq \lfloor \frac{t}{2} \rfloor$.

If t is greater than 3 and not divisible by 2 then we can apply Lemma IV.18 to

get

$$\begin{aligned}\int_{\Psi}\sum_{i=1}^{l}\beta_i f_i^{\times t-3}f_i(x)d\pi(x) &= \int_{\Psi}\sum_{j=l+1}^{t}\beta_j f_j^{\times t-3}f_j(y)d\pi(y)\\ \Rightarrow\sum_{i=1}^{l}\beta_i f_i^{\times t-3} &= \sum_{j=l+1}^{t}\beta_j f_j^{\times t-3}.\end{aligned}$$

Applying Lemma IV.18 again we get

$$\begin{aligned}\sum_{i=1}^{l}\beta_i f_i^{\otimes t-3} &= \sum_{j=l+1}^{t}\beta_j f_j^{\otimes t-3}\\ \Rightarrow\sum_{i=1}^{l}\beta_i f_i^{\otimes\frac{t-1}{2}-1}\otimes f_i^{\otimes\frac{t-1}{2}-1} &= \sum_{j=l+1}^{t}\beta_j f_j^{\otimes\frac{t-1}{2}-1}\otimes f_j^{\otimes\frac{t-1}{2}-1}.\end{aligned}$$

Recall that $\left\lfloor\frac{t}{2}\right\rfloor\geq l$ so we also have that

$$\begin{aligned}\frac{t}{2}-l &\geq -\frac{1}{2}\\ \Rightarrow t-l &\geq \frac{t-1}{2}.\end{aligned}$$

From Lemma IV.20 and Lemma IV.21 we have that the RHS of (4.6) has rank at least $\frac{t-1}{2}$ and thus $l\geq\frac{t-1}{2}$. From this we have that $t-l, l\geq\left\lfloor\frac{t}{2}\right\rfloor$ once again. So $l, t-l\geq\lfloor\frac{t}{2}\rfloor$ for any $t\geq 3$. Applying Lemma IV.18 to (4.5) we have

$$\sum_{i=1}^{l}\beta_i f_i^{\times t-2}=\sum_{j=l+1}^{t}\beta_j f_j^{\times t-2}.$$

From Lemma IV.24 we have

$$\sum_{i=1}^{l}\beta_i\left(\varepsilon_i\gamma+(1-\varepsilon_i)\gamma'\right)^{\times t-2} = \sum_{j=l+1}^{t}\beta_j\left(\varepsilon_j\gamma+(1-\varepsilon_j)\gamma'\right)^{\times t-2}.$$

□

Proof of Theorem IV.9. Let $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ and $\mathscr{P}' = \sum_{j=1}^{l} b_j \delta_{\nu_j}$ be mixtures of measures such that $\mathscr{P}' \neq \mathscr{P}$. We will proceed by contradiction. Suppose that $\sum_{i=1}^{m} a_i \mu_i^{\times 2m} = \sum_{j=1}^{l} b_j \nu_j^{\times 2m}$. From Theorem IV.5 we know that $\mathscr{P}$ is $2m-1$-identifiable and therefore $2m$-identifiable by Lemma IV.7. It follows that $l > m$. From Lemma IV.23 there exists a finite measure ξ and non-negative functions $p_1, \ldots, p_m, q_1, \ldots, q_l \in L^1(\Omega, \mathcal{F}, \xi) \cap L^2(\Omega, \mathcal{F}, \xi)$ such that, for all $B \in \mathcal{F}$, $\mu_i(B) = \int_B p_i d\xi$ and $\nu_j(B) = \int_B q_j d\xi$ for all i, j. Using Lemmas IV.24 and IV.25 we have

$$\sum_{i=1}^{m} a_i p_i^{\times 2m} = \sum_{j=1}^{l} b_j q_j^{\times 2m}.$$

By Lemma IV.18 we have

$$\sum_{i=1}^{m} a_i p_i^{\otimes 2m} = \sum_{j=1}^{l} b_j q_j^{\otimes 2m},$$

and therefore

$$\sum_{i=1}^{m} a_i p_i^{\otimes m} \otimes p_i^{\otimes m} = \sum_{j=1}^{l} b_j q_j^{\otimes m} \otimes q_j^{\otimes m}.$$

Consider the elements in the last inequality as order two tensors in $L^2(\Omega, \mathcal{F}, \xi)^{\otimes m} \otimes L^2(\Omega, \mathcal{F}, \xi)^{\otimes m}$. Since no pair of vectors in $p_1, \ldots, p_m$ are collinear, from Lemma IV.20 and Lemma IV.21 we know that the LHS has rank m. On the other hand, no pair of vectors $q_1, \ldots, q_l$ are collinear either, so Lemma IV.20 says that there is a subset of $\left\{q_1^{\otimes m}, \ldots, q_l^{\otimes m}\right\}$ which contains at least $m+1$ linearly independent elements. By Lemma IV.21 it follows that the RHS has rank at least $m+1$, a contradiction. □

Proof of Theorem IV.10. To prove this theorem we will construct a pair of mixture of measures, $\mathscr{P} \neq \mathscr{P}'$ which contain m and $m+1$ components respectively and satisfy $V_{2m-1}(\mathscr{P}) = V_{2m-1}(\mathscr{P}')$. From our definition of $(\Omega, \mathcal{F})$ we know there exists $F \in \mathcal{F}$

such that F, F^C are nonempty. Let $x \in F$ and $x' \in F^C$. It follows that δ_x and $\delta_{x'}$ are different probability measures on $(\Omega, \mathcal{F})$. Let $\varepsilon_1, \ldots, \varepsilon_{2m+1}$ be distinct values in $[0,1]$. Applying Lemma IV.26 with $t = 2m+1$ and letting $\mu_i = \varepsilon_i \delta_x + (1-\varepsilon_i)\,\delta_{x'}$, there exists a permutation $\sigma : [2m+1] \to [2m+1]$ and $\beta_1, \ldots, \beta_{2m+1}$, with $\beta_i > 0$ for all i and $\sum_{i=1}^{m} \beta_i = \sum_{j=m+1}^{2m+1} \beta_j = 1$, such that

$$\sum_{i=1}^{m} \beta_i \mu_{\sigma(i)}^{\times 2m-1} = \sum_{j=m+1}^{2m+1} \beta_j \mu_{\sigma(j)}^{\times 2m-1}.$$

If we let $\mathscr{P} = \sum_{i=1}^{m} \beta_i \delta_{\mu_{\sigma(i)}}$ and $\mathscr{P}' = \sum_{j=m+1}^{2m+1} \beta_j \delta_{\mu_{\sigma(j)}}$, we have that $V_{2m-1}(\mathscr{P}) = V_{2m-1}(\mathscr{P}')$. □

To prove the remaining theorems we will need to make use of bounded linear operators on Hilbert spaces. Given a pair of Hilbert spaces H, H' we define $\mathcal{L}(H, H')$ as the space of *bounded linear operators* from H to H'. An operator, T, is in this space if there exists a nonnegative number C such that $\|Tx\|_{H'} \leq C \|x\|_H$ for all $x \in H$. The space of bounded linear operators is a Banach space when equipped with the norm

$$\|T\| \triangleq \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|}.$$

We will also need to employ Hilbert-Schmidt operators which are a subspace of the bounded linear operators.

Definition IV.27. Let H, H' be Hilbert spaces and $T \in \mathcal{L}(H, H')$. T is called a *Hilbert-Schmidt operator* if $\sum_{x \in J} \|Tx\|^2 < \infty$ for an orthonormal basis $J \subset H$. We denote the set of Hilbert-Schmidt operators in $\mathcal{L}(H, H')$ by $\mathscr{HS}(H, H')$.

This definition does not depend on the choice of orthonormal basis: the sum $\sum_{x \in J} \|T(x)\|^2$ will always yield the same value regardless of the choice of orthonormal basis J.

The following properties of Hilbert-Schmidt operators will not be used in the next proof, but they will be useful later. The set of Hilbert-Schmidt operators is itself a Hilbert space when equipped with the inner product

$$\sum_{x \in J} \langle Tx, Sx \rangle$$

where J is an orthonormal basis. The Hilbert-Schmidt norm will be denoted as $\|\cdot\|_{\mathscr{HS}}$ and the standard operator norm will have no subscript. There is a well known bound relating the two norms: for a Hilbert-Schmidt operator T we have that

$$\|T\| \leq \|T\|_{\mathscr{HS}} .$$

Proof of Theorem IV.13. Let $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ be a mixture of measures with linearly independent components. Let $\mathscr{P}' = \sum_{j=1}^{l} b_j \delta_{\nu_j}$ be a mixture of measures with $V_3(\mathscr{P}) = V_3(\mathscr{P}')$ and $l \leq m$. From Lemma IV.23 there exists a finite measure ξ and non-negative functions $p_1, \ldots, p_m, q_1, \ldots, q_l \in L^1(\Omega, \mathcal{F}, \xi) \cap L^2(\Omega, \mathcal{F}, \xi)$ such that, for all $B \in \mathcal{F}$, $\int_B p_i d\xi = \mu_i(B)$ and $\int_B q_j d\xi = \nu_j$ for all i, j. Using Lemma IV.22, IV.24 , and IV.25 as we did in the previous theorem proofs it follows that

$$\sum_{i=1}^{m} a_i p_i^{\times 2} = \sum_{j=1}^{l} b_j q_j^{\times 2}.$$

From Lemma IV.18 we have

$$\sum_{i=1}^{m} a_i p_i^{\otimes 2} = \sum_{j=1}^{l} b_j q_j^{\otimes 2}.$$

By Lemma IV.21 we now know that $l = m$ and $q_1, \ldots, q_m$ are linearly independent. We will now show that $q_j \in \operatorname{span}(\{p_1, \ldots, p_m\})$ for all j. Suppose that $q_t \notin \operatorname{span}(\{p_1, \ldots, p_m\})$. Then there exists $z \in L^2(\Omega, \mathcal{F}, \xi)$ such that $z \perp p_1, \ldots, p_m$

but $z \not\perp q_t$. Now we have

$$\begin{aligned}
\sum_{i=1}^{m} a_i p_i^{\otimes 2} &= \sum_{j=1}^{m} b_j q_j^{\otimes 2} \\
\Rightarrow \left\langle \sum_{i=1}^{m} a_i p_i \otimes p_i, z \otimes z \right\rangle &= \left\langle \sum_{j=1}^{m} b_j q_j \otimes q_j, z \otimes z \right\rangle \\
\Rightarrow \sum_{i=1}^{m} a_i \langle p_i \otimes p_i, z \otimes z \rangle &= \sum_{j=1}^{m} b_j \langle q_j \otimes q_j, z \otimes z \rangle \\
\Rightarrow \sum_{i=1}^{m} a_i \langle p_i, z \rangle^2 &= \sum_{j=1}^{m} b_j \langle q_j, z \rangle^2 .
\end{aligned}$$

We know that the LHS of the last equation is zero but the RHS is not, a contradiction.

We will find the following well known property of tensor products to be useful for continuing the proof (*Kadison and Ringrose* (1983) Proposition 2.6.9).

Lemma IV.28. *Let H, H' be Hilbert spaces. There exists a unitary operator $U : H \otimes H' \to \mathscr{HS}(H, H')$ such that, for any simple tensor $h \otimes h' \in H \otimes H'$, $U(h \otimes h') = \langle h, \cdot \rangle h'$.*

Because $p_1, \ldots, p_m$ are linearly independent we can do the following: for each $k \in [m]$ let $z_k \in \operatorname{span}(\{p_1, \ldots, p_m\})$ be such that $z_k \perp \{p_i : i \neq k\}$ and $\langle z_k, p_k \rangle = 1$. By considering elements of $L^2(\Omega, \mathcal{F}, \xi)^{\otimes 3}$ as elements of $L^2(\Omega, \mathcal{F}, \xi) \otimes L^2(\Omega, \mathcal{F}, \xi)^{\otimes 2}$, we can use Lemma IV.28 to transform elements in $L^2(\Omega, \mathcal{F}, \xi)^{\otimes 3}$ into elements of $\mathscr{HS}\left(L^2(\Omega, \mathcal{F}, \xi), L^2(\Omega, \mathcal{F}, \xi)^{\otimes 2}\right)$,

$$\begin{aligned}
\sum_{i=1}^{m} a_i p_i^{\otimes 3} &= \sum_{j=1}^{m} b_j q_j^{\otimes 3} \\
\Rightarrow \sum_{i=1}^{m} a_i p_i^{\otimes 2} \langle p_i, \cdot \rangle &= \sum_{j=1}^{m} b_j q_j^{\otimes 2} \langle q_j, \cdot \rangle .
\end{aligned}$$

It now follows that

$$\begin{aligned}\sum_{i=1}^{m} a_i p_i^{\otimes 2} \langle p_i, z_k \rangle &= \sum_{j=1}^{m} b_j q_j^{\otimes 2} \langle q_j, z_k \rangle \\ \Rightarrow a_k p_k^{\otimes 2} &= \sum_{j=1}^{m} b_j q_j^{\otimes 2} \langle q_j, z_k \rangle .\end{aligned}$$

Using Lemma IV.28 we have

$$a_k p_k \langle p_k, \cdot \rangle = \sum_{j=1}^{m} b_j \langle q_j, z_k \rangle q_j \langle q_j, \cdot \rangle . \tag{4.6}$$

The LHS of (4.6) is a rank one operator and thus the RHS must have exactly one nonzero summand, since $q_1, \ldots, q_m$ are linearly independent. Let $\varphi : [m] \to [m]$ be a function such that, for all k,

$$a_k p_k^{\otimes 2} = \left\langle q_{\varphi(k)}, z_k \right\rangle b_{\varphi(k)} q_{\varphi(k)}^{\otimes 2}.$$

From Lemma IV.24 we have

$$a_k \mu_k^{\times 2} = \left\langle q_{\varphi(k)}, z_k \right\rangle b_{\varphi(k)} \nu_{\varphi(k)}^{\times 2},$$

for all k. By Lemma IV.22 we have that $a_k \mu_k = \left\langle q_{\varphi(k)}, z_k \right\rangle b_{\varphi(k)} \nu_{\varphi(k)}$ for all k and thus $\mu_k = \nu_{\varphi(k)}$. Because $\mu_i \neq \mu_j$ for all i, j we have that φ must be a bijection. Let $\sigma = \varphi^{-1}$. By Lemma IV.22 we have that

$$\sum_{i=1}^{m} a_i \mu_i = \sum_{j=1}^{m} b_j \mu_{\sigma(j)}.$$

Since $\mu_1, \ldots, \mu_m$ are linearly independent the last equation only has one solution for

$b_1, \ldots, b_m$, which is $b_k = a_{\sigma(k)}$, for all k. Thus

$$\mathscr{P}' = \sum_{i=1}^{m} a_{\sigma(i)} \delta_{\mu_{\sigma(i)}}$$

which is equal to $\mathscr{P}$. □

Proof of Theorem IV.14. Let $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ be a mixture of measures with linearly independent components. We will proceed by contradiction: let $\mathscr{P}' = \sum_{j=1}^{l} b_j \delta_{\nu_j} \neq \mathscr{P}$ be a mixture of measures with $V_4(\mathscr{P}) = V_4(\mathscr{P}')$. From Theorem IV.5 we know that $\mathscr{P}$ is 3-identifiable. By Lemma IV.7 it follows that $\mathscr{P}$ is 4-identifiable and thus $l > m$. From Lemma IV.23 there exists a finite measure ξ and non-negative functions $p_1, \ldots, p_m, q_1, \ldots, q_l \in L^1(\Omega, \mathcal{F}, \xi) \cap L^2(\Omega, \mathcal{F}, \xi)$ such that, for all $B \in \mathcal{F}$, $\int_B p_i d\xi = \mu_i(B)$ and $\int_B q_j d\xi = \nu_j(B)$ for all i, j.

Proceeding as we did in the proof of Theorem IV.13 we have that

$$\sum_{i=1}^{m} a_i p_i^{\otimes 4} = \sum_{j=1}^{l} b_j q_j^{\otimes 4}.$$

Suppose that there exists k such that $\nu_k \notin \operatorname{span}(\{\mu_1, \ldots, \mu_m\})$. From this it would follow that there exists z such that $z \perp \{p_1, \ldots, p_m\}$ and $z \not\perp q_k$. Then we would have that

$$\begin{aligned} \left\langle \sum_{i=1}^{m} a_i p_i^{\otimes 4}, z^{\otimes 4} \right\rangle &= \left\langle \sum_{j=1}^{l} b_j q_j^{\otimes 4}, z^{\otimes 4} \right\rangle \\ \Rightarrow \sum_{i=1}^{m} a_i \langle p_i, z \rangle^4 &= \sum_{j=1}^{l} b_j \langle q_j, z \rangle^4, \end{aligned}$$

but the LHS of the last equation is 0 and the RHS is positive, a contradiction. Thus we have that $q_k \in \operatorname{span}(\{p_1, \ldots, p_m\})$ for all k.

Since $l > m$ and no pair of elements in $q_1, \ldots, q_m$ are collinear, there must a vector in $q_1, \ldots, q_l$ which is a nontrivial linear combination of $p_1, \ldots, p_m$. Without

loss of generality we will assume that $q_1 = \sum_{i=1}^m c_i p_i$ with c_1 and c_2 nonzero. By the linear independence of $p_1, \ldots, p_m$ there must exist vectors z_1, z_2 such that $\langle z_1, p_1 \rangle = 1$, $z_1 \perp \{p_i : i \neq 1\}$, $\langle z_2, p_2 \rangle = 1$, and $z_2 \perp \{p_i : i \neq 2\}$. Now consider

$$\begin{aligned} \left\langle \sum_{i=1}^m a_i p_i^{\otimes 4}, z_1^{\otimes 2} \otimes z_2^{\otimes 2} \right\rangle &= \left\langle \sum_{j=1}^l b_j q_j^{\otimes 4}, z_1^{\otimes 2} \otimes z_2^{\otimes 2} \right\rangle \\ \Rightarrow \sum_{i=1}^m a_i \langle p_i, z_1 \rangle^2 \langle p_i, z_2 \rangle^2 &= \sum_{j=1}^l b_j \langle q_j, z_1 \rangle^2 \langle q_j, z_2 \rangle^2 . \end{aligned}$$

The LHS of the last equation is 0 and the RHS is positive, a contradiction. □

Proof of Theorem IV.16. Let $\mathscr{P} = \sum_{i=1}^m a_i \delta_{\mu_i}$ be a mixture of measures with jointly irreducible components. Consider a mixture of measures $\mathscr{P}' = \sum_{j=1}^l b_j \delta_{\nu_j}$ with $V_2(\mathscr{P}) = V_2(\mathscr{P}')$. From Lemma IV.23 there exists a finite measure ξ and non-negative functions $p_1, \ldots, p_m, q_1, \ldots, q_l \in L^1(\Omega, \mathcal{F}, \xi) \cap L^2(\Omega, \mathcal{F}, \xi)$ such that, for all $B \in \mathcal{F}$, $\int_B p_i d\xi = \mu_i(B)$ and $\int_B q_j d\xi = \nu_j(B)$ for all i, j. From Lemmas IV.24 and IV.25 we have

$$\sum_{i=1}^m a_i p_i \times p_i = \sum_{j=1}^l b_j q_j \times q_j.$$

From Lemma IV.18 we have

$$\sum_{i=1}^m a_i p_i \otimes p_i = \sum_{j=1}^l b_j q_j \otimes q_j. \tag{4.7}$$

Suppose for a moment that $\mathscr{P}'$ contains a mixture component which does not lie in $\operatorname{span}(\{\mu_1, \ldots, \mu_m\})$. Without loss of generality we will assume that $\nu_1 \notin \operatorname{span}(\{\mu_1, \ldots, \mu_m\})$. Recall that joint irreducibility implies linear independence so $\nu_1, \mu_1, \ldots, \mu_m$ are a linearly independent set of measures and thus $q_1, p_1, \ldots, p_m$ are linearly independent. It follows that we can find some $z \in L^2(\Omega, \mathcal{F}, \xi)$ such that

$\langle z, q_1 \rangle \neq 0$ and $z \perp \{p_i : i \in [m]\}$ for all i. From (4.7) we have the following

$$\begin{aligned}
\left\langle \sum_{i=1}^{m} a_i p_i \otimes p_i, z \otimes z \right\rangle &= \left\langle \sum_{j=1}^{l} b_j q_j \otimes q_j, z \otimes z \right\rangle \\
\Rightarrow \sum_{i=1}^{m} a_i \langle p_i \otimes p_i, z \otimes z \rangle &= \sum_{j=1}^{l} b_j \langle q_j \otimes q_j, z \otimes z \rangle \\
\Rightarrow \sum_{i=1}^{m} a_i \langle p_i, z \rangle^2 &= \sum_{j=1}^{l} b_j \langle q_j, z \rangle^2 .
\end{aligned}$$

All the summands on both sides of the last equation are nonnegative. By our construction of z the LHS of the previous equation is zero and the first summand on the RHS is positive, a contradiction. Thus, each component in $\mathscr{P}'$ must lie in the span of the components of $\mathscr{P}$.

Now we have, for all j, $q_j = \sum_{i=1}^{m} c_i^j p_i$. From joint irreducibility we have that $c_i^j \geq 0$ for all i and j. Now suppose that there exists r, s, s' such that $c_s^r, c_{s'}^r > 0$. From the linear independence of $p_1, \ldots, p_m$ we can find a z such that $\langle p_s, z \rangle = 1$ and $z \perp \{p_q : q \in [m] \setminus \{s\}\}$. Applying Lemma IV.28 to (4.7) we have

$$\begin{aligned}
\sum_{i=1}^{m} a_i p_i \langle p_i, \cdot \rangle &= \sum_{j=1}^{l} b_j q_j \langle q_j, \cdot \rangle \\
\Rightarrow \sum_{i=1}^{m} a_i p_i \langle p_i, z \rangle &= \sum_{j=1}^{l} b_j q_j \langle q_j, z \rangle \\
\Rightarrow a_s p_s &= \sum_{j=1}^{l} b_j \left[\sum_{t=1}^{m} c_t^j p_t \right] \left\langle \sum_{u=1}^{m} c_u^j p_u, z \right\rangle \\
\Rightarrow a_s p_s &= \sum_{j=1}^{l} b_j \left[\sum_{t=1}^{m} c_t^j p_t \right] c_s^j \\
&= \sum_{t=1}^{m} \sum_{j=1}^{l} b_j c_t^j c_s^j p_t \\
&= \sum_{t=1}^{m} p_t \sum_{j=1}^{l} b_j c_t^j c_s^j .
\end{aligned}$$

Let $\alpha_t = \sum_{j=1}^{l} b_j c_t^j c_s^j$ for all t and note that each summand is nonnegative. Now we have

$$a_s p_s = \sum_{t=1}^{m} \alpha_t p_t.$$

We know that $\alpha_{s'} > 0$ since $b_r c_s^r c_{s'}^r > 0$. This violates the linear independence of $p_1, \ldots, p_m$. Now we have that for all i there exists j such that $p_i = q_j$. From the minimality of the representation of mixtures of measures it follows that $l = m$ and without loss of generality we can assert that $p_i = q_i$ for all i and thus $\mu_i = \nu_i$ for all i. Because $p_1, \ldots, p_m$ are linearly independent it follows that $p_1 \otimes p_1, \ldots p_m \otimes p_m$ are linearly independent. We can show this by the contrapositive, suppose $p_1 \otimes p_1, \ldots, p_m \otimes p_m$ are not linearly independent then there exists a nontrivial linear combination such that $\sum_{i=1}^{m} \kappa_i p_i \otimes p_i = 0$. Assume without loss of generality that $\kappa_1 \neq 0$. Applying Lemma IV.28 we get that

$$\begin{aligned} \sum_{i=1}^{m} \kappa_i p_i \langle p_i, \cdot \rangle &= 0 \\ \Rightarrow \sum_{i=1}^{m} \kappa_i p_i \langle p_i, p_1 \rangle &= 0 \\ \Rightarrow \kappa_1 p_1 \|p_1\|_{L^2}^2 + \sum_{i=2}^{m} \kappa_i p_i \langle p_i, p_1 \rangle &= 0 \end{aligned}$$

and thus $p_1, \ldots, p_m$ are not linearly independent.

Since $p_1 \otimes p_1, \ldots, p_m \otimes p_m$ are linearly independent it follows that $a_i = b_i$ for all i and thus $\mathscr{P} = \mathscr{P}'$.

□

4.5 Identifiability and Determinedness of Mixtures of Multinomial Distributions

Using the previous results we can show analogous identifiability and determinedness results for mixtures of multinomial distributions. The identifiability of mixtures of multinomial distributions was originally studied in *Kim* (1984) which contains a proof of Corollary IV.30 from this paper. An alternative proof of this corollary can be found in *Elmore and Wang* (2003). These results are analogous to identifiability results presented in this paper. Our proofs use techniques which are very different from those used in *Kim* (1984); *Elmore and Wang* (2003). These techniques can also be used to prove a determinedness style result, Corollary IV.31, which we have not seen addressed elsewhere in the multinomial mixture model literature.

Before our proof we must first introduce some definitions and notation. Any multinomial distribution is completely characterized by positive integers n and q and a probability vector in $\mathbb{R}^q$, $p = [p_1, \ldots, p_q]^T$. A multinomial random variable can be thought of as totalling the outcomes of repeated iid sampling from a categorical distribution. With this view the value q represents the number of possible outcomes of a trial, p is the likelihood of each outcome on a trial, and n is the number of trials. For whole numbers k, l we define $C_{k,l} = \left\{ x \in \mathbb{N}^{\times l} : \sum_{i=1}^{l} x_i = k \right\}$. These are vectors of the form $[x_1, \ldots, x_l]$ where $\sum_{i=1}^{l} x_i = k$. Using the values n and q above, the multinomial distribution is a probability measure over $C_{n,q}$. If Q is a multinomial distribution with parameters n, p, q as defined above then its probability mass function is

$$Q\left(\left\{[x_1, \ldots, x_q]^T\right\}\right) = \frac{n!}{x_1! \cdots x_q!} p_1^{x_1} \cdots p_q^{x_q}$$

for $x \in C_{n,q}$. We will denote this measure as $Q_{n,p,q}$. Let

$$\mathscr{M}(n,q) \triangleq \{Q_{n,p,q} : p \text{ is a probability vector in } \mathbb{R}^q\},$$

i.e. the space of all multinomial distributions with n and q fixed.

To show identifiability and determinedness of mixtures of multinomial distributions we will construct a linear operator $T_{n,q}$ from $\operatorname{span}\left(\mathcal{D}\left(C_{n,q}, 2^{C_{n,q}}\right)\right)$ to $\operatorname{span}\left(\mathcal{D}\left([q]^{\times n}, 2^{[q]^{\times n}}\right)\right)$ and use it to show that non-identifiable mixtures of multinomial distributions yield non-identifiable mixtures of measures, and likewise for non-determined mixtures of multinomial distributions.

Since $C_{n,q}$ is a finite set, the vector space of finite signed measures on $\left(C_{n,q}, 2^{C_{n,q}}\right)$ is a finite dimensional space and the set $\{\delta_x : x \in C_{n,q}\}$ is a basis for this space. Note that $\{\delta_x : x \in C_{n,q}\}$ is the set of all *point masses* on $C_{n,q}$, not vectors in the ambient space of $C_{n,q}$. Thus, to completely define the operator $T_{n,q}$, we need only define $T_{n,q}(\delta_x)$ for all $x \in C_{n,q}$. To this end let $x = C_{n,q}$. We define the function $F_{n,q} : C_{n,q} \to [q]^{\times n}$ as $F_{n,q}(x) = 1^{\times x_1} \times \cdots \times q^{\times x_q}$, where the exponents represent Cartesian powers. The definition of $F_{n,q}$ is a bit dense so we will do a simple example. Suppose $n = 6, q = 4$ and $x = [1,0,3,2]^T$ then $F_{n,q}(x) = [1,3,3,3,4,4]^T$. Let S_n be the symmetric group on n symbols. We define our linear operator as follows

$$T_{n,q}(\delta_x) = \frac{1}{n!} \sum_{\sigma \in S_n} \delta_{\sigma(F_{n,q}(x))},$$

where σ is permuting the entries of $F_{n,q}(x)$. This operator is similar to the projection operator onto the set of order n symmetric tensors *Comon et al.* (2008). The following lemma makes the crucial connection between the space of multinomial distributions and the probability measures of grouped samples.

Lemma IV.29. *Let $Q_{n,p,q} \in \mathscr{M}(n,q)$, then*

$$T_{n,q}(Q_{n,p,q}) = V_n\left(\delta_{\sum_{i=1}^{q} p_i \delta_i}\right).$$

Proof of Lemma IV.29. For brevity's sake let

$$Q = T_{n,q}(Q_{n,p,q})$$

and

$$R = V_n\left(\delta_{\sum_{i=1}^{q} p_i \delta_i}\right).$$

Let $y \in [q]^{\times n}$ be arbitrary. We will prove that $Q(\{y\}) = R(\{y\})$ which, since y is arbitrary, clearly generalizes to $Q = R$. From the definition of V_n we have that $R(\{y\}) = \left(\sum_{i=1}^{q} p_i \delta_i\right)^{\times n}(\{y\}) = \prod_{i=1}^{n} p_{y_i}$.

Let $\check{y} \in C_{n,q}$ be the element such that $\check{y}_i = |\{j : y_j = i\}|$ for all i, i.e. the ith index of $\check{y}$ contains the number of times the value i occurs in y. We define χ to be the indicator function, which is equal to 1 if its subscript is true and 0 otherwise. Consider some $z \neq \check{y}$. We have

$$\begin{aligned} T_{n,q}(\delta_z)(\{y\}) &= \frac{1}{n!}\sum_{\sigma \in S_n} \delta_{\sigma(F_{n,q}(z))}(\{y\}) \\ &= \frac{1}{n!}\sum_{\sigma \in S_n} \chi_{\sigma(F_{n,q}(z))=y}. \end{aligned}$$

From our definition of $F_{n,q}$ and $\check{y}$ it is clear that, there must exist some r such that the number of entries of $F_{n,q}(z)$ which equal r is different from the number of indices of y which equal r. Because of this no permutation of $F_{n,q}(z)$ can equal y and thus $T_{n,q}(\delta_z)(\{y\}) = 0$. From this it follows that $T_{n,q}(\delta_z)(\{y\}) = 0$ for all $z \neq \check{y}$.

Now we will consider $T_{n,q}\left(\delta_{\check{y}}\right)\left(\{y\}\right)$. Again we have

$$T_{n,q}\left(\delta_{\check{y}}\right)\left(\{y\}\right) = \frac{1}{n!}\sum_{\sigma\in S_n}\chi_{\sigma(F_{n,q}(\check{y}))=y},$$

so we need only determine how many permutations of $F_{n,q}\left(\check{y}\right)$ are equal to y. Basic combinatorics tells us that there are $\check{y}_1!\cdots\check{y}_q!$ such permutations. The coefficient of $\delta_{\check{y}}$ in $Q_{n,p,q}$ is $\frac{n!}{\check{y}_1!\cdots\check{y}_q!}p_1^{\check{y}_1}\cdots p_n^{\check{y}_n}$ so we have that $Q(\{y\}) = R(\{y\})$ by direct evaluation. □

This lemma allows us to make some assertions about the identifiability of mixtures of multinomial distributions.

In the following we will assume that all multinomial mixture models under consideration have only nonzero summands and distinct components. In the context of multinomial mixture models, a multinomial mixture model $\sum_{i=1}^{m} a_i Q_{n,p_i,q}$ is identifiable if it being equal to a different multinomial mixture model,

$$\sum_{i=1}^{m} a_i Q_{n,p_i,q} = \sum_{j=1}^{s} b_j Q_{n,r_j,q},$$

with $s \leq m$ implies that $s = m$ and there exists some permutation σ such that $a_i = b_{\sigma(i)}$ and $Q_{n,p_i,q} = Q_{n,r_{\sigma(i)},q}$ for all i. The mixture model is determined if the previous statement holds without the restriction $s \leq m$.

Multinomial mixture models are identifiable if the number of components m and the number of trials in each component n satisfy $n \geq 2m-1$.

Corollary IV.30. *Let $m \in \mathbb{N}_+$, $n \geq 2m-1$, and fix $q \in \mathbb{N}_+$. Let*

$$Q_{n,p_1,q},\ldots,Q_{n,p_m,q},Q_{n,r_1,q},\ldots,Q_{n,r_s,q} \in \mathscr{M}\left(n,q\right)$$

with $Q_{n,p_1,q}, \ldots, Q_{n,p_m,q}$ *distinct,* $Q_{n,r_1,q}, \ldots, Q_{n,r_s,q}$ *distinct, and* $s \leq m$*. If*

$$\sum_{i=1}^{m} a_i Q_{n,p_i,q} = \sum_{j=1}^{s} b_j Q_{n,r_j,q}$$

with $a_i > 0, b_i > 0$ *for all* i *and* $\sum_{i=1}^{m} a_i = \sum_{j=1}^{s} b_j = 1$*, then* $s = m$ *and there exists some permutation* σ *such that* $a_i = b_{\sigma(i)}$ *and* $p_i = r_{\sigma(i)}$*.*

Alternatively this corollary says that, given two different finite mixtures with components in $\mathscr{M}(n, q)$, one mixture with m components and the other with s components, if $n \geq 2m - 1$ and $n \geq 2s - 1$ then the mixtures induce different measures.

Proof of Corollary IV.30. We will proceed by contradiction and assume that there exists two mixtures of the form above,

$$\sum_{i=1}^{m} a_i Q_{n,p_i,q} = \sum_{j=1}^{s} b_j Q_{n,r_j,q}$$

but $s \neq m$ or $s = m$ and there exists no permutation such that $a_i Q_{n,p_i,q} = b_{\sigma(i)} Q_{n,r_{\sigma(i)},q}$. If we apply $T_{n,q}$ defined earlier, from Lemma IV.29 it follows that

$$V_n\left(\sum_{i=1}^{m} a_i \delta_{\sum_{k=1}^{q} p_{i,k}\delta_k}\right) = V_n\left(\sum_{j=1}^{s} b_j \delta_{\sum_{l=1}^{q} r_{j,l}\delta_l}\right).$$

We have that $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\sum_{k=1}^{q} p_{i,k}\delta_k}$ and $\mathscr{P}' = \sum_{j=1}^{s} b_j \delta_{\sum_{l=1}^{q} r_{j,l}\delta_l}$ are mixtures of measures which are not n-identifiable. Our contradiction hypothesis implies that $\mathscr{P} \neq \mathscr{P}'$. From Lemma IV.8 we have that

$$V_{2m-1}\left(\sum_{i=1}^{m} a_i \delta_{\sum_{k=1}^{q} p_{i,k}\delta_k}\right) = V_{2m-1}\left(\sum_{j=1}^{s} b_j \delta_{\sum_{l=1}^{q} r_{j,l}\delta_l}\right),$$

which contradicts Theorem IV.5. □

Additionally multinomial mixture models are determined if the number of com-

ponents m and the number of trials in each component n satisfy $n \geq 2m$.

Corollary IV.31. *Let $n \geq 2m$ and fix $q \in \mathbb{N}$. Let $Q_{n,p_1,q}, \ldots, Q_{n,p_m,q}$ and $Q_{n,r_1,q}, \ldots, Q_{n,r_s,q}$ be elements of $\mathscr{M}(n,q)$ with $Q_{n,p_1,q}, \ldots, Q_{n,p_m,q}$ distinct and $Q_{n,r_1,q}, \ldots, Q_{n,r_s,q}$ distinct. If*

$$\sum_{i=1}^{m} a_i Q_{n,p_i,q} = \sum_{j=1}^{s} b_j Q_{n,r_j,q}$$

with $a_i > 0, b_i > 0$ for all i and $\sum_{i=1}^{m} a_i = \sum_{j=1}^{m} b_i = 1$, then $m = s$ and there exists some permutation σ such that $a_i = b_{\sigma(i)}$ and $p_i = r_{\sigma(i)}$.

The proof is almost identical to the proof of Corollary IV.30, so we will omit it. Using these proof techniques one could establish additional identifiability/determinedness style results for multinomial mixture models along the lines of Theorems IV.13, IV.14, and IV.16. Furthermore it seems likely that one could use the algorithms described in the next section or from *Anandkumar et al.* (2014); *Arora et al.* (2012); *Rabani et al.* (2014) to recover these components, using the transform $T_{n,q}$.

4.6 Meta-Algorithms

Here we will present a few algorithms for the recovery of mixture components and proportions from data. The algorithms are quite general and can be applied to any measurable space. Unfortunately, due to the generality of the proposed algorithms, some of the implementation details are setting specific which makes in-depth theoretical analysis difficult. As one concrete illustration, we will show consistency for categorical measures.

Let $\sum_{i=1}^{m} w_i \delta_{\mu_i}$ be an arbitrary mixture of measures on some measurable space $(\Omega, \mathcal{F})$, which we are interested in recovering. Let $p_1, \ldots, p_m$ be square integrable

densities with respect to a dominating measure ξ, with $\int_A p_i d\xi = \mu_i(A)$ for all $i \in [m]$ and $A \in \mathcal{F}$. A measure ξ and densities $p_1, \ldots, p_m$ satisfying these properties are guaranteed to exist as a consequence of Lemma IV.23.

We will initially consider the situation where we have $2m$ samples per random group and have access to the tensors $\sum_{i=1}^m w_i p_i^{\otimes 2m}$ and $\sum_{i=1}^m w_i p_i^{\otimes 2m-2}$. In a finite discrete space, estimating these tensors is equivalent to estimating moment tensors of order $2m$ and $2m-2$. For measures over $\mathbb{R}^d$ dominated by the Lebesgue measure, one could estimate these tensors using a kernel density estimator in $\mathbb{R}^{d(2m)}$ and $\mathbb{R}^{d(2m-2)}$ using each sample group as a kernel center. We will also assume that $p_1, \ldots, p_m$ have distinct norms. We will need to introduce tensor products of bounded linear operators. The following lemma is exactly proposition 2.6.12 from *Kadison and Ringrose* (1983).

Lemma IV.32. *Let $H_1, \ldots, H_n, H_1', \ldots, H_n'$ be Hilbert spaces and let $U_i \in \mathcal{L}(H_i, H_i')$ for all $i \in [n]$. There exists a unique*

$$U \in \mathcal{L}(H_1 \otimes \cdots \otimes H_n, H_1' \otimes \cdots \otimes H_n'),$$

such that $U(h_1 \otimes \cdots \otimes h_n) = U_1(h_1) \otimes \cdots \otimes U_n(h_n)$ for all $h_1 \in H_1, \ldots, h_n \in H_n$.

Definition IV.33. The operator constructed in Lemma IV.32 is called the *tensor product of $U_1, \ldots, U_n$* and is denoted $U_1 \otimes \cdots \otimes U_n$.

The following equality is mentioned in *Kadison and Ringrose* (1983).

Lemma IV.34. *Let $U_1, \ldots, U_n$ be defined as in Lemma IV.32. Then*

$$\|U_1 \otimes \cdots \otimes U_n\| = \|U_1\| \|U_2\| \cdots \|U_n\|.$$

Before we introduce the meta-algorithms we will discuss an important point regarding computational implementation and Lemmas IV.28 and IV.32. For the remainder of this paragraph we will assume that Euclidean spaces are equipped with

the standard inner product. Vectors in a space of tensor products of Euclidean space, for example $\mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_s}$ are easily represented on computers as elements of $\mathbb{R}^{d_1 \times \cdots \times d_s}$ *Comon et al.* (2008). Linear operators from some Euclidean tensor space to another can also be easily represented. Furthermore the transformation in Lemma IV.28 and the construction of new operators from Lemma IV.32 can be implemented in computers by "unfolding" the tensors into matrices, applying common linear algebraic manipulations and "folding" them back into tensors. The inner workings of these manipulations are beyond the scope of this paper and we refer the reader to *Golub and Van Loan* (1996) for details. Practically speaking this means the manipulations mentioned in Lemmas IV.28 and IV.32 are straightforward to implement with a bit of tensor programming knowhow. Implementation may also be streamlined by using programming libraries that assist with these tensor manipulations such as the NumPy library for Python.

Because of the points mentioned in the previous paragraph the following algorithms are readily implementable for estimating categorical distributions, where the measures can be represented as probability vectors on a Euclidean space. We will go into this point in more detail later. Similarly, we expect that these techniques could be extended to probability densities on Euclidean space using kernel density estimators with a kernel function that can be evaluated in closed form (although implementation may be significantly more involved).

To begin our analysis we will apply the transform from Lemma IV.28 to get the operator

$$C = \sum_{i=1}^{m} w_i p_i^{\otimes m-1} \left\langle p_i^{\otimes m-1}, \cdot \right\rangle = \sum_{i=1}^{m} \sqrt{w_i} p_i^{\otimes m-1} \left\langle \sqrt{w_i} p_i^{\otimes m-1}, \cdot \right\rangle .$$

Here C is a positive semi-definite (PSD) operator in $\mathcal{L}\left(L^2\left(\Omega, \mathcal{F}, \xi\right)^{\otimes m-1}\right)$. Let $C^\dagger$ be the (Moore-Penrose) pseudoinverse of C and $W = \sqrt{C^\dagger}$. Now W is an operator

that whitens $\sqrt{w_1}p_1^{\otimes m-1}, \ldots, \sqrt{w_m}p_m^{\otimes m-1}$. That is, $W\sqrt{w_1}p_1^{\otimes m-1}, \ldots, W\sqrt{w_m}p_m^{\otimes m-1}$ are orthonormal vectors. Using the operator construction from Lemma IV.32 we can construct $I \otimes W \otimes I \otimes W$ where, for all simple tensors in $L^2(\Omega, \mathcal{F}, \xi)^{\otimes 2m}$ we have,

$$\begin{aligned}
&(I \otimes W \otimes I \otimes W)(x_1 \otimes \cdots \otimes x_{2m}) \\
&\quad = x_1 \otimes W(x_2 \otimes \cdots \otimes x_m) \otimes x_{m+1} \otimes W(x_{m+1} \otimes \cdots \otimes x_{2m}).
\end{aligned}$$

Applying $I \otimes W \otimes I \otimes W$ to $\sum_{i=1}^{2m} w_i p_i^{\otimes 2m}$ yields

$$\sum_{i=1}^{m} w_i p_i \otimes W p_i^{\otimes m-1} \otimes p_i \otimes W p_i^{\otimes m-1},$$

which can again be represented as a PSD operator

$$\begin{aligned}
S &\triangleq \sum_{i=1}^{m} w_i p_i \otimes W p_i^{\otimes m-1} \left\langle p_i \otimes W p_i^{\otimes m-1}, \cdot \right\rangle \\
&= \sum_{i=1}^{m} p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1} \left\langle p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1}, \cdot \right\rangle.
\end{aligned}$$

For $i \neq j$ it follows that $p_i \otimes \sqrt{w_i} W p_i^{\otimes m-1} \perp p_j \otimes W\sqrt{w_j} p_j^{\otimes m-1}$. To see this

$$\begin{aligned}
&\left\langle p_i \otimes W\sqrt{w_i} p_i^{\otimes m-1}, p_j \otimes W\sqrt{w_j} p_j^{\otimes m-1} \right\rangle \\
&\qquad = \langle p_i, p_j \rangle \left\langle W\sqrt{w_i} p_i^{\otimes m-1}, W\sqrt{w_j} p_j^{\otimes m-1} \right\rangle \\
&\qquad = \langle p_i, p_j \rangle 0 \\
&\qquad = 0.
\end{aligned}$$

Also note that

$$\begin{aligned}\left\|p_i \otimes W\sqrt{w_i}p_i^{\otimes m-1}\right\|^2 &= \left\langle p_i \otimes W\sqrt{w_i}p_i^{\otimes m-1}, p_i \otimes W\sqrt{w_i}p_i^{\otimes m-1}\right\rangle \\ &= \langle p_i, p_i\rangle \left\langle W\sqrt{w_i}p_i^{\otimes m-1}, W\sqrt{w_i}p_i^{\otimes m-1}\right\rangle \\ &= \|p_i\|^2 .\end{aligned}$$

If $p_1, \ldots, p_m$ have distinct norms then it follows that

$$\sum_{i=1}^{m} p_i \otimes W\sqrt{w_i}p_i^{\otimes m-1} \left\langle p_i \otimes W\sqrt{w_i}p_i^{\otimes m-1}, \cdot\right\rangle$$

is the unique spectral decomposition of S since the vectors $p_1 \otimes W\sqrt{w_1}p_1^{\otimes m-1}, \ldots, p_m \otimes W\sqrt{w_m}p_m^{\otimes m-1}$ are orthogonal, have distinct norms, and thus distinct positive eigenvalues. Given an eigenvector of S, $p_i \otimes W\sqrt{w_i}p_i^{\otimes m-1}$, we need only view it as a linear operator $p_i \left\langle W\sqrt{w_i}p_i^{\otimes m-1}, \cdot\right\rangle$ and apply this operator to some vector z which is not orthogonal to $W\sqrt{w_i}p_i^{\otimes m-1}$, thus yielding p_i scaled by $\left\langle W\sqrt{w_i}p_i^{\otimes m-1}, z\right\rangle$.

Were the norms of $p_1, \ldots, p_m$ not distinct, then there would not be a spectral gap between some of the eigenvalues in S, and a spectral decomposition of S may contain some eigenvectors that are not $p_1 \otimes W\sqrt{w_1}p_1^{\otimes m-1}, \ldots, p_m \otimes W\sqrt{w_m}p_m^{\otimes m-1}$, but are instead linear combinations of these vectors.

Once the mixture components $p_1, \ldots, p_m$ are recovered form the spectral decomposition we can estimate the mixture proportions. From these mixture components we can construct the tensors $p_1^{\otimes 2m-2}, \ldots, p_m^{\otimes 2m-2}$. These tensors are linearly independent by Lemma IV.20. The tensor $\sum_{i=1}^{m} w_i p_i^{\otimes 2m-2}$ is known. By the linear independence of the components there is exactly one solution for $a_1, \ldots, a_m$ in the equation

$$\sum_{i=1}^{m} w_i p_i^{\otimes 2m-2} = \sum_{j=1}^{m} a_j p_j^{\otimes 2m-2},$$

so simply minimizing $\left\|\sum_{i=1}^m w_i p_i^{\otimes 2m-2} - \sum_{j=1}^m a_j p_j^{\otimes 2m-2}\right\|$ over $a_1, \ldots, a_m$ will give us the mixture proportions. We could also use a different tensor power $\left\|\sum_{i=1}^m w_i p_i^{\otimes r} - \sum_{j=1}^m a_j p_j^{\otimes r}\right\|$, so long as $r \geq m-1$ to guarantee independence of the components.

We can construct a similar algorithm with 4 samples per group when the mixture components are known to be linearly independent. The details of this algorithm are in Appendix B.2. In such a setting it would be advisable to use the algorithms from *Anandkumar et al.* (2014); *Song et al.* (2014) since they better studied. We mention our algorithm for purely theoretical interest. There are likely a multitude of possible algorithms for the recovery of mixture components whose necessary group size depends on the geometry of the mixture components.

Taking inspiration from *Anandkumar et al.* (2014) and *Song et al.* (2014) we can suggest yet another algorithm. The previous papers demonstrate algorithms for recovering mixture components which are measures on discrete spaces and $\mathbb{R}^d$, from random groups of size 3, provided the mixture components are linearly independent. Given a mixture of measures $\mathscr{P} = \sum_{i=1}^m w_i \delta_{\mu_i}$ with density functions $p_1, \ldots, p_m$, the tensors $p_1^{\otimes m-1}, \ldots, p_m^{\otimes m-1}$ are linearly independent. Thus, with $3m-3$ samples per random group, we can estimate the tensors $\sum_{i=1}^m w_i p_i^{\otimes 3m-3}$ and we can use the algorithms from the previous papers to recover $p_1^{\otimes m-1}, \ldots, p_m^{\otimes m-1}$ from which it is straightforward to recover $p_1, \ldots, p_m$.

We can also recover the components with $2m-1$ samples per group. We will adopt the same setting as in our first algorithm, but with $2m-1$ samples per group in stead of $2m$. Let W be as before. Using Lemma IV.32 we can construct the operator $I \otimes W \otimes W$ on the space $L^2(\Omega, \mathcal{F}, \xi)^{\otimes 2m-1}$ which maps simple tensors in the following way: $(I \otimes W \otimes W)(x_1 \otimes \cdots \otimes x_{2m-1}) = x_1 \otimes W(x_2 \otimes \cdots \otimes x_m) \otimes W(x_{m+1} \otimes \cdots \otimes x_{2m-1})$. Applying this operator to $\sum_{i=1}^m w_i p_i^{\otimes 2m-1}$ gives us the tensor

$$
\begin{aligned}
A &\triangleq \sum_{i=1}^{m} w_i p_i \otimes W\left(p_i^{\otimes m-1}\right) \otimes W\left(p_i^{\otimes m-1}\right) \\
&= \sum_{i=1}^{m} p_i \otimes W\left(\sqrt{w_i} p_i^{\otimes m-1}\right) \otimes W\left(\sqrt{w_i} p_i^{\otimes m-1}\right).
\end{aligned}
$$

From Lemma IV.28 we can transform the tensor A into the operator T,

$$
T = \sum_{i=1}^{m} p_i \otimes W\left(\sqrt{w_i} p_i^{\otimes m-1}\right) \left\langle W\left(\sqrt{w_i} p_i^{\otimes m-1}\right), \cdot \right\rangle. \tag{4.8}
$$

Now the operator TT^H is

$$
\begin{aligned}
TT^H =& \sum_{i=1}^{m} p_i \otimes W\left(\sqrt{w_i} p_i^{\otimes m-1}\right) \left\langle W\left(\sqrt{w_i} p_i^{\otimes m-1}\right), \cdots \right. \\
& \left. \sum_{j=1}^{m} W\left(\sqrt{w_j} p_j^{\otimes m-1}\right) \left\langle p_j \otimes W\left(\sqrt{w_j} p_j^{\otimes m-1}\right), \cdot \right\rangle \right\rangle \\
=& \sum_{i=1}^{m} p_i \otimes W\left(\sqrt{w_i} p_i^{\otimes m-1}\right) \left\langle p_i \otimes W\left(\sqrt{w_i} p_i^{\otimes m-1}\right), \cdot \right\rangle
\end{aligned}
$$

which is simply the operator S from the previous section. The last step is justified since the vectors $W\left(\sqrt{w_1} p_1^{\otimes m-1}\right), \ldots, W\left(\sqrt{w_m} p_m^{\otimes m-1}\right)$ are orthonormal. This tensor is precisely the tensor from which we recovered the mixture components in the first algorithm.

4.6.1 Spreading the eigenvalue gaps for categorical distributions

Here we will introduce a trick to guarantee that the norms of the mixture component distributions are distinct. Let $\left(\Omega, 2^{\Omega}\right)$ be a finite discrete measurable space with $\Omega = \{\omega_1, \ldots, \omega_d\}$. Let $\mu_1, \ldots, \mu_m$ be distinct measures on this space. Let $y_1, \ldots, y_d \overset{iid}{\sim} \operatorname{unif}(1, 2)$ and let ξ be a random measure on $\left(\Omega, 2^{\Omega}\right)$ defined by $\xi\left(\{\omega_i\}\right) = y_i$ for all i. Clearly ξ dominates all $\mu_1, \ldots, \mu_m$ and thus we can define

Radon-Nikodym derivatives $p_i = \frac{d\mu_i}{d\xi}$ for all i. We will treat these Radon-Nikodym derivatives as being elements in $L^2\left(\Omega, 2^{\Omega}, \xi\right)$. We have the following lemma

Lemma IV.35. *With probability one*

$$\int p_i(\omega)^2 d\xi(\omega) \neq \int p_j(\omega)^2 d\xi(\omega)$$

for all $i \neq j$.

Proof. Observe that, for all i, j,

$$\int\limits_{\{\omega_j\}} p_i d\xi = p_i(\omega_j)\xi\left(\{\omega_j\}\right) = p_i(\omega_j)y_j = \mu_i\left(\{\omega_j\}\right)$$

and thus $p_i\left(\omega_j\right) = \frac{\mu_i(\{w_j\})}{y_j}$. We will show that $\|p_1\|^2_{\ell^2\left(\mathbb{R}^d\right)} \neq \|p_2\|^2_{\ell^2\left(\mathbb{R}^d\right)}$ with probability one, which implies $\|p_i\|_{\ell^2\left(\mathbb{R}^d\right)} \neq \|p_j\|_{\ell^2\left(\mathbb{R}^d\right)}$ for all $i \neq j$ with probability one (here and for the rest of the paper $\|\cdot\|_{\ell^2\left(\mathbb{R}^d\right)}$ will denote the standard Euclidean norm on $\mathbb{R}^d$ and $\langle\cdot,\cdot\rangle_{\ell^2\left(\mathbb{R}^d\right)}$ the standard inner product).

Because $\mu_1 \neq \mu_2$ it follows that there exists some j such that $\mu_1\left(\{\omega_j\}\right) \neq \mu_2\left(\{\omega_j\}\right)$. Without loss of generality we will assume that $j = 1$ in the previous statement. Now we have

$$\begin{aligned}
& P\left(\int p_1\left(\omega\right)^2 d\xi\left(\omega\right) = \int p_2\left(\omega\right)^2 d\xi\left(\omega\right)\right) \\
= \; & P\left(\sum_{i=1}^{d} \frac{\mu_1\left(\{\omega_i\}\right)^2}{y_i} = \sum_{j=1}^{d} \frac{\mu_2\left(\{\omega_j\}\right)^2}{y_j}\right) \\
= \; & P\left(\left(\frac{\mu_1\left(\{\omega_1\}\right)^2}{y_1} - \frac{\mu_2\left(\{\omega_1\}^2\right)}{y_1}\right) = \left(\sum_{i=2}^{d} \frac{\mu_1\left(\{\omega_i\}\right)^2}{y_i} - \sum_{j=2}^{d} \frac{\mu_2\left(\{\omega_j\}\right)^2}{y_j}\right)\right)
\end{aligned}$$

which is clearly zero since $\left(\mu_1\left(\{\omega_1\}\right)\right)^2 - \left(\mu_2\left(\{\omega_1\}\right)\right)^2 \neq 0$ and $y_1, \ldots, y_d$ are all independent random variables and from a non-atomic measure. □

Applying the previous trick with the recovery algorithm for groups of size $2m-1$ we have an algorithm for recovering mixtures on finite measure spaces with m components. The paper *Rabani et al.* (2014) recovers the mixture components given a setting almost identical to ours, but we feel that our algorithm is more straightforward and easily extended to non-discrete spaces.

4.6.2 Recovery Algorithm For Discrete Spaces

Let $\left(\Omega, 2^{\Omega}\right)$ be a finite measurable space with $|\Omega| = d$. To simplify exposition we will assume that Ω is simply the set of d dimensional indicator vectors in $\mathbb{R}^d$, $e_1, \ldots, e_d$. Note that Euclidean space with the standard inner product is $L^2\left(\Omega, 2^{\Omega}, \sum_{i=1}^{d} \delta_{e_i}\right) = \ell^2\left(\mathbb{R}^d\right)$. Let $\mu_1, \ldots, \mu_m$ be distinct probability measures on Ω. Let $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ be a mixture of measures. Let $\tilde{p}_i \triangleq \mathbb{E}_{x \sim \mu_i}[x]$ for all i. Note that $\tilde{p}_{i,j} = \mu_i(\{e_j\})$ for all i, j. Let $X_1, X_2, \ldots \overset{iid}{\sim} V_{2m-1}(\mathscr{P})$ with $X_i = [X_{i,1}, \ldots, X_{i,2m-1}]$.

To begin we construct the random dominating measure described in Section 4.6.1. Let $y_1, \ldots, y_d \overset{iid}{\sim} \text{unif}(1,2)$. The random dominating measure ξ is defined by $\xi(\{e_i\}) = y_i$ for all i. Let $p_i = \frac{d\mu_i}{d\xi}$, i.e. $p_i(e_j) = \frac{\tilde{p}_{i,j}}{y_j}$ for all i and j. There is a bit of a computational issue with this representation for the densities $p_1, \ldots, p_m$ since the new dominating measure changes the inner product from the standard inner product. We can remedy this with the following lemma.

Lemma IV.36. *Let $x, v \in \ell^2\left(\mathbb{R}^d\right)$, ξ be as above, and*

$$B = \begin{bmatrix} \frac{1}{\sqrt{y_1}} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{y_2}} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{\sqrt{y_d}} \end{bmatrix}.$$

Then $\langle Bx, Bv \rangle_{L^2(\Omega, 2^{\Omega}, \xi)} = \langle x, v \rangle_{\ell^2\left(\mathbb{R}^d\right)}$.

Proof of Lemma IV.36. We have

$$\begin{aligned}
\langle Bx, Bv\rangle_{L^2(\Omega,2^\Omega,\xi)} &= \int (Bx)(i)(Bv)(i)d\xi(i) \\
&= \sum_{i=1}^{d}(Bx)(i)(Bv)(i)y_i \\
&= \sum_{i=1}^{d}\frac{x(i)}{\sqrt{y_i}}\frac{v(i)}{\sqrt{y_i}}y_i \\
&= \sum_{i=1}^{d}x(i)y(i) \\
&= \langle x, y\rangle_{\ell^2\left(\mathbb{R}^d\right)}.
\end{aligned}$$

□

From this lemma we have that B, when considered as an operator in $\mathcal{L}\left(\ell^2\left(\mathbb{R}^d\right), L^2\left(\Omega, 2^\Omega, \xi\right)\right)$, is a unitary transform. We are interested in estimating the tensor $\sum_{i=1}^{m} w_i p_i^{\otimes 2m-1}$, but in order to keep the algorithm operating in standard Euclidean space we will instead transform it into $\ell^2\left(\mathbb{R}^d\right)$. To this end consider an arbitrary i. We have

$$\begin{aligned}
B^{-1}p_i &= B^{-1}\left[p_{i,1}, \ldots, p_{i,d}\right]^T \\
&= B^{-1}\left[\frac{\tilde{p}_{i,1}}{y_1}, \ldots, \frac{\tilde{p}_{i,d}}{y_d}\right]^T \\
&= \left[\frac{\tilde{p}_{i,1}}{\sqrt{y_1}}, \ldots, \frac{\tilde{p}_{i,d}}{\sqrt{y_d}}\right]^T,
\end{aligned}$$

and thus $B^{-1}p_j = B\tilde{p}_j$ for all j.

We will use the following lemma to find the expected value of

$$\mathbb{E}\left[BX_{i,1} \otimes \cdots \otimes BX_{i,2m-1}\right]$$

Lemma IV.37. *Let $n > 1$ and $Z_1, \ldots, Z_n$ be independent random vectors in*

$\mathbb{R}^{d_1}, \ldots, \mathbb{R}^{d_n}$ *such that* $\mathbb{E}[Z_i]$ *exists for all* i. *Then* $\mathbb{E}[Z_1 \otimes \cdots \otimes Z_n] = \mathbb{E}[Z_1] \otimes \cdots \otimes \mathbb{E}[Z_n]$.

Proof of Lemma IV.37. Let $[i_1, \ldots, i_n] \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$ be arbitrary. We have that

$$\begin{aligned} \mathbb{E}[Z_1 \otimes \cdots \otimes Z_n]_{i_1,\ldots,i_n} &= \mathbb{E}[Z_{1,i_1} \cdots Z_{n,i_n}] \\ &= \mathbb{E}[Z_{1,i_1}] \cdots \mathbb{E}[Z_{n,i_n}]. \end{aligned}$$

Since $i_1, \ldots, i_n$ were arbitrary it implies that all entries of $\mathbb{E}[Z_1 \otimes \cdots \otimes Z_n]$ and $\mathbb{E}[Z_1] \otimes \cdots \otimes \mathbb{E}[Z_n]$ are equal. □

Recall that $X_{i,1}, \ldots, X_{i,2m-1} \overset{iid}{\sim} \mu$ with $\mu \sim \mathscr{P}$. From the previous lemma and the definition of $\tilde{p}_i$ it follows that

$$\begin{aligned} &\mathbb{E}[BX_{i,1} \otimes \cdots \otimes BX_{i,2m-1}] \\ &= \mathbb{E}_{\mu\sim\mathscr{P}}[\mathbb{E}[BX_{i,1} \otimes \cdots \otimes BX_{i,2m-1}|\,\mu]] \\ &= \mathbb{E}_{\mu\sim\mathscr{P}}[\mathbb{E}[BX_{i,1}|\mu] \otimes \cdots \otimes \mathbb{E}[BX_{i,2m-1}|\mu]] \\ &= \mathbb{E}_{\mu\sim\mathscr{P}}[B\mathbb{E}[X_{i,1}|\mu] \otimes \cdots \otimes B\mathbb{E}[X_{i,2m-1}|\mu]] \\ &= \sum_{i=1}^{m} w_i B\mathbb{E}[X_{i,1}|\mu = \mu_i] \otimes \cdots \otimes B\mathbb{E}[X_{i,2m-1}|\mu = \mu_i] \\ &= \sum_{i=1}^{m} w_i (B\tilde{p}_i)^{\otimes 2m-1}. \end{aligned}$$

Let $Y_{i,j} = BX_{i,j}$. Now we will construct the whitening operator. To do this first construct the operator

$$\begin{aligned} \widehat{C} = \frac{1}{(2m-1)!}\frac{1}{n}\sum_{i=1}^{n}\sum_{\sigma\in S_{2m-1}} & Y_{i,\sigma(1)} \otimes \cdots \otimes Y_{i,\sigma(m-1)} \\ & \left\langle Y_{i,\sigma(m)} \otimes \cdots \otimes Y_{i,\sigma(2m-2)}, \cdot \right\rangle. \end{aligned}$$

There are some repeated terms in the previous summation, which is not an issue.

Instead we could have set $\widehat{C}$ to be equal to

$$\frac{1}{(2m-2)!}\frac{1}{n}\sum_{i=1}^{n}\sum_{\sigma\in S_{2m-2}} Y_{i,\sigma(1)}\otimes\cdots\otimes Y_{i,\sigma(m-1)}\left\langle Y_{i,\sigma(m)}\otimes\cdots\otimes Y_{i,\sigma(2m-2)},\cdot\right\rangle,$$

but this would not utilize all the data, specifically $Y_{1,2m-1},\ldots,Y_{n,2m-1}$. In the second operator the average over S_{2m-2} functions as a projection onto the space of symmetric tensors and the summation over S_{2m-1} in the definition of $\widehat{C}$ serves a similar purpose. Viewed alternatively, the distribution of $[Y_{i,1},\ldots,Y_{i,2m-1}]^T$ does not change if we reorder the entries of the vector, so the summation is considering all possible orderings of random groups. This symmetrization conveniently assures that $\widehat{C}$ is a Hermitian operator. This $\widehat{C}$ is estimating the C mentioned in the meta-algorithm. Let $\lambda_{\widehat{C},1},\ldots,\lambda_{\widehat{C},m}$ be the top m eigenvalues of $\widehat{C}$ and $v_{\widehat{C},1},\ldots,v_{\widehat{C},m}$ be their associated eigenvectors. We can now construct the whitening operator

$$\widehat{W}=\sum_{i=1}^{m}\lambda_{\widehat{C},i}^{-\frac{1}{2}}v_{\widehat{C},i}\left\langle v_{\widehat{C},i},\cdot\right\rangle.$$

Now construct the tensor

$$\widehat{A}=\frac{1}{(2m-1)!}\frac{1}{n}\sum_{i=1}^{n}\sum_{\sigma\in S_{2m-1}} Y_{i,\sigma(1)}\otimes\widehat{W}\left(Y_{i,\sigma(2)}\otimes\cdots\otimes Y_{i,\sigma(m)}\right)\otimes\cdots \widehat{W}\left(Y_{i,\sigma(m+1)}\otimes\cdots\otimes Y_{i,\sigma(2m-1)}\right).$$

Using simple unfolding techniques we can transform $\widehat{A}$ in to the operator $\widehat{T}$:

$$\widehat{T}=\frac{1}{(2m-1)!}\frac{1}{n}\sum_{i=1}^{n}\sum_{\sigma\in S_{2m-1}} Y_{i,\sigma(1)}\otimes\widehat{W}\left(Y_{i,\sigma(2)}\otimes\cdots\otimes Y_{i,\sigma(m)}\right)\cdots \left\langle\widehat{W}\left(Y_{i,\sigma(m+1)}\otimes\cdots\otimes Y_{i,\sigma(2m-1)}\right),\cdot\right\rangle,$$

as well as its Hermitian, $\widehat{T}^H$:

$$\frac{1}{(2m-1)!}\frac{1}{n}\sum_{i=1}^{n}\sum_{\sigma\in S_{2m-1}}\widehat{W}\left(Y_{i,\sigma(m+1)}\otimes\cdots\otimes Y_{i,\sigma(2m-1)}\right)\cdots \\ \left\langle Y_{i,\sigma(1)}\otimes\widehat{W}\left(Y_{i,\sigma(2)}\otimes\cdots\otimes Y_{i,\sigma(m)}\right),\cdot\right\rangle.$$

Let $v_1,\ldots,v_m$ be the top m eigenvectors of $\widehat{T}\widehat{T}^H$ (4.8), which will be elements of $\ell^2\left(\mathbb{R}^d\right)^{\otimes m}$. These vectors are estimates of $\|B\tilde{p}_1\|_2^{-1}B\tilde{p}_1\otimes\widehat{W}\sqrt{w_1}\left(B\tilde{p}_1\right)^{\otimes m-1},\ldots,\|B\tilde{p}_m\|_2^{-1}B\tilde{p}_m\otimes\widehat{W}\sqrt{w_m}\left(B\tilde{p}_m\right)^{\otimes m-1}$ (possibly multiplied by -1). The factors in front of the tensors normalize the tensors to have norm 1.

Using a transform of the form in Lemma IV.28, we can implement a transform

$$U:\ell^2\left(\mathbb{R}^d\right)^{\otimes m}\to\mathscr{HS}\left(\ell^2\left(\mathbb{R}^d\right)^{\otimes m-1},\ell^2\left(\mathbb{R}^d\right)\right)$$

which maps simple tensors $x_1\otimes\cdots\otimes x_m$ to $x_1\left\langle x_2\otimes\cdots\otimes x_m,\cdot\right\rangle$. Applying this transform to $v_1,\ldots,v_m$ yields estimates of $\|B\tilde{p}_i\|_{\ell^2\left(\mathbb{R}^d\right)}^{-1}B\tilde{p}_i\left\langle\widehat{W}\sqrt{w_i}\left(B\tilde{p}_i\right)^{\otimes m-1},\cdot\right\rangle$, for all i. At this point one simply needs to find vectors $q_1,\ldots,q_m$ which are not orthogonal to $\widehat{W}\sqrt{w_1}\left(B\tilde{p}_1\right)^{\otimes m-1},\ldots,\widehat{W}\sqrt{w_m}\left(B\tilde{p}_m\right)^{\otimes m-1}$ to get $\|B\tilde{p}_i\|_{\ell^2\left(\mathbb{R}^d\right)}^{-1}B\tilde{p}_i\left\langle\widehat{W}\sqrt{w_i}\left(B\tilde{p}_i\right)^{\otimes m-1},q_i\right\rangle$, which is $B\tilde{p}_i,\ldots,B\tilde{p}_i$ up to scaling. Such vectors can be found by simply using a tensor populated by iid standard normal random variables. After this we can recover $\tilde{p}_1,\ldots,\tilde{p}_m$, up to scaling, by simply applying B^{-1}, which we would then want to normalize to sum to one. Alternatively we could take the largest left singular vector of these operators. We will call these estimates $\widehat{p}_1,\ldots,\widehat{p}_m$.

Using the data we can estimate the tensor $\sum_{i=1}^m w_i\tilde{p}_i^{\otimes m-1}$ with the estimator

$$\widehat{E}=\frac{1}{2m-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{\sigma\in S_{2m-1}}X_{i,\sigma(1)}\otimes\cdots\otimes X_{i,\sigma(m-1)}$$

To estimate the mixture proportions we find the value of $\alpha=(\alpha_1,\ldots,\alpha_m)$ which

minimizes

$$\left\| \widehat{E} - \sum_{i=1}^{m} \alpha_i \widehat{p_i}^{\otimes m-1} \right\|.$$

4.6.3 Consistency of Recovery Algorithm

We will now show that the recovery algorithm for categorical distributions is consistent. Let $C, \widehat{C}, T, \widehat{T}, W$, and $\widehat{W}$ be as they were defined in the first part of this section. The crux of our algorithm is the recovery of the eigenvectors of TT^H, from which we then recover the mixture components through the application of linear and continuous transforms to the eigenvectors. In order to simplify the notation in our explanation we will assume that the norms of $\tilde{p}_1, \ldots, \tilde{p}_m$ are distinct. We do this so that there are gaps in the spectral decomposition of TT^H thus making the random dominating measure trick unnecessary. Were this not the case, we could simply represent the probability vectors as densities with respect to some dominating measure which makes their norms distinct, as we did in the previous section. Because of this assumption we can simply set B to be the identity operator. From this we have that $p_i = \tilde{p}_i$ for all i and $X_{i,j} = Y_{i,j}$ for all i and j. The following theorem demonstrates that the algorithm does indeed recover the eigenvectors of TT^H.

Theorem IV.38. *With T and $\widehat{T}$ defined as above, as $n \to \infty$ then*

$$\left\| TT^H - \widehat{T}\widehat{T}^H \right\|_{\mathscr{HS}} \xrightarrow{p} 0.$$

Proof of Theorem IV.38. Let

$$Q = \sum_{i=1}^{m} w_i p_i^{\otimes 2m-1}$$

and

$$\widehat{Q} = \frac{1}{(2m-1)!}\frac{1}{n}\sum_{i=1}^{n}\sum_{\sigma\in S_{2m-1}} X_{i,\sigma(1)}\otimes\cdots\otimes X_{i,\sigma(2m-1)}.$$

Note that

$$\left(I\otimes W\otimes W\right)(Q) \;=\; \sum_{i=1}^{m} w_i p_i \otimes W\left(p_i^{\otimes m-1}\right)\otimes W\left(p_i^{\otimes m-1}\right)$$

and

$$\begin{aligned}
&\left(I\otimes\widehat{W}\otimes\widehat{W}\right)(\widehat{Q})\\
&= \frac{1}{(2m-1)!}\frac{1}{n}\sum_{i=1}^{n}\sum_{\sigma\in S_{2m-1}} X_{i,\sigma(1)}\otimes\widehat{W}\left(X_{i,\sigma(2)}\otimes\cdots\otimes X_{i,\sigma(m)}\right)\otimes\cdots\\
&\quad \widehat{W}\left(X_{i,\sigma(m+1)}\otimes\cdots\otimes X_{i,\sigma(2m-1)}\right).
\end{aligned}$$

Since the transform in Lemma IV.28 is unitary, we have that

$$\left\|T-\widehat{T}\right\|_{\mathscr{HS}} = \left\|\left(I\otimes W\otimes W\right)(Q) - \left(I\otimes\widehat{W}\otimes\widehat{W}\right)\left(\widehat{Q}\right)\right\|_{\ell^2\left(\mathbb{R}^d\right)^{\otimes 2m-1}}.$$

We will now show that $\left\|T-\widehat{T}\right\| \xrightarrow{p} 0$.

$$\begin{aligned}
\left\|T-\widehat{T}\right\| &\leq \left\|T-\widehat{T}\right\|_{\mathscr{HS}}\\
&= \left\|(I\otimes W\otimes W)(Q) - \left(I\otimes\widehat{W}\otimes\widehat{W}\right)\left(\widehat{Q}\right)\right\|_{\ell^2\left(\mathbb{R}^d\right)^{\otimes 2m-1}}\\
&\leq \left\|(I\otimes W\otimes W)(Q) - (I\otimes W\otimes W)(\widehat{Q})\right\|_{\ell^2\left(\mathbb{R}^d\right)^{\otimes 2m-1}}\\
&\quad + \left\|(I\otimes W\otimes W)(\widehat{Q}) - \left(I\otimes\widehat{W}\otimes\widehat{W}\right)\left(\widehat{Q}\right)\right\|_{\ell^2\left(\mathbb{R}^d\right)^{\otimes 2m-1}}\\
&\leq \|I\otimes W\otimes W\|\left\|Q-\widehat{Q}\right\|_{\ell^2\left(\mathbb{R}^d\right)^{\otimes 2m-1}}\\
&\quad + \left\|I\otimes W\otimes W - I\otimes\widehat{W}\otimes\widehat{W}\right\|\left\|\widehat{Q}\right\|_{\ell^2\left(\mathbb{R}^d\right)^{\otimes 2m-1}}.
\end{aligned}$$

We have that $\mathbb{E}\left[\widehat{Q}\right] = Q$ so the first summand goes to zero in probability by the law of large numbers. All we need to show is that $\left\|I \otimes W \otimes W - I \otimes \widehat{W} \otimes \widehat{W}\right\| \xrightarrow{p} 0$.

From Lemma IV.34 we have that

$$\begin{aligned}
\left\|I \otimes W \otimes W - I \otimes \widehat{W} \otimes \widehat{W}\right\| &\leq \|I\| \left\|W \otimes W - \widehat{W} \otimes \widehat{W}\right\| \\
&= \left\|W \otimes W - \widehat{W} \otimes \widehat{W}\right\| \\
&\leq \left\|W \otimes W - W \otimes \widehat{W}\right\| + \cdots \\
&\qquad \left\|W \otimes \widehat{W} - \widehat{W} \otimes \widehat{W}\right\| \\
&= \|W\| \left\|W - \widehat{W}\right\| + \left\|\widehat{W}\right\| \left\|W - \widehat{W}\right\| \\
&= \left(\|W\| + \left\|\widehat{W}\right\|\right) \left\|W - \widehat{W}\right\|.
\end{aligned}$$

The left factor converges in probability to $2\|W\|$ and the right factor converges to 0 in probability and so we have that $\left\|T - \widehat{T}\right\| \xrightarrow{p} 0$. From this we also have that $\left\|\widehat{T}\widehat{T}^H - TT^H\right\| \xrightarrow{p} 0$. □

As demonstrated earlier in this section the mixture components are recovered by applying a composition of linear and continuous operators to the eigenvectors of TT^H, thus consistent estimation of the eigenvectors of TT^H gives us consistent estimation of the mixture components.

4.6.4 Experiments

Here we will present some experimental results of our algorithm applied to a simple synthetic dataset. The sample space for the experiments is $\Omega = \{0, 1, 2\}$. The mixture components of our dataset are μ_1, μ_2, μ_3 with μ_1 distributed according to a binomial distribution with $n = 2$ and $p = 0.2$, μ_2 is similar with $p = 0.8$ and $\mu_3 = \frac{1}{3}\mu_1 + \frac{2}{3}\mu_2$. The component weights are $w_1 = 0.5, w_2 = 0.3, w_3 = 0.2$. So our mixture of measures is $\mathscr{P} = \sum_{i=1}^{3} w_i \delta_{\mu_i}$. Our samples come from $V_5(\mathscr{P})$, and we will apply the algorithm

described previously with a random dominating measure and using the left singular value estimator rather than the application of a random vector.

When considered as vectors in $\mathbb{R}^3$, μ_1 and μ_2 have the same norm. The mixture components are also not linearly independent. We will construct our own performance measure which measures the recovery of all the components jointly. Let $\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\mu}_3$ be the three estimates of the mixture components from some algorithm. We will view these estimates as vectors in $\mathbb{R}^3$. Our metric is $\min_{\sigma \in S_3} \frac{1}{3} \sum_{i=1}^{3} \left\| \mu_i - \widehat{\mu}_{\sigma(i)} \right\|_{\ell^1(\mathbb{R}^3)}$. That is, we take the sum of total variations of the best matching of the estimated mixture components to the true components.

4.6.5 Proposed Algorithm Experiments

We include two different implementations of our proposed algorithm with two different sample sizes. For our first experiment we generate the dominating measure from the square of Gaussian random variables with mean 0 and standard deviation 0.03 This experiment was performed with a sample size of 50,000 random groups. We used the Gaussian random variables instead of a uniform distribution for the random dominating measure because the Gaussian random measure performed better. For our second experiment we fix our dominating measure ξ as $\xi(\{0\}) = 3^2, \xi(\{1\}) = 2^2$ and $\xi(\{2\}) = 1$ with 50,000 random groups. For our third and fourth experiments repeated the previous two experiments but increased the number of random groups to 10,000,000. The purpose of the last two experiments is to demonstrate that a well chosen dominating measure can significantly affect performance. For our proposed algorithm we repeat the experiment 20 times and report relevant statistics. We make one additional adjustment to the algorithm described earlier. If the estimator yields a component which has a negative entry, we simply set the negative entry to zero and renormalise.

Table 4.1: Experimental Results

Method	Performance
Random Dominating Measure, 50,000 samples	Mean:0.1407, Variance:0.0169
Fixed Dominating Measure, 50,000 samples	Mean:0.0524, Variance:0.0011
Random Dominating Measure, 10,000,000 samples	Mean:0.0433, Variance:0.0062
Fixed Dominating Measure, 10,000,000 samples	Mean:0.0037, Variance: $4e{-}6$
Randomly Selected Measures	Mean:0.5323, Variance:0.0203
Anandkumar, et al. *Anandkumar et al.* (2014)	0.3214 or 0.1758

4.6.6 Competing Algorithms

We compare our algorithm to the algorithm from *Anandkumar et al.* (2014) as well as simply choosing 3 measures uniformly at random from the probabilistic simplex. The randomly selected components algorithm was repeated 1000 times. The algorithm in *Anandkumar et al.* (2014) is executed on the true population measures. Note that this algorithm is not intended to be used on mixtures of measures with linearly dependent components.

4.6.7 Results

The Results are summarized in Table 1. As expected the algorithm from *Anandkumar et al.* (2014) is not capable of recovering components which are not linearly independent. We chose the initial vector for tensor power iteration in *Anandkumar et al.* (2014) randomly and the performance of this algorithm seems to depend on the choice of initial vector.

CHAPTER V

Future Work, Discussion, and Conclusion

This chapter contains possible directions for future research related to the results presented in this thesis and concluding remarks.

5.1 Robust Kernel Density Estimator Consistency

In this work we have shown that the limit of the RKDE, as $n \to \infty$ and $\sigma \to 0$, is the distribution f. Therefore the robustness of the RKDE is not manifested in its asymptotic limit, at least for the class of strictly convex losses we study. Rather, the robustness of the RKDE is manifested for finite sample sizes as demonstrated by *Kim and Scott* (2012).

A key feature of our work is our nonstandard analysis. Standard analysis proceeds by the decomposition, $\|f - f_\sigma^n\|_1 \leq \|f - f_\sigma\|_1 + \|f_\sigma - f_\sigma^n\|_1$, where f_σ is the minimizer of J_σ (defined in Eqn. (2.2)). Using proof techniques from *Kim and Scott* (2012) it is easy to show that there exists a pdf, p_σ, satisfying

$$f_\sigma = \int p_\sigma(x)\, \Phi_\sigma(x)\, dx$$

and

$$p_\sigma(x) = \frac{\varphi\left(\|\Phi_\sigma(x) - f_\sigma\|_{\mathcal{H}_\sigma}\right) f(x)}{\int \varphi\left(\|\Phi_\sigma(y) - f_\sigma\|_{\mathcal{H}_\sigma}\right) f(y)\, dy}.$$

In the case of the classic KDE, φ is a constant so $p_\sigma = f$. For a robust loss however, φ is a non-constant function so p_σ does not have a closed form expression. The fact that f_σ and f_σ^n do not have closed form expressions makes the standard analysis difficult.

The function R_σ^n is of some interest of its own. It is mentioned in *Kim and Scott* (2012) that the IRWLS algorithm converges to the RKDE after very few iterations. This phenomenon may be explained by the small contraction constant exhibited by R_σ^n in Lemma II.9. It is also worth noting that the density estimator generated by applying the IRWLS algorithm a fixed number of times is also consistent. More precisely, let $f_\sigma^{n,k} = R_\sigma^n(\cdots R_\sigma^n(R_\sigma^n(0))\cdots)$, where R_σ^n is applied k times, then, given the same consistency requirements for the RKDE, $\left\|f_\sigma^{n,k} - f\right\|_1 \xrightarrow{p} 0$.

The last line of the proof for Theorem II.5 allows us to say something about the RKDE rate of convergence. From the proof, if $n\sigma^d \to \infty$, there exists $C > 0$ such that, with probability going to one, $\left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_{\mathcal{H}_\sigma} \leq C\sigma^{d/2}$. Letting $\sigma^{d/2} = \frac{\log(n)}{\sqrt{n}}$ gives us $\left\|\bar{f}_\sigma^n - f_\sigma^n\right\|_{\mathcal{H}_\sigma} \frac{\sqrt{n}}{\log(n)} \leq C$, a rate of convergence of the RKDE to the KDE. We anticipate that this result can be extended to L^1 convergence of the RKDE to f and will be a focus of future work.

We also note that just as f_σ^n is a robust version of $\bar{f}_\sigma^n$ so is f_σ a robust version of $\bar{f}_\sigma$. To see this consider the expression for p_σ. For the traditional KDE φ is a constant, yielding $p_\sigma = f$. When using a robust loss φ is a decreasing function causing $p_\sigma(x)$ to be smaller for more outlying x. We can consider p_σ to be a robust version of f since it suppresses low density regions of f.

The primary thrust of future work will focus on extending this result to nonconvex functions. In *Kim and Scott* (2012) it was demonstrated that the RKDE performed

well when using a Hampel loss. Though optimization in a nonconvex setting is typically difficult to analyze, a proof using the fixed point techniques from this work seem to be a promising tool for tackling such a challenge.

5.2 Scale and Project Kernel Density Estimator

Previous works have demonstrated that adaptive kernels can significantly improve the performance of the KDE (*Terrell and Scott*, 1992; *Liu et al.*, 2007; *Mahapatruni and Gray*, 2011). Considering that the SPKDE already outperforms standard KDE, it would be interesting to see if coupling the SPKDE with some adaptive kernel technique could yield performance superior to both the SPKDE and adaptive kernel methods.

There are a few interesting theoretical questions regarding the SPKDE. Most obviously one would want to know why the performance of the SPKDE is generally superior to that of the standard KDE, even with no contaminating data. In the case of no contamination the SPKDE does not converge to the true density, which seems to imply that the bias of the SPKDE is somehow superior to that of the standard KDE. A better understanding of this could lead to even better nonparametric density estimators. A second question would be that of the sample complexity. How does the scaling and projection affect the convergence of the density estimator?

Algorithmically it would be desirable the accelerate the optimization for finding the kernel weights, but since the optimization is a simple quadratic program this avenue for research is subsumed by standard optimization theory.

5.3 An Operator Theoretic Approach to Nonparametric Mixture Models

This chapter in particular presents many possibilities for future work.

5.3.1 Future Work Related to the Recovery Algorithm

We feel that there is significant room left for improving our proposed algorithm. Though we do not include these experiments, we observed a phenomena that having a large separation between the norms of the components significantly improves the ability for the algorithm to recover the mixture components. As the experiments demonstrate, choosing a good dominating measure which separates the norms can improve performance. An avenue for possible improvement is intelligent selection of a dominating measure. One possible disadvantage of choosing the dominating measure with iid random variables is that a sort of central limit type of effect occurs which draws the norms together. Perhaps there is some way to select the dominating measure from the data which will improve performance.

A second improvement may come from better estimates of the C and T operators in the algorithm. Principally, estimating these depends on good estimates of symmetric tensors which represent categorical distributions. It has been shown that the estimation of discrete distributions can be improved by not simply using the frequencies of each occurrence of each category (*Lehmann and Casella*, 2003; *Valiant and Valiant*, 2016; *Orlitsky and Suresh*, 2015; *Kamath et al.*, 2015; *Han et al.*, 2014; *Paninski*, 2005). It seems possible that leveraging the techniques used for estimating categorical distributions with the structure of symmetric tensors can yield improved estimates of the symmetric tensors we use and thus improve the performance of the algorithm.

Additionally it would be desirable to find some sample complexity bounds and convergence rates for our recovery algorithms.

5.3.2 Additional Identifiability Results

While we have derived several core identifiability and determinedness there are still many possibilities for other such results. For example, can the techniques presented

here be extended to identifiability results for mixture models which are not "finite" mixture models? What happens if a mixture of measures $\mathscr{P}$ has an infinite number of components or is non-atomic? Additionally it would be interesting to see if we can derive similar results for hidden Markov models, which are essentially the stochastic stochastic version of finite mixture models.

Returning to the realm of finite nonparametric mixture models, there are a couple of questions worth investigating. One of these is the notion of "identifiable subspaces." Given a mixture of measures $\mathscr{P}$ and access to $V_n(\mathscr{P})$, is it possible that some mixture components are identifiable while others are not? We can pose a similar question. Given some mixture of measures, $\mathscr{P}$, what mixture components can we add to $\mathscr{P}$ so that these new components are identifiable? What does this subspace of identifiable components look like? Given data, can we hope to recover components in these identifiable subspaces and know with certainty that we are indeed recovering a true mixture component? Finally we would like to completely characterize the n-identifiability and n-determinedness of a mixture of measures based on the geometry of its components.

5.3.3 Potential Statistical Test and Estimator

The results on determinedness suggest the possibility of a goodness of fit test. Suppose we have grouped samples from some mixture of measures $\mathscr{P}' = \sum_{i=1}^{m'} w_i' \delta_{\mu_i'}$. Further suppose some null hypothesis

$$H_0 : \mathscr{P}' = \mathscr{P} \triangleq \sum_{i=1}^{m} w_i \delta_{\mu_i}.$$

We may be able to reject the null hypothesis provided we have $2m$ samples per group if we have some way of consistently estimating $M \triangleq \sum_{i=1}^{m} w_i \mu_i^{\times 2m}$ from the groups of samples. We will call such an estimator $\widehat{M}$. If $\widehat{M}$ does not converge to M then we

can reject the null hypothesis.

One interesting observation from the proof of Theorem IV.9 is that, if $\mathscr{P} = \sum_{i=1}^{m} w_i \delta_{\mu_i}$ is a mixture of measures, p_i is a pdf for μ_i for all i, and $n > m$, then the rank of $\sum_{i=1}^{m} a_i p_i^{\otimes n} \otimes p_i^{\otimes n}$ will be exactly m. This suggests a statistical estimator for the number of mixture components. The form of this tensor is amenable to spectral methods since it is a positive semi-definite tensor of order 2, which is akin to a positive semi-definite matrix. Embedding the data with the kernel mean mapping, using a universal kernel *Micchelli et al.* (2006), seems like a promising approach to constructing such a test or estimator.

5.3.4 Identifiability and the Value $2n - 1$

The value $2n - 1$ seems to carry some significance for identifiability beyond the setting we proposed. This value can also be found in results concerning metrics on trees *Pachter and Speyer* (2004), hidden Markov models *Paz* (1971), and frame theory, with applications to signal processing *Balan et al.* (2006). All of these results are related to identifiability of an object or the injectivity of an operator. We can offer no further insight as to why this value recurs, but it appears to be an algebraic phenomenon.

5.4 Conclusion

This work has presented results concerning two extensions of the question of non-parametric density estimation. The first extension was adapting kernel density estimation to be robust to contamination and outliers. To this end we demonstrated the asymptotic behavior of a proposed robust kernel density estimator, with optimal rate on bandwidth. We also proposed a new robust kernel density estimator, the SPKDE. The asymptotic behavior of this estimator was analysed, and was shown to converge to a transformed version of the sample distribution. Provided certain assumptions

on contaminating data, this transform will converge to the uncontaminated distribution. This estimator was also shown to perform well experimentally, oftentimes outperforming the standard KDE, even with no contamination.

The second extension was concerned with nonparametric mixture modelling. In this setting we had access to groups of samples which were known to come from the same density. Using measure theoretic techniques we could embed this setting into a Hilbert space and apply functional theoretic techniques. We demonstrated several tight bounds for identifiability and determinedness as well as a highly general algorithm, with a proof of concept experiment, for recovering the densities. These techniques relied heavily on the Hilbert space embedding and demonstrated the power of this technique.

APPENDICES

APPENDIX A

Chapter III Additional Proofs and Experimental Results

A.1 Proofs

Proof of Lemma III.2 and III.5. We will prove these lemmas simultaneously. The f in lemmas III.2 and III.5 are the same and all notation is consistent between the two lemmas. First we will show that $\|g_{\alpha,\beta}\|_{L^1}$ is continuous in α. Let $\{a_i\}_1^\infty$ be a non-negative sequence in $\mathbb{R}$ converging to arbitrary $a \geq 0$. Since $g_{a_i,\beta}$ is dominated by βf and $g_{a_i,\beta}$ converges to $g_{a,\beta}$ pointwise, by the dominated convergence theorem we know $\|g_{a_i,\beta}\|_{L^1} \to \|g_{a,\beta}\|_{L^1}$, thus proving the continuity of $\|g_{\alpha,\beta}\|_{L^1}$. Since $\|g_{0,\beta}\|_{L^1} = \beta > 1$ and $\|g_{\alpha,\beta}\|_{L^1} \to 0$ as $\alpha \to \infty$, by the intermediate value theorem there exists α' such that $\|g_{\alpha',\beta}\|_{L^1} = 1$. This proves the existence part of Lemma III.2. Let $\widetilde{f}_\beta = g_{\alpha',\beta}$. Clearly $\mathcal{D}$ is convex so the closure (in L^2) $\bar{\mathcal{D}}$ is also convex. Since $\bar{\mathcal{D}}$ is a closed and convex set in a Hilbert space, $\arg\min_{g\in\bar{\mathcal{D}}} \|g - \beta f\|_{L^2}$ admits a unique minimizer. Note that $\widetilde{f}_\beta$ being the unique minimizer is equivalent to showing that, for all c in $\bar{\mathcal{D}}$

(Theorem 3.14 in *Bauschke and Combettes* (2011))

$$\left\langle c - \widetilde{f}_\beta, \beta f - \widetilde{f}_\beta \right\rangle \leq 0.$$

Because this is continuous over the c term and $\mathcal{D}$ is dense in $\bar{\mathcal{D}}$ we need only show that the inequality holds over all $c \in \mathcal{D}$. To this end, note that for all x,

$$\beta f(x) - \max\{0, \beta f(x) - \alpha'\} \leq \alpha'$$

and that if $\widetilde{f}_\beta(x) > 0$ then

$$\widetilde{f}_\beta(x) = \beta f(x) - \alpha'.$$

From this we have

$$\begin{aligned}
&\left\langle c - \widetilde{f}_\beta, \beta f - \widetilde{f}_\beta \right\rangle \\
&= \left\langle c, \beta f - \widetilde{f}_\beta \right\rangle - \left\langle \widetilde{f}_\beta, \beta f - \widetilde{f}_\beta \right\rangle \\
&= \int c(x) \left(\beta f(x) - \widetilde{f}_\beta(x)\right) dx \\
&\quad - \int \widetilde{f}_\beta(x) \left(\beta f(x) - \widetilde{f}_\beta(x)\right) dx \\
&\leq \int c(x) \alpha' dx \\
&\quad - \int \widetilde{f}_\beta(x) \left(\beta f(x) - (\beta f(x) - \alpha')\right) dx \\
&= \alpha' - \alpha' \\
&= 0.
\end{aligned}$$

From this we get that $\widetilde{f}_\beta$ is the unique minimizer. If there existed $\alpha'' \neq \alpha'$ such that $g_{\alpha'',\beta}$ was also a pdf, then there would be two minimizers of $\arg\min_{g \in \bar{\mathcal{D}}} \|g - \beta f\|_{L^2}$, which is impossible since the minimizer is unique, thus proving the uniqueness of

α'. $\square$

Proof of Proposition III.4. In this proof we will be working with a hypothetical f_{tar} and f_{con} in $\mathcal{D}$. Define "Assumption B" to be that there exists two sets $S \subset \text{supp}(f_{tar})$ and $T \subset \mathbb{R}^d$, which have nonzero Lebesgue measure, such that $f_{con}(T) > f_{con}(S)$. We will now show that Assumption A not holding is equivalent to Assumption B.

A⇒not B: Let $S \subset \text{supp}(f_{tar})$ and $T \subset \mathbb{R}^d$ both have nonzero Lebesgue measure. From Assumption A we know for Lebesgue almost all $s \in S$ that $f_{con}(s) = u$, for some u and $f_{con}(T) \leq u$ Lebesgue almost everywhere.

not A⇒B: If Assumption A is not satisfied either f_{con} is not almost Lebesgue everywhere uniform over $\text{supp}(f_{tar})$ or f_{con} is Lebesgue almost everywhere uniform on $\text{supp}\,(f_{tar})$ with value u but there exists some set $Q \subset \mathbb{R}^d$ of nonzero Lebesgue measure such that $f_{con}\,(Q) > u$. Both of these situations clearly imply Assumption B.

This proves that the negation of Assumption A is Assumption B.

Let f_{con} and f_{tar} satisfy Assumption B and $\varepsilon \in (0,1)$ be arbitrary. By Lemma III.2 we know there exists a unique α such that $\max\left\{ \frac{1}{1-\varepsilon}\left((1-\varepsilon)f_{tar}(\cdot) + \varepsilon f_{con}\right) - \alpha, 0 \right\}$ is a pdf. First we will show that $\alpha < \text{ess sup}_x \frac{\varepsilon}{1-\varepsilon} f_{con}(x)$. If $\text{ess sup}_x \frac{\varepsilon}{1-\varepsilon} f_{con}(x) = \infty$ then clearly $\alpha < \text{ess sup}_x \frac{\varepsilon}{1-\varepsilon} f_{con}(x)$. Let $r = \text{ess sup}_x \frac{\varepsilon}{1-\varepsilon} f_{con}(x) < \infty$. Let $S, T \subset \mathbb{R}^d$ satisfy the properties in the definition of Assumption B. Observe that

$$\begin{aligned}
&\int \max\left\{ \frac{1}{1-\varepsilon}\left((1-\varepsilon)f_{tar}(x) + \varepsilon f_{con}(x)\right) - r, 0 \right\} dx \\
&= \int \max\left\{ f_{tar}(x) + \frac{\varepsilon}{1-\varepsilon} f_{con}(x) - r, 0 \right\} dx \\
&= \int_S \max\left\{ f_{tar}(x) + \frac{\varepsilon}{1-\varepsilon} f_{con}(x) - r, 0 \right\} dx \ldots \\
&\quad + \int_{S^C} \max\left\{ f_{tar}(x) + \frac{\varepsilon}{1-\varepsilon} f_{con}(x) - r, 0 \right\} dx.
\end{aligned}$$

Note that on the set S we have that $\max\left\{f_{tar}(x) + \frac{\varepsilon}{1-\varepsilon} f_{con}(x) - r, 0\right\} < f_{tar}$. Now we have

$$\begin{aligned}
&\int_S \max\left\{f_{tar}(x) + \frac{\varepsilon}{1-\varepsilon} f_{con}(x) - r, 0\right\} dx \ldots \\
&\qquad + \int_{S^C} \max\left\{f_{tar}(x) + \frac{\varepsilon}{1-\varepsilon} f_{con}(x) - r, 0\right\} dx \\
&< \int_S f_{tar}(x)dx + \int_{S^C} f_{tar}(x)dx \\
&< 1
\end{aligned}$$

and thus $\alpha < r$ (i.e. the cutoff value for $R_\varepsilon^A(f_{obs})$ is lower than the essential supremum of f_{con}). Because $\alpha < \operatorname{ess\,sup}_x \frac{\varepsilon}{1-\varepsilon} f_{con}(x)$, on the set for which $\frac{\varepsilon}{1-\varepsilon} f_{con}(\cdot) > \alpha$ (which has nonzero Lebesgue measure) we have that $\max\left\{f_{tar}(\cdot) + \frac{\varepsilon}{1-\varepsilon} f_{con} - \alpha, 0\right\} > f_{tar}$, so $\max\left\{f_{tar}(\cdot) + \frac{\varepsilon}{1-\varepsilon} f_{con} - \alpha, 0\right\} \neq f_{tar}$. □

Proof of Theorem III.6. Given a set $S \subset L^2\left(\mathbb{R}^d\right)$ let P_S be the projection operator onto S. Consider the following decomposition

$$\begin{aligned}
\left\|f_{\sigma,\beta}^n - f_\beta'\right\|_{L^2} &= \left\|P_{\mathcal{D}_\sigma^n} \beta \bar{f}_\sigma^n - P_{\bar{\mathcal{D}}} \beta f\right\|_{L^2} \\
&\leq \left\|P_{\mathcal{D}_\sigma^n} \beta \bar{f}_\sigma^n - P_{\mathcal{D}_\sigma^n} \beta f\right\|_{L^2} + \left\|P_{\mathcal{D}_\sigma^n} \beta f - P_{\bar{\mathcal{D}}} \beta f\right\|_{L^2}
\end{aligned}$$

Note that we are projecting onto $\bar{\mathcal{D}}$ rather than $\mathcal{D}$ does not matter as was shown in the proof of Lemma III.2 and III.5. Furthermore note that $f_\beta' = P_{\bar{\mathcal{D}}} \beta f$. The projection operator onto a closed convex set is Lipschitz continuous with constant 1 (Proposition 4.8 in *Bauschke and Combettes* (2011)) so the first term goes to zero by standard KDE consistency (which we prove later). Convergence of the second term is a bit more involved. First we will show that $\left\|P_{\mathcal{D}_\sigma^n} \beta f - \beta f\right\|_{L^2} \xrightarrow{p} \left\|P_{\bar{\mathcal{D}}} \beta f - \beta f\right\|_{L^2}$,

and then we will show that this implies $\left\|P_{\mathcal{D}_\sigma^n}\beta f - f'_\beta\right\|_{L^2} \xrightarrow{p} 0$.

We know $\mathcal{D}_\sigma^n \subset \bar{\mathcal{D}}$ so $\left\|P_{\mathcal{D}_\sigma^n}\beta f - \beta f\right\|_{L^2} \geq \left\|P_{\bar{\mathcal{D}}}\beta f - \beta f\right\|_{L^2}$. We also know that for all $\delta \in \mathcal{D}_\sigma^n$, $\left\|P_{\mathcal{D}_\sigma^n}\beta f - \beta f\right\|_{L^2} \leq \left\|\delta - \beta f\right\|_{L^2}$. Because of these two facts, in order to show $\left\|P_{\mathcal{D}_\sigma^n}\beta f - \beta f\right\|_{L^2} \xrightarrow{p} \left\|P_{\bar{\mathcal{D}}}\beta f - \beta f\right\|_{L^2}$, it is sufficient to find a sequence $\{g_\sigma^n\} \subset \mathcal{D}_\sigma^n$ such that $\left\|g_\sigma^n - f'_\beta\right\|_{L^2} \xrightarrow{p} 0$. Since $\beta f > f'_\beta$ we can generate g_σ^n by applying rejection sampling to $X_1, \ldots, X_n$ to generate a subsample $X'_1, \ldots, X'_{m_n}$ which are iid from f'_β. For all i the event of X_i getting rejected is independent with equal probability. The probability of a sample not being rejected is greater than zero so there exists a $b > 0$ such that $\mathbb{E}[m_n] > bn$. From this and the strong law of large numbers we have that $\mathbb{P}\left(m_n\sigma^d \to \infty\right) = 1$. Using this subsample we can construct $g_\sigma^n \triangleq \frac{1}{m_n}\sum_1^{m_n} k_\sigma(\cdot, X'_i) \in \mathcal{D}_\sigma^n$ which is a KDE of f'_β, so by standard KDE consistency $\left\|f'_\beta - g_\sigma^n\right\|_{L^2} \xrightarrow{p} 0$, and thus $\left\|P_{\mathcal{D}_\sigma^n}\beta f - \beta f\right\|_{L^2} \xrightarrow{p} \left\|P_{\bar{\mathcal{D}}}\beta f - \beta f\right\|_{L^2}$.

Let $\widetilde{f}_{\sigma,\beta}^n \triangleq P_{\mathcal{D}_\sigma^n}\beta f$. Finally we are going to show that

$$\left\|P_{\mathcal{D}_\sigma^n}\beta f - \beta f\right\|_{L^2} \xrightarrow{p} \left\|P_{\bar{\mathcal{D}}}\beta f - \beta f\right\|_{L^2}$$

implies that

$$\left\|\widetilde{f}_{\sigma,\beta}^n - f'_\beta\right\|_{L^2} \xrightarrow{p} 0.$$

The functional $\left\|\beta f - \cdot\right\|_{L^2}^2$ is strongly convex with convexity constant 2 (Example *Bauschke and Combettes* (2011)). This means that for any $a \in (0,1)$, we have

$$\begin{aligned}&\left\|\beta f - \left(a\widetilde{f}_{\sigma,\beta}^n + (1-a)\, f'_\beta\right)\right\|_{L^2}^2 + a(1-a)\left\|\widetilde{f}_{\sigma,\beta}^n - f'_\beta\right\|_{L^2}^2 \\ &\quad\leq a\left\|\beta f - \widetilde{f}_{\sigma,\beta}^n\right\|_{L^2}^2 + (1-a)\left\|\beta f - f'_\beta\right\|_{L^2}^2.\end{aligned}$$

Letting $a = 1/2$ gives us

$$\left\|\beta f - \frac{\widetilde{f}_{\sigma,\beta}^n + f'_\beta}{2}\right\|_{L^2}^2 + \frac{1}{4}\left\|\widetilde{f}_{\sigma,\beta}^n - f'_\beta\right\|_{L^2}^2 \leq \frac{1}{2}\left\|\beta f - \widetilde{f}_{\sigma,\beta}^n\right\|_{L^2}^2 + \frac{1}{2}\left\|\beta f - f'_\beta\right\|_{L^2}^2$$

Since

$$\left\|\beta f - f'_\beta\right\|_{L^2}^2 \leq \left\|\beta f - \widetilde{f}^n_{\sigma,\beta}\right\|_{L^2}^2$$

and

$$\left\|\beta f - f'_\beta\right\|_{L^2}^2 \leq \left\|\beta f - \frac{\widetilde{f}^n_{\sigma,\beta} + f'_\beta}{2}\right\|_{L^2}^2$$

we have

$$\left\|\beta f - f'_\beta\right\|_{L^2}^2 + \frac{1}{4}\left\|\widetilde{f}^n_{\sigma,\beta} - f'_\beta\right\|_{L^2}^2 \leq \left\|\beta f - \widetilde{f}^n_{\sigma,\beta}\right\|_{L^2}^2$$

or equivalently

$$\left\|\widetilde{f}^n_{\sigma,\beta} - f'_\beta\right\|_{L^2}^2 \leq 4\left(\left\|\beta f - \widetilde{f}^n_{\sigma,\beta}\right\|_{L^2}^2 - \left\|\beta f - f'_\beta\right\|_{L^2}^2\right).$$

The right side of the last equation goes to zero in probability, thus finishing our proof. □

Proof of KDE L^2 consistency. Let $\bar{f}_\sigma = \mathbb{E}\left[k_\sigma\left(\cdot, X_i\right)\right] = \int k_\sigma\left(\cdot, x\right) g\left(x\right) dx$. Using the triangle inequality we have

$$\left\|f - \bar{f}^n_\sigma\right\|_{L^2} \leq \left\|f - \bar{f}_\sigma\right\|_{L^2} + \left\|\bar{f}_\sigma - \bar{f}^n_\sigma\right\|_{L^2}.$$

The left summand goes to zero as $\sigma \to 0$ by elementary analysis (see Theorem 8.14 in *Folland* (1999)). To take care of the right side with use the following lemma which is a Hilbert space version of Hoeffding's inequality from *Steinwart and Christmann* (2008), Corollary 6.15.

Lemma A.1 (Hoeffding's inequality in Hilbert space). *Let $(\Omega, \mathcal{A}, P)$ be a probability space, H be a separable Hilbert space, and $B > 0$. Furthermore, let $\xi_1, \ldots \xi_n : \Omega \to H$ be independent H-valued random variables satisfying $\left\|\xi_i\right\|_\infty \leq B$*

for all i. Then, for all $\tau > 0$, we have

$$P\left(\left\|\frac{1}{n}\sum_{1}^{n}\left(\xi_i - \mathbb{E}\left[\xi_i\right]\right)\right\|_H \geq B\sqrt{\frac{2\tau}{n}} + B\sqrt{\frac{1}{n}} + \frac{4B\tau}{3n}\right) \leq e^{-\tau}..$$

Note that $\|\xi_i\|_\infty = \operatorname{ess\,sup}_{\omega\in\Omega}\|\xi_i(\omega)\|_H$. Plugging in $\xi_i = k_\sigma(\cdot, X_i)$ we get

$$\begin{aligned} P\Bigg(&\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{L^2} \geq \|k_\sigma(\cdot, X_i)\|_{L^2}\sqrt{\frac{2\tau}{n}} \\ &+ \|k_\sigma(\cdot, X_i)\|_{L^2}\sqrt{\frac{1}{n}} + \frac{4\|k_\sigma(\cdot, X_i)\|_{L^2}\tau}{3n}\Bigg) \leq e^{-\tau}. \end{aligned}$$

It is straightforward to show that there exists $Q > 0$ such that $\|k_\sigma(\cdot, X_i)\|_{L^2} = Q\sigma^{-d/2}$, giving us

$$\begin{aligned} P\Bigg(&\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{L^2} \geq Q\sigma^{-d/2}\sqrt{\frac{2\tau}{n}} \\ &+Q\sigma^{-d/2}\sqrt{\frac{1}{n}} + \frac{4Q\sigma^{-d/2}\tau}{3n}\Bigg) \leq e^{-\tau}. \end{aligned}$$

Letting $n\sigma^d \to \infty$ sends all of the summands in the previous expression to zero for fixed τ. Because of this there exists a positive sequence $\{\tau_i\}_1^\infty$ such that $\tau_i \to \infty$ and but increases slowly enough that $Q\sigma^{-d/2}\sqrt{\frac{2\tau_n}{n}} + Q\sigma^{-d/2}\sqrt{\frac{1}{n}} + \frac{4Q\sigma^{-d/2}\tau_n}{3n} \to 0$ as $n \to \infty$, where σ depends implicitly on n. From this it is clear that $\left\|\bar{f}_\sigma^n - \bar{f}_\sigma\right\|_{L^2} \xrightarrow{p} 0$. □

Proof of Corollary III.7. Let λ be the Lebesgue measure. Let $S \subset \mathbb{R}^d$ be such that $\lambda(S) < \infty$. By Hölders inequality we have

$$\begin{aligned} \left\|\left(f'_\beta - f^n_{\sigma,\beta}\right)\chi_S\right\|_{L^1} &< \left\|f'_\beta - f^n_{\sigma,\beta}\right\|_{L^2}\|\chi_S\|_{L^2} \\ &= \left\|f'_\beta - f^n_{\sigma,\beta}\right\|_{L^2}\sqrt{\lambda(S)}. \end{aligned}$$

From this we have that, that $f^n_{\sigma,\beta}$ converges in probability to f'_β in L^1 norm, when restricted to a set of finite Lebesgue measure. Let $\delta > 0$ be arbitrary. Choose S to be

a set of finite measure large enough that $\int_{S^C} f'_\beta(x)\,dx < \delta/8$. Note that this implies $\left\|f'_\beta \chi_S\right\|_{L^1} \geq \frac{7}{8}\delta$, a fact we will use later. Notice that

$$\left\|f'_\beta - f^n_{\sigma,\beta}\right\|_{L^1} = \left\|\left(f'_\beta - f^n_{\sigma,\beta}\right)\chi_S\right\|_{L^1} + \left\|\left(f'_\beta - f^n_{\sigma,\beta}\right)\chi_{S^C}\right\|_{L^1}.$$

We have already shown that the left summand in the converges in probability to zero, so it becomes bounded by $\delta/8$ with probability going to one. To finish the proof we need only show that the right summand is bounded by $\frac{7}{8}\delta$ with probability going to one. Using the triangle inequality we have

$$\begin{aligned}\left\|\left(f'_\beta - f^n_{\sigma,\beta}\right)\chi_{S^C}\right\|_{L^1} &\leq \left\|f'_\beta \chi_{S^C}\right\|_1 + \left\|f^n_{\sigma,\beta}\chi_{S^C}\right\|_{L^1} \\ &< \delta/8 + \left\|f^n_{\sigma,\beta}\chi_{S^C}\right\|_{L^1}.\end{aligned}$$

Now it is sufficient to show that $\left\|f^n_{\sigma,\beta}\chi_{S^C}\right\|_1$ becomes bounded by $\frac{3}{4}\delta$ with probability going to one. To finish the proof,

$$\left\|f^n_{\sigma,\beta}\chi_S\right\|_{L^1} + \left\|f^n_{\sigma,\beta}\chi_{S^C}\right\|_{L^1} = 1$$

therefore

$$\left\|f^n_{\sigma,\beta}\chi_{S^C}\right\|_{L^1} = 1 - \left\|f^n_{\sigma,\beta}\chi_S\right\|_{L^1}$$

and we know that $\left\|f^n_{\sigma,\beta}\chi_S\right\|_{L^1} \xrightarrow{p} \left\|f'_\beta\chi_S\right\|_{L^1} \geq \frac{7}{8}\delta$ so with probability going to one $\left\|f^n_{\sigma,\beta}\chi_S\right\|_{L^1} \geq \delta/2$ and thus $\left\|f^n_{\sigma,\beta}\chi_{S^C}\right\|_{L^1} < \delta/2$. □

Proof of Theorem III.8. By the triangle inequality we have

$$\left\|f^n_{\sigma,\beta} - f_{tar}\right\|_{L^1} \leq \left\|f^n_{\sigma,\beta} - f'_\beta\right\|_{L^1} + \left\|f'_\beta - f_{tar}\right\|_{L^1}.$$

The left summand in the previous inequality goes to zero by Corollary III.7, so it is sufficient to show that the right term is zero. The rest of this proof will effectively prove Proposition III.3. Again let $g_{\alpha,\beta}(\cdot) = \max\{0, \beta f_{obs}(\cdot) - \alpha\}$. From Assumption A we know that Lebesgue almost everywhere on the support of f_{tar}, that f_{con} is equal to some value u and that f_{con} is less than or equal to u Lebesgue almost everywhere on $\mathbb{R}^d$. We will show that, $\alpha' = \frac{\varepsilon u}{1-\varepsilon}$, gives us $g_{\alpha',\beta} = f_{tar}$ which, by Lemma III.2, implies $f_{tar} = f'_\beta$. Let K be the support of f_{tar}.

First consider $x \in K^C$. Almost everywhere on K^C have

$$\begin{aligned} g_{\alpha',\beta}(x) &= \max\left\{0, \beta f_{obs}(x) - \frac{\varepsilon u}{1-\varepsilon}\right\} \\ &= \max\left\{0, \frac{1}{1-\varepsilon} f_{con}(x)\,\varepsilon - \frac{\varepsilon u}{1-\varepsilon}\right\} \\ &\leq \max\left\{0, \frac{1}{1-\varepsilon} u\varepsilon - \frac{\varepsilon u}{1-\varepsilon}\right\} \\ &= 0. \end{aligned}$$

So $g_{\alpha',\beta}$ is zero almost everywhere not on the support of f_{tar}. Now let $x \in K$, then Lebesgue almost everywhere in K we have

$$\begin{aligned} & g_{\alpha',\beta}(x) \\ &= \max\left\{0, \beta f_{obs}(x) - \frac{\varepsilon u}{1-\varepsilon}\right\} \\ &= \max\left\{0, \frac{1}{1-\varepsilon}\left((1-\varepsilon) f_{tar}(x) + f_{con}(x)\,\varepsilon\right) - \frac{\varepsilon u}{1-\varepsilon}\right\} \\ &= \max\left\{0, \frac{1}{1-\varepsilon}\left((1-\varepsilon) f_{tar}(x) + u\varepsilon\right) - \frac{\varepsilon u}{1-\varepsilon}\right\} \\ &= \max\left\{0, f_{tar}(x) + \frac{\varepsilon u}{1-\varepsilon} - \frac{\varepsilon u}{1-\varepsilon}\right\} \\ &= f_{tar}(x). \end{aligned}$$

From this we have that $g_{\alpha',\beta} = f_{tar}$ which is a pdf, which by Lemma III.2 is therefore equal to f'_β. □

A.2 Experimental Results

Table A.1: Mean and Standard Deviation of $D_{KL}\left(\widehat{f}||f_0\right)$

Dataset	Algorithm	ε						
		0.00	0.05	0.10	0.15	0.20	0.25	0.30
banana	SPKDE	0.19±0.04	0.15±0.03	0.14±0.03	0.17±0.07	0.23±0.08	0.35±0.1	0.51±0.2
	KDE	0.19±0.1	0.32±0.1	0.53±0.2	0.66±0.2	0.84±0.2	1.1±0.2	1.2±0.2
	RKDE	0.81±0.3	0.78±0.3	0.77±0.3	0.71±0.4	0.61±0.3	0.63±0.3	0.66±0.3
	rejKDE	0.19±0.2	0.35±0.2	0.52±0.2	0.7±0.2	0.84±0.2	1.1±0.2	1.3±0.2
breast-cancer	SPKDE	3.2±0.7	3.4±0.8	3.2±0.8	3.5±0.9	3.7±1	3.9±1	4.2±1
	KDE	4±0.9	4.1±1	4±1	4.3±1	4.6±1	4.8±1	5±1
	RKDE	3.1±0.7	3.2±0.7	3±0.5	3.2±0.6	3.5±0.8	3.7±0.9	4±0.9
	rejKDE	4±0.8	4.1±1	4.1±1	4.3±1	4.6±1	4.8±1	4.9±1
diabetis	SPKDE	0.8±0.05	0.84±0.09	0.8±0.1	0.84±0.1	0.87±0.1	0.91±0.08	0.89±0.09
	KDE	1.5±0.2	1.6±0.3	1.8±0.3	1.8±0.4	1.9±0.4	2±0.3	2±0.4
	RKDE	0.99±0.1	1±0.1	0.96±0.1	0.98±0.1	1±0.1	1±0.1	0.98±0.1
	rejKDE	1.5±0.2	1.6±0.2	1.8±0.4	1.9±0.5	1.9±0.5	2±0.4	2.1±0.5
german	SPKDE	6.6±0.9	6.8±1	6.9±0.9	7±0.9	6.9±1	7.2±0.7	7.4±0.7
	KDE	7±1	7±1	7.3±0.9	7.4±1	7.4±1	7.6±0.8	7.8±0.8
	RKDE	5.4±0.7	5.6±0.8	5.8±0.7	5.8±0.8	5.9±0.8	6±0.7	6.2±0.6
	rejKDE	7±1	7.2±1	7.4±1	7.5±1	7.5±1	7.7±0.8	7.8±0.7
heart	SPKDE	4±0.7	4±0.9	4.2±0.7	4.5±0.8	4.8±1	5.1±1	5.1±1
	KDE	4.7±1	5.1±1	5.3±1	5.6±1	5.8±1	6.2±1	6.6±1
	RKDE	3.8±0.9	3.8±0.8	3.9±0.6	4.2±0.8	4.2±0.9	4.5±1	4.9±1
	rejKDE	4.8±0.9	5.3±1	5.2±1	5.6±1	5.6±1	6.3±1	6.4±1
ionosphere scale	SPKDE	13±2	13±2	13±2	13±2	12±2	11±2	11±1
	KDE	15±2	14±2	14±2	15±2	14±2	13±2	14±2
	RKDE	10±2	10±2	9.9±2	9.2±2	8±3	6.7±2	7.5±3
	rejKDE	16±2	15±2	15±2	14±1	14±2	14±2	14±2
ringnorm	SPKDE	4.8±0.4	5.3±0.9	6.3±1	7.3±1	8±1	9.2±1	9±0.9
	KDE	4.9±0.4	5.7±0.9	7.4±1	8.6±1	11±2	13±2	14±0.7
	RKDE	4.4±0.2	3.8±0.6	4±0.6	4.1±0.6	4.7±1	5.7±0.6	6.1±0.5
	rejKDE	5±0.3	5.8±0.8	7.3±1	8.5±1	10±2	13±1	14±0.8
sonar scale	SPKDE	30±7	31±8	30±8	33±7	33±7	33±7	35±7
	KDE	31±6	31±9	31±8	32±8	34±7	35±8	35±8
	RKDE	32±9	32±7	32±7	31±7	33±8	34±7	35±7
	rejKDE	31±9	32±8	32±9	34±7	33±8	33±7	36±8
splice	SPKDE	21±0.3	21±0.2	21±0.3	21±0.3	21±0.2	21±0.2	20±0.4
	KDE	21±0.3	21±0.2	21±0.2	21±0.3	21±0.3	21±0.2	20±0.2
	RKDE	21±0.5	21±0.5	21±0.6	21±0.4	21±0.4	20±0.6	20±0.6
	rejKDE	21±0.3	21±0.3	21±0.2	21±0.2	21±0.3	21±0.2	20±0.2
thyroid	SPKDE	0.59±0.2	0.69±0.4	1.1±0.8	1.3±0.8	1.2±0.7	1.1±0.7	1.3±0.6
	KDE	0.6±0.2	4.5±3	11±7	16±7	20±7	22±5	32±8
	RKDE	0.56±0.1	0.88±0.5	1.3±0.9	1.6±1	1.5±0.8	1.3±0.6	1.4±0.8
	rejKDE	0.59±0.2	4.9±3	8.6±5	17±6	22±9	25±7	33±8
twonorm	SPKDE	4.8±0.4	4.6±0.5	4.6±0.5	4.8±0.7	5±0.9	5.4±0.9	6.2±1
	KDE	4.8±0.4	4.8±0.5	4.9±0.5	5.1±0.6	5.2±0.9	5.7±0.9	6.6±1
	RKDE	4.2±0.4	3.8±0.4	3.9±0.5	4±0.5	4.1±0.7	4.7±0.9	5.5±0.8
	rejKDE	4.9±0.5	4.7±0.6	4.9±0.5	5±0.7	5.2±0.8	5.7±0.9	6.6±1
waveform	SPKDE	4.8±0.8	4.8±0.8	5.2±1	5.6±0.9	6.1±0.8	6.2±0.8	6.7±0.5
	KDE	5±0.7	4.9±0.7	5.3±1	5.7±1	6.3±0.9	6.2±0.8	6.8±0.4
	RKDE	4.5±0.7	4.4±0.6	4.7±0.9	5.2±1	5.6±0.8	5.7±0.7	6.1±0.4
	rejKDE	4.9±0.7	4.9±0.7	5.4±1	5.8±0.9	6.2±0.9	6.3±0.8	6.8±0.4

Table A.2: Mean and Standard Deviation of $D_{KL}\left(f_0||\widehat{f}\right)$

Dataset	Algorithm	ε						
		0.00	0.05	0.10	0.15	0.20	0.25	0.30
banana	SPKDE	-0.57±0.2	-0.69±0.2	-0.73±0.2	-0.78±0.2	-0.81±0.2	-0.79±0.2	-0.75±0.2
	KDE	-0.85±0.2	-0.83±0.2	-0.8±0.1	-0.8±0.1	-0.8±0.1	-0.77±0.1	-0.74±0.1
	RKDE	15±1e+01	12±9	11±9	8.6±9	5.7±7	6.5±9	7.1±9
	rejKDE	-0.73±0.2	-0.8±0.2	-0.8±0.2	-0.82±0.1	-0.82±0.1	-0.79±0.1	-0.75±0.1
breast-cancer	SPKDE	-1.7±0.7	-1.8±0.7	-2±0.6	-2±0.6	-2.2±0.6	-2.4±0.6	-2.6±0.7
	KDE	-1.8±0.7	-1.9±0.6	-2.1±0.6	-2.1±0.6	-2.3±0.6	-2.4±0.6	-2.6±0.7
	RKDE	2.2±2	1.8±3	1.4±2	0.77±2	0.29±2	-0.025±2	-0.43±2
	rejKDE	0.4±2	0.1±2	-0.35±2	-0.69±1	-1±1	-1.2±1	-1.4±1
diabetis	SPKDE	-3.4±0.8	-3.7±0.7	-4±0.6	-4.2±0.6	-4.5±0.5	-4.6±0.4	-4.8±0.5
	KDE	-3.9±0.5	-4.1±0.5	-4.3±0.4	-4.4±0.3	-4.6±0.4	-4.7±0.3	-5±0.3
	RKDE	-1.3±1	-1.7±2	-1.7±1	-2±1	-2.1±2	-2.6±2	-2.5±1
	rejKDE	-3.7±0.7	-3.9±0.6	-4.2±0.5	-4.3±0.4	-4.5±0.4	-4.6±0.4	-4.9±0.4
german	SPKDE	-0.067±0.4	-0.15±0.4	-0.21±0.4	-0.26±0.4	-0.32±0.4	-0.41±0.4	-0.48±0.4
	KDE	-0.043±0.4	-0.12±0.4	-0.19±0.4	-0.23±0.4	-0.29±0.4	-0.38±0.4	-0.45±0.4
	RKDE	0.71±0.5	0.62±0.5	0.56±0.7	0.52±0.6	0.45±0.6	0.35±0.6	0.29±0.6
	rejKDE	0.26±0.5	0.16±0.5	0.07±0.5	0.039±0.5	-0.026±0.5	-0.12±0.5	-0.2±0.5
heart	SPKDE	0.7±0.7	0.44±0.9	0.17±0.7	0.071±0.7	-0.044±0.8	-0.21±0.8	-0.32±0.8
	KDE	0.71±0.7	0.46±0.8	0.2±0.7	0.12±0.7	0.0049±0.8	-0.15±0.8	-0.26±0.7
	RKDE	2.4±1	1.9±0.9	1.5±0.8	1.4±1	1.2±0.8	1±0.9	0.82±0.8
	rejKDE	1.3±0.9	1±0.9	0.68±0.9	0.6±0.9	0.42±0.9	0.23±0.9	0.12±0.8
ionosphere scale	SPKDE	7.5±1	7.3±1	7.2±1	7.1±1	7±1	7±1	7.5±2
	KDE	7.8±1	7.6±1	7.5±1	7.3±1	7.3±1	7.3±1	7.7±2
	RKDE	7.6±1	7.5±1	7.4±1	7.4±2	7.6±2	8.9±4	9.9±4
	rejKDE	7.7±1	7.6±1	7.4±1	7.2±1	7.2±1	7.2±1	7.6±2
ringnorm	SPKDE	-3±0.4	-8±1	-10±0.8	-12±0.8	-13±0.7	-13±0.4	-14±0.4
	KDE	-3±0.4	-7.8±1	-9.8±0.8	-11±0.8	-12±0.7	-13±0.4	-14±0.4
	RKDE	-3.2±0.4	-8.1±1	-10±0.8	-12±0.8	-13±0.7	-13±0.4	-14±0.4
	rejKDE	-3.1±0.4	-7.9±1	-9.9±0.8	-12±0.8	-12±0.7	-13±0.4	-14±0.4
sonar scale	SPKDE	-16±6	-16±5	-17±5	-17±5	-18±5	-19±5	-19±5
	KDE	-16±6	-16±5	-17±5	-17±5	-18±5	-19±5	-19±5
	RKDE	-16±6	-16±5	-17±5	-16±7	-18±5	-19±5	-19±5
	rejKDE	-8.2±9	-9.4±8	-9.6±8	-10±8	-11±8	-11±8	-11±8
splice	SPKDE	34±0.3	34±0.3	34±0.3	34±0.2	34±0.2	34±0.2	34±0.2
	KDE	34±0.3	34±0.3	34±0.3	34±0.2	34±0.2	34±0.2	34±0.2
	RKDE	34±0.3	34±0.3	34±0.2	34±0.2	34±0.2	34±0.2	34±0.2
	rejKDE	34±0.3	34±0.3	34±0.3	34±0.2	34±0.2	34±0.2	34±0.2
thyroid	SPKDE	-0.86±0.9	-4.1±0.9	-5.1±1	-5.9±0.5	-6.4±0.4	-6.7±0.2	-6.8±0.2
	KDE	-0.89±0.7	-4±0.7	-5±0.8	-5.6±0.4	-6.1±0.3	-6.3±0.2	-6.4±0.2
	RKDE	-0.71±0.9	-3.9±0.9	-5±1	-5.8±0.4	-6.3±0.3	-6.6±0.2	-6.8±0.2
	rejKDE	-0.88±0.8	-4.1±0.7	-5.1±0.8	-5.7±0.4	-6.1±0.3	-6.4±0.2	-6.5±0.2
twonorm	SPKDE	-3.2±0.6	-3.8±0.5	-4±0.5	-4.4±0.4	-4.6±0.3	-4.8±0.4	-5.1±0.4
	KDE	-3.1±0.6	-3.7±0.5	-3.9±0.4	-4.3±0.4	-4.5±0.3	-4.7±0.4	-5±0.5
	RKDE	-3.3±0.6	-3.9±0.5	-4.1±0.5	-4.5±0.4	-4.7±0.3	-4.9±0.4	-5.2±0.5
	rejKDE	-3.2±0.6	-3.8±0.5	-4±0.5	-4.3±0.4	-4.6±0.3	-4.8±0.4	-5.1±0.5
waveform	SPKDE	-7.6±0.3	-7.7±0.3	-7.9±0.3	-8±0.4	-8.1±0.3	-8.3±0.3	-8.3±0.3
	KDE	-7.5±0.3	-7.7±0.4	-7.8±0.4	-8±0.4	-8.1±0.4	-8.2±0.4	-8.3±0.3
	RKDE	-7.6±0.3	-7.8±0.3	-8±0.4	-8.1±0.4	-8.2±0.4	-8.4±0.4	-8.4±0.3
	rejKDE	-7.6±0.3	-7.8±0.4	-7.9±0.4	-8±0.4	-8.2±0.4	-8.3±0.4	-8.4±0.3

APPENDIX B

Chapter IV Additional Proofs and Algorithm

B.1 Additional Proofs

Some of the proofs use Hilbert-Schmidt operators. See Definition IV.27 for the definition of Hilbert-Schmidt operator.

Proof of Lemma IV.1. Because both representations are minimal it follows that $\alpha'_i \neq 0$ for all i and $\nu'_i \neq \nu'_j$ for all $i \neq j$. From this we know $\mathscr{Q}(\{\nu'_i\}) \neq 0$ for all i. Because $\mathscr{Q}(\{\nu'_i\}) \neq 0$ for all i it follows that for any i there exists some j such that $\nu'_i = \nu_j$. Let $\psi : [r] \to [r]$ be a function satisfying $\nu'_i = \nu_{\psi(i)}$. Because the elements $\nu_1, \ldots, \nu_r$ are also distinct, ψ must be injective and thus a permutation. Again from this distinctness we get that, for all i, $\mathscr{Q}(\{\nu'_i\}) = \alpha'_i = \alpha_{\psi(i)}$ and we are done. □

Proof of Lemma IV.7 and IV.11. We will proceed by contradiction. Let $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ be n-identifiable/determined, let $\mathscr{P}' = \sum_{j=1}^{l} b_j \delta_{\nu_j}$ be a different mixture of measures, with $l \leq m$ for the n-identifiable case, and

$$\sum_{i=1}^{m} a_i \mu_i^{\times q} = \sum_{j=1}^{l} b_j \nu_j^{\times q}$$

for some $q > n$. Let $A \in \mathcal{F}^{\times n}$ be arbitrary. We have

$$\begin{aligned} \sum_{i=1}^{m} a_i \mu_i^{\times q} &= \sum_{j=1}^{l} b_j \nu_j^{\times q} \\ \Rightarrow \sum_{i=1}^{m} a_i \mu_i^{\times q} \left(A \times \Omega^{\times q-n}\right) &= \sum_{j=1}^{l} b_j \nu_j^{\times q} \left(A \times \Omega^{\times q-n}\right) \\ \Rightarrow \sum_{i=1}^{m} a_i \mu_i^{\times n} (A) &= \sum_{j=1}^{l} b_j \nu_j^{\times n} (A) . \end{aligned}$$

This implies that $\mathscr{P}$ is not n-identifiable/determined, a contradiction. □

Proof of Lemma IV.8 and IV.12. Let a mixture of measures $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ not be n-identifiable/determined. It follows that there exists a different mixture of measures $\mathscr{P}' = \sum_{j=1}^{l} b_j \delta_{\nu_j}$, with $l \leq m$ for the n-identifiability case, such that

$$\sum_{i=1}^{m} a_i \mu_i^{\times n} = \sum_{j=1}^{l} b_j \nu_j^{\times n}.$$

Let $A \in \mathcal{F}^{\times q}$ be arbitrary, we have

$$\begin{aligned} \sum_{i=1}^{m} a_i \mu_i^{\times n} \left(A \times \Omega^{\times n-q}\right) &= \sum_{j=1}^{l} b_j \nu_j^{\times n} \left(A \times \Omega^{\times n-q}\right) \\ \Rightarrow \sum_{i=1}^{m} a_i \mu_i^{\times q} (A) &= \sum_{j=1}^{l} b_j \nu_j^{\times q} (A) \end{aligned}$$

and therefore $\mathscr{P}$ is not q-identifiable/determined. □

Proof of Lemma IV.18. Example 2.6.11 in *Kadison and Ringrose* (1983) states that for any two σ-finite measure spaces $(S, \mathscr{S}, m), (S', \mathscr{S}', m')$ there exists a unitary operator $U : L^2(S, \mathscr{S}, m) \otimes L^2(S', \mathscr{S}', m') \to L^2(S \times S', \mathscr{S} \times \mathscr{S}', m \times m')$ such that, for all f, g,

$$U(f \otimes g) = f(\cdot) g(\cdot).$$

Because $(\Psi, \mathcal{G}, \gamma)$ is a σ-finite measure space it follows that $(\Psi^{\times m}, \mathcal{G}^{\times m}, \gamma^{\times m})$ is a σ-finite measure space for all $m \in \mathbb{N}$. We will now proceed by induction. Clearly the lemma holds for $n = 1$. Suppose the lemma holds for $n-1$. From the induction hypothesis we know that there exists a unitary transform $U_{n-1} : L^2(\Psi, \mathcal{G}, \gamma)^{\otimes n-1} \to L^2(\Psi^{\times n-1}, \mathcal{G}^{\times n-1}, \gamma^{\times n-1})$ such that for all simple tensors $f_1 \otimes \cdots \otimes f_{n-1} \in L^2(\Psi, \mathcal{G}, \gamma)^{\otimes n-1}$ we have $U_{n-1}(f_1 \otimes \cdots \otimes f_{n-1}) = f_1(\cdot) \cdots f_{n-1}(\cdot)$. Combining U_{n-1} with the identity map via Lemma IV.19 we can construct a unitary operator $T_n : L^2(\Psi, \mathcal{G}, \gamma)^{\otimes n-1} \otimes L^2(\Psi, \mathcal{G}, \gamma) \to L^2(\Psi^{\times n-1}, \mathcal{G}^{\times n-1}, \gamma^{\times n-1}) \otimes L^2(\Psi, \mathcal{G}, \gamma)$, which maps $f_1 \otimes \cdots \otimes f_{n-1} \otimes f_n \mapsto f_1(\cdot) \cdots f_{n-1}(\cdot) \otimes f_n$.

From the aforementioned example there exists a unitary transform $K_n : L^2(\Psi^{\times n-1}, \mathcal{G}^{\times n-1}, \gamma^{\times n-1}) \otimes L^2(\Psi, \mathcal{G}, \gamma) \to L^2(\Psi^{\times n-1} \times \Psi, \mathcal{G}^{\times n-1} \times \mathcal{G}, \gamma^{\times n-1} \times \gamma)$ which maps simple tensors $g \otimes g' \in L^2(\Psi^{\times n-1}, \mathcal{G}^{\times n-1}, \gamma^{\times n-1}) \otimes L^2(\Psi, \mathcal{G}, \gamma)$ as $K_n(g \otimes g') = g(\cdot)g'(\cdot)$. Defining $U_n(\cdot) = K_n(T_n(\cdot))$ yields our desired unitary transform. □

Proof of Lemma IV.19. Lemma IV.32 states that there exists a continuous linear operator $\tilde{U} : H_1 \otimes \cdots \otimes H_n \to H_1' \otimes \cdots \otimes H_n'$ such that $\tilde{U}(h_1 \otimes \cdots \otimes h_n) = U_1(h_1) \otimes \cdots \otimes U_n(h_n)$ for all $h_1 \in H_1, \cdots, h_n \in H_n$. Let $\widehat{H}$ be the set of simple tensors in $H_1 \otimes \cdots \otimes H_n$ and $\widehat{H}'$ be the set of simple tensors in $H_1' \otimes \cdots \otimes H_n'$. Because U_i is surjective for all i, clearly $\tilde{U}(\widehat{H}) = \widehat{H}'$. The linearity of $\tilde{U}$ implies that $\tilde{U}(\text{span}(\widehat{H})) = \text{span}(\widehat{H}')$. Because $\text{span}(\widehat{H}')$ is dense in $H_1' \otimes \cdots \otimes H_n'$ the continuity of $\tilde{U}$ implies that $\tilde{U}(H_1 \otimes \cdots \otimes H_n) = H_1' \otimes \cdots \otimes H_n'$ so $\tilde{U}$ is surjective. All that remains to be shown is that $\tilde{U}$ preserves the inner product (see Theorem 4.18 in *Young* (1988)). By the continuity of inner product we need only show that $\langle h, g \rangle = \left\langle \tilde{U}(h), \tilde{U}(g) \right\rangle$ for $h, g \in \text{span}(\widehat{H})$. With this in mind let $h_1, \ldots, h_N, g_1, \ldots, g_M$ be simple tensors in

$H_1 \otimes \cdots \otimes H_n$. We have the following

$$\begin{aligned}
\left\langle \tilde{U}\left(\sum_{i=1}^{N} h_i\right), \tilde{U}\left(\sum_{j=1}^{M} g_j\right)\right\rangle &= \left\langle \sum_{i=1}^{N} \tilde{U}\left(h_i\right), \sum_{j=1}^{M} \tilde{U}\left(g_j\right)\right\rangle \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M} \left\langle \tilde{U}\left(h_i\right), \tilde{U}\left(g_j\right)\right\rangle \\
&= \sum_{i=1}^{N}\sum_{j=1}^{M} \langle h_i, g_j\rangle \\
&= \left\langle \sum_{i=1}^{N} h_i, \sum_{j=1}^{M} g_j\right\rangle.
\end{aligned}$$

We have now shown that $\tilde{U}$ is unitary which completes our proof. □

Proof of Lemma IV.20. We will proceed by induction. For $n = 2$ the lemma clearly holds. Suppose the lemma holds for $n-1$ and let $h_1, \ldots, h_n$ satisfy the assumptions in the lemma statement. Let $\alpha_1, \ldots, \alpha_n$ satisfy

$$\sum_{i=1}^{n} \alpha_i h_i^{\otimes n-1} = 0. \tag{B.1}$$

To finish the proof we will show that α_1 must be zero which can be generalized to any α_i. Applying Lemma IV.28 to (B.1) we get

$$\sum_{i=1}^{n} \alpha_i h_i^{\otimes n-2} \langle h_i, \cdot\rangle = 0. \tag{B.2}$$

Because h_1 and h_n are linearly independent we can choose z such that $\langle h_1, z\rangle \neq 0$ and $z \perp h_n$. Plugging z into (B.2) yields

$$\sum_{i=1}^{n-1} \alpha_i h_i^{\otimes n-2} \langle h_i, z\rangle = 0$$

and therefore $\alpha_1 = 0$ by the inductive hypothesis. □

Proof of Lemma IV.21. Let $\dim(\operatorname{span}(h_1, \ldots, h_m)) = l$ and let $h = \sum_{i=1}^{m} h_i^{\otimes 2}$. Without loss of generality assume that $h_1, \ldots, h_l$ are linearly independent and nonzero. From Lemma IV.28 there exists a unitary transform $U : H \otimes H \to \mathscr{HS}(H, H)$ which, for any simple tensor $x \otimes y$, we have $U(x \otimes y) = x \langle y, \cdot \rangle$.

First we will show that the rank is greater than or equal to l by contradiction. Suppose that $g = \sum_{i=1}^{l'} x_i \otimes y_i = h$ with $l' < l$. Since $l' < l$ there must exist some j such that $h_j \notin \operatorname{span}(x_1, \ldots, x_{l'})$. Let $z \perp x_1, \ldots, x_{l'}$ and $z \not\perp h_j$. Now we have

$$\langle z \otimes z, h \rangle = \sum_{i=1}^{m} \langle z, h_i \rangle^2 \geq \langle z, h_j \rangle^2 > 0,$$

but

$$\langle z \otimes z, g \rangle = \sum_{i=1}^{l'} \langle z, x_i \rangle \langle z, y_i \rangle = 0,$$

a contradiction.

For the other direction, observe that $U(h)$ is a compact Hermitian operator and thus admits an spectral decomposition (*Young* (1988) Theorem 8.15). From this we have that $U(h) = \sum_{i=1}^{m} h_i \langle h_i, \cdot \rangle = \sum_{i=1}^{\infty} \lambda_i \langle \psi_i, \cdot \rangle \psi_i$ with $(\psi_i)_{i=1}^{\infty}$ orthonormal and $\lambda_i \geq 0$ for all i since $U(h)$ is PSD. Clearly the dimension of the span of $U(h)$ is less than or equal to l and thus this decomposition has exactly l nonzero terms. From this we can let $U(h) = \sum_{i=1}^{l} \lambda_i \langle \psi_i, \cdot \rangle \psi_i$ and applying U^{-1} we have that $h = \sum_{i=1}^{l} \lambda_i \psi_i^{\otimes 2}$. From this it follows that the rank of h is less than or equal to l and we are done. □

Proof of Lemma IV.22. The lemma is obvious when $n = n'$. Assume that $n' < n$.

Let $A \in \mathcal{G}^{\times n'}$ be arbitrary. We have that

$$\begin{aligned}
\sum_{i=1}^{m} a_i \gamma_i^{\times n}\left(A \times \Psi^{\times n-n'}\right) &= \sum_{j=1}^{l} b_j \pi_j^{\times n}\left(A \times \Psi^{\times n-n'}\right) \\
\Rightarrow \sum_{i=1}^{m} a_i \gamma_i^{\times' n}(A)\, \gamma_i^{\times n-n'}\left(\Psi^{\times n-n'}\right) &= \sum_{j=1}^{l} b_j \pi_j^{\times n'}(A)\, \pi_j^{\times n-n'}\left(\Psi^{\times n-n'}\right) \\
\Rightarrow \sum_{i=1}^{m} a_i \gamma_i^{\times n'}(A) &= \sum_{j=1}^{l} b_j \pi_j^{\times n'}(A).
\end{aligned}$$

Since A was chosen arbitrarily we have that $\sum_{i=1}^{m} a_i \gamma_i^{\times n'} = \sum_{j=1}^{l} b_j \pi_j^{\times n'}$. □

Proof of Lemma IV.23. Let $\pi = \sum_{i=1}^{n} \gamma_i$. Because π is σ-finite for all i we can define $f_i = \frac{d\gamma_i}{d\pi}$, where the derivatives are Radon-Nikodym derivatives. Let f_k be arbitrary. We will first show that $f_k \leq 1$ π-almost everywhere. Suppose there exists a non π-null set $A \in \mathcal{G}$ such that $f_i(A) > 1$. Then we would have

$$\begin{aligned}
\gamma_k(A) &= \int_A f_k d\pi \\
&> \int_A 1 d\pi \\
&= \sum_{i=1}^{n} \gamma_i(A) \\
&\geq \gamma_k(A)
\end{aligned}$$

a contradiction. From this we have

$$\begin{aligned}
\int f_k^2 d\pi &\leq \int 1 d\pi \\
&\leq \sum_{i=1}^{n} \gamma_i(\Psi) \\
&< \infty.
\end{aligned}$$

From our construction it is clear that $f_i \geq 0$ ξ-almost everywhere so we can assert

$f_i \geq 0$ without issue. □

Proof of Lemma IV.24. The fact that f is non-negative and integrable implies that the map $S \mapsto \int_S f^{\times n} d\pi^{\times n}$ is a bounded measure on $(\Psi^{\times n}, \mathcal{G}^{\times n})$ (see *Folland* (1999) Exercise 2.12).

Let $R = R_1 \times \cdots \times R_n$ be a rectangle in $\mathcal{G}^{\times n}$. Let $\mathbb{1}_S$ be the indicator function for a set S. Integrating over R and using Tonelli's theorem we get

$$\begin{aligned}
\int_R f^{\times n} d\pi^{\times n} &= \int \mathbb{1}_R f^{\times n} d\pi^{\times n} \\
&= \int \left(\prod_{i=1}^{n} \mathbb{1}_{R_i}(x_i) \right) \left(\prod_{j=1}^{n} f(x_j) \right) d\pi^{\times n}(x_1, \ldots, x_n) \\
&= \int \cdots \int \left(\prod_{i=1}^{n} \mathbb{1}_{R_i}(x_i) \right) \left(\prod_{j=1}^{n} f(x_j) \right) d\pi(x_1) \cdots d\pi(x_n) \\
&= \int \cdots \int \left(\prod_{i=1}^{n} \mathbb{1}_{R_i}(x_i) f(x_i) \right) d\pi(x_1) \cdots d\pi(x_n) \\
&= \prod_{i=1}^{n} \left(\int \mathbb{1}_{R_i}(x_i) f(x_i) d\pi(x_i) \right) \\
&= \prod_{i=1}^{n} \gamma(R_i) \\
&= \gamma^{\times n}(R).
\end{aligned}$$

Any product probability measure is uniquely determined by its measure over the rectangles (this is a consequence of Lemma 1.17 in *Kallenberg* (2002) and the definition of product σ-algebra) therefore, for all $B \in \mathcal{G}^{\times n}$,

$$\gamma^{\times n}(B) = \int_B f^{\times n} d\pi^{\times n}.$$

□

B.2 Spectral Algorithm for Linearly Independent Components

Let $p_1, \ldots, p_m \in L^2(\Omega, \mathcal{F}, \xi)$ be linearly independent pdfs with distinct norms. Their associated mixture proportions are $w_1, \ldots, w_m$. With four samples per random group we will have access to the tensors

$$\sum_{i=1}^{m} w_i p_i^{\otimes 4} \tag{B.3}$$

and

$$\sum_{i=1}^{m} w_i p_i^{\otimes 2}. \tag{B.4}$$

We can transform the tensor in (B.4) to an operator

$$\begin{aligned} C &\triangleq \sum_{i=1}^{m} w_i p_i \langle p_i, \cdot \rangle \\ &= \sum_{i=1}^{m} \sqrt{w_i} p_i \langle \sqrt{w_i} p_i, \cdot \rangle . \end{aligned}$$

Letting $W = \sqrt{C^{\dagger}}$ we have that $W\sqrt{w_1}p_1, \ldots, W\sqrt{w_m}p_m$ are orthonormal. Applying $I \otimes W \otimes I \otimes W$ to the tensor in (B.3) we can construct the tensor

$$\sum_{i=1}^{m} w_i p_i \otimes W p_i \otimes p_i \otimes W p_i = \sum_{i=1}^{m} p_i \otimes W\sqrt{w_i} p_i \otimes p_i \otimes W\sqrt{w_i} p_i.$$

which can be transformed into the operator

$$\sum_{i=1}^{m} p_i \otimes W\sqrt{w_i} p_i \langle p_i \otimes W\sqrt{w_i} p_i, \cdot \rangle . \tag{B.5}$$

Note that for $i \neq j$ we have

$$\left\langle p_i \otimes W\sqrt{w_i}p_i, p_j \otimes W\sqrt{w_j}p_j \right\rangle = \langle p_i, p_j \rangle \left\langle W\sqrt{w_i}p_i, W\sqrt{w_j} \right\rangle = 0.$$

We also have that, for all i

$$\begin{aligned} \|p_i \otimes W\sqrt{w_i}p_i\| &= \sqrt{\langle p_i \otimes W\sqrt{w_i}p_i, p_i \otimes W\sqrt{w_i}p_i \rangle} \\ &= \sqrt{\langle p_i, p_i \rangle \langle W\sqrt{w_i}p_i, W\sqrt{w_i}p_i \rangle} \\ &= \sqrt{\langle p_i, p_i \rangle} \\ &= \|p_i\| \end{aligned}$$

and thus the tensors $p_1 \otimes W\sqrt{w_1}p_1, \ldots, p_m \otimes W\sqrt{w_m}p_m$ have distinct norms. Because of this the spectral decomposition of the operator in (B.5) will yield the eigenvectors $p_1 \otimes W\sqrt{w_1}p_1, \ldots, p_m \otimes W\sqrt{w_m}p_m$. Then, using the techniques from Section 4.6, we can recover the mixture components and mixture proportions.

BIBLIOGRAPHY

BIBLIOGRAPHY

Allman, E. S., C. Matias, and J. A. Rhodes (2009), Identifiability of parameters in latent structure models with many observed variables, *Ann. Statist.*, *37*(6A), 3099–3132, doi:10.1214/09-AOS689.

Anandkumar, A., R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky (2014), Tensor decompositions for learning latent variable models, *Journal of Machine Learning Research*, *15*, 2773–2832.

Anderson, J., M. Belkin, N. Goyal, L. Rademacher, and J. Voss (2014), The more, the merrier: The blessing of dimensionality for learning large Gaussian mixtures, in *Proceedings of The 27th Conference on Learning Theory*, pp. 1135–1164.

Aronszajn, N. (1950), Theory of reproducing kernels, *Transactions of the American Mathematical Society*, *68*.

Arora, S., R. Ge, R. Kannan, and A. Moitra (2012), Computing a nonnegative matrix factorization – provably, in *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pp. 145–162, ACM, New York, NY, USA, doi:10.1145/2213977.2213994.

Balan, R., P. Casazza, and D. Edidin (2006), On signal reconstruction without phase, *Applied and Computational Harmonic Analysis*, *20*(3), 345 – 356, doi: http://dx.doi.org/10.1016/j.acha.2005.07.001.

Bauer, F., S. Pereverzev, and L. Rosasco (2007), On regularization algorithms in learning theory, *J. Complex.*, *23*(1), 52–72, doi:10.1016/j.jco.2006.07.001.

Bauschke, H., and P. Combettes (2011), *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics, Ouvrages de mathématiques de la SMC, Springer New York.

Berry, D., K. Chaloner, J. Geweke, and A. Zellner (1996), *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, A Wiley Interscience publication, Wiley.

Blanchard, G., and C. Scott (2014), Decontamination of mutually contaminated models, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pp. 1–9.

Brucker, P. (1984), An o(n) algorithm for quadratic knapsack problems, *Operations Research Letters*, *3*(3), 163 – 166.

Bruni, C., and G. Koch (1985), Identifiability of continuous mixtures of unknown Gaussian distributions, *Ann. Probab.*, *13*(4), 1341–1357, doi: 10.1214/aop/1176992817.

Caponnetto, A., and E. D. Vito (2007), Optimal rates for the regularized least-squares algorithm, *Foundations of Computational Mathematics*, *7*(3), 331–368.

Comon, P., G. Golub, L.-H. Lim, and B. Mourrain (2008), Symmetric tensors and symmetric tensor rank, *SIAM Journal on Matrix Analysis and Applications*, *30*(3), 1254–1279, doi:10.1137/060661569.

Deheuvels, P. (2000), Uniform limit laws for kernel density estimators on possi-

bly unbounded interval, in *Recent Advances in Reliability Theory*, pp. 477–492, Birkhäuser.

Devroye, L., and G. Lugosi (2001), *Combinatorial Methods in Density Estimation*, Springer, New York.

Donoho, D., and V. Stodden (2003), When does non-negative matrix factorization give a correct decomposition into parts?, in *Advances in neural information processing systems*, p. None.

Duchi, J. C., S. Shalev-Shwartz, Y. Singer, and T. Chandra (2008), Efficient projections onto the l_1-ball for learning in high dimensions, in *ICML*, pp. 272–279.

Einmahl, U., and D. Mason (2000), An empirical process approach to the uniform consistency of kernel-type function estimators, *J. Theoret. Probab.*, *13*, 1–37.

El-Yaniv, R., and M. Nisenson (2007), Optimal single-class classification strategies, in *Adv. in Neural Inform. Proc. Systems 19*, edited by B. Schölkopf, J. Platt, and T. Hoffman, MIT Press, Cambridge, MA.

Elmore, R., and S. Wang (2003), Identifiability and estimation in finite mixture models with multinomial components.

Folland, G. B. (1999), *Real analysis: modern techniques and their applications*, Pure and applied mathematics, Wiley.

Giné, E., and A. Guillou (2002), Rates of strong uniform consistency for multivariate kernel density estimators, *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, *38*(6), 907 – 921, doi:10.1016/S0246-0203(02)01128-7.

Gine, E., V. Koltchinskii, and J. Zinn (2004), Weighted uniform consistency of kernel density estimators, *Ann. Probab.*, *32*, 2570–2605.

Golub, G. H., and C. F. Van Loan (1996), *Matrix Computations (3rd Ed.)*, Johns Hopkins University Press, Baltimore, MD, USA.

Han, Y., J. Jiao, and T. Weissman (2014), Minimax estimation of discrete distributions under ℓ_1 loss, *CoRR*, *abs/1411.1467*.

Kadison, R., and J. Ringrose (1983), *Fundamentals of the theory of operator algebras. V1: Elementary theory*, Pure and Applied Mathematics, Elsevier Science.

Kallenberg, O. (2002), *Foundations of Modern Probability*, Probability and Its Applications, Springer New York.

Kamath, S., A. Orlitsky, D. Pichapati, and A. T. Suresh (2015), On learning distributions from their samples., in *COLT*, edited by P. Grnwald, E. Hazan, and S. Kale, JMLR Proceedings.

Kim, B. S. (1984), Studies of multinomial mixture models, Ph.D. thesis, The University of North Carolina at Chapel Hill.

Kim, J., and C. Scott (2012), Robust kernel density estimation, *J. Machine Learning Res.*, *13*, 2529–2565.

Kruskal, J. B. (1977), Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, *Linear Algebra and its Applications*, *18*(2), 95 – 138.

Lanckriet, G., L. E. Ghaoui, and M. I. Jordan (2003), Robust novelty detection with single-class mpm, in *Advances in Neural Information Processing Systems 15*, edited by S. T. S. Becker and K. Obermayer, pp. 905–912, MIT Press, Cambridge, MA.

Lehmann, E., and G. Casella (2003), *Theory of Point Estimation*, Springer Texts in Statistics, Springer New York.

Liu, H., J. D. Lafferty, and L. A. Wasserman (2007), Sparse nonparametric density estimation in high dimensions using the rodeo, in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07)*, vol. 2, pp. 283–290, Journal of Machine Learning Research - Proceedings Track.

Mahapatruni, R. S. G., and A. G. Gray (2011), Cake: Convex adaptive kernel density estimation., in *AISTATS*, *JMLR Proceedings*, vol. 15, pp. 498–506, JMLR.org.

Micchelli, C. A., Y. Xu, H. Zhang, and G. Lugosi (2006), Universal kernels, *J. Machine Learning Research*, *7*, 2651–2667.

Orlitsky, A., and A. T. Suresh (2015), Competitive distribution estimation: Why is good-turing good, in *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, pp. 2143–2151, Curran Associates, Inc.

Pachter, L., and D. E. Speyer (2004), Reconstructing trees from subtree weights, *Applied Mathematics Letters*, *17*, 615–621.

Paninski, L. (2005), Variational minimax estimation of discrete distributions under kl loss, in *Advances in Neural Information Processing Systems 17*, edited by L. K. Saul, Y. Weiss, and L. Bottou, pp. 1033–1040, MIT Press.

Pardalos, P., and N. Kovoor (1990), An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds, *Mathematical Programming*, *46*(1-3), 321–328.

Paz, A. (1971), *Introduction to Probabilistic Automata (Computer Science and Applied Mathematics)*, Academic Press, Inc., Orlando, FL, USA.

Pinelis, I. (1994), Optimum bounds for the distributions of martingales in banach spaces, *The Annals of Probability*, *22*(4), pp. 1679–1706.

Rabani, Y., L. J. Schulman, and C. Swamy (2014), Learning mixtures of arbitrary distributions over large discrete domains, in *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, pp. 207–224, ACM, New York, NY, USA, doi:10.1145/2554797.2554818.

Schölkopf, B., J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson (2001), Estimating the support of a high-dimensional distribution, *Neural Computation*, *13*(7), 1443–1472.

Scott, D. W. (1992), *Multivariate Density Estimation*, Wiley, New York.

Scovel, C., D. Hush, I. Steinwart, and J. Theiler (2010), Radial kernels and their reproducing kernel Hilbert spaces, *Journal of Complexity*, *26*(6), 641 – 660, doi: 10.1016/j.jco.2010.03.002.

Silverman, B. W. (1978), Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, *Ann. Statist.*, *6*(1), 177–184, doi: 10.1214/aos/1176344076.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Smale, S., and D.-X. Zhou (2007), Learning theory estimates via integral operators and their approximations, *Constructive Approximation*, *26*, 153–172, 10.1007/s00365-006-0659-y.

Song, L., A. Anandkumar, B. Dai, and B. Xie (2014), Nonparametric estimation of multi-view latent variable models, in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 640–648.

Sricharan, K., and A. Hero (2011), Efficient anomaly detection using bipartite k-nn graphs, in *Advances in Neural Information Processing Systems 24*, edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, pp. 478–486.

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

Steinwart, I., D. Hush, and C. Scovel (2005), A classification framework for anomaly detection, *JMLR*, *6*, 211–232.

Stute, W. (1982), A law of the logarithm for kernel density estimators, *Ann. Probab.*, *10*, 414–422.

Teicher, H. (1963), Identifiability of finite mixtures, *Ann. Math. Statist.*, *34*(4), 1265–1269, doi:10.1214/aoms/1177703862.

Terrell, G. R., and D. W. Scott (1992), Variable kernel density estimation, *Ann. Statist.*, *20*(3), 1236–1265, doi:10.1214/aos/1176348768.

Theiler, J., and D. M. Cai (2003), Resampling approach for anomaly detection in multispectral images, in *Proc. SPIE*, vol. 5093, pp. 230–240.

Valiant, G., and P. Valiant (2016), Instance optimal learning of discrete distributions, in *STOC, 2016 (to appear)*.

Vert, R., and J.-P. Vert (2006), Consistency and convergence rates of one-class SVM and related algorithms, *JMLR*, pp. 817–854.

Wied, D., and R. Weissbach (2012), Consistency of the kernel density estimator: a survey, *Statistical Papers*, *53*, 1–21.

Wilcoxon, F. (1945), Individual comparisons by ranking methods, *Biometrics Bulletin*, *1*(6), 80–83.

Yakowitz, S. J., and J. D. Spragins (1968), On the identifiability of finite mixtures, *Ann. Math. Statist.*, *39*(1), 209–214, doi:10.1214/aoms/1177698520.

Young, N. (1988), *An Introduction to Hilbert Space*, Cambridge mathematical textbooks, Cambridge University Press.