Covariate Structure as Weak Supervision in Estimation Problems

by

Alexander Ritchie

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in the University of Michigan
2026

Doctoral Committee:

Professor Clayton Scott, Co-Chair
Professor Laura Balzano, Co-Chair
Collegiate Professor Jeffrey Fessler
Associate Professor Yuekai Sun

Alexander Ritchie
aritch@umich.edu
ORCID iD: 0000-0002-8599-8222

© Alexander Ritchie 2026

DEDICATION

To Michelle, whose love and support sustained me through the most difficult times.

ACKNOWLEDGMENTS

I have been extraordinarily fortunate in my advisors, Clay Scott and Laura Balzano, who guided this work in equal measure. Beyond the technical, both of them modeled a patience and generosity that I will carry with me. There are few people in this world who have extended to me such kindness and grace. For this and everything else they have done, I am grateful.

I extend my most heartfelt thanks to the other members of my committee, Jeffrey Fessler and Yuekai Sun, for their time, their careful reading, and their pointed questions. Their feedback and perspectives helped make this work more clear and precise than it would otherwise have been.

This dissertation reflects the contributions of several collaborators. I thank Robert Vandermeulen, Jessica Jiang, and Daniel Kessler for the work we did together, and for many conversations that clarified my own thinking along the way. I am also grateful to Chandra Sripada, whose perspective broadened the questions I was willing to ask and connected this methodological work to problems that matter in practice.

I want to thank my labmates and the broader community at Michigan, whose camaraderie made the long stretches of this work far more bearable.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGMENTS iii

LIST OF FIGURES vii

LIST OF TABLES viii

ABSTRACT ix

1 Introduction 1

1.1 Nonparametric Mixture Models 1
1.2 Supervised Principal Component Analysis 2
1.3 Organization of Document 2

2 Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations 4

2.1 Introduction 4
2.2 Background and Previous Work 5
2.3 Notation 7
2.4 Problem Statement 7
2.5 A Weighted Kernel Density Estimator 8
2.6 Theoretical Results 10
2.7 Optimization 11
2.8 Experiments 12
2.9 Conclusion 15
2.10 Supplement 15
2.10.1 Additional Experimental Details 15
2.10.2 Optimization Details 17
2.10.3 Proof and General Form of Theorem 1 24
2.10.4 Proof and General Form of Theorem 2 29
2.10.5 Background on the Grouped Sample Setting and Proof of Theorem 3 . . 34
2.10.6 General Version of Corollary 1 43

3 Consistent Estimation of Nonparametric Mixtures of Regressions from Grouped Observations 44

3.1 Introduction 44
3.2 Problem Setup 45

3.2.1 Estimation Problem . . . 46
3.2.2 EM Algorithm . . . 46
3.2.3 Kernel Method . . . 48
3.3 Statistical Theory . . . 51
3.3.1 Uniform deviation bound . . . 51
3.3.2 Consistency for the kernel method . . . 53
3.3.3 Consistent Recovery of Mixture Components . . . 54
3.3.4 Identifiability of NMR . . . 54
3.4 Conclusion . . . 55
3.5 Supplement . . . 55
3.5.1 Proof of Theorem 5 . . . 55
3.5.2 Proof of Theorem 6 . . . 57
3.5.3 Proof of Theorem 7 . . . 70
3.5.4 Proof of Theorem 5 . . . 77
3.5.5 Proof of Theorem 8 . . . 78
3.5.6 Proof of Identifiability Results . . . 80

4 Supervised PCA: A Multiobjective Approach . . . 85

4.1 Introduction . . . 85
4.1.1 Notation . . . 86
4.2 Background and Related Work . . . 86
4.2.1 PCA . . . 87
4.2.2 Supervised Dimension Reduction Methods . . . 87
4.2.3 Theory . . . 89
4.3 Method . . . 90
4.3.1 Problem Reformulation . . . 90
4.3.2 Manifold Geometry Necessities . . . 91
4.3.3 Riemannian Gradient Descent . . . 92
4.3.4 Kernel SPCA . . . 93
4.4 Theoretical Results . . . 93
4.4.1 Convergence to a Stationary Point . . . 94
4.4.2 Characterization of Critical Points . . . 95
4.4.3 Results for Squared-Error . . . 95
4.5 Experiments . . . 97
4.5.1 Prediction Performance . . . 97
4.5.2 Visualizing the Learned Features . . . 99
4.5.3 HCP . . . 100
4.5.4 Evaluating Pareto Optimality . . . 102
4.6 Conclusion . . . 103
4.7 Supplement . . . 104
4.7.1 Theorem Proofs . . . 104
4.8 Related Statistical Generative Model . . . 112
4.8.1 Derivation of NLL . . . 112
4.8.2 Kernel Supervised Dimension Reduction . . . 117
4.8.3 Links to Datasets . . . 118

5 Conclusions and Future Work 119

BIBLIOGRAPHY 122

LIST OF FIGURES

FIGURE

2.1 Example cluster assignments of four synthetic datasets by each method. Mean ARI (standard deviation) over 10 runs is shown at the bottom left of each clustering (larger is better).The datasets are overlapping moons (top), Olympic rings (middle), and half-disks (bottom). Half-disks has been annotated to show the true components. 13

2.2 Component density estimate contours produced NDIGO. From left to right: overlapping moons, Olympic rings, half-disks 14

2.3 Receiver operating characteristic for MAGIC gamma ray detection dataset. 15

4.1 Comparison of features learned by (from left to right) PCA, LRPCA, and LFDA, LDA, and PLS on the task of classifying (top) ones and sevens from MNIST and (bottom) dresses and shirts from FMNIST. Shown below each embedding is the resulting classification accuracy on test data. 101

4.2 (4.2a) Correlation between predicted and actual GE as a function of subspace dimension. (4.2b) Comparison of components produced by PCA and LSPCA on HCP data, with the components reshaped to reflect ICN assignments of Power [P$^+$11]. Blue (red) denote component entries that are two standard deviations above (below) the component mean. 102

4.3 Comparison of Pareto optimality of competing methods on the residential dataset in terms of prediction error and variation explained, with subspace dimension $r = 2$. . . . 103

5.1 Stereographic projection of the LSPCA loss landscape for a planted regression model with $r = q = 1$, where the optimal regression subspace is $L_y = 0.25v_1 + 0.75v_2$ and v_i is the ith principal direction of the data matrix X. 121

LIST OF TABLES

TABLE

2.1 Description of datasets. *Quantities after preprocessing. 14
2.2 Russian-troll talking points learned from Twitter dataset. * Censored (racial epithet) . 15
2.3 Mean and standard deviation of topic coherence on Twitter dataset over five experiments. 15
2.4 Out-of-sample ARI and standard deviation over 10 runs on synthetic datasets. 16
2.5 Some popular kernel functions and their associated $\widetilde{k}$. Here $\|\cdot\|_2$ is the Euclidean norm. 18

4.1 Datasets used in the experiments. For classification, q denotes the number of classes; for regression, the dimension of the response. The Music dataset was subsampled to $n = 100$ to obtain an $n<p$ setting (full dataset size in parentheses). 98
4.2 Comparison of mean squared error (regression) or error rate (classification) of competing methods, with standard error. Subspace dimension ($r = 2$) was held fixed for results in the first column of each dataset. For results in the second column, subspace dimension was chosen by 10-fold CV. SPCA methods have names shown in **bold**. For each experiment, the best linear method is shown in ***red***, the best kernel method is shown in ***blue*** 99

ABSTRACT

This dissertation develops methods for recovering latent structure from data under minimal assumptions and with provable guarantees, spanning two settings: the nonparametric estimation of mixture models and supervised dimension reduction.

Mixture models capture heterogeneity arising from sub-populations within a larger population, but their identifiability typically requires restrictive parametric assumptions that real data rarely satisfy, and general nonparametric mixtures are not identifiable from i.i.d. data. Building on the result that any mixture model becomes identifiable given grouped observations of sufficient size, we develop an estimator for nonparametric mixture models in this setting and prove consistent recovery of arbitrary components. We then extend the approach to mixtures of nonlinear regressions and show that the proposed method consistently recovers mixing weights and mean functions.

Turning to supervised dimension reduction, we revisit supervised principal component analysis (SPCA). Prior methods evaluate only prediction error and ignore variation explained, the quantity PCA inherently maximizes. We formulate SPCA as an optimization problem over a product manifold that simultaneously maximizes variation explained and minimizes prediction error, prove convergence of the proposed algorithm to a stationary point, and characterize its local minima. Empirically, the method dominates existing approaches in both prediction error and variation explained.

CHAPTER 1

Introduction

1.1 Nonparametric Mixture Models

Mixture models are a common tool in statistics and machine learning to account for group-level heterogeneity, i.e., the existence of sub-populations within a larger population. In particular, mixture models are used in situations where the properties of sub-populations are of interest, but only measurements of the pooled population are available.

A key issue in mixture modeling is identifiability, i.e., the existence of a unique best representation among the class of models considered. In this document, usage of the term identifiable ignores the trivial nonidentifiability associated with permutations of the component labels, which cannot be eliminated. For reasons of identifiability, the majority of the work in mixture modeling concerns mixtures of parametric probability distributions. For example, the well known Gaussian mixture model has long been known to be identifiable when the cumulative distribution functions of the components are linearly independent [YS68]. However, real-world data is unlikely to satisfy these parametric assumptions. Unfortunately, the general mixture model with nonparametric components is not identifiable from iid data without additional assumptions.

Given access to a type of grouped observations rather than iid observations, it has been shown that *any* mixture model is identifiable, without additional assumptions [VS19]. In Chapter 2, we develop an estimator for nonparametric mixture models in this setting and show consistent recovery of arbitrary mixture components from grouped observations of sufficient size.

Mixtures of regressions are a special case of mixture models where a response variable follows a mixture distribution, while no distributional assumptions are made on the covariates. Such models arise naturally in applications where measurements may correspond to one of several latent groups, each with a different response model. The mixtures of linear regressions problem is a classical one. The corresponding nonlinear problem has received little attention in the literature, but the work introduced in Chapter 2 suggests an approach based on grouped observations. In Chapter 3, we present work on the nonparametric mixtures regressions problem and show theoretically that

the proposed approach consistently recovers the mixture components.

1.2 Supervised Principal Component Analysis

Principal component analysis (PCA) was first formulated by Karl Pearson in 1901 [Pea01] and later reinvented by Harold Hotelling [Hot33]. Geometrically, it can be thought of as the problem of finding an affine subspace of best fit to a collection of points in the squared error sense. PCA is a ubiquitous technique for unsupervised dimension reduction and is often used to alleviate the curse of dimensionality in downstream supervised learning problems or for exploratory data analysis and visualization. In the former case, it is implicitly assumed that directions of high variation in the data contain all or most of the information necessary for the prediction problem of interest. Of course, this is not always the case.

There have been a number of attempts to create a method for supervised PCA (SPCA). Unfortunately, there is no obvious or single-best way to add supervision to the inherently unsupervised problem that is PCA. Importantly, prior work on SPCA has only measured performance in terms of prediction error and has not evaluated the proportion of variation explained, the metric inherently maximized by PCA. In chapter 4, we propose formulating SPCA as an optimization problem over a product manfiold that maximizes variation explained while minimizing prediction error. For the proposed algorithm, we prove convergence to a stationary point and characterize local minima. Experimentally, our method is found to dominate existing approaches, i.e., it outperforms them in terms of both prediction error and variation explained.

1.3 Organization of Document

This dissertation consists of three major chapters. Chapters 2 and 3 cover my work on nonparametric mixture models and nonparametric mixtures of regressions, while chapter 4 covers my work on SPCA. In particular, these chapters are based on the following academic works, either published or in submission:

Chapter 2:

- Alexander Ritchie*, Robert A. Vandermeulen*, and Clayton Scott. Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations. In NeurIPS 2020.

*Equal contribution.

Chapter 3:

- Alexander Ritchie, Jessica Jiang, Laura Balzano, and Clayton Scott. Consistent Estimation of Nonparametric Mixtures of Regressions from Grouped Observations. In preparation for submission.

Chapter 4:

- Alexander Ritchie, Clayton Scott, Laura Balzano, Daniel Kessler, and Chandra S. Sripada. Supervised Principal Component Analysis via Manifold Optimization. In IEEE Data Science Workshop 2019.
- Alexander Ritchie, Clayton Scott, Laura Balzano, Daniel Kessler, and Chandra S. Sripada. Supervised PCA: A Multiobjective Approach. In preparation for submission.

CHAPTER 2

Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations

2.1 Introduction

In statistics and machine learning, finite mixture models are often used to describe the distribution of subpopulations within a larger population. A finite mixture model can be written

$$p = \sum_{m=1}^{M} w_m^* p_m^*, \tag{2.1}$$

where $w_m^* > 0$ are mixing weights such that $\sum_{m=1}^{M} w_m^* = 1$, and p_m^* are probability densities. Without additional assumptions, the mixture model p is not identifiable from iid data. Typically, identifiability is ensured by restricting the p_m^* to some family of parametric distributions. Restricting the p_m^* to be Gaussian yields the Gaussian mixture model (GMM) which is identifiable when the cumulative distribution functions (CDFs) of the mixture components are linearly independent [ABG+14, YS68], e.g., when the component means are distinct.

Most work on estimating mixture models assumes an iid sampling scheme. In this work we examine an alternative sampling scheme where observations occur in iid groups. Each group is generated by sampling a component $m \in [M]$ according to w_m^*, and then drawing N iid observations from p_m^*.

Recent work has shown that *any* finite mixture model is identifiable given grouped observations of sufficient size [VS19]. In the worst case, any finite mixture model with M components is

The work in this chapter is joint with Robert A. Vandermeulen (equal contribution) and Clayton Scott. Specifically, the algorithmic and experimental work is my own, Theorems 1 and 2 are joint work between myself and Clayton, and Theorem 3 is due to Robert. This work is published as *Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations* at NeurIPS 2020.

identifiable given groups of size $N \geq 2M-1$. It was also shown that, if the underlying components of the mixture model are *jointly irreducible* [BS14], then the mixture is identifiable given paired observations ($N = 2$). This framework provides a setting where the potential exists to recover nonparametric and highly overlapping mixture components. As of yet, however, no general theory or algorithms are known for this estimation problem.

This work makes the following contributions. We introduce a novel variant of the kernel density estimator that yields statistically consistent estimates of *any identifiable* nonparametric mixture model (NoMM) from grouped observations. To prove this result, we establish an oracle inequality for weighted kernel density estimators. We also establish a general result showing that consistent estimation (with an estimator possessing a natural factored form) of the distribution on groups implies consistent estimates of the underlying components when the NoMM is identifiable. The only additional condition imposed by our theory is that the p_m^* be square integrable. In the case of $N = 2$, we offer an efficient algorithm and demonstrate its effectiveness on several datasets.

We study two applications where paired observations naturally arise. The first is nuclear source detection, where nuclear particles interact with a detector to produce some form of measurement. A critical challenge in this application is to classify incoming particles as belonging to source or background. Because of changing environments, training data are typically not available, and these two classes also have substantial overlap. By positioning two detectors side-by-side, it is possible to simultaneously measure two particles from the same (unknown) class.

We also apply our method to topic modeling of Twitter data. Since tweets usually express a small set of very closely related ideas, words in tweets contain common underlying semantic information. The pairing of words has the potential to encode this semantic information in a way that accounts for context. The proposed method, which operates on continuous word embeddings, allows for flexible modeling of the distributions of topics over words using static word embeddings [PSM14]. Furthermore our method does not require anchor words, allows for substantial overlap of topics without loss of identifiability, and can be trained using documents with as few as two words *without* any document aggregation [GPW$^+$19, QCWW17]. While other works have explored topic modeling with word embeddings, which we call continuous topic modeling, most either impose parametric assumptions or are not suited for very short texts. To our knowledge, this is the first work to consider a nonparametric approach to continuous topic modeling of very short texts.

2.2 Background and Previous Work

Much of the literature concerning NoMMs falls in the category of Bayesian nonparametrics, a thorough summary of which is given in [XLZ19]. Typically, mixture models in this setting do not assume that the number of mixture components is known, and instead assume that the mix-

ture components are from a known parametric family of distributions. An in-depth treatment of Bayesian NoMMs (BNoMMs) can be found in [Lin95]. The parametric assumptions on the mixture components have been relaxed in [ADRX18], but the identifiability results impose regularity and separation conditions on the components. We mention BNoMMs only for completeness and emphasize that our work considers an alternative setting where the number of mixture components is known, but few to no assumptions are made on the mixture components themselves.

Mixture models are often utilized to solve the clustering problem. Parametric mixture models, such as GMMs, are able to capture overlapping clusters. Most clustering algorithms, however, such as k-means [For65, Llo82], DBSCAN [EKSX96], and spectral clustering [JM00, Lux07], assume clusters are non-overlapping and hence fail when clusters overlap. The grouped observation setting considered in this work is known in the clustering literature as *clustering with instance-level constraints* [WC00, WCRS01, WBD06]. A survey of constrained clustering is given in [GCFL20]. Grouped observations correspond to so-called must-link constraints, where two or more observations are known, through expert knowledge or some other means, to belong to the same cluster. Most constrained clustering approaches cannot model overlapping clusters effectively [ST10].

There is relatively little work on mixture modeling with nonparametric components, and to our knowledge no prior work addresses the incorporation of instance-level constraints in the NoMM setting. Mallapragada et al. [MJJ10] use a mixture of kernel density estimators to estimate a NoMM, but do not address identifiability or provide statistical guarantees. Aragam et al. [ADRX18] prove identifiability of NoMMs under regularity and separation conditions on the components. They provide a simple algorithm that gives Bayes optimal cluster assignments in the limit, but they does not guarantee on recovery of the underlying components. Schiebinger et al. [SWY15] study kernelized spectral clustering and characterize recoverability of components with small overlap. Zheng and Wu [ZW19] establish consistent estimation of NoMMs under the assumption that mixture components have independent marginals. Bao et al. [BNS18] consider the related problem of "similar-unsupervised" binary classification, which assumes access to unlabeled data in addition to must-link constraints.

In the grouped observation setting, previous works on multi-view models can be adapted to prove identifiability results and give algorithms to recover mixture model components. When the mixture components are linearly independent it has been shown that three observations per group is sufficient to yield identifiability as well as an algorithm to provably recover the components [AGH$^+$14, AMR09]. We note that these approaches require three observations per group, while the proposed method works with as few as two observations per group. This difference amounts to performing kernel density estimation in three times the ambient dimension versus two. With the instability of KDEs in high dimension, the reduction to groups of size two can be very meaningful in practice. Furthermore, in applications like nuclear particle classification, triples may be

exceedingly rare or difficult to measure. For discrete data, similar results from nonnegative matrix factorization exist under joint irreducibility with two observations per group [AGKM12], and algorithms have been proposed to recover arbitrary mixture models with M components given $2M-1$ observations per group [RSS14, VS19].

2.3 Notation

For $1 \leq p < \infty$ denote $\|f\|_p := (\int_{\mathbb{R}^d} |f(x)|^p dx)^{1/p}$, and $L^p := \{f : \mathbb{R}^d \to \mathbb{R} : \|f\|_p < \infty\}$. The transpose of a matrix A will be written A'. Random variables will be referred to by capital letters, and instances of random variables will be referred to by the corresponding lowercase letter. We represent the set of positive integers $\{1, 2, \ldots, M\}$ by $[M]$. We let Δ^R be the probability simplex in $\mathbb{R}^R$. We denote the M-fold Cartesian product of a set with a subscript, e.g., $\Delta_M^R = \underbrace{\Delta^R \times \cdots \times \Delta^R}_{M}$.

2.4 Problem Statement

We precisely introduce the grouped observation setting, review known identifiability results, and formalize the estimation problem. The standard sampling procedure for a mixture model of the form $p = \sum_{m=1}^M w_m^* p_m^*$ can be viewed as a two step process wherein one samples a mixture component p_m^* with probability w_m^* and then observes one draw from that distribution $X \sim p_m^*$. The *grouped observation setting* considers an alternative sampling scheme where, after selecting a mixture component p_m^*, instead of only drawing a single observation, a group of observations $\mathbf{X} = (X_1, \ldots, X_N)$ are drawn iid from p_m^*. As in a standard mixture model, one does not know *a priori* from which mixture component a grouped observation is sampled. Repeating this n times, one's data consists of n groups of N observations per group $\mathbf{X}_1 = (X_{1,1}, \ldots, X_{1,N}), \ldots, \mathbf{X}_n = (X_{n,1}, \ldots, X_{n,N})$. The distribution on groups is $\mathbf{X} \overset{iid}{\sim} \sum_{m=1}^M w_m^* {p_m^*}^{\times N}$, where ${p_m^*}^{\times N} : \mathbb{R}_N^d \to \mathbb{R}$ denotes the product density such that ${p_m^*}^{\times N}(y_1, y_2, \ldots, y_N) = p_m^*(y_1) p_m^*(y_2) \ldots p_m^*(y_N)$. Note that when $N = 1$ this is simply a standard mixture model.

Vandermeulen and Scott [VS19] characterized identifiability from grouped observations for mixtures of general probability measures. A mixture model $p = \sum_{m=1}^M w_m^* p_m^*$ is said to be N-*identifiable* if p cannot be expressed $p = \sum_{m=1}^{M'} w_m' p_m'$ for some distinct mixture model such $M' \leq M$ and $\sum_{m=1}^M w_m^* {p_m^*}^{\times N} = \sum_{m=1}^{M'} w_m' {p_m'}^{\times N}$. In words, N-identifiability of p means there is no other mixture model with M or fewer components that induces the same distribution on groups. They show that a general mixture model is N-identifiable from grouped observations provided $N \geq 2M - 1$, and that this cannot be improved without imposing restrictions on the components. The result places no assumptions whatsoever on the components.

In practice, the bound of $2M - 1$ is probably pessimistic, and the most useful cases are likely when N is small, say two or three. The authors of [VS19] also show that if the p_m^* are *jointly irreducible* (*linearly independent*), then the mixture is N-identifiable for $N = 2$ ($N = 3$). A collection of probability densities $\mu_1, \mu_2, \ldots, \mu_M$ is said to be *jointly irreducible* (JI) if $\sum_{m=1}^{M} c_m \mu_m$ is never a valid density whenever some $c_m < 0$. JI is satisfied, for example, if the support of each mixture component has some subset of positive measure that does not intersect the supports of the other mixture components (a continuous analogue of the anchor word assumption). This is not necessary, however; JI is still possible if all densities have the same support. In the remainder of the paper we focus on the setting of $N = 2$, not only because JI provides a flexible nonparametric condition where paired observations suffice, but also because the notation for our estimator becomes cumbersome when $N > 2$. Our theory generalizes easily to $N > 2$, and these details are described in Appendices 2.10.3, 2.10.4, and 2.10.5.

The paired observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with $\mathbf{X}_i = (X_{i,1}, X_{i,2}) \in \mathbb{R}^d \times \mathbb{R}^d$ are iid and have density

$$q(x, x') \coloneqq \sum_{m=1}^{M} w_m^* p_m^*(x) p_m^*(x') \quad x, x' \in \mathbb{R}^d. \tag{2.2}$$

We assume M is known. Our goal is to consistently estimate w_m^* and p_m^* when p is identifiable.

2.5 A Weighted Kernel Density Estimator

Our overall strategy is to first devise a consistent estimator of q, the density on pairs, where the estimator has a factorized form reflecting the group sampling scheme. In the next section we prove that if an estimator for q is consistent, and p is identifiable, then the components comprising our estimator converge to the true components.

Let $k : \mathbb{R}^d \to \mathbb{R}$ be a function, called a *kernel*, such that $k \geq 0$ and $\int k(x)dx = 1$. An example is the Gaussian kernel $k(x) = (2\pi)^{-1/2} \exp(-\|x\|^2/2)$. For $\sigma > 0$, define $k_\sigma(x, x') := \sigma^{-d} k((x - x')/\sigma)$. We refer to the second argument of k_σ as the *center* of the kernel. A *weighted kernel density estimator* (wKDE) for a density on $\mathbb{R}^d$, but constructed from the paired observations $\mathbf{X}_i$, has the form

$$p(x; \theta) = \sum_{r=1}^{n} \sum_{r'=1}^{2} \theta_{r,r'} k_\sigma(x, X_{r,r'}),$$

where $\theta_{r,r'}$ is the element of $\theta = [\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{n,1}, \theta_{n,2}]'$ corresponding to the weight of the kernel centered at $X_{r,r'}$. We propose to model the mixture components as wKDEs. Specifically, given n

paired observations, we consider estimators of q of the form

$$q_{w,\alpha}(x, x') = \sum_{m=1}^{M} w_m p(x; \alpha_m) p(x'; \alpha_m), \tag{2.3}$$

where $w = [w_1, w_2, \ldots, w_M]' \in \Delta^M$, $\alpha_m = [\alpha_{m,1,1}, \alpha_{m,1,2}, \ldots, \alpha_{m,n,1}, \alpha_{m,n,2}]' \in \Delta^{2n}$ for all $m \in [M]$, with $\alpha_{m,r,r'}$ corresponding to the weight of the kernel centered at $X_{r,r'}$ in the estimate of the mth mixture component, and $\alpha \coloneqq (\alpha_1,\ \alpha_2, \ldots,\ \alpha_M) \in \Delta_M^{2n}$.

To select the parameters (w, α), we propose to minimize the integrated square error (ISE) of $q_{w,\alpha}$ given by$\|q - q_{w,a}\|_2^2 \coloneqq \int [q(x, x') - q_{w,a}(x, x')]^2 dx dx'$. Expanding the ISE gives

$$\|q - q_{w,a}\|_2^2 = \int q_{w,\alpha}^2(x, x') dx dx' - 2 \int q_{w,\alpha}(x, x') q(x, x') dx dx' + \cancel{\int q^2(x, x') dx dx'}^{\text{const.}}.$$

Since the final term is constant with respect to w and α, we focus on minimizing the first two terms which we call the truncated ISE (TISE) and denote by $J(w, \alpha)$. Substituting the definition of $q_{w,\alpha}$ in the TISE yields

$$J(w, \alpha) \coloneqq \int q_{w,\alpha}^2(x, x') dx dx' - 2 \sum_{m=1}^{M} \sum_{r=1}^{n} \sum_{r'=1}^{2} \sum_{s=1}^{n} \sum_{s'=1}^{2} w_m \alpha_{m,r,r'} \alpha_{m,s,s'} h(r, r', s, s'), \tag{2.4}$$

where $h(r, r', s, s') \coloneqq \int k_\sigma(x, X_{r,r'}) k_\sigma(x', X_{s,s'}) q(x, x') dx dx'$. Since q is unknown, the ISE and therefore $J(w, a)$ cannot be calculated directly. Noting that $h(r, r', s, s')$ is an expectation, we estimate this term using a hybrid leave-one-out/leave-two-out (LOO/LTO) estimator

$$\hat{h}(r, r', s, s') \coloneqq \begin{cases} \frac{1}{n-2} \sum_{i \in [n] \setminus \{r,s\}} k_\sigma(X_{i,1}, X_{r,r'}) k_\sigma(X_{i,2}, X_{s,s'}) & r \neq s \\ \frac{1}{n-1} \sum_{i \in [n] \setminus \{r\}} k_\sigma(X_{i,1}, X_{r,r'}) k_\sigma(X_{i,2}, X_{s,s'}) & r = s \end{cases}.$$

In this manner we have the empirical TISE

$$\hat{J}(w, \alpha) \coloneqq \int q_{w,\alpha}^2(x, x') dx dx' - 2 \sum_{m=1}^{M} \sum_{r=1}^{n} \sum_{r'=1}^{2} \sum_{s=1}^{n} \sum_{s'=1}^{2} w_m \alpha_{m,r,r'} \alpha_{m,s,s'} \hat{h}(r, r', s, s'). \tag{2.5}$$

With all the notation in place, our estimate of the nonparametric mixture model is determined by

$$(\hat{w}, \hat{\alpha}) \coloneqq \underset{w \in \Delta^M,\ \alpha \in \Delta_M^{2n}}{\arg\min} \hat{J}(w, \alpha), \tag{2.6}$$

where $\hat{w}_m$ are the mixing weights and $p(x; \hat{\alpha}_m)$ are the mixture components for $m \in [M]$. The theoretical results presented in Section 2.6 concern the behavior of the minimizer of (2.6). We

show not only that the empirical TISE minimizing estimator $\hat{q} \coloneqq q_{\hat{w},\hat{\alpha}}$ consistently estimates q, but its components also consistently estimate the underlying mixture model if it is identifiable.

2.6 Theoretical Results

In this section we state our assumptions and main results. Formal proofs are given in the Appendices 2.10.3, 2.10.4, and 2.10.5. Our overall approach is to first show that the proposed $\hat{q}$ is a consistent estimate of q (Theorems 1 and 2). We then show that if p in (2.1) is identifiable, then the components $p(x;\hat{\alpha}_m)$ defining $\hat{q}$ are consistent estimates of p^*_m, as are the $\hat{w}_m$ for w^*_m (Theorem 3).

We assume throughout this section that $p^*_m \in L^2$ for all m. We also require that the kernel k satisfy two additional conditions: $k \in L^2$ and $k \leq C_k$ for some constant $C_k < \infty$.

We begin with an oracle inequality, which shows that our estimator selects an approximately optimal member of our model class.

Theorem 1. *Let $\epsilon > 0$ and set $\delta = 8(n^2-n)\exp\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\} + 8n\exp\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\}$. With probability at least $1-\delta$ the following holds: $\|q - q_{\hat{w},\hat{\alpha}}\|_2^2 \leq \inf_{w\in\Delta^M,\, \alpha\in\Delta_M^{2n}} \|q - q_{w,\alpha}\|_2^2 + \epsilon$.*

Proof Sketch. The estimators $\hat{h}$ are constructed so that they are sums of independent random variables, allowing us to apply Hoeffding's inequality to show that each $\hat{h}$ concentrates around its h. Then using basic inequalities (triangle inequality, union bound) and the simplex constraints on w and α, we show that $\hat{J}(w,\alpha)$ concentrates around $J(w,\alpha)$ uniformly over the parameter space. □

The next result uses Theorem 1 to establish that $\hat{q}$ is a consistent estimate of q in the L^1 norm.

Theorem 2. *If $\sigma \to 0$ and $\frac{n\sigma^{4d}}{\log n} \to \infty$ as $n \to \infty$, then $\|q - q_{\hat{w},\hat{\alpha}}\|_1 \xrightarrow{a.s.} 0$.*

Proof Sketch. We appeal to a result of [GM90] showing that if $\int \hat{q} = 1$, which it does in our case, then strong consistency (i.e., a.s. convergence) of a density estimator in L^2 implies strong consistency in L^1. To show strong consistency in L^2, from Theorem 1 it suffices to exhibit $w \in \Delta^M$ and $\alpha \in \Delta_M^{2n}$ such that $\|q - q_{w,\alpha}\|_1 \xrightarrow{a.s.} 0$. For this we take $w = w^*$ and $\alpha = \alpha^*$ such that each α^*_m is uniform on the data points drawn from p^*_m. This makes $p(\cdot\,;\alpha^*_m)$ the usual (uniformly weighted) KDE for p^*_m, which is known to be a strongly consistent estimator. Strong consistency of $\hat{q}$ then easily follows. □

The preceding results hold regardless of whether p in (2.1) is identifiable. The next result states that if p is identifiable, then the estimates $p(\cdot\,;\hat{\alpha}_m)$ comprising $\hat{q}$ are consistent estimates of the true components p^*_m, as are the $\hat{w}_m$ of w^*_m. The result is stated for $N \geq 2$.

Theorem 3. *Let $\sum_{m=1}^{M} w_m p_m$ be an N-identifiable mixture model, and $\sum_{m=1}^{M} \hat{w}_{m,j} \hat{p}_{m,j}$ be a sequence of mixture models such that $\left\| \sum_{m=1}^{M} \hat{w}_{m,j} \hat{p}_{m,j}^{\times N} - \sum_{m=1}^{M} w_m p_m^{\times N} \right\|_1 \to 0$. Then there is a sequence of permutations σ_j so that $\hat{w}_{\sigma_j(m),j} \to w_m$ and $\left\| \hat{p}_{\sigma_j(m),j} - p_m \right\|_1 \to 0$ for all m.*

Proof Sketch. We show that if $\left\| \sum_{m=1}^{M} \hat{w}_{m,j} \hat{p}_{m,j}^{\times N} - \sum_{m=1}^{M} w_m p_m^{\times N} \right\|_1 \to 0$ then the components $\hat{p}_{m,j}$ admit some convergent subsequence, and therefore so do $\hat{p}_{m,j}^{\times N}$. If a subsequence $\hat{p}_{m,j}^{\times N}$ stays away from the components $p_m^{\times N}$ then some subsequence would converge to a component other than some $p_m^{\times N}$. This allows us to construct a mixture model violating N-identifiability, a contradiction. □

This result has been stated in terms of densities for readability, but Appendix 2.10.5 contains a general measure-theoretic version. We may combine Theorems 2 and 3 to establish the following (returning to the setting of $N = 2$). To our knowledge, this is the first result to establish consistent estimation, under any sampling scheme, of NoMMs with substantial overlap.

Corollary 1. *If $\sigma \to 0$ and $\frac{n\sigma^{4d}}{\log n} \to \infty$ as $n \to \infty$, and p is 2-identifiable (e.g., the p_m^* are jointly irreducible), then $\hat{w}_m \overset{a.s.}{\to} w_m^*$ and $\|p(\cdot; \hat{\alpha}_m) - p_m^*\|_1 \overset{a.s.}{\to} 0$, up to a permutation.*

The significance of the result is that joint irreducibility is both a flexible nonparametric assumption, while ensuring identifiability in the case $N = 2$ for which a practical implementation of $\hat{q}$ is possible. We include an analogous result for *all identifiable NoMMs* in Appendix 2.10.6.

2.7 Optimization

In this section we suggest an approach for solving (2.6). We first consider the problem as presented up to this point, which we call the *full problem*. We then consider an approach for speeding up optimization by heuristically choosing a coreset as the kernel centers, which we call the *coreset approach*. In what follows, we assume that $\widetilde{k}_\sigma(z_r, z_u) \coloneqq \int k_\sigma(x, z_r) k_\sigma(x, z_u) dx$ has a closed-form expression or can otherwise be computed efficiently. This assumption is satisfied by many common kernels such as the Gaussian, Cauchy, and Laplacian kernels.

Form of the Optimization Problem. The optimization problem (2.6) can be written

$$\min_{w \in \Delta^M,\, \alpha \in \Delta_M^R} \sum_{k=1}^{M} \sum_{\ell=1}^{M} w_k w_\ell \left(\alpha_k' G \alpha_\ell \right)^2 - 2 \sum_{m=1}^{M} w_m \left(\alpha_m' C \alpha_m \right), \tag{2.7}$$

where the matrices $G, C \in \mathbb{R}^{R \times R}$ will be defined shortly. Details are given in Appendix 2.10.2. In particular, both the full problem and the coreset approach can be written in the form of (2.7),

differing only in the definitions of R and G, C. We therefore propose to use the same optimization approach for both problems. For the full problem, $R = 2n$ and the matrices G and C have the form

$$G_{a,b} = \widetilde{k}_\sigma(X_{\lfloor \frac{a}{2} \rfloor, a \bmod 2}, X_{\lfloor \frac{b}{2} \rfloor, b \bmod 2}) \quad C_{a,b} = \hat{h}(\lfloor \frac{a}{2} \rfloor, a \bmod 2, \lfloor \frac{b}{2} \rfloor, b \bmod 2).$$

Though the problem (2.7) is nonconvex, we observe that a properly initialized alternating projected stochastic gradient descent (APSGD) procedure produces good solutions in practice.

Pseudocode for the APSGD algorithm for solving (2.7) is given in Appendix 2.10.2. We mention that the projections are onto the probability simplex, a decaying step size is used, and stochasticity is introduced via the matrix $C^{(t)}$, which is a mini-batch version of C defined by $C^{(t)}_{a,b} = \frac{1}{|\Omega^{(t)} \setminus \{a,b\}|} \sum_{i \in |\Omega^{(t)} \setminus \{a,b\}|} k_\sigma(X_{i,1}, X_{\lfloor \frac{a}{2} \rfloor, a \bmod 2}) k_\sigma(X_{i,2}, X_{\lfloor \frac{b}{2} \rfloor, b \bmod 2})$, where $\Omega^{(t)}$ is the index set corresponding to the tth mini-batch.

Coreset Approach. KDEs traditionally center kernels at the location of each observation, i.e., $k_\sigma(\cdot, x_{i,i'})$, where $x_{i,i'}$ is the kernel center. Rather than constraining the wKDE to have kernels centered at the observations, we can formulate the optimization problem with R kernel centers $z_r \in \mathbb{R}^d$ for some suitably chosen z_r, which we take to be our coreset. Further details are given in Appendix 2.10.2. We note the per-batch computational complexity for our APSGD algorithm is dominated by the gradient calculations and calculating $C^{(t)}$. If we assume $R > M$, the total complexity is $\mathcal{O}(n_e n(M + d)R^2)$ where n_e is the number of training epochs. Thus, choosing $R \ll 2n$ offers a substantial speed-up.

Initialization. We adopt a spectral initialization scheme. We focus on the full problem for concision, but the coreset approach is similar; further details for both are provided in Appendix 2.10.2. By Lemmas 5.1 and 8.2 of Vandermeulen and Scott [VS19], one can view the standard KDE on the full sample as a symmetric linear operator $T : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$. We use the eigenvectors of T, which are wKDEs on $\mathbb{R}^d$, to form a low-rank approximation of the standard KDE initialize our algorithm. This initialization is a low-rank approximation of the standard KDE.

2.8 Experiments

In this section we compare our coreset approach against several competing methods on a number of real and highly overlapping synthetic datasets. Datasets are described in Table 2.1. We call the proposed method Nonparametric Density estimation of Identifiable mixture models from Grouped Observations (NDIGO). All code and synthetic datasets are publicly available. The MAGIC gamma ray detection dataset [BCG$^+$04] is publicly available via the UCI machine learning repository. The

Available Online: https://github.com/aritchie9590/NDIGO

Russian-troll-tweets Twitter dataset is publicly available through FiveThirtyEight. For NDIGO and MVLVM, we used a Gaussian kernel in all experiments and Scott's rule [Sco15] was used for bandwidth selection. For synthetic experiments, R was selected to yield the initialization with the lowest empirical TISE. R was chosen from $\{10, 20, 30, 40, 50\}$ for both moons datasets, and from $\{60, 70, 80, 90, 100\}$ for the Olympic rings and half-disks datasets. We used $R = 200$ for the MAGIC and Twitter datasets.

Several of the methods we compare against do not produce density estimates, so we evaluate the clustering induced by each method. For constrained clustering methods, we compare against constrained spectral clustering (CSC) [WQD14], and constrained GMM (CGMM) [BDW08]. We also compare against the NoMM methods NPMIX of Aragam et al. [ADRX18] and MVLVM of Song et al. [SADX14]. MVLVM is our most similar competitor as it considers groups of size three. Each constrained clustering algorithm was given access to all pair information. MVLVM was supplied triplets from the training data. NPMIX does not utilize the pair information in any way. Following the literature, we report the clustering results for the training sample. Out-of-sample results are provided in Appendix 2.10.1, but we mention NPMIX is the best performer. Parameters for CSC and NPMIX were optimized w.r.t. a separately generated holdout dataset. Average results over ten runs on the synthetic datasets are shown in Figure 2.1. NDIGO outperforms all methods considered. The synthetic datasets were constructed to have clusters that are non-ellipsoidal in shape with

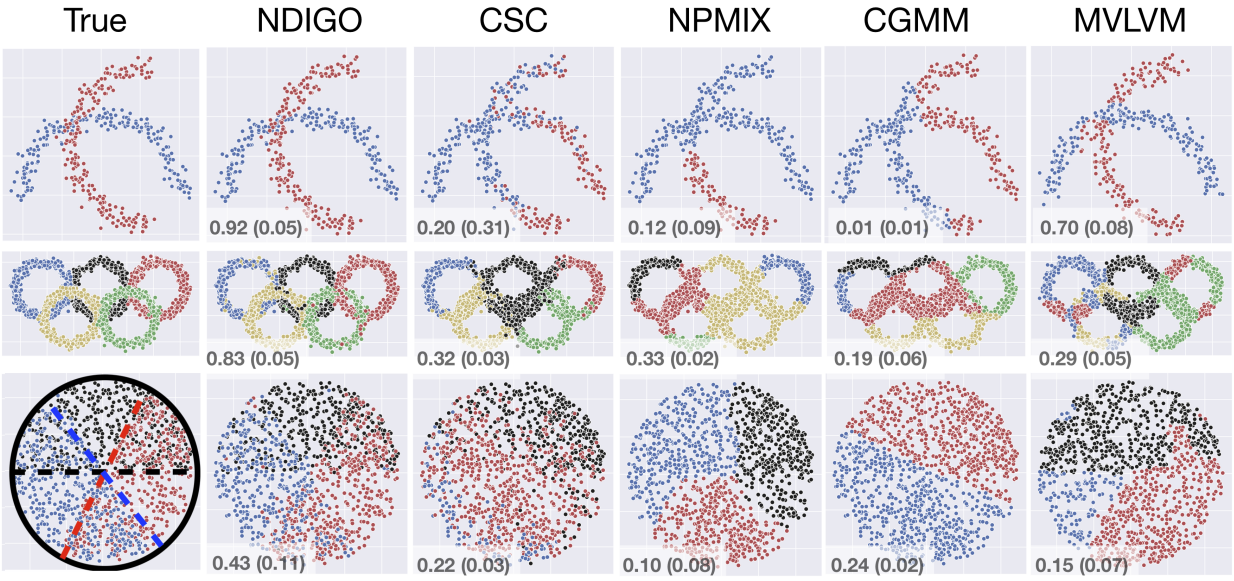


Figure 2.1: Example cluster assignments of four synthetic datasets by each method. Mean ARI (standard deviation) over 10 runs is shown at the bottom left of each clustering (larger is better).The datasets are overlapping moons (top), Olympic rings (middle), and half-disks (bottom). Half-disks has been annotated to show the true components.

Available Online: https://github.com/fivethirtyeight/russian-troll-tweets

substantial overlap between clusters. The clusterings induced by each method are shown in Figure 2.1. Performance is measured in terms of the adjusted Rand index (ARI) [HA85]. We observe that NDIGO gives superior performance across all experiments, especially when clusters have substantial overlap. Density estimates produced by our method for synthetic datasets are shown in Figure 2.2.

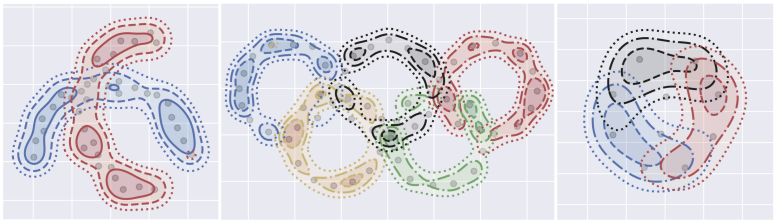


Figure 2.2: Component density estimate contours produced NDIGO. From left to right: overlapping moons, Olympic rings, half-disks

Table 2.1: Description of datasets. *Quantities after preprocessing.

Dataset (2n)	M/d
Ovlp. Moons (400)	2/2
Olympic Rings (2000)	5/2
Half-disks (1200)	3/2
MAGIC $(19,020)$	2/10
Twitter $(3,382,162^*)$	-/10*

Results on the MAGIC dataset are shown in Figure 2.3. The task is to detect gamma radiation events among background radiation. When detecting rare events, the proper performance indicator is given by the receiver operating characteristic (ROC) curve, which plots the true positive rate vs. the false positive rate, parameterized by the threshold of a likelihood ratio test (LRT). Each method was trained using 80% of the available data, and the ROC curve was generated from the remaining 20%. CSC was excluded from this test because it does not produce a density estimate, so a LRT cannot be applied. As an upper bound on possible unsupervised performance, we trained KDEs on each class and plugged the resulting density estimates into an LRT. Previous studies concluded this method, which we call KDE-plugin, is the best approach [BCG+04]. We find NDIGO and CGMM perform very similarly in this experiment, outperforming other methods and approaching KDE-plugin.

We applied NDIGO to topic modeling on the Twitter dataset. Results are shown in Tables 2.2 and 2.3. Details of data preprocessing are deferred to Appendix 2.10.1. After preprocessing, the dataset consisted of $1,691,081$ pairs of 10-dimensional embedded words where each element in a pair comes from the same tweet. Algorithms for competing methods, as described by their respective authors, could not scale to this experiment. Therefore, we compare to recent methods designed for continuous topic modeling of short texts: LF-DMM [NBDJ15], and GPU-DMM[LWZ+16] as implemented by Qiang et al. [QZL+19]. A selection of the top 10 words of topics uncovered by NDIGO is given in Table 2.2. We find that the discovered topics correspond well to other analyses of the dataset [LW20]. Using topic coherence (pointwise mutual information) as an evaluation metric [NLGB10], we observe that NDIGO is competitive with the competing methods.

available online: https://github.com/qiang2100/STTM

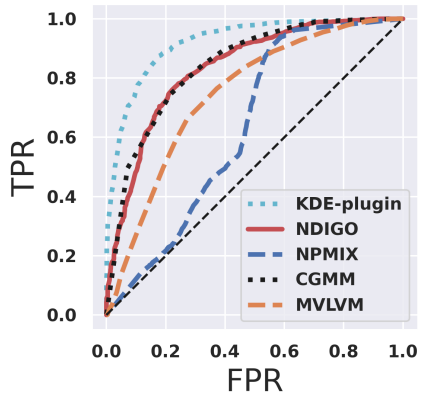


Figure 2.3: Receiver operating characteristic for MAGIC gamma ray detection dataset.

Table 2.2: Russian-troll talking points learned from Twitter dataset. * Censored (racial epithet)

Topic	Selected Top 10 Words
1	dead, man, kill, missing, families, young
2	make, good, better, enough, yet, even, get
3	politics, inside, news, local, police, new, state
4	trial, a*, gentrified, wk, deport, b*
5	businesses, competitive, strength, people, white

Table 2.3: Mean and standard deviation of topic coherence on Twitter dataset over five experiments.

NDIGO	LF-DMM	GPU-DMM
0.521 ± 0.084	0.493 ± 0.018	0.435 ± 0.009

2.9 Conclusion

In this work we introduced a novel variant of the kernel density estimator that yields consistent estimates of any identifiable nonparametric mixture model from grouped observations. We established an oracle inequality for weighted kernel density estimators, and a general consistency result for estimators of the form $q_{w,\alpha}$. Namely, consistent estimation of q implies consistent estimates of the underlying components when the NoMM is identifiable. In the case of $N = 2$, we offer an efficient algorithm and demonstrate its effectiveness on several datasets where traditional approaches fail. Additionally, we show our approach has practical applications in topic modeling with very small documents and nuclear source detection.

2.10 Supplement

2.10.1 Additional Experimental Details

In this section we provide details of the Twitter experiment and out-of-sample results for experiments on synthetic datasets.

2.10.1.1 Out-of-sample Results

Here we provide out-of-sample results for the synthetic experiments shown in the main paper. These are shown in Table 2.4. We make the realistic assumption that pair information is not avail-

able for out-of-sample data. We generate a test dataset 20% the size of the training set according to the distribution of the training data.

Table 2.4: Out-of-sample ARI and standard deviation over 10 runs on synthetic datasets.

Dataset	NDIGO	CSC	NPMIX	CGMM	MVLVM
Overlapping moons	0.705 ± 0.091	0.405 ± 0.283	0.131 ± 0.075	0.002 ± 0.011	0.593 ± 0.229
Olympic Rings	0.607 ± 0.058	0.387 ± 0.033	0.367 ± 0.063	0.127 ± 0.012	0.290 ± 0.053
Half-disks	0.221 ± 0.036	0.215 ± 0.038	0.102 ± 0.095	0.185 ± 0.076	0.127 ± 0.062

2.10.1.2 Preprocessing of Twitter Dataset

Twitter dataset is publicly available through FiveThirtyEight. The data consist of tweets, from a variety Russian-troll twitter accounts, tweeted between 2015 and 2018. We considered all tweets from 2016, a total of $878,878$. We pre-processed the tweets by removing stop words, punctuation, and hyperlinks, followed by a lemmatization step and the removal of any words that were not contained in the vocabulary of the six billion token GloVe word vectors [PSM14]. For completeness, we mention that lemmatization is a common pre-processing step in natural language processing that removes inflectional differences from words by mapping each inflection to a common base form called the lemma. For example, lemmatization will map each of the words dog, dogs, dog's, dogs', and doggy to the word dog. For each tweet, we paired the constituent words uniformly at random without replacement, resulting in $1,691,081$ pairs of words where the words of a given pair come from the same tweet. No other information was retained. We emphasize that a given word from a given tweet will *not* be assigned to more than one pair. However, if a given word appears in multiple tweets (which may not all be about the same topic), it will show up in multiple pairs.

For the embedding step we performed PCA on the pre-trained GloVe 50-dimensional embeddings to obtain 10-dimensional vectors which were used to encode the paired words. The kernel centers were obtained by running mini-batch k-means with $R = 200$ on a uniform random sample of $10,000$ of the $1,691,081$ word pairs, from which the matrix G was also calculated. We then trained on the full $1,691,081$ word pairs, which are utilized in mini-batches through the matrix $C^{(t)}$.

Algorithms for competing methods, as described by their respective authors, could not scale to this experiment. Therefore, we compare to recent methods designed for continuous topic modeling of short texts: LF-DMM [NBDJ15], and GPU-DMM[LWZ+16] as implemented by Qiang et al.

https://github.com/fivethirtyeight/russian-troll-tweets

[QZL+19]. LF-DMM and GPU-DMM were trained on the same preprocessed data as NDIGO, where each of the $1,691,081$ word pairs is considered a unique document. Each of these methods were run with default hyperparameters, as described in the documentation for GPU-DMM and LF-DMM.[2] After training, UCI topic coherence [NLGB10] (which measures pointwise mutual information) was used to evaluate performance. UCI topic coherence uses a reference dataset to estimate word co-occurrence probabilities, which is more robust in the short text setting as very common words in a given topic may never be observed to co-occur. A recent Wikipedia article dump was used for the reference dataset, and is provided with our code.

2.10.2 Optimization Details

In this section we provide details of our algorithm, including the form of the objective function and initialization, for both the full problem and the coreset approach.

Recall the expression for the ETISE

$$\hat{J}(w,\alpha) \triangleq \int q_{w,\alpha}^2(x,x')dxdx' - 2\sum_{m=1}^{M}\sum_{r=1}^{n}\sum_{r'=1}^{2}\sum_{s=1}^{n}\sum_{s'=1}^{2} w_m \alpha_{mrr'}\alpha_{mss'}\hat{h}(r,r',s,s'),$$

where

$$\hat{h}(r,r',s,s') \triangleq \begin{cases} \hat{h}_{\text{LTO}}(r,r',s,s'), & r \neq s \\ \hat{h}_{\text{LOO}}(r,r',s'), & r = s \end{cases}$$

$$\hat{h}_{\text{LOO}}(r,r',r'') \triangleq \frac{1}{n-1}\sum_{i\in[n]\setminus\{r\}} k_\sigma(x_{i,1},x_{r,r'})k_\sigma(x_{i,2},x_{r,r''})$$

$$\hat{h}_{\text{LTO}}(r,r',s,s') \triangleq \frac{1}{n-2}\sum_{i\in[n]\setminus\{r,s\}} k_\sigma(x_{i,1},x_{r,r'})k_\sigma(x_{i,2},x_{s,s'}).$$

For ease of computation, we will rewrite $\hat{J}(w,\alpha)$ in terms of matrix operations. In what follows, we assume that $\widetilde{k}_\sigma(z_r,z_u) \coloneqq \int k_\sigma(x,z_r)k_\sigma(x,z_u)dx$ has a closed-form expression or can otherwise be computed efficiently. Some examples [KS12] are give in Table 2.5.

[2] https://github.com/qiang2100/STTM

Table 2.5: Some popular kernel functions and their associated $\widetilde{k}$. Here $\|\cdot\|_2$ is the Euclidean norm.

Kernel	$k_\sigma(x,x')$	$\widetilde{k}_\sigma(x,x')$
Gaussian	$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$	$k_{\sqrt{2}\sigma}(x,x')$
Cauchy	$\left(\frac{1}{\sqrt{\pi}\sigma}\right)^d \left(\frac{\Gamma((1+d)/2)}{\Gamma(1/2)}\right)\left(\frac{\sigma^2+\|x-x'\|_2^2}{\sigma^2}\right)^{-\frac{1+d}{2}}$	$k_{2\sigma}(x,x')$
Laplacian	$\frac{c_d}{\sigma^d}\exp\left(-\frac{\|x-x'\|_1}{\sigma}\right)$	$\frac{1}{(4\sigma)^d}\prod_{l=1}^d\left(\frac{\sigma+\|x_l-x_l'\|}{\sigma}\right)\exp\left(-\frac{\|x-x'\|_1}{\sigma}\right)$

2.10.2.1 Full Optimization Problem

We begin by examining the first term of $\hat{J}(w,\alpha)$

$$\begin{aligned}
\int q_{w,\alpha}(x,x')^2 dxdx' &= \int \left(\sum_m w_m \sum_r\sum_{r'} \alpha_{m,r,r'} k_\sigma(x,x_{r,r'}) \sum_s\sum_{s'} \alpha_{m,s,s'} k_\sigma(x,x_{s,s'})\right) \\
&\qquad \times \left(\sum_j w_j \sum_u\sum_{u'} \alpha_{j,u,u'} k_\sigma(x,x_{u,u'}) \sum_v\sum_{v'} \alpha_{j,v,v'} k_\sigma(x,x_{v,v'})\right) dxdx' \\
&= \sum_{m,j} w_m w_j \sum_{r,r',u,u'}\sum_{s,s',v,v'} \alpha_{m,r,r'}\alpha_{m,s,s'}\alpha_{j,u,u'}\alpha_{j,v,v'} \\
&\qquad \times \int k_\sigma(x,x_{r,r'})k_\sigma(x,x_{u,u'})dx \int k_\sigma(x',x_{s,s'})k_\sigma(x',x_{v,v'})dx' \\
&= \sum_{m,j} w_m w_j \sum_{r,r',u,u'} \alpha_{m,r,r'}\alpha_{j,u,u'}\widetilde{k}_\sigma(x_{r,r'},x_{u,u'}) \sum_{s,s',v,v'} \alpha_{m,s,s'}\alpha_{j,v,v'}\widetilde{k}_\sigma(x_{s,s'},x_{v,v'}) \\
&= \sum_{m,j} w_m w_j \left(\alpha_m' G a_j\right)^2,
\end{aligned}$$

where $\times$ in the first line is scalar multiplication and G is the kernel matrix of the data and is given by

$$G = \begin{bmatrix}
\widetilde{k}_\sigma(x_{1,1},x_{1,1}) & \widetilde{k}_\sigma(x_{1,1},x_{1,2}) & \cdots & \cdots & \widetilde{k}_\sigma(x_{1,1},x_{n,1}) & \widetilde{k}_\sigma(x_{1,1},x_{n,2}) \\
\widetilde{k}_\sigma(x_{1,2},x_{1,1}) & \widetilde{k}_\sigma(x_{1,2},x_{1,2}) & \cdots & \cdots & \widetilde{k}_\sigma(x_{1,2},x_{n,1}) & \widetilde{k}_\sigma(x_{1,2},x_{n,2}) \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
\widetilde{k}_\sigma(x_{n,1},x_{1,1}) & \widetilde{k}_\sigma(x_{n,1},x_{1,2}) & \cdots & \cdots & \widetilde{k}_\sigma(x_{n,1},x_{n,1}) & \widetilde{k}_\sigma(x_{n,1},x_{n,2}) \\
\widetilde{k}_\sigma(x_{n,2},x_{1,1}) & \widetilde{k}_\sigma(x_{n,2},x_{1,2}) & \cdots & \cdots & \widetilde{k}_\sigma(x_{n,2},x_{n,1}) & \widetilde{k}_\sigma(x_{n,2},x_{n,2})
\end{bmatrix}.$$

Examining the second term of the ETISE yields

$$\begin{aligned}
\mathbb{E}_q\left[q_{w,\alpha}\right] &\approx \sum_{m=1}^{M}\sum_{r=1}^{n}\sum_{r'=1}^{2}\sum_{s=1}^{n}\sum_{s'=1}^{2} w_m \alpha_{m,r,r'}\alpha_{m,s,s'}\hat{h}(r,r',s,s') \\
&= \begin{cases} \frac{1}{n-2}\sum_m \sum_{r,r'}\sum_{s,s'} w_m\alpha_{m,r,r'}\alpha_{m,s,s'}\sum_{i\in[n]\setminus\{r,s\}} k_\sigma(x_{i,1},x_{r,r'})k_\sigma(x_{i,2},x_{s,s'}), & r\neq s \\ \frac{1}{n-1}\sum_m \sum_{r,r'}\sum_{s,s'} w_m\alpha_{m,r,r'}\alpha_{m,s,s'}\sum_{i\in[n]\setminus\{r\}} k_\sigma(x_{i,1},x_{r,r'})k_\sigma(x_{i,2},x_{r,s'}), & r=s \end{cases} \\
&= \sum_{m=1}^{M} w_m\left(\alpha_m' C \alpha_m\right),
\end{aligned}$$

where C is given by

$$C = \begin{bmatrix} \hat{h}(1,1,1,1) & \hat{h}(1,1,1,2) & \cdots & \cdots & \hat{h}(1,1,n,1) & \hat{h}(1,1,n,2) \\ \hat{h}(1,2,1,1) & \hat{h}(1,2,1,2) & \cdots & \cdots & \hat{h}(1,2,n,1) & \hat{h}(1,2,n,2) \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \hat{h}(n,1,1,1) & \hat{h}(n,1,1,2) & \cdots & \cdots & \hat{h}(n,1,n,1) & \hat{h}(n,1,n,2) \\ \hat{h}(n,2,1,1) & \hat{h}(n,2,1,2) & \cdots & \cdots & \hat{h}(n,2,n,1) & \hat{h}(n,2,n,2) \end{bmatrix}.$$

The diagonal blocks of size two of the matrix C use the leave one out estimator, while the other entries use the leave two out estimator.

2.10.2.2 Coreset Approach

In the full problem the matrices $G, C \in \mathbb{R}^{2n\times 2n}$ grow linearly with the data. This can make the proposed optimization problem (2.6) costly to solve, as the complexity of gradient calculations are quadratic in the dimensions of G, C. Additionally, the complexity of evaluating out-of-sample data is quadratic in n for general KDEs. The motivation of the coreset approach is to reduce this complexity.

KDEs traditionally center kernels at the location of each observation, i.e., $k_\sigma(\cdot, x_{i,i'})$, where we call $x_{i,i'}$ the kernel center. Rather than constraining the wKDE to have kernels centered at the observations, we can formulate the optimization problem with R kernel centers $z_r \in \mathbb{R}^d$ for some suitably chosen z_r. Additionally, choosing $R \ll n$ will substantially reduce the complexity of gradient calculations and out-of-sample evaluation. The collection of kernel centers z_r will be our coreset. We don't provide guarantees for the optimality of any particular coreset. The coreset could potentially be chosen as the cluster centers output by some clustering algorithm, some suitable subset of the data, or perhaps via some more principled scheme. In all of our experiments, we chose the coreset to be cluster centers output by mini-batch k-means, where the number of clusters

was chosen to be $R > M$.

For the coreset approach, the ETISE has the same form but the matrices G and C have the form

$$G = \begin{bmatrix} \widetilde{k}_\sigma(z_1, z_1) & \cdots & \cdots & \widetilde{k}_\sigma(z_1, z_R) \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \widetilde{k}_\sigma(z_R, z_1) & \cdots & \cdots & \widetilde{k}_\sigma(z_R, z_R) \end{bmatrix},$$

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} k_\sigma(x_{i,1}, z_1)k_\sigma(x_{i,2}, z_1) & \cdots & \cdots & k_\sigma(x_{i,1}, z_1)k_\sigma(x_{i,2}, z_R) \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ k_\sigma(x_{i,1}, z_R)k_\sigma(x_{i,2}, z_1) & \cdots & \cdots & k_\sigma(x_{i,1}, z_R)k_\sigma(x_{i,2}, z_R) \end{bmatrix}$$

This is derived in the same way as the full problem, replacing the kernel centers $x_{r,r'}, x_{s,s'}$ with the coreset z_r, and using the "leave-none-out" estimator in place of the LOO/LTO estimator $\hat{h}$.

2.10.2.3 Algorithm

Though the problem (2.7) is nonconvex, we observe that a properly initialized alternating projected stochastic gradient descent (APSGD) procedure produces good solutions in practice. Pseudocode for the APSGD algorithm for solving (2.7) is given in Algorithm 1. We mention that the projections Π_Δ are onto the probability simplex, a decaying step size $\eta^{(t)}$ is used, and stochasticity is introduced via the matrix $C^{(t)}$, which is a mini-batch version of C defined by $C^{(t)}_{a,b} = \frac{1}{|\Omega^{(t)} \setminus \{a,b\}|} \sum_{i \in |\Omega^{(t)} \setminus \{a,b\}|} k_\sigma(X_{i,1}, X_{\lfloor \frac{a}{2} \rfloor, a \bmod 2}) k_\sigma(X_{i,2}, X_{\lfloor \frac{b}{2} \rfloor, b \bmod 2})$, where $\Omega^{(t)}$ is the index set corresponding to the tth mini-batch.

Algorithm 1 Alternating Projected SGD

init: $\alpha^{(0)}, w^{(0)}, \eta^{(0)}$
procedure APSGD($\alpha^{(0)}, w^{(0)}, \eta^{(0)}$)
    Form G
    for $t = 1, 2, \ldots$ **do**
        Take a minibatch of paired observations indexed by $\Omega^{(t)}$
        Form $C^{(t)}$ from the minibatch according to the definition of C
        $w^{(t)} = \Pi_\Delta(w^{(t-1)} - \eta^{(t)} \nabla_w \hat{J}(w^{(t-1)}, \alpha^{(t-1)}))$
        for $j = 1, \ldots, M$ **do**
            $\alpha_j^{(t)} = \Pi_\Delta(\alpha_j^{(t-1)} - \eta^{(t)} \nabla_{\alpha_j} \hat{J}(w^{(t)}, \alpha_j^{(t-1)}))$

2.10.2.4 Spectral Initialization

We adopt a spectral initialization scheme. First, the initialization is presented for the full problem, then we adapt it to the coreset approach. The idea here is, given some estimator of q, lets say $\tilde{q}$, to find a low rank approximation of $\tilde{q}$

$$\tilde{q}(x, y) \approx \sum_{i=1}^{M} \lambda_i \psi_i(x) \psi_i(y)$$

and then to use λ_i and ψ_i as starting points for our mixture weights and components. We do this by using the full grouped sample data as an estimate of q which we transform into a linear operator and decompose using a functional eigenvector decomposition.

We begin with a standard KDE applied to our full samples using a product kernel:

$$f_\sigma(y, y') = \frac{1}{2n} \sum_{i=1}^{n} k_\sigma(y, x_{i,1}) k_\sigma(y', x_{i,2}) + k_\sigma(y, x_{i,2}) k_\sigma(y', x_{i,1}).$$

Note that we include centers at both $(x_{i,1}, x_{i,2})$ and $(x_{i,2}, x_{i,1})$ so our KDE is symmetric in y, y'.

By Lemmas 5.1 and 8.2 of Vandermeulen and Scott [VS19], f_σ can be viewed as an element of a tensor product space $L^2(\mathbb{R}^d) \otimes L^2(\mathbb{R}^d)$ as follows

$$f_\sigma = \frac{1}{2n} \sum_{i=1}^{n} k_\sigma(\cdot, x_{i,1}) \otimes k_\sigma(\cdot, x_{i,2}) + k_\sigma(\cdot, x_{i,2}) \otimes k_\sigma(\cdot, x_{i,1}).$$

By the Lemmas referenced above, there is a unitary transformation on the KDE f_σ such that it can be viewed as a linear operator $T : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ given by

$$T(g) \coloneqq \sum_{i=1}^{n} k_\sigma(\cdot, x_{i,1}) \langle k_\sigma(\cdot, x_{i,2}), g(\cdot) \rangle_{L^2} + k_\sigma(\cdot, x_{i,2}) \langle k_\sigma(\cdot, x_{i,1}), g(\cdot) \rangle_{L^2}, \quad \forall g \in L^2(\mathbb{R}^d)$$

which is symmetric since it includes $k_\sigma(\cdot, x_{i,1}) \langle k_\sigma(\cdot, x_{i,2}), g(\cdot) \rangle_{L^2}$ and $k_\sigma(\cdot, x_{i,2}) \langle k_\sigma(\cdot, x_{i,1}), g(\cdot) \rangle_{L^2}$ terms. We have removed the $1/(2n)$ coefficient since it will not affect the spectral decomposition. For any $g \in L^2$ the quantity $\langle k_\sigma(\cdot, x_{i,i'}), g(\cdot) \rangle_{L^2}$ will be a finite scalar, so $T(g)$ will be a linear combination of the $k_\sigma(\cdot, x_{i,i'})$. Therefore, eigenvectors of the above linear operator will have the form $g(\cdot) = \sum_{j,j'} \beta_{j,j'} k_\sigma(\cdot, x_{j,j'})$ since T applied to any vector must lie in the span of $k_\sigma(\cdot, x_{j,j'})$.

Evaluating T on vectors of this form (not necessarily an eigenvector) will yield

$$\begin{aligned} T(g) &:= \sum_{i=1}^{n} \Big\{ k_\sigma(\cdot, x_{i,1}) \langle k_\sigma(\cdot, x_{i,2}), \sum_{j,j'} \beta_{j,j'} k_\sigma(\cdot, x_{j,j'}) \rangle_{L^2} \\ &\qquad + k_\sigma(\cdot, x_{i,2}) \langle k_\sigma(\cdot, x_{i,1}), \sum_{j,j'} \beta_{j,j'} k_\sigma(\cdot, x_{j,j'}) \rangle_{L^2} \Big\} \\ &= \sum_i \zeta_{i,1} k_\sigma(\cdot, x_{i,1}) + \zeta_{i,2} k_\sigma(\cdot, x_{i,2}) \end{aligned}$$

where $\zeta_{i,1} = \sum_{j,j'} \beta_{j,j'} \tilde{k}_\sigma(x_{i,2}, x_{j,j'})$, $\zeta_{i,2} = \sum_{j,j'} \beta_{j,j'} \tilde{k}_\sigma(x_{i,1}, x_{j,j'})$ and $\tilde{k}_\sigma(y, y')$ is defined in Appendix 2.10.2.

Define the ordering of the elements of β and ζ by

$$\begin{aligned} \beta &= [\beta_{1,1}, \beta_{1,2}, \beta_{2,1}, \beta_{2,2}, \ldots, \beta_{n,1}, \beta_{n,2}]', \\ \zeta &= [\zeta_{1,1}, \zeta_{1,2}, \zeta_{2,1}, \zeta_{2,2}, \ldots, \zeta_{n,1}, \zeta_{n,2}]'. \end{aligned}$$

Then we have

$$\zeta = \bar{G}\beta,$$

where

$$\bar{G} = \begin{bmatrix} \widetilde{k}_\sigma(x_{1,2}, x_{1,1}) & \widetilde{k}_\sigma(x_{1,2}, x_{1,2}) & \widetilde{k}_\sigma(x_{1,2}, x_{2,1}) & \cdots & \widetilde{k}_\sigma(x_{1,2}, x_{n,1}) & \widetilde{k}_\sigma(x_{1,2}, x_{n,2}) \\ \widetilde{k}_\sigma(x_{1,1}, x_{1,1}) & \widetilde{k}_\sigma(x_{1,1}, x_{1,2}) & \widetilde{k}_\sigma(x_{1,1}, x_{2,1}) & \cdots & \widetilde{k}_\sigma(x_{1,1}, x_{n,1}) & \widetilde{k}_\sigma(x_{1,1}, x_{n,2}) \\ \widetilde{k}_\sigma(x_{2,2}, x_{1,1}) & \widetilde{k}_\sigma(x_{2,2}, x_{1,2}) & \widetilde{k}_\sigma(x_{2,2}, x_{2,1}) & \cdots & \widetilde{k}_\sigma(x_{2,2}, x_{n,1}) & \widetilde{k}_\sigma(x_{2,2}, x_{n,2}) \\ \widetilde{k}_\sigma(x_{2,1}, x_{1,1}) & \widetilde{k}_\sigma(x_{2,1}, x_{1,2}) & \widetilde{k}_\sigma(x_{2,1}, x_{2,1}) & \cdots & \widetilde{k}_\sigma(x_{2,1}, x_{n,1}) & \widetilde{k}_\sigma(x_{2,1}, x_{n,2}) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \widetilde{k}_\sigma(x_{n,2}, x_{1,1}) & \widetilde{k}_\sigma(x_{n,2}, x_{1,2}) & \widetilde{k}_\sigma(x_{n,2}, x_{2,1}) & \cdots & \widetilde{k}_\sigma(x_{n,2}, x_{n,1}) & \widetilde{k}_\sigma(x_{n,2}, x_{n,2}) \\ \widetilde{k}_\sigma(x_{n,1}, x_{1,1}) & \widetilde{k}_\sigma(x_{n,1}, x_{1,2}) & \widetilde{k}_\sigma(x_{n,1}, x_{2,1}) & \cdots & \widetilde{k}_\sigma(x_{n,1}, x_{n,1}) & \widetilde{k}_\sigma(x_{n,1}, x_{n,2}) \end{bmatrix}$$

is a row permutation of G defined for the full problem, obtained by exchanging rows corresponding to the first and second elements of each paired observation. The takeaway is that the coefficients of the eigenvectors of T are given by the right eigenvectors of $\bar{G}$. In particular, we will take the right eigenvectors of $\bar{G}$ corresponding to the M largest real eigenvalues with the intuition that they will capture the dominant modes of T. Note that these eigenvectors will contain real-valued entries by the spectral theorem since T is symmetric. We call these eigenvectors the first, second, and so on. It should be noted that $\bar{G}$ is not a symmetric matrix and so eigenvectors should be found according to, for example, the power iteration or orthogonal iteration. In general the eigenvectors of $\bar{G}$ will have negative entries and not sum to one, so we project these eigenvectors onto the

probability simplex to obtain the non-negative weights for initialization where we take α_1 to the be the projection of the first eigenvector of $\bar{G}$, α_2 to the be the second, and so on. For the initial w_i, We take w_1 to be the first eigenvalue of $\bar{G}$, w_2 to be the second and so on, projecting the resulting $w = [w_1, w_2, \ldots, w_M]'$ onto the probability simplex.

Coreset Approach Initialization Initialization for the coreset approach is similar. Since we assume no relationship between z_i, we don't have the same notion of using paired kernel centers even though our data is still paired. However, we can still write the KDE using a product kernel over the coreset as

$$f_\sigma(y, y') = \frac{1}{R}\sum_{i=1}^{R} k_\sigma(y, z_i)k_\sigma(y', z_i).$$

This KDE is symmetric since the same kernel center is used in each term of the product kernel. Again appealing to Lemmas 5.1 and 8.2 of Vandermeulen and Scott [VS19], f_σ can be viewed as an element of a tensor product space $L^2(\mathbb{R}^d) \otimes L^2(\mathbb{R}^d)$ as follows

$$f_\sigma = \frac{1}{R}\sum_{i=1}^{R} k_\sigma(\cdot, z_i) \otimes k_\sigma(\cdot, z_i).$$

By the Lemmas referenced above, there is a unitary transformation on the KDE f_σ such that it can be viewed as a linear operator $T : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ given by

$$T(g) = \sum_{i=1}^{R} k_\sigma(\cdot, z_i)\langle k_\sigma(\cdot, z_i), g(\cdot)\rangle_{L^2}, \quad \forall g \in L^2.$$

By the same argument as the full problem, the eigenvectors of the above operator have the form $g(\cdot) = \sum_j \beta_j k_\sigma(\cdot, z_j)$. Applying T to a vector of this form (not necessarily an eigenvector), we have

$$\begin{aligned} T(\sum_j \beta_j k_\sigma(\cdot, z_j)) &= \sum_{i=1}^{R} k_\sigma(\cdot, z_i)\langle k_\sigma(\cdot, z_i), \sum_j \beta_j k_\sigma(\cdot, z_j)\rangle_{L^2} \\ &= \sum_{i=1}^{R} \zeta_i k_\sigma(\cdot, z_i), \end{aligned}$$

where $\zeta_i = \sum_j \beta_j k_\sigma(\cdot, z_i)$. In this setting we have the standard ordering, $\beta = [\beta_1, \beta_2, \ldots, \beta_R]'$ and $\zeta = [\zeta_1, \zeta_2, \ldots, \zeta_R]'$. In matrix form, the relationship between β and ζ is given by

$$\zeta = G\beta,$$

where G is as previously defined for the coreset approach. No row permutation is needed in this setting as both centers of our product kernel are the same. From this point, the initialization scheme is essentially the same as for the full problem, but using the eigenvectors and eigenvalues of G. One key difference is that G *is* a symmetric matrix, so the eigenvalues and eigenvectors of G can be found using any standard solver.

2.10.3 Proof and General Form of Theorem 1

Theorem 1 was presented in the main paper for groups of size $N = 2$. Here, we provide the proof for groups of size two, as well as the proof for the more general case of arbitrary group size $N > 2$. One tool we will use is Hoeffding's inequality for independent bounded random variables, which we state here for completeness.

Theorem. *Hoeffding's Inequality: Let $V_1, V_2, \ldots, V_n$ be independent bounded random variables such that $a_i \leq V_i \leq b_i$ with probability one. If $S_n = \sum_{i=1}^n V_i$, then for all $t > 0$*

$$P\left\{\left|S_n - \mathbb{E}\{S_n\}\right| \geq t\right\} \leq 2\exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

2.10.3.1 Proof of Theorem 1: Groups of Size Two

We restate Theorem 1 for convenience.

Theorem 1. *Let $\epsilon > 0$ and set $\delta = 8(n^2 - n)\exp\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\} + 8n\exp\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\}$. With probability at least $1 - \delta$ the following holds:*

$$\|q - q_{\hat{w},\hat{\alpha}}\|_2^2 \leq \inf_{w \in \Delta^M,\ \alpha \in \Delta_M^{2n}} \|q - q_{w,\alpha}\|_2^2 + \epsilon.$$

Proof. Our goal is to bound $|J(w,a) - \hat{J}(w,a)|$ uniformly over $w \in \Delta^M$, $\alpha \in \Delta_M^{2n}$. Recall the following definitions

$$\begin{aligned}
h(r, r', s, s') &\coloneqq \int k_\sigma(x, x_{r,r'}) k_\sigma(x', x_{s,s'}) q(x, x') dx dx' \\
\hat{h}(r, r', s, s') &\coloneqq \begin{cases} \hat{h}_{\text{LTO}}(r, r', s, s'), & r \neq s \\ \hat{h}_{\text{LOO}}(r, r', s'), & r = s \end{cases} \\
\hat{h}_{\text{LOO}}(r, r', r'') &\coloneqq \frac{1}{n-1} \sum_{i \in [n] \setminus \{r\}} k_\sigma(x_{i,1}, x_{r,r'}) k_\sigma(x_{i,2}, x_{r,r''}) \\
\hat{h}_{\text{LTO}}(r, r', s, s') &\coloneqq \frac{1}{n-2} \sum_{i \in [n] \setminus \{r,s\}} k_\sigma(x_{i,1}, x_{r,r'}) k_\sigma(x_{i,2}, x_{s,s'}).
\end{aligned}$$

The use of the leave one out (LOO) and leave two out (LTO) estimators above is to ensure independence so that we will be able to apply Hoeffding's inequality. We have

$$
\begin{aligned}
&P_q\{\sup_{\substack{w\in\Delta_M^M\\ \alpha\in\Delta_M^{2n}}}|J(w,\alpha)-\hat{J}(w,\alpha)|>\frac{\epsilon}{2}\}\\
&\qquad=P_q\Big\{\sup_{\substack{w\in\Delta_M^M\\ \alpha\in\Delta_M^{2n}}}\Big|\sum_{m=1}^{M}w_m\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{r'=1}^{2}\sum_{s'=1}^{2}\alpha_{m,r,r'}\alpha_{m,s,s'}h(r,r',s,s')\\
&\qquad\qquad\qquad-\sum_{m=1}^{M}w_m\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{r'=1}^{2}\sum_{s'=1}^{2}\alpha_{m,r,r'}\alpha_{m,s,s'}\hat{h}(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}\\
&\qquad\le P_q\Big\{\sup_{\substack{w\in\Delta_M^M\\ \alpha\in\Delta_M^{2n}}}\sum_{m}\sum_{r,s}\sum_{r',s'}w_m\alpha_{m,r,r'}\alpha_{m,s,s'}\Big|h(r,r',s,s')-\hat{h}(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}\\
&\qquad\le P_q\Big\{\max_{r,s,r',s'}\Big|h(r,r',s,s')-\hat{h}(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}\\
&\qquad\le\sum_{r,s}\sum_{r',s'}P_q\Big\{\Big|h(r,r',s,s')-\hat{h}(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}.
\end{aligned}
$$

The second step above is due to the triangle inequality, and the penultimate step is due to simplex constraints on w,α. Let $k_i(r,r',s,s') \coloneqq k_\sigma(x_{i,1},x_{r,r'})k_\sigma(x_{i,2},x_{s,s'})$. Noting that $h(r,r',s,s') = \mathbb{E}_{(x_{i,1},x_{i,2})\sim q}\{k_i(r,r',s,s')\}$,

$$
\begin{aligned}
&P_q\Big\{\Big|h(r,r',s,s')-\hat{h}(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}\\
&=\begin{cases}P_q\Big\{\Big|\mathbb{E}_{(x_{i,1},x_{i,2})\sim q}\{k_i(r,r',s,s')\}-\frac{1}{n-2}\sum_{i\in[n]\setminus\{r,s\}}k_i(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}, & r\neq s\\ P_q\Big\{\Big|\mathbb{E}_{(x_{i,1},x_{i,2})\sim q}\{k_i(r,r',s,s')\}-\frac{1}{n-1}\sum_{i\in[n]\setminus\{r\}}k_i(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}, & r=s\end{cases}\\
&=\begin{cases}P_q\Big\{\Big|\frac{1}{n-2}\sum_{i\in[n]\setminus\{r,s\}}\mathbb{E}_{(x_{i,1},x_{i,2})\sim q}\{k_i(r,r',s,s')\}-k_i(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}, & r\neq s\\ P_q\Big\{\Big|\frac{1}{n-1}\sum_{i\in[n]\setminus\{r\}}\mathbb{E}_{(x_{i,1},x_{i,2})\sim q}\{k_i(r,r',s,s')\}-k_i(r,r',s,s')\Big|>\frac{\epsilon}{4}\Big\}, & r=s\end{cases}\\
&=\begin{cases}P_q\Big\{\Big|\sum_{i\in[n]\setminus\{r,s\}}\mathbb{E}_{(x_{i,1},x_{i,2})\sim q}\{k_i(r,r',s,s')\}-k_i(r,r',s,s')\Big|>\frac{(n-2)\epsilon}{4}\Big\}, & r\neq s\\ P_q\Big\{\Big|\sum_{i\in[n]\setminus\{r\}}\mathbb{E}_{(x_{i,1},x_{i,2})\sim q}\{k_i(r,r',s,s')\}-k_i(r,r',s,s')\Big|>\frac{(n-1)\epsilon}{4}\Big\}, & r=s\end{cases}
\end{aligned}
$$

The terms $k_i(r,r',s,s')$ are independent random variables due to use of the LOO/LTO estimator. By assumption, $0\le k_i(r,r',s,s')\le C_k^2\sigma^{-2d}$ so the k_i are bounded for fixed $\sigma>0$. We apply

Hoeffding's inequality

$$P_q\left\{\left|h(r,r',s,s')-\hat{h}(r,r',s,s')\right|>\frac{\epsilon}{4}\right\}\leq\begin{cases}2\exp\{-\frac{2(n-2)^2\epsilon^2}{16(n-2)C_k^4\sigma^{-4d}}\}, & r\neq s\\ 2\exp\{-\frac{2(n-1)^2\epsilon^2}{16(n-1)C_k^4\sigma^{-4d}}\}, & r=s\end{cases}$$
$$\leq\begin{cases}2\exp\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\}, & r\neq s\\ 2\exp\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\}, & r=s\end{cases}. \tag{2.8}$$

Substituting backward we obtain the desired upper bound

$$P_q\{\sup_{\substack{w\in\Delta^M\\ \alpha\in\Delta_M^{2n}}}|J(w,\alpha)-\hat{J}(w,\alpha)|>\frac{\epsilon}{2}\}\leq\sum_{r,s}\sum_{r',s'}\begin{cases}2\exp\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\}, & r\neq s\\ 2\exp\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\}, & r=s\end{cases}$$
$$=8(n^2-n)\exp\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\}+8n\exp\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\}.$$

Letting $\delta=8(n^2-n)\exp\{-\frac{\sigma^{4d}(n-2)\epsilon^2}{8C_k^4}\}+8n\exp\{-\frac{\sigma^{4d}(n-1)\epsilon^2}{8C_k^4}\}$, we have

$$J(w,\alpha)-\frac{\epsilon}{2}\leq\hat{J}(w,\alpha)\leq J(w,\alpha)+\frac{\epsilon}{2}\quad\forall w,\alpha \tag{2.9}$$

with probability at least $1-\delta$. Thus, with probability at least $1-\delta$, for any $w\in\Delta^M,\alpha\in\Delta_M^{2n}$

$$\begin{aligned}J(\hat{w},\hat{\alpha})&\leq\hat{J}(\hat{w},\hat{\alpha})+\frac{\epsilon}{2}\\ &\leq\hat{J}(w,\alpha)+\frac{\epsilon}{2}\\ &\leq J(w,\alpha)+\epsilon,\end{aligned}$$

where $\hat{w},\hat{\alpha}$ are defined in (2.6). Then with probability at least $1-\delta$

$$\hat{J}(\hat{w},\hat{\alpha})\leq\inf_{\substack{w\in\Delta^M\\ \alpha\in\Delta_M^{2n}}}J(w,\alpha)+\epsilon. \tag{2.10}$$

Combining (2.10) with the definition of the ISE shows, with probability at least $1-\delta$,

$$\|q-q_{\hat{w},\hat{\alpha}}\|_2^2\leq\inf_{\substack{w\in\Delta^M\\ \alpha\in\Delta_M^{2n}}}\|q-q_{w,\alpha}\|_2^2+\epsilon.$$

□

2.10.3.2 Theorem 1: Arbitrary Group Size

Preliminaries Before beginning the proof, we start by redefining q, $q_{w,a}$, J, and $\hat{J}$ for arbitrary group size. Once this is done, the proof will follow the same basic steps as the proof for groups of size two.

Suppose we change the problem setup only in the size of the grouped observations. Consider grouped observations of size N. Consider a set of n grouped observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,N}) \in \mathbb{R}_N^d \coloneqq \underbrace{\mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{N}$ drawn i.i.d. from

$$q(y_1, y_2, \ldots, y_N) = \sum_{m=1}^{M} w_m^* p_m^*(y_1) p_m^*(y_2) \ldots p_m^*(y_N), \quad y_1, y_2, \ldots, y_N \in \mathbb{R}^d. \tag{2.11}$$

Similar to the paired observation setting, a wKDE in this setting will have the form

$$p(y; \theta) = \sum_{r=1}^{n} \sum_{r'=1}^{N} \theta_{r,r'} k_\sigma(y, x_{r,r'}).$$

We may write the corresponding estimator of q

$$q_{w,\alpha}(y_1, y_2, \ldots, y_N) = \sum_{m=1}^{M} w_m p(y_1; \alpha_m) p(y_2; \alpha_m) \ldots p(y_N; \alpha_m)$$

where $\alpha_m = [\alpha_{m,1,1} \ldots \alpha_{m,1,N} \ldots\ldots \alpha_{m,n,1} \ldots \alpha_{m,n,N}]' \in \Delta^{Nn}$ for $m = 1, \ldots, M$, with $\alpha_{m,r,r'}$ corresponding to the weight of the kernel centered at $x_{r,r'}$ in the estimate of the mth mixture component.

In what follows we use $\sum_{r,r'} \coloneqq \sum_{r_1,r_1'} \cdots \sum_{r_N,r_N'}$ to ease notation. Similar to the paired sample case, we define

$$J(w, \alpha) \coloneqq \int q_{w,\alpha}^2(y_1, \ldots, y_N) dy_1 \ldots dy_N - 2 \sum_{m,r,r'} \left(\prod_{i \in [N]} w_m \alpha_{m,r_i,r_i'} \right) h(r_1, r_1', \ldots, r_N, r_N')$$

$$\hat{J}(w, \alpha) \coloneqq \int q_{w,\alpha}^2(y_1, \ldots, y_N) dy_1 \ldots dy_N - 2 \sum_{m,r,r'} \left(\prod_{i \in [N]} w_m \alpha_{m,r_i,r_i'} \right) \hat{h}(r_1, r_1', \ldots, r_N, r_N'),$$

where

$$h(r_1, r_1', \ldots, r_N, r_N') \coloneqq \int k_\sigma(y_1, x_{r_1,r_1'}) \ldots k_\sigma(y_N, x_{r_N,r_N'}) q(y_1, \ldots, y_N) dy_1 \ldots dy_N,$$

$$\hat{h} \coloneqq \hat{h}_{\text{LNO}}(r_1, r_1', \ldots, r_N, r_N') \coloneqq \frac{1}{n-N} \sum_{i \in [n] \setminus L(r_1, r_2, \ldots, r_N)} k_\sigma(x_{i,1}, x_{r_1,r_1'}) \ldots k_\sigma(x_{i,N}, x_{r_N,r_N'}),$$

where $L(r_1, r_2, \ldots, r_N)$ is any subset of $[n]$ containing $\{r_1, r_2, \ldots, r_N\}$ and having cardinality N. If $r_1, r_2, \ldots, r_N$ are not distinct, the additional indices can be chosen arbitrarily. For simplicity, we use a leave-N-out (LNO) estimator $\hat{h}_{\text{LNO}}$ rather than a hybrid estimator like we used in the case of paired observations. As in the paired observation setting, we define

$$(\hat{w}, \hat{\alpha}) \coloneqq \operatorname*{arg\,min}_{w \in \Delta^M,\ \alpha \in \Delta_M^{Nn}} \hat{J}(w, \alpha),$$

and similarly define $\hat{q} \coloneqq q_{\hat{w},\hat{\alpha}}$. Whenever $\hat{w}$, $\hat{\alpha}$, q, or $q_{\hat{w},\hat{\alpha}}$ are referenced in the arbitrary group size setting, we will be referring to these estimators.

Proof of Theorem 1: Arbitrary Group Size We now state Theorem 1 for arbitrary group size.

Theorem 1a. *Given grouped observations of size N, let $\epsilon > 0$ and $\delta = 2(Nn)^N \exp\left\{-\frac{\sigma^{2Nd}(n-N)\epsilon^2}{8C_k^{2N}}\right\}$. With probability at least $1-\delta$ the following holds:*

$$\|q - q_{\hat{w},\hat{\alpha}}\|_2^2 \leq \inf_{w \in \Delta^M,\ \alpha \in \Delta_M^{Nn}} \|q - q_{w,\alpha}\|_2^2 + \epsilon.$$

Proof. The proof proceeds as in the paired observation setting. In particular,

$$\begin{aligned}
&P_q\{ \sup_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} |J(w,\alpha) - \hat{J}(w,\alpha)| > \frac{\epsilon}{2}\} \\
&\quad \leq P_q\Big\{ \sup_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} \sum_{m,r,r'} \prod_{i \in [N]} w_m \alpha_{m,r_i,r_i'} \Big| h(r_1, r_1', \ldots, r_N, r_N') - \hat{h}(r_1, r_1', \ldots, r_N, r_N') \Big| > \frac{\epsilon}{4} \Big\} \\
&\quad \leq P_q\Big\{ \max_{r_1, r_1', \ldots, r_N, r_N'} \Big| h(r_1, r_1', \ldots, r_N, r_N') - \hat{h}(r_1, r_1', \ldots, r_N, r_N') \Big| > \frac{\epsilon}{4} \Big\} \\
&\quad \leq \sum_{r,r'} P_q\Big\{ \Big| h(r_1, r_1', \ldots, r_N, r_N') - \hat{h}(r_1, r_1', \ldots, r_N, r_N') \Big| > \frac{\epsilon}{4} \Big\}
\end{aligned}$$

The first step above is due to the triangle inequality, and the penultimate step is due to simplex constraints on w, α. Let $k_i(r_1, r_1', \ldots, r_N, r_N') \coloneqq k_\sigma(x_{i,1}, x_{r_1,r_1'}) k_\sigma(x_{i,2}, x_{r_2,r_2'}) \cdots k_\sigma(x_{i,N}, x_{r_N,r_N'})$. Noting that $h(r_1, r_1', \ldots, r_N, r_N') =$

$\mathbb{E}_q\{k_\sigma(x_{i,1}, x_{r_1,r'_1}) \cdots k_\sigma(x_{i,N}, x_{r_N,r'_N})\}$, we have

$$P_q\left\{\left|h(r_1,r'_1,\ldots,r_N,r'_N) - \hat{h}(r_1,r'_1,\ldots,r_N,r'_N)\right| > \frac{\epsilon}{4}\right\}$$
$$= P_q\Bigg\{\Big|\sum_{i\in[n]\setminus L(r_1,\ldots,r_N)} \mathbb{E}_{(x_{i,1},\ldots,x_{i,N})\sim q}\{k_i(r_1,r'_1,\ldots,r_N,r'_N)\}$$
$$- k_i(r_1,r'_1,\ldots,r_N,r'_N)\Big| > \frac{(n-N)\epsilon}{4}\Bigg\}$$

The terms $k_i(r_1,r'_1,\ldots,r_N,r'_N)$ are independent random variables due to use of the LNO estimator. By assumption, $0 \leq k_i(r_1,r'_1,\ldots,r_N,r'_N) \leq C_k^N \sigma^{-Nd}$ so the k_i are bounded for fixed $\sigma > 0$. We apply Hoeffding's inequality

$$P_q\left\{\left|h(r_1,r'_1,\ldots,r_N,r'_N) - \hat{h}(r_1,r'_1,\ldots,r_N,r'_N)\right| > \frac{\epsilon}{4}\right\} \leq 2\exp\{-\frac{2(n-N)^2\epsilon^2}{16(n-N)C_k^{2N}\sigma^{-2Nd}}\}$$
$$= 2\exp\{-\frac{\sigma^{2Nd}(n-N)\epsilon^2}{8C_k^{2N}}\}.$$

Substituting backward we obtain the desired upper bound

$$P_q\{\sup_{\substack{w\in\Delta_M^M \\ \alpha\in\Delta_M^{Nn}}} |J(w,\alpha) - \hat{J}(w,\alpha)| > \frac{\epsilon}{2}\} \leq \sum_{r_1,r'_1}\cdots\sum_{r_N,r'_N} 2\exp\{-\frac{\sigma^{2Nd}(n-N)\epsilon^2}{8C_k^{2N}}\}$$
$$= 2(Nn)^N \exp\{-\frac{\sigma^{2Nd}(n-N)\epsilon^2}{8C_k^{2N}}\}$$

From here the proof is identical to the paired observation case, but with

$$\delta = 2(Nn)^N \exp\{-\frac{\sigma^{2Nd}(n-N)\epsilon^2}{8C_k^{2N}}\}.$$

□

2.10.4 Proof and General Form of Theorem 2

In this section we give the proof of Theorem 2 for groups of size two, before extending it to groups of arbitrary size. For readability, we first present some intermediate results to be used in the main proofs.

2.10.4.1 Intermediate Results

We first prove two supporting results.

Lemma 1. *For any $1 \leq p < \infty$, any $f, g \in L^p$, and any integer $a \geq 2$,*

$$\|f^{\times a} - g^{\times a}\|_p \leq \|f\|_p^{a-1}\|f - g\|_p + \|g\|_p\|f^{\times(a-1)} - g^{\times(a-1)}\|_p,$$

where $f^{\times a}(y_1, y_2, \ldots, y_a) := f(y_1)f(y_2)\cdots f(y_a)$.

Proof. Let $f, g \in L^p$, $1 \leq p < \infty$. Then

$$\begin{aligned}
\|f^{\times a} - g^{\times a}\|_p &= \|f^{\times a} - f^{\times(a-1)} \times g + f^{\times(a-1)} \times g - g^{\times a}\|_p \\
&\leq \|f^{\times a} - f^{\times(a-1)} \times g\|_p + \|f^{\times(a-1)} \times g - g^{\times a}\|_p \\
&= \left(\int |f(x_1)\cdots f(x_{a-1})(f(x_a) - g(x_a))|^p dx_1 \ldots dx_a\right)^{\frac{1}{p}} \\
&\quad + \left(\int |g(x_a)(f(x_1)\cdots f(x_{a-1}) - g(x_1)\cdots g(x_{a-1}))|^p dx_1 \ldots dx_a\right)^{\frac{1}{p}} \\
&= \|f^{\times(a-1)}\|_p\|f - g\|_p + \|g\|_p\|f^{\times(a-1)} - g^{\times(a-1)}\|_p \\
&= \|f\|_p^{a-1}\|f - g\|_p + \|g\|_p\|f^{\times(a-1)} - g^{\times(a-1)}\|_p.
\end{aligned}$$

□

We have the following corollary.

Corollary 2. *For any $1 \leq p < \infty$, any $f, g \in L^p$, and any integer $a \geq 2$,*

$$\|f^{\times a} - g^{\times a}\|_p \leq \left(\sum_{b=1}^{a} \|f\|_p^{a-b}\|g\|_p^{b-1}\right)\|f - g\|_p.$$

Proof. The proof is by induction. Lemma 1 provides the base of the recursion for $a = 2$. Now suppose the statement is true for $a \geq 2$. To prove the statement for $a+1$, we apply Lemma 1 again, together with the induction hypothesis, to get

$$\begin{aligned}
\|f^{\times(a+1)} - g^{\times(a+1)}\|_p &\leq \|f\|_p^a\|f - g\|_p + \|g\|_p\|f^{\times a} - g^{\times a}\|_p \\
&\leq \|f\|_p^a\|f - g\|_p + \|g\|_p\left(\sum_{b=1}^{a} \|f\|_p^{a-b}\|g\|_p^{b-1}\right)\|f - g\|_p \\
&= \left(\sum_{b=1}^{a+1} \|f\|_p^{a+1-b}\|g\|_p^{b-1}\right)\|f - g\|_p.
\end{aligned}$$

This completes the proof. □

2.10.4.2 Proof of Theorem 2: Groups of Size Two

We restate Theorem 2 for convenience.

Theorem 2. *If* $\sigma \to 0$ *and* $\frac{n\sigma^{4d}}{\log n} \to \infty$ *as* $n \to \infty$*, then* $\|q - q_{\hat{w},\hat{\alpha}}\|_1 \xrightarrow{a.s.} 0$.

Proof. Lemma 3.1 of [GM90] states that if $\int \hat{q} = 1$ and $\|\hat{q} - q\|_2 \xrightarrow{a.s.} 0$, then $\|\hat{q} - q\|_1 \xrightarrow{a.s.} 0$. Since $\int \hat{q} = 1$ in our case, our strategy is to show $\|\hat{q} - q\|_2 \xrightarrow{a.s.} 0$. To do this it suffices to show that

$$\|q - \hat{q}\|_2^2 - \inf_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2^2 \xrightarrow{a.s.} 0 \tag{2.12}$$

and

$$\inf_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2 \xrightarrow{a.s.} 0. \tag{2.13}$$

To show (2.12), by the Borel-Cantelli lemma, it suffices to show that for all $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P_q \left(\|q - \hat{q}\|_2^2 - \inf_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{2n}}} \|q - q_{w,\alpha}\|_2^2 \geq \epsilon \right) < \infty.$$

Thus let $\epsilon > 0$. By Theorem 1a, the probability in question is at most

$$\begin{aligned} \delta &= 2(nN)^N \exp\left\{ -\frac{(n-N)\sigma^{2Nd}\epsilon^2}{8C_k^{2N}} \right\} \\ &= 8 \exp\left\{ -2\log n \left(\frac{(n-2)\sigma^{4d}\epsilon^2}{16C_k^4 \log n} - 1 \right) \right\}. \end{aligned}$$

By assumption on the growth of n and σ, there exists N_ϵ such that for all $n \geq N_\epsilon$,

$$\frac{(n-2)\sigma^{4d}\epsilon^2}{16C_k^4 \log n} \geq 2.$$

For such n we have

$$\delta \leq 8 \exp\{-2\log n\} = \frac{8}{n^2}$$

which is summable.

To show (2.13), let w^* be the true mixing weights from (2.2). For $i = 1, \ldots, n$ let e_i be the

$m \in [M]$ such that $X_i = (x_{i,1}, x_{i,2}) \overset{i.i.d.}{\sim} p_m^*$. Define

$$n_m = \big|\{i : e_i = m\}\big|, \qquad m = 1, \dots, M$$
$$\alpha^*_{m,i,1} = \alpha^*_{m,i,2} = \begin{cases} \frac{1}{2n_m}, & e_i = m \\ 0, & \text{otherwise} \end{cases}, \qquad m = 1, \dots, M$$

With this "oracle" assignment of weights, $p(x; \alpha_m^*)$ is just the regular KDE for p_m^*. Therefore, we may apply known results for consistency of standard KDEs. In particular, we will apply Theorem 3.1 of [GM90] which implies

$$\|p(\cdot\,; \alpha_m^*) - p_m^*\|_2 \xrightarrow{a.s.} 0 \text{ as } n_m \to \infty \tag{2.14}$$

provided $k \in L^2$ and $\sum_n \frac{1}{n^2 \sigma_n^d} < \infty$. Both of these conditions are satisfied by assumption in our setting. Furthermore, as $n \to \infty$ we have $\frac{n_m}{n} \to w_m^*$ almost surely, and therefore $n_m \to \infty$ almost surely.

Finally, we have

$$\begin{aligned}
\inf_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{2n}}} \|q - q_{w,\alpha}\|_2 &\le \|q - q_{w^*,\alpha^*}\|_2 \\
&= \left\| \sum_{m=1}^M w_m^* (p_m^* \times p_m^* - p(\cdot\,; \alpha_m^*) \times p(\cdot; \alpha_m^*)) \right\|_2 \\
&\le \sum_{m=1}^M w_m^* \|p_m^* \times p_m^* - p(\cdot\,; \alpha_m^*) \times p(\cdot\,; \alpha_m^*))\|_2 \\
&\le \sum_{m=1}^M w_m^* (\|p_m^*\|_2 + \|p(\cdot\,; \alpha_m^*)\|_2) \|p_m^* - p(\cdot\,; \alpha_m^*)\|_2 \\
&\le \sum_{m=1}^M w_m^* 3 \|p_m^*\|_2 \|p_m^* - p(\cdot\,; \alpha_m^*)\|_2 \\
&\xrightarrow{a.s.} 0 \text{ as } n \to \infty,
\end{aligned}$$

where the fourth step uses Lemma 1 and the fifth step holds for n sufficiently large (a.s.). This completes the proof. □

2.10.4.3 Proof of Theorem 2: Arbitrary Group Size

We consider the problem for arbitrary group size as described in Section 2.10.3.2 of this document. The proof of Theorem 2 for arbitrary group size is similar to the proof for groups of size two. The

main difference will be in use of Theorem 1a rather than Theorem 1 to invoke the Borel-Cantelli lemma.

Theorem 2a. *Given grouped observations of size $N \in \mathbb{Z}^+$, if $\sigma \to 0$ and $\frac{n\sigma^{2Nd}}{\log n} \to \infty$ as $n \to \infty$ then $\|q - \hat{q}\|_1 \xrightarrow{a.s.} 0$ as $n \to \infty$.*

Proof. We will appeal to Lemma 3.1 of [GM90] as we did for groups of size two. Namely, if $\int \hat{q} = 1$ and $\|\hat{q} - q\|_2 \xrightarrow{a.s.} 0$, then $\|\hat{q} - q\|_1 \xrightarrow{a.s.} 0$. Our strategy again is to show $\|\hat{q} - q\|_2 \xrightarrow{a.s.} 0$. To do this it suffices to show that

$$\|q - \hat{q}\|_2^2 - \inf_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2^2 \xrightarrow{a.s.} 0 \tag{2.15}$$

and

$$\inf_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2 \xrightarrow{a.s.} 0. \tag{2.16}$$

To show (2.15), by the Borel-Cantelli lemma, it suffices to show that for all $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P_q \left(\|q - \hat{q}\|_2^2 - \inf_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2^2 \geq \epsilon \right) < \infty.$$

Thus let $\epsilon > 0$. By Theorem 1a, the probability in question is at most

$$\begin{aligned}
\delta &= 2(Nn)^N \exp\{-\frac{\sigma^{2Nd}(n-N)\epsilon^2}{8C_k^{2N}}\} \\
&= 2N^N \exp\left\{ N \log n + \left(\frac{(n-N)\sigma^{2Nd}\epsilon^2}{8C_k^{2N}} \right) \right\} \\
&= 2N^N \exp\left\{ -N \log n \left(\frac{(n-N)\sigma^{2Nd}\epsilon^2}{8C_k^{2N} N \log n} - 1 \right) \right\}.
\end{aligned}$$

By assumption on the growth of n and σ, there exists N_ϵ such that for all $n \geq N_\epsilon$,

$$\frac{(n-N)\sigma^{2Nd}\epsilon^2}{8C_k^{2N} N \log n} \geq 2.$$

For such n we have

$$\delta \leq 2N^N \exp\{-N \log n\} = \frac{2N^N}{n^N}$$

which is summable for $N > 1$.

To show (2.16), let w^* be the true mixing weights from (2.11). For $i = 1, \ldots, n$ let e_i be the

$m \in [M]$ such that $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,N}) \overset{i.i.d.}{\sim} p^*_m$. Define

$$n_m = \big|\{i : e_i = m\}\big|, \qquad m = 1, \ldots, M$$
$$\alpha^*_{m,i,j} = \begin{cases} \frac{1}{n_m N}, & e_i = m \\ 0, & \text{otherwise} \end{cases}, \qquad m = 1, \ldots, M,\ j = 1, \ldots, N$$

We are again using an "oracle" assignment of weights, so $p(x; \alpha^*_m)$ is just the regular KDE for p^*_m. Therefore, we may again apply Theorem 3.1 of [GM90] which implies

$$\|p(\cdot\,; \alpha^*_m) - p^*_m\|_2 \xrightarrow{a.s.} 0 \text{ as } n_m \to \infty \tag{2.17}$$

provided $k \in L^2$ and $\sum_n \frac{1}{n^2 \sigma_n^d} < \infty$. Both of these conditions are satisfied by assumption in our setting. Furthermore, as $n \to \infty$ we have $\frac{n_m}{n} \to w^*_m$ almost surely, and therefore $n_m \to \infty$ almost surely.

Finally, we have

$$\begin{aligned}
\inf_{\substack{w \in \Delta^M \\ \alpha \in \Delta_M^{Nn}}} \|q - q_{w,\alpha}\|_2 &\le \|q - q_{w^*,\alpha^*}\|_2 \\
&= \left\| \sum_{m=1}^M w^*_m (p^{*\times N}_m - p(\cdot\,; \alpha^*_m)^{\times N}) \right\|_2 \\
&\le \sum_{m=1}^M w^*_m \|p^{*\times N}_m - p(\cdot\,; \alpha^*_m)^{\times N}\|_2 \\
&\le \sum_{m=1}^M w^*_m \left(\sum_{b=1}^N \|p^*_m\|_2^{N-b} \|p(\cdot\,; \alpha^*_m)\|_2^{b-1} \right) \|p^*_m - p(\cdot\,; \alpha^*_m)\|_2 \\
&\le \sum_{m=1}^M w^*_m \left(\sum_{b=1}^N 2^{b-1} \|p^*_m\|_2^{N-1} \right) \|p^*_m - p(\cdot\,; \alpha^*_m)\|_2 \\
&\xrightarrow{a.s.} 0 \text{ as } n \to \infty,
\end{aligned}$$

where the penultimate step uses (2.17) and Corollary 2, and the final step holds for n sufficiently large (a.s.). This completes the proof. □

2.10.5 Background on the Grouped Sample Setting and Proof of Theorem 3

Here we prove Theorem 3. We will be proving a general and more technical version of this theorem, Theorem 4, from which Theorem 3 is a direct consequence. First we will introduce some background to the problem setting which was introduced in [VS19]. **This section uses its own**

notation which does not extend to other appendices or main text.

2.10.5.1 Identifiability in the Grouped Sample Setting

We will be concerned with probability measures on a measurable space $(\Omega, \mathcal{F})$. Let δ be the Dirac measure. Let $\mathcal{D}$ be the set of probability measures on $(\Omega, \mathcal{F})$. We call a probability measure on $\mathcal{D}$ of the form

$$\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$$

a *mixture of measures* [VS19]. For all mixtures of measures we will assume that $a_i > 0$ for all i and $\mu_i \neq \mu_j$ when $i \neq j$ so that m is the number of distinct mixture components. The *grouped sample* setting from [VS19] considers the situation where samples come in groups of size n by first sampling a random measure component from a mixture of measures $\gamma \sim \mathscr{P}$, which is then sampled iid n times. So one has access to samples of the form $\mathbf{X} = (X_1, \ldots, X_n)$ with $X_1, \ldots, X_n \overset{iid}{\sim} \gamma$. In this situation the identifiability of $\mathscr{P}$ depends on whether the distribution of $\mathbf{X}$ is uniquely determined by $\mathscr{P}$ and the number of samples per group n. To this end [VS19] introduced the V_n operator which maps a mixture of measures to the distribution of $\mathbf{X}$:

$$V_n\left(\sum_{i=1}^{m} a_i \delta_{\mu_i}\right) = \sum_{i=1}^{m} a_i \mu_i^{\times n},$$

where $\mu^{\times n}$ denotes the product measure n times. We note that $n = 1$ corresponds to a typical mixture model where each mixture component is sampled once after being selected and there is no grouped sample structure. For the grouped sample setting [VS19] introduces the following notion of identifiability.

Definition 1. *A mixture of measures, $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$, is called* n-identifiable *if there does not exist a different mixture of measures $\mathscr{Q} = \sum_{j=1}^{m'} b_j \delta_{\nu_j}$, with $m' \leq m$, such that $V_n(\mathscr{P}) = V_n(\mathscr{Q})$.*

A *completely* rigorous mathematical treatment of the previous notions is a bit involved and can be found in [VS19]. In [VS19] it is shown that if the mixture components are jointly irreducible then a mixture of measures is 2-identifiable, if they are linearly independent then they are 3-identifiable, and that any mixture of measures with m components is $(2m-1)$-identifiable.

2.10.5.2 Notation

Before we state and prove the main theorem of this section we need to first introduce some notation.

Let S_m be the symmetric group over m symbols. Abusing notation slightly we will let the

elements of S_m be a group action on $[m]$ as well as $\mathbb{R}^m$. On $\mathbb{R}^m$ it is defined as the following

$$\sigma\left(\left[x_1, \ldots, x_m\right]^T\right) = \left[x_{\sigma(1)}, \ldots, x_{\sigma(m)}\right]^T. \tag{2.18}$$

We also let S_m be an operator where $S_m \cdot x$ is the orbit of x, i.e.

$$S_m \cdot x \triangleq \{\sigma(x) : \sigma \in S_m\}. \tag{2.19}$$

Recall that for a pair of Hilbert spaces H, H' the direct sum $H \oplus H'$ is a Hilbert space with elements of the form $x \oplus x'$ and inner product defined as $\langle x \oplus x', y \oplus y' \rangle = \langle x, y \rangle + \langle x', y' \rangle$. For a pair of Banach spaces B, B' we define the direct sum via the norm $\|b \oplus b'\|_{B \oplus B'} \triangleq \|b\|_B + \|b'\|_{B'}$ which is itself a Banach space ([Ban32] p. 183).

For a pair of Hilbert spaces H, H' let $H \otimes H'$ be the tensor product of these two spaces and $h \otimes h'$ be the tensor product of vectors $h \in H$ and $h' \in H'$. For a vector in a Hilbert space h let $h^{\otimes n}$ denote the tensor power, i.e. $\underbrace{h \otimes \cdots \otimes h}_{n \text{ times}}$.

In the following the space of finite signed measures is equipped with the total variation topology and unadorned norms refer to the total variation norm on finite signed measures, which forms a Banach space. Norms for various Lebesgue spaces will have the associated subscript. Finally we note that for two Hilbert spaces of square-integrable functions over σ-finite measure spaces $L^2(\Omega, \mathcal{F}, \mu)$ and $L^2(\Omega', \mathcal{F}', \mu')$ we have that $L^2(\Omega, \mathcal{F}, \mu) \otimes L^2(\Omega', \mathcal{F}', \mu') \cong L^2(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mu \times \mu')$ via an isomorphism $f \otimes f' \mapsto f \times f'$ ([KR83] Example 2.6.11) and we will use a 2 subscript for both norms.

2.10.5.3 Full Theorem Statement and Proof

The following is the full general version of Theorem 3 and the main result of this section.

Theorem 4. *Let $(\Omega, \mathcal{F})$ be a measurable space, $\mathscr{P} = \sum_{j=1}^{m} a_j \delta_{\mu_j}$ a mixture of measures on that space which is n-identifiable, and $\mathscr{P}_i = \sum_{j=1}^{m_i'} b_{i,j} \delta_{\nu_{i,j}}$ a sequence of mixtures of measures with $m_i' \leq m$ for all i, such that $V_n(\mathscr{P}_i) \to V_n(\mathscr{P})$. Then $m_i' \to m$ and there exists a sequence of permutations σ_i such that $\sigma_i(b_i) \to a$ and $\nu_{i,\sigma_i(j)} \to \mu_j$ for all j.*

Essentially this says that as one finds grouped sample distributions $V_n(\mathscr{Q}_i)$ which approach the true grouped sample distribution $V_n(\mathscr{P})$ the mixture of measures $\mathscr{Q}_i$ will automatically recover the true mixing weights and components from $\mathscr{P}$ so long as $\mathscr{P}$ is n-identifiable. In other words, one simply needs to fit the grouped distribution $V_n(\mathscr{P})$ well to get a good estimate of the mixture components. Theorem 3 from the main text is a direct consequence of Theorem 4.

Corollary 3 (Theorem 3)**.** *Let* $\sum_{m=1}^{M} w_m p_m$ *be an* N*-identifiable mixture model, and* $\sum_{m=1}^{M} \hat{w}_{m,j}\hat{p}_{m,j}$ *be a sequence of mixture models such that* $\left\|\sum_{m=1}^{M} \hat{w}_{m,j}\hat{p}_{m,j}^{\times N} - \sum_{m=1}^{M} w_m p_m^{\times N}\right\|_1 \to 0$*. Then there is a sequence of permutations* σ_j *so that* $\hat{w}_{\sigma_j(m),j} \to w_m$ *and* $\left\|\hat{p}_{\sigma_j(m),j} - p_m\right\|_1 \to 0$ *for all* m*.*

We introduce some preliminary results before proving Theorem 4. The following lemma will be needed for our proof.

Lemma 2. *Let* $\mathscr{P}$ *and* $\mathscr{Q}$ *be mixtures of measures, then* $\|V_{n'}(\mathscr{P}) - V_{n'}(\mathscr{Q})\| \leq \|V_n(\mathscr{P}) - V_n(\mathscr{Q})\|$ *for all* $n' \leq n$*.*

Proof of Lemma 2. From [Fol99] (Section 3.1 Exercise 7a) we have the following

$$
\begin{aligned}
&\|V_n(\mathscr{P}) - V_n(\mathscr{Q})\| \\
&= \sup\left\{ \sum_{i=1}^{k} |(V_n(\mathscr{P}) - V_n(\mathscr{Q}))(E_i)| : \right. \\
&\qquad\qquad \left. k \in \mathbb{N}, E_1, \ldots, E_k \in \mathcal{F}^{\times n} \text{ are disjoint, and } \bigcup_{i=1}^{k} E_i = \Omega^{\times n} \right\} \\
&\geq \sup\left\{ \sum_{i=1}^{k} \left|(V_n(\mathscr{P}) - V_n(\mathscr{Q}))(E_i \times \Omega^{\times n-n'})\right| : \right. \\
&\qquad\qquad \left. k \in \mathbb{N}, E_1, \ldots, E_k \in \mathcal{F}^{\times n'} \text{ are disjoint, and } \bigcup_{i=1}^{k} E_i = \Omega^{\times n'} \right\} \\
&= \sup\left\{ \sum_{i=1}^{k} |(V_{n'}(\mathscr{P}) - V_{n'}(\mathscr{Q}))(E_i)| : \right. \\
&\qquad\qquad \left. k \in \mathbb{N}, E_1, \ldots, E_k \in \mathcal{F}^{\times n'} \text{ are disjoint, and } \bigcup_{i=1}^{k} E_i = \Omega^{\times n'} \right\} \\
&= \|V_{n'}(\mathscr{P}) - V_{n'}(\mathscr{Q})\| .
\end{aligned}
$$

□

The following lemma is the main workhorse in the proof of Theorem 4.

Lemma 3. *Let* $(\Omega, \mathcal{F})$ *be a measurable space,* $\mathscr{P} = \sum_{i=1}^{m} a_i \delta_{\mu_i}$ *a mixture of measures on that space,* $n \in \mathbb{N}$*, and* $\mathscr{P}_i = \sum_{j=1}^{m'} b_j \delta_{\nu_{i,j}}$ *a sequence of mixtures of measures (*m' *is fixed) with such that* $V_n(\mathscr{P}_i) \to V_n(\mathscr{P})$ *(*b *does not depend on* i*). Then there exists a subsequence* i_k *and a collection of probability measures* $\nu_1, \ldots, \nu_{m'}$ *such that* $\nu_{i_k,j} \to \nu_j$ *for all* j *and* $V_n(\mathscr{P}) = V_n\left(\sum_{j=1}^{m'} b_j \delta_{\nu_j}\right)$*.*

Proof of Lemma 3. We will use bold symbols to represent elements that depend on i, e.g. $\boldsymbol{\nu}_j = \nu_{i,j}$. Let $\bar{\mu} = \sum_{k=1}^{m} a_k \mu_k$. By the Lebesgue-Radon-Nikodym Theorem ([Fol99] Theorem 3.8) there exists series of measures $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{m'}$ and $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{m'}$ such that $\boldsymbol{\nu}_k = \boldsymbol{\lambda}_k + \boldsymbol{\rho}_k$ with $\boldsymbol{\lambda}_k \perp \bar{\mu}$ and $\boldsymbol{\rho}_k \ll \bar{\mu}$ for all $k \in [m']$.

For some fixed ℓ let $\mathbf{A}_\ell$ be the sequence of measurable sets such that $\boldsymbol{\lambda}_\ell(\cdot \cap \mathbf{A}_\ell) = \boldsymbol{\lambda}_\ell$ and $\bar{\mu}(\mathbf{A}_\ell) = 0$, this is possible since $\boldsymbol{\lambda}_\ell \perp \bar{\mu}$. From Lemma 2 we have that

$$\left\| \sum_{k=1}^{m} a_k \mu_k - \sum_{j=1}^{m'} b_j \boldsymbol{\nu}_j \right\| \to 0 \Rightarrow \left| \sum_{k=1}^{m} a_k \mu_k(\mathbf{A}_\ell) - \sum_{j=1}^{m'} b_j \boldsymbol{\nu}_j(\mathbf{A}_\ell) \right| \to 0 \tag{2.20}$$

$$\Rightarrow \left| b_\ell \boldsymbol{\rho}_\ell(\mathbf{A}_\ell) + b_\ell \boldsymbol{\lambda}_\ell(\mathbf{A}_\ell) + \sum_{j \in [m'] \setminus \{\ell\}} b_j \boldsymbol{\nu}_j(\mathbf{A}_\ell) \right| \to 0 \tag{2.21}$$

$$\Rightarrow \left| b_\ell \boldsymbol{\lambda}_\ell(\mathbf{A}_\ell) + \sum_{j \in [m'] \setminus \{\ell\}} b_j \boldsymbol{\nu}_j(\mathbf{A}_\ell) \right| \to 0. \tag{2.22}$$

Because all of the summands inside the absolute value on the last line are positive we have that $\|\boldsymbol{\lambda}_\ell\| \to 0$ and thus $\|\boldsymbol{\rho}_\ell\| \to 1$. Eventually in our sequence we must have that $\|\boldsymbol{\rho}_\ell\| > 0$, so eventualy in our subsequence we can define $\boldsymbol{\nu}'_\ell = \boldsymbol{\rho}_\ell / \|\boldsymbol{\rho}_\ell\|$ which is now a sequence of probability measures which are absolutely continuous with respect to $\bar{\mu}$ and $\|\boldsymbol{\nu}'_\ell - \boldsymbol{\nu}_\ell\| \to 0$.

From this we have that there exists sequences of probability measures $\boldsymbol{\nu}'_1, \ldots, \boldsymbol{\nu}'_{m'}$ such that $\|\boldsymbol{\nu}_k - \boldsymbol{\nu}'_k\| \to 0$ and $\boldsymbol{\nu}'_k \ll \bar{\mu}$ for all $k \in [m']$. Lemma 3.3.7 in [Rei89] states that, for probablity measures over the same domain $\xi_1, \ldots \xi_d, \gamma_1, \ldots, \gamma_d$ that $\left\| \prod_{j=1}^{d} \xi_j - \prod_{k=1}^{d} \gamma_k \right\| \leq \sum_{k=1}^{d} \|\xi_k - \gamma_k\|$. It follows therefore that $\left\| \boldsymbol{\nu}'^{\times n}_k - \boldsymbol{\nu}^{\times n}_k \right\| \to 0$ for all k and

$$\left\| \sum_{k=1}^{m} a_k \mu_k^{\times n} - \sum_{j=1}^{m'} b_j \boldsymbol{\nu}'^{\times n}_j \right\| \to 0. \tag{2.23}$$

For some fixed ℓ let $\mathbf{q}'_\ell$ be the Radon-Nikodym derivative of $\boldsymbol{\nu}'_\ell$ with respect to $\bar{\mu}$. Let $\mathbf{B}_\ell = \mathbf{q}'^{-1}_\ell([2/b_\ell, \infty))$. We have the following

$$\sum_{k=1}^{m'} b_k \boldsymbol{\nu}'_k(\mathbf{B}_\ell) \geq b_\ell \boldsymbol{\nu}'_\ell(\mathbf{B}_\ell) \tag{2.24}$$

$$\geq b_\ell \int_{\mathbf{B}_1} 2/b_\ell d\bar{\mu} \tag{2.25}$$

$$\geq 2\bar{\mu}(\mathbf{B}_\ell). \tag{2.26}$$

From Lemma 2 applied to (2.23) we have that $\left|\sum_{k=1}^{m'} b_k \boldsymbol{\nu}'_k(\mathbf{B}_\ell) - \bar{\mu}(\mathbf{B}_\ell)\right| \to 0$, and it follows that $\bar{\mu}(\mathbf{B}_\ell) \to 0$ because $\left|\sum_{k=1}^{m'} b_k \boldsymbol{\nu}'_k(\mathbf{B}_\ell) - \bar{\mu}(\mathbf{B}_\ell)\right| \geq \bar{\mu}(\mathbf{B}_\ell)$. Now we have that $\sum_{k=1}^{m'} b_k \boldsymbol{\nu}'_k(\mathbf{B}_\ell) \to 0$ and thus $\boldsymbol{\nu}'_\ell(\mathbf{B}_\ell) \to 0$.

Because $\boldsymbol{\nu}'_\ell\left(\mathbf{q}'^{-1}_\ell\left([2/b_\ell, \infty)\right)\right) \to 0$ and therefore $\boldsymbol{\nu}'_\ell\left(\mathbf{B}^C_\ell\right) \to 1$, for sufficiently large i we can now define a sequence of probability measures $\boldsymbol{\nu}''_\ell$ via $\boldsymbol{\nu}''_\ell(A) = \boldsymbol{\nu}'_\ell(A \cap \mathbf{B}^C_\ell)/\boldsymbol{\nu}'_\ell(\mathbf{B}^C_\ell)$. We have that

$$\begin{aligned}
\|\boldsymbol{\nu}'_\ell - \boldsymbol{\nu}''_\ell\| &= \left\|\left(\boldsymbol{\nu}'_\ell\left(\mathbf{B}_\ell \cap \cdot\right) + \boldsymbol{\nu}'_\ell\left(\mathbf{B}^C_\ell \cap \cdot\right)\right) - \left(\boldsymbol{\nu}''_\ell\left(\mathbf{B}_\ell \cap \cdot\right) + \boldsymbol{\nu}''_\ell\left(\mathbf{B}^C_\ell \cap \cdot\right)\right)\right\| && (2.27)\\
&\leq \left\|\boldsymbol{\nu}'_\ell\left(\mathbf{B}_\ell \cap \cdot\right) - \boldsymbol{\nu}''_\ell\left(\mathbf{B}_\ell \cap \cdot\right)\right\| + \left\|\boldsymbol{\nu}'_\ell\left(\mathbf{B}^C_\ell \cap \cdot\right) - \boldsymbol{\nu}''_\ell\left(\mathbf{B}^C_\ell \cap \cdot\right)\right\| && (2.28)\\
&= \boldsymbol{\nu}'_\ell\left(\mathbf{B}_\ell\right) + \left\|\boldsymbol{\nu}'_\ell\left(\mathbf{B}^C_\ell \cap \cdot\right) - \boldsymbol{\nu}'_\ell\left(\mathbf{B}^C_\ell \cap \cdot\right)/\boldsymbol{\nu}'_\ell\left(\mathbf{B}^C_\ell\right)\right\| && (2.29)\\
&= \boldsymbol{\nu}'_\ell\left(\mathbf{B}_\ell\right) + \left|1 - 1/\boldsymbol{\nu}'_\ell\left(\mathbf{B}^C_\ell\right)\right| \left\|\boldsymbol{\nu}'_\ell\left(\mathbf{B}_\ell \cap \cdot\right)\right\| && (2.30)
\end{aligned}$$

which goes to zero, so $\|\boldsymbol{\nu}''_\ell - \boldsymbol{\nu}_\ell\| \to 0$. Note that $\boldsymbol{\nu}''_\ell$ is a sequence of probability measures with Radon-Nikodym derivatives $\mathbf{q}''_\ell \triangleq \mathbf{q}'_\ell \mathbf{1}_{\mathbf{B}^C_\ell}/\boldsymbol{\nu}'_\ell\left(\mathbf{B}^C_\ell\right)$ ($\mathbf{1}$ is the indicator function) and thus

$$\sup_x \mathbf{q}''_\ell(x) = \sup_x \mathbf{q}'_\ell(x)\mathbf{1}_{\mathbf{B}^C_\ell}(x)/\boldsymbol{\nu}_\ell\left(\mathbf{B}^C_\ell\right) \leq 2/(b_\ell \boldsymbol{\nu}_\ell\left(\mathbf{B}^C_\ell\right))$$

and since $\boldsymbol{\nu}_\ell\left(\mathbf{B}^C_\ell\right) \to 1$ eventually $\|\mathbf{q}''_\ell\|_\infty \leq 3/b_\ell$. From this we have that $\mathbf{q}''_\ell \in L^1(\Omega, \mathcal{F}, \bar{\mu}) \cap L^\infty(\Omega, \mathcal{F}, \bar{\mu})$ and $\|\mathbf{q}''_\ell\|_\infty$ is a bounded sequence. From Hölders's Inequality we have that

$$= \|\mathbf{q}''_\ell\|_2^2 = \|\mathbf{q}''_\ell \mathbf{q}''_\ell\|_1^2 \leq \|\mathbf{q}''_\ell\|_1 \|\mathbf{q}''_\ell\|_\infty = \|\mathbf{q}''_\ell\|_\infty$$

so $\mathbf{q}''_\ell$ is a bounded sequence in $L^2(\Omega, \mathcal{F}, \bar{\mu})$.

We now define $\boldsymbol{\nu}''_1, \ldots, \boldsymbol{\nu}''_{m'}$ $\mathbf{q}''_1, \ldots, \mathbf{q}''_{m'}$ similarly. There exists β such that $\left\|\mathbf{q}''_j\right\|_\infty \leq \beta$ and $\left\|\mathbf{q}''_j\right\|_2 \leq \beta$ along the whole series and for all j. Let $p_1, \ldots, p_m$ be the radon Nikodym derivatives for $\mu_1, \ldots, \mu_m$ with respect to $\bar{\mu}$, again these are in $L^1(\Omega, \mathcal{F}, \bar{\mu}) \cap L^2(\Omega, \mathcal{F}, \bar{\mu}) \cap L^\infty(\Omega, \mathcal{F}, \bar{\mu})$. To see this note that $p_i \leq 1/a_i$ otherwise we have that

$$\begin{aligned}
\mu_i\left(p_i^{-1}\left((1/a_i, \infty)\right)\right) &= \int_{p_i^{-1}((1/a_i,\infty))} p_i d\bar{\mu}\\
&> \int_{p_i^{-1}((1/a_i,\infty))} 1/a_i d\bar{\mu}\\
&> \sum_j \int_{p_i^{-1}((1/a_i,\infty))} 1/a_i a_j d\mu_j\\
&\geq \mu_i\left(p_i^{-1}\left((1/a_i, \infty)\right)\right)
\end{aligned}$$

a contradiction. Now we have

$$\left\| \sum_{k=1}^{m} a_k p_k^{\times n} - \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\times n} \right\|_1 \to 0. \tag{2.31}$$

and Lemma 2 implies

$$\left\| \sum_{k=1}^{m} a_k p_k^{\times 2} - \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\times 2} \right\|_1 \to 0. \tag{2.32}$$

From Hölder's Inequality ($\|f\|_2^2 \leq \|f\|_1 \|f\|_\infty$) we have that

$$\left\| \sum_{k=1}^{m} a_k p_k^{\times 2} - \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\times 2} \right\|_2 \to 0 \tag{2.33}$$

and

$$\left\| \sum_{k=1}^{m} a_k p_k^{\otimes 2} - \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\otimes 2} \right\|_2 \to 0. \tag{2.34}$$

Let $S = \operatorname{span}(\{p_1, \ldots, p_m\})$ and $\ell \in [m']$ be arbitrary. We have that $\mathbf{q}_\ell'' = \operatorname{proj}_S(\mathbf{q}_\ell'') + \operatorname{proj}_{S^\perp}(\mathbf{q}_\ell'')$, noting that the summands in the decomposition are both L^2 bounded sequences. So now we have that

$$\left\langle \sum_{k=1}^{m'} b_k \mathbf{q}_k''^{\otimes 2} - \sum_{j=1}^{m} a_j p_j^{\otimes 2}, \operatorname{proj}_{S^\perp}(\mathbf{q}_\ell'')^{\otimes 2} \right\rangle \to 0 \tag{2.35}$$

$$\Rightarrow \left\langle \sum_{j=1}^{m'} b_j \mathbf{q}_j''^{\otimes 2}, \operatorname{proj}_{S^\perp}(\mathbf{q}_\ell'')^{\otimes 2} \right\rangle \to 0 \tag{2.36}$$

$$\Rightarrow b_\ell \left\langle \operatorname{proj}_{S^\perp}(\mathbf{q}_\ell'')^{\otimes 2}, \operatorname{proj}_{S^\perp}(\mathbf{q}_\ell'')^{\otimes 2} \right\rangle + \sum_{j \in [m'] \setminus \{\ell\}} b_j \left\langle \mathbf{q}_j'', \operatorname{proj}_{S^\perp}(\mathbf{q}_j'') \right\rangle^2 \to 0 \tag{2.37}$$

$$\Rightarrow b_\ell \left\| \operatorname{proj}_{S^\perp}(\mathbf{q}''_\ell) \right\|_2^4 \to 0. \tag{2.38}$$

From this we have that $\|\operatorname{proj}_S(\mathbf{q}_k'') - \mathbf{q}_k''\|_2 \to 0$ for all k. Since $\bigoplus_{j=1}^{m'} \operatorname{proj}_S(\mathbf{q}_j'')$ is a L^2 bounded sequence on a finite dimensional space by the Bolzano-Weierstrass theorem it has a convergent subsequence which converges to $\bigoplus_{j=1}^{m'} q_j''$ so $\mathbf{q}_j'' \to q_j''$ in L^2. From Hölder's Inequality we have

that, along this subsequence

$$\|\mathbf{q}_k'' - q_k''\|_1 \leq \|\mathbf{q}_k'' - q_k''\|_2 \|1\|_2 \leq \|\mathbf{q}_k'' - q_k''\|_2 \sqrt{\int 1^2 d\bar{\mu}} = \|\mathbf{q}_k'' - q_k''\|_2 \to 0 \tag{2.39}$$

so q_k'' is a probability density for all k, since they must be nonnegative to converge and integrate to one. Now we have that

$$\sum_{j=1}^{m} a_j p_j^{\times n} = \sum_{k=1}^{m'} b_k q_k''^{\times n}. \tag{2.40}$$

And defining ν_k as the probability measure associated with q_k'' we have that there exists a subsequence such that $\|\boldsymbol{\nu}_k - \nu_k\| \to 0$ for all k and

$$\sum_{j=1}^{m} a_j \mu_j^{\times n} = \sum_{k=1}^{m'} b_k {\nu_k}^{\times n}. \tag{2.41}$$

□

We can now prove Theorem 4.

Proof of Theorem 4. To help lighten notation we will simply bold some elements which depend on the sequence $\mathscr{P}_i$. Let $\mathscr{P}_i = \sum_{j=1}^{m_i'} \mathbf{b}_j \delta_{\boldsymbol{\nu}_j}$ be a sequence of mixtures of measures ($\mathbf{b}_j$, $\boldsymbol{\nu}_j$ are functions of i) such that $V_n(\mathscr{P}_i) \to V_n(\mathscr{P})$.

We define $\widetilde{\mathbf{b}}$ a sequence in Δ^m so that $\widetilde{\mathbf{b}}_j = \mathbf{b}_j$ for $j \leq m_i'$ and $\widetilde{\mathbf{b}}_k = 0$ for $k > m_i'$. Consider the case where there exists no sequence of permutations such that $\boldsymbol{\sigma}(\widetilde{\mathbf{b}}) \to a$. From this it would follow that there exists a subsequence on i and $\varepsilon > 0$ such that $\left\|\widetilde{\mathbf{b}} - \sigma(a)\right\| > \varepsilon$ for all $\sigma \in S_m$. The space

$$\Delta^m \cap \left(\bigcap_{\sigma \in S_m} \text{ball}\left(\sigma(a), \varepsilon\right)^C \right) \tag{2.42}$$

is compact (the ball is open) so there exists a sub-subsequence of i where $\widetilde{\mathbf{b}}$ converges to a point $b \notin S_m \cdot a$. Let $I \subset [m]$ be the indices of b which are nonzero and $m' = \max(I)$. For sufficiently

large i along our sub-subsequence we have that $m_i' \geq m'$ and furthermore

$$\left\| \sum_{j=1}^{m_i'} \mathbf{b}_j \boldsymbol{\nu}_j^{\times n} - \sum_{k \in I} b_k \boldsymbol{\nu}_k^{\times n} \right\| \leq \left\| \sum_{j \in I} \mathbf{b}_j \boldsymbol{\nu}_j^{\times n} - \sum_{k \in I} b_k \boldsymbol{\nu}_k^{\times n} \right\| + \left\| \sum_{j \in I^C} \mathbf{b}_j \boldsymbol{\nu}_j^{\times n} \right\| \tag{2.43}$$

$$\leq \left\| \sum_{j \in I} (\mathbf{b}_j - b_j) \boldsymbol{\nu}_j^{\times n} \right\| + \left| \sum_{j \in I^C} \mathbf{b}_j \right| \tag{2.44}$$

$$\leq \sum_{j \in I} |\mathbf{b}_j - b_j| + \left| \sum_{j \in I^C} \mathbf{b}_j \right| \to 0 \tag{2.45}$$

and therefore

$$\left\| \sum_{k=1}^{m} a_k \mu_k^{\times n} - \sum_{j \in I} b_j \boldsymbol{\nu}_j^{\times n} \right\| \to 0. \tag{2.46}$$

From Lemma 3 we have that there exists a subsequence of this sub-subsequence such that for $k \in I$ there exists probability measures ν_k with $\|\nu_k - \boldsymbol{\nu}_k\| \to 0$ and

$$\sum_{k=1}^{m} a_k \mu_k^{\times n} = \sum_{j \in I} b_j \nu_j^{\times n}. \tag{2.47}$$

If $|I| < m$ or $\nu_j = \nu_k$ for any $k \neq j$ and $j, k \in I$ then we have clearly violated identifiability since we can construct a mixture of measures $\mathscr{P}'$ with fewer components than $\mathscr{P}$ and $V_n(\mathscr{P}') = V_n(\mathscr{P})$. If $|I| = m$ (i.e. $I = [m]$) and ν_j are all distinct we have also arrived at a contradiction since letting $\mathscr{P}' = \sum_{j=1}^{m} b_j \delta_{\nu_j} \neq \mathscr{P}$ because there exists no σ such that $\sigma(b) = a$ and $V_n(\mathscr{P}') = V_n(\mathscr{P})$, contradicting identifiability.

So we have that for sufficiently large i that $m_i' = m$ and there exists at least one sequence $\boldsymbol{\sigma}$ such that $\boldsymbol{\sigma}(\mathbf{b}) \to a$. So let $\left\| \sum_{k=1}^{m} a_k \mu_k^{\times n} - \sum_{j=1}^{m} \mathbf{b}_j \boldsymbol{\nu}_j^{\times n} \right\| \to 0$. From what we have just shown, we can permute the indices and, without loss of generality, we can assume that $\mathbf{b} \to a$. So now we have that $\left\| \sum_{i=1}^{m} a_i \mu_i^{\times n} - \sum_{j=1}^{m} a_j \boldsymbol{\nu}_j^{\times n} \right\| \to 0$.

Let $\widetilde{S}_m \subset S_m$ be the subgroup of permutations such that $\sigma(a) = a$ for $\sigma \in \widetilde{S}_m$ (also known as the *stabilizer* of a). Note that if $a_1, \ldots, a_m$ are distinct then $\widetilde{S}_m$ only contains the identity. We proceed by contradiction: suppose there exists no sequence of permutations $\boldsymbol{\sigma} \in \mathbb{N}^{\widetilde{S}_m}$ such that $\boldsymbol{\nu}_{\boldsymbol{\sigma}(k)} \to \mu_k$ for all k. From this it follows that there exists a subsequence and a $\varepsilon > 0$, such that $\bigoplus_{k=1}^{m} \boldsymbol{\nu}_k$ does not lie in $\bigcap_{\sigma \in \widetilde{S}_m} \left(\text{ball} \left(\bigoplus_{k=1}^{m} \mu_{\sigma(k)} \right), \varepsilon \right)^C$. From Lemma 3 there exists probability

measures, $\nu_1, \ldots, \nu_m$ such that for some subsequence $\|\boldsymbol{\nu}_k - \nu_k\| \to 0$ for all k and

$$\sum_{j=1}^{m} a_j \mu_j^{\times n} = \sum_{k=1}^{m} a_k \nu_k^{\times n}.$$

Because $\bigcap_{\sigma \in \widetilde{S}_m} \left(\text{ball}\left(\bigoplus_{k=1}^{m} \mu_{\sigma(k)}\right), \varepsilon\right)^C$ is closed we have $\bigoplus_{j=1}^{m} \nu_j \in \bigcap_{\sigma \in \widetilde{S}_m} \left(\text{ball}\left(\bigoplus_{k=1}^{m} \mu_{\sigma(k)}\right), \varepsilon\right)^C$ and there exists no $\sigma \in \widetilde{S}_m$ such that $\nu_{\sigma(k)} = \mu_k$ for all k so. Setting $\mathscr{P}' = \sum_{k=1}^{m} a_k \delta_{\nu_k}$ we have that $\mathscr{P}' \neq \mathscr{P}$ but $V_n(\mathscr{P}') = V_n(\mathscr{P})$, a contradiction.

□

2.10.6 General Version of Corollary 1

Here we present the general version of Corollary 1 which guarantees recovery of the true mixture components using our estimator for any mixture model, provided there are a sufficient number of samples per group. For a mixture model $p = \sum_{m=1}^{M} w_m^* p_m^*$, using the estimator $\hat{q}$ from Section 2.10.3.2 to estimate (2.11):

$$q(y_1, y_2, \ldots, y_N) = \sum_{m=1}^{M} w_m^* p_m^*(y_1) p_m^*(y_2) \ldots p_m^*(y_N), \quad y_1, y_2, \ldots, y_N \in \mathbb{R}^d.$$

combining Theorem 2a and Theorem 3 gives the following result.

Corollary 4. *If $\sigma \to 0$ and $\frac{n\sigma^{2Nd}}{\log n} \to \infty$ as $n \to \infty$, and p is N-identifiable (e.g. $N = 2M - 1$), then $\hat{w}_m \xrightarrow{a.s.} w_m^*$ and $\|p(\cdot; \hat{\alpha}_m) - p_m^*\|_1 \xrightarrow{a.s.} 0$, up to a permutation.*

CHAPTER 3

Consistent Estimation of Nonparametric Mixtures of Regressions from Grouped Observations

3.1 Introduction

Mixtures of regressions differ from the standard mixture model in that no distributional assumptions are made on the covariates, while the response variable follows a mixture distribution. The mixtures of linear regressions (MLR) problem was first formalized by DeVeaux [DV89] partially motivated by a musical perception task [Coh84]. Such a model can be applied whenever one of multiple processes may be responsible for the measured response. For example, consider regressing students' exam scores against the time taken to complete the exam. We may expect that, for many students, scores will increase if extra time and care are taken on the exam. However, students that are very familiar with the material may finish the exam quickly and score very well. Further still, some students who are struggling with the material may become frustrated and hand in an incomplete exam early. In general, the mixtures of regressions framework can be applied whenever a dataset is well fit only by a function of very high complexity, while partitioning data into groups which are each well fit by functions of low complexity.

Much of the existing work in this space considers mixtures of linear regressions. Also called hierarchical mixtures of experts [JJ94], there are known results regarding local and global convergence of expectation maximization (EM) in the noiseless [YCS14] and noisy [KC20, BWY17] settings, sample complexity [KHC21], and other properties of the problem. Much less is known for nonparametric mixtures of regressions (NMR). Identifiability of the NMR model has been shown, and a consistent estimator developed, in the univariate case under restrictive assumptions on the

This chapter presents ongoing and unpublished work with Clayton Scott, Laura Balzano, and Jessica Jiang. We extend the approach in Chapter 2 to the mixtures of nonlinear regressions problem.

mean functions [HLW13]. Again in the univariate case, consistency and convergence of EM can be shown for NMR if the mean and variance functions are assumed constant [XY18].

The current work provides considers the NMR problem in the multivariate setting. In Section 3.3 we show consistency results for a kernel-based estimator under minimal assumptions on the mean functions. We do this by considering the grouped observation setting, analogous to Chapter 2. Furthermore, we characterize the conditions under which mixtures of regressions are identifiable from groups of size 2 (pairs) and 3 (triples). Notably, we show that mixtures of regressions are identifiable from (at most) triples in nearly all problems of practical interest, while pairs suffice in an interesting class of problems. These results are shown in Section 3.3.4.

3.2 Problem Setup

Throughout, let $\mathcal{X}$ denote the design space and $[M] = \{1, 2, \ldots, M\}$ denotes the index set of mixture components, with M known and fixed. Let $\mathcal{X} \subseteq \mathbb{R}^d$. Let $M \geq 2$. Let $f : \mathcal{X} \to \mathbb{R}^M$ and write $f = (f_1, \ldots, f_M)$, where $f_m : \mathcal{X} \to \mathbb{R}$ for $m = 1, \ldots, M$. Let π^* belong to the probability simplex Δ^{M-1} and have all nonzero entries: $\sum_m \pi_m^* = 1$ and $\pi_m^* > 0$ for all m. Denote $\pi_{\text{min}}^* = \min_m \pi_m^* > 0$. Let π^* and $f_1^*, \ldots, f_M^*$ by fixed. Define the joint distribution of

$$(X, Y, X', Y')$$

as follows. X and X' are drawn independently according to the Borel probability measure Q_X. Conditioned on $X = x$ and $X' = x'$, Y and Y' have conditional density

$$q(y, y' \mid x, x'; \pi^*, f^*),$$

where

$$q(y, y' \mid x, x'; \pi, f) := \sum_{m=1}^{M} \pi_m \phi(y; f_m(x), \sigma^2)\phi(y'; f_m(x'), \sigma^2),$$

ϕ is the univariate normal pdf, and $\sigma^2 > 0$ is fixed and known.

Equivalently, to generate a realization of (X, Y, X', Y'), X and X' are drawn independently according to Q_X, and then, conditioned on X and X', Y and Y' are generated by first selecting component m with probability π_m^*, and then setting $Y = f_m^*(X) + \sigma W$ and $Y' = f_m^*(X') + \sigma W'$ where $W, W' \sim \mathcal{N}(0, 1)$ are independent.

3.2.1 Estimation Problem

For π in the probability simplex and $t, t' \in \mathbb{R}^M$, define the loss function, equal to the negative conditional log-likelihood,

$$\ell(y, y', \pi, t, t') = -\log\left(\sum_{m=1}^{M} \pi_m \phi(y; t_m, \sigma^2)\phi(y'; t'_m, \sigma^2)\right).$$

Given a probability vector π and $f : \mathcal{X} \to \mathbb{R}^M$, define the risk

$$J(\pi, f) = \mathbb{E}[\ell(Y, Y', \pi, f(X), f(X')]$$

and the empirical risk

$$\widehat{J}(\pi, f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, Y'_i, \pi, f(X_i), f(X'_i))$$

where $(X_i, Y_i, X'_i, Y'_i)_{i=1}^n$ are i.i.d. copies of (X, Y, X', Y'). The excess risk for any (π, f) is the expected KL divergence

$$J(\pi, f) - J(\pi^*, f^*) = \mathbb{E}_{X,X'} D_{\mathrm{KL}}(q(\cdot, \cdot|X, X'; \pi, f) \,\|\, q(\cdot, \cdot|X, X'; \pi^*, f^*)).$$

where the expectation is taken with respect to the true distribution of (X, X'). We note that (π^*, f^*) is a global minimizer of J, since for any (π, f), nonnegativity of the KL divergence implies $J(\pi, f) - J(\pi^*, f^*) \geq 0$ Furthermore, equality holds iff the two conditional densities agree almost surely (Q_X).

This work considers the regularized maximum likelihood estimation problem,

$$(\hat{\pi}, \hat{f}) = \underset{\pi, f}{\arg\min}\ \widehat{J}(\pi, f) + \lambda\psi(f), \tag{3.1}$$

where $\lambda \geq 0$ and $\psi(\cdot)$ is a regularizer controlling the complexity of the estimated mean functions $f_1, \ldots, f_M$. In general, the form of $\psi(\cdot)$ will depend on the parameterization of the estimator used to represent the mean functions. These considerations are covered in more detail in section 4.4. First, however, we discuss optimization of problem 3.1.

3.2.2 EM Algorithm

In this section, we derive an EM algorithm for solving problem 3.1 in a form that is agnostic to the parameterization of the mean functions. Introduce the latent class variables $Z_1, \ldots, Z_n$ following

a categorical distribution parameterized by π^*. Denote the observed data log-likelihood

$$\ell(\pi, f) \coloneqq \sum_{i=1}^{n} \log \left[\sum_{m=1}^{M} \pi_m \, \phi(y_i;\, f_m(x_i),\, \sigma^2)\, \phi(y_i';\, f_m(x_i'),\, \sigma^2) \right], \tag{3.2}$$

and the complete data log-likelihood

$$\ell_c(\pi, f) \coloneqq \sum_{i=1}^{n} \sum_{m=1}^{M} \mathbf{1}_{\{z_i = m\}} \left[\log \pi_m + \log \phi(y_i;\, f_m(x_i),\, \sigma^2) + \log \phi(y_i';\, f_m(x_i'),\, \sigma^2) \right]. \tag{3.3}$$

Let $f^{t-1} = (f_1^{t-1}, \ldots, f_M^{t-1})$ denote the current estimate of f^*, parameterized by some parameter vector (kernel coefficients, neural network weights, etc.). Similarly, let π^{t-1} be the current estimate of π^*.

E-Step. The E-step computes the posterior responsibility of the the component m for pair i, given the current parameter estimates (π^{t-1}, f^{t-1}):

$$\begin{aligned} w_{i,m}^t &= \mathbb{E}_{Z|X,Y,X',Y'} \left[\mathbf{1}_{\{z_i = m\}} \right] \\ &= \Pr\left(z_i = m | X = x_i, Y = y_i, X' = x_i', Y' = y_i' \right) \\ &= \frac{\pi_m^{t-1}\, \phi(y_i;\, f_m^{t-1}(x_i),\, \sigma^2)\, \phi(y_i';\, f_m^{t-1}(x_i'),\, \sigma^2)}{\displaystyle\sum_{s=1}^{M} \pi_s^{t-1}\, \phi(y_i;\, f_s^{t-1}(x_i),\, \sigma^2)\, \phi(y_i';\, f_s^{t-1}(x_i'),\, \sigma^2)}. \end{aligned} \tag{3.4}$$

This is the standard posterior computation, with the crucial difference that the likelihood for each pair involves *two* evaluations of f_m (at x_i and x_i'), coupled through the assignment of the shared component z_i. This means that even if one observation in a pair is ambiguous (equidistant from two components), the other observation can disambiguate the assignment. Intuitively, this is the mechanism by which grouped observations provide identifiability.

M-Step The M-step forms the expected complete data log-likelihood and computes the mixing weights and mean functions given the current estimates of the posterior responsibilities $w_{i,m}^t$. The expected complete data log-likelihood is

$$\sum_{i=1}^{n} \sum_{m=1}^{M} w_{i,m}^t \left[\log \pi_m - \frac{1}{2\sigma^2} \left((y_i - f_m(x_i))^2 + (y_i' - f_m(x_i'))^2 \right) \right] + C, \tag{3.5}$$

where C collects terms that do not depend on (π, f). Maximizing over $\pi \in \Delta^{M-1}$ gives the standard update

$$\pi_m^t = \frac{1}{n} \sum_{i=1}^{n} w_{i,m}^t. \tag{3.6}$$

For the mean functions, maximizing over f decouples across components. For each $m \in [M]$, we solve

$$f_m^t = \arg\min_{f_m \in \mathcal{F}} \sum_{i=1}^{n} w_{i,m}^t \left[(y_i - f_m(x_i))^2 + (y_i' - f_m(x_i'))^2 \right] + \lambda \psi(f_m), \tag{3.7}$$

where $\mathcal{F}$ is the function class (RKHS, family of neural networks, etc.). This is a regularized weighted least squares problem where the weight for both observations in pair i is $w_{i,m}^t$ (the *same* weight for both members of the pair, reflecting their shared component assignment).

We note that the grouped observation model does not change the functional form of the M-step optimization. It simply provides a specific weighting pattern where paired observations share weights. The function f_m is evaluated pointwise and does not need to "know" about the pairing.

3.2.3 Kernel Method

Up to this point, the problem description and approach presented has remained general enough to accommodate different parameterizations mean function estimator $\hat{f}$. In what follows, we limit our attention to a particular form of kernel-based estimator. Specifically, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric, positive definite kernel with reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. Let $\nu \in (0, \pi_{\min}^*)$ and define $\Delta_\nu^{M-1} := \{\pi = (\pi_1, \ldots, \pi_M) \mid \sum_k \pi_k = 1, \pi_k \geq \nu\}$. We now set up the particular form of problem 3.1 applied to the kernel estimator. Let the regularizer $\psi(f_m) = \|f_m\|_{\mathcal{H}}^2$ be the RKHS norm. For $\lambda > 0$, problem 3.1 becomes

$$(\widehat{\pi}_n, \widehat{f}_n) \in \underset{(\pi, f) \in \Delta_\nu^{M-1} \times \mathcal{H}^M}{\arg\min} \widehat{J}(\pi, f) + \lambda \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}}^2, \tag{3.8}$$

which is an infinite dimensional optimization problem. Fortunately, a representer theorem holds for problem (3.8):

Theorem 5. *If $(\widehat{\pi}, \widehat{f})$ is a global minimizer of 3.8, then for each m, there exist $\alpha_m, \alpha_m' \in \mathbb{R}^n$ such that*

$$\widehat{f}_m(x) = \sum_{i=1}^{n} \alpha_{m,i} k(x, X_i) + \sum_{i=1}^{n} \alpha_{m,i}' k(x, X_i').$$

This reduces the infinite dimensional optimization problem to the finite dimensional problem

$$(\widehat{\pi}, \widehat{\alpha}) \in \underset{(\pi,\alpha)\in\Delta_\nu^{M-1}\times\mathbb{R}^{M\times 2n}}{\arg\min} \widehat{J}(\pi,\alpha) + \lambda \sum_{m=1}^{M} \alpha_m^\top K \alpha_m, \tag{3.9}$$

where we define

$$\widehat{J}(\pi,\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \log q_{\pi,\alpha}(y_{i,1}, y_{i,2}|x_{i,1}, x_{i,2})$$

which can be solved by a particular instantiation of the EM algorithm presented in section 3.2.2.

EM Algorithm for Kernel Method. To set up the M-step, we first introduce some additional notation. For convenience, we define an alternative indexing of the observed data

$$(x_{i,1}, y_{i,1}, x_{i,2}, y_{i,2}) \coloneqq (x_i, y_i, x_i', y_i,), \ \forall i \in [n],$$

so that, appealing to Theorem 5, we consider mean function estimators of the form

$$f_m^\alpha(\cdot) = \sum_{i=1}^{n}\sum_{i'=1}^{2} \alpha_{m,i,i'}\, k(\cdot, x_{i,i'}). \tag{3.10}$$

Further denote the mth row of α (as a column vector) by

$$\alpha_m = \left[\alpha_{m,1,1}, \alpha_{m,1,2}, \alpha_{m,2,1}, \alpha_{m,2,2}, \ldots, \alpha_{m,n,1}, \alpha_{m,n,2}\right]^\top.$$

Now define

$$Y = [y_{1,1},\, y_{1,2},\, y_{2,1},\, y_{2,2},\, \ldots,\, y_{n,1},\, y_{n,2}]^\top \in \mathbb{R}^{2n},$$

the kernel matrix $K \in \mathbb{R}^{2n\times 2n}$ with entries

$$K_{(2i-2+i',\, 2j-2+j')} = k(x_{i,i'},\, x_{j,j'}), \qquad i,j \in [n], \quad i',j' \in \{1,2\},$$

and the diagonal weight matrix

$$W_m^t = \mathrm{diag}(w_{1,m}^t,\, w_{1,m}^t,\, w_{2,m}^t,\, w_{2,m}^t,\, \ldots,\, w_{n,m}^t,\, w_{n,m}^t) \in \mathbb{R}^{2n\times 2n}.$$

Note that each responsibility $w_{i,m}^t$ appears twice on the diagonal, once for each element of the pair. The M-step objective (3.7) becomes

$$\min_{\alpha_m} (Y - K\alpha_m)^\top W_m^t (Y - K\alpha_m) + \lambda \alpha_m^\top K \alpha_m,$$

with closed-form solution given by

$$\alpha_m^t = (W_m^t K + \lambda I)^{-1} W_m^t Y. \tag{3.11}$$

To derive the E-step, we must take into account the lower bound ν. The conditions of Theorem 7 imply ν should be selected as a function of n such that $\frac{1}{\nu}$ grows slower than $O(n)$, e.g., $\nu = \frac{c \log(n)}{n}$.

Let $a_1, \ldots, a_M > 0$, and let $0 < \nu \leq 1/M$. With ν chosen, the E-step is equivalent to the problem

$$\min_{\pi \in \mathbb{R}^M} \left\{ -\sum_{i=1}^n \sum_{m=1}^M a_m \log \pi_m \right\}$$

subject to

$$\sum_{m=1}^M \pi_m = 1, \qquad \pi_m \geq \nu, \quad m = 1, \ldots, M,$$

where $a_m = \sum_{i=1}^n w_{i,m}$.

Note that both the objective and feasible set are convex. The KKT conditions imply that the solution has the form

$$\pi_m^\star = \max\left\{\nu, \frac{a_m}{\lambda}\right\}, \qquad m = 1, \ldots, M,$$

where $\lambda > 0$ is chosen so that the weights sum to one:

$$\sum_{m=1}^M \max\left\{\nu, \frac{a_m}{\lambda}\right\} = 1.$$

Equivalently, λ is the unique solution of

$$g(\lambda) = \sum_{m=1}^M \max\left\{\nu, \frac{a_m}{\lambda}\right\} - 1 = 0.$$

The function g is monotone decreasing in λ, so λ can be found by one-dimensional bisection. Once λ has been found, the optimizer is given by

$$\pi_m^\star = \max\left\{\nu, \frac{a_m}{\lambda}\right\}.$$

In the special case $\nu = 0$, this reduces to the familiar closed-form solution

$$\pi_m^\star = \frac{a_m}{\sum_{\ell=1}^M a_\ell}.$$

3.3 Statistical Theory

In this section we state our assumptions and main results. Formal proofs are given in the Supplement 3.5. Our overall approach mirrors that of section 2.6. First, we show that the minimizer $(\widehat{\pi}, \widehat{\alpha})$ of (3.9) produces a consistent estimate $q_{\hat{\pi},\hat{\alpha}}$ of the mixture of regressions q (Theorems 6, 7, and Corollary 5). Second, we show that if q is 2-identifiable, then the mean functions $\widehat{f}_m$ and mixing weights $\widehat{\pi}_m$ of $q_{\hat{\pi},\hat{\alpha}}$ are consistent estimates of f_m^* and π_m^*, respectively (Theorem 8). We remark that the extension of our consistency results to the case of N-identifiability is straightforward. Finally, we characterize the conditions for 2-identifiability (Lemma 5) and 3-identifiability (Lemma 6) of the mixture of regressions q, showing that nearly all problems of practical interest fall under these two cases. Taken together, these results characterize consistent nonparametric estimation of mixtures of regressions in the grouped observation setting.

Throughout, we assume that $\mathcal{X}$ is a compact metric space, and that Q_X is a Borel probability measure. Additionally, we make the following assumptions.

Assumption 1. *The kernel function is positive definite and bounded, i.e.,* $0 < k(x, x') \leq C_k$ *for all* $x, x' \in \mathcal{X}$.

Assumption 2 (Universal kernel)**.** *The kernel* k *is universal on the compact metric space* $\mathcal{X}$*, i.e.,* $\mathcal{H}$ *is dense in* $C(\mathcal{X})$ *in the supremum norm.*

Assumption 3. *The true mixing weights are strictly positive, i.e.,* $\pi_{min}^* = \min_m \pi_m^* > 0$.

Remark 1. *(Lower bound of* π_{min}^**) Certain parts of the proof machinery consider a restriction of* $\widehat{\pi}$ *to the set* Δ_ν^{M-1} *defined previously. Ultimately, we show consistency when* $\nu \to 0$ *as* $n \to \infty$ *at an appropriate rate (see Theorem 7), so Assumption 3 is sufficient for our purposes.*

Assumption 4. *The true mean functions* $f_m^* \in L^2(Q_X)$ *for all* $m \in [M]$.

Assumption 5. *There exists a finite constant* $C_* \in (0, \infty)$ *such that* $|f_m^*(x)| \leq C_*$ *for all* $x \in \mathcal{X}, m \in [M]$.

3.3.1 Uniform deviation bound

We make use of the following assumption, introduced by Scott and Zhang [SZ20], concerning the boundedness and Rademacher complexity of a function class $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$.

(BR) There exist finite constants A and B such that $\sup_{f\in\mathcal{F}}\sup_{x\in\mathcal{X}}|f(x)|\le A$, and for all $N\ge 2$ and $x_1,\dots,x_N\in\mathcal{X}$,

$$\mathbb{E}_{(\epsilon_i)}\left[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}\epsilon_i f(x_i)\right]\le\frac{B}{\sqrt{N}}.$$

Lemma 4. *If $\sup_{x\in\mathcal{X}}\sqrt{k(x,x)}=K<\infty$ and $\mathcal{F}$ is the ball in $\mathcal{H}$ with radius R and centered at 0, then $\mathcal{F}$ satisfies* **(BR)** *with $A=RK$ and $B=RK$.*

Lemma 4 is a restatement of a result by Scott and Zhang [SZ20], who also note that the result holds for the class of two-layer neural networks with ReLU activation.

We now state a uniform deviation bound for $|J(\pi,f)-\widehat{J}(\pi,f)|$, where $f_m\in\mathcal{F}$ for all $m\in[M]$, for a function class $\mathcal{F}$ satisfying assumption **(BR)**.

Theorem 6 (Uniform deviation bound)**.** *Let Assumptions 3 and 5 hold. Further assume that $f_m^*\in\mathcal{F}$ for all $m\in[M]$ for $\mathcal{F}$ satisfying assumption* **(BR)**. *For $n\ge 2$ and $\delta\in(0,1)$, define*

$$u_{n,\delta}=\sqrt{2\log\frac{8n}{\delta}},$$

$$L_{n,\delta}=\frac{A+C_*+\sigma u_{n,\delta}}{\sigma^2},$$

$$C_{n,\delta}=-\log\nu+\max\{0,\log(2\pi\sigma^2)\}+\frac{(A+C_*+\sigma u_{n,\delta})^2}{\sigma^2}.$$

Also define

$$C_0=-\log\nu+|\log(2\pi\sigma^2)|+\frac{2A^2}{\sigma^2}+\frac{4C_*^2}{\sigma^2}.$$

Then, with probability at least $1-\delta$,

$$\sup_{\pi\in\Delta_\nu^{M-1},\,f\in\mathcal{F}^M}|J(\pi,f)-\widehat{J}(\pi,f)|\le r_n(\delta),$$

where

$$r_n(\delta)=\frac{8M^{3/2}L_{n,\delta}B}{\sqrt{n}}+C_{n,\delta}\sqrt{\frac{(M-1)\log(2Mn+1)+\log(4/\delta)}{2n}}+\frac{2}{\nu n}+\frac{\delta}{n}\left(\frac{C_0}{2}+1+\sqrt{2\log\frac{8n}{\delta}}\right).$$

Consequently, if

$$\delta_\epsilon(n) = \Pr\left(\sup_{\pi\in\Delta_\nu^{M-1},\, f\in\mathcal{F}^M} |J(\pi,f) - \widehat{J}(\pi,f)| > \epsilon\right),$$

then for every fixed $\epsilon > 0$*,*

$$\sum_{n=1}^{\infty} \delta_\epsilon(n) < \infty.$$

3.3.2 Consistency for the kernel method

Critically, the probabilities $\delta_\epsilon(n)$ in Theorem 6 are summable, allowing us to establish almost sure convergence of $J(\widehat{\pi}, \widehat{f}) \to J(\pi^*, f^*)$ as $n \to \infty$, by applying the first Borel-Cantelli lemma with appropriate growth rates (as a function of sample-size n) on ν and λ. The result is established for kernels k, such that the corresponding RKHS is dense in the space of continuous functions on $\mathcal{X}$. Such kernels are referred to as *universal* on $\mathcal{X}$.

Theorem 7 (Consistency of the kernel method)**.** *Let kernel* k *satisfy assumptions 1 and 2 with RKHS* $\mathcal{H}$*. Let the mean functions* f_m^* *satisfy assumptions 4 and 5. Let* π^* *satisfy assumption 3. For each* $n \geq 2$*, Let* $\nu^{(n)} \in (0, 1/M)$ *and* $\lambda^{(n)} > 0$ *satisfy*

$$\nu^{(n)} \to 0, \qquad \lambda^{(n)} \to 0,$$

as $n \to \infty$*, and*

$$\frac{1}{\nu^{(n)} n} \to 0, \qquad \frac{1}{\lambda^{(n)}}\sqrt{\frac{\log n}{n}} \to 0.$$

For each n*, let*

$$(\widehat{\pi}^{(n)}, \widehat{f}^{(n)}) \in \operatorname*{arg\,min}_{(\pi,f)\in\Delta_{\nu^{(n)}}^{M-1}\times\mathcal{H}^M} \left\{\widehat{J}(\pi,f) + \lambda^{(n)} \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}}^2\right\}.$$

Then

$$J(\widehat{\pi}^{(n)}, \widehat{f}^{(n)}) \to J(\pi^*, f^*) \qquad \textit{almost surely.}$$

The following consistency result for $q_{\hat{\pi},\hat{\alpha}}$ follows almost immediately from Theorem 7 via Pinsker's and Jensen's inequalities.

Corollary 5 (L^1-consistency)**.** *Under the conditions of Theorem 7,*

$$\mathbb{E}_{X,X'}\left[\|q(\cdot,\cdot \mid x,x') - q_{\hat{\pi},\hat{\alpha}}(\cdot,\cdot \mid x,x')\|_{L^1(\mathbb{R}^2)}\right] \overset{a.s.}{\to} 0.$$

3.3.3 Consistent Recovery of Mixture Components

This section states our main result, showing that consistent estimation of q implies consistent estimation of the true mixing weights π^* and mean functions f_m^*. For now, we assume the distribution Q is 2-identifiable. If it is instead N-identifiable, we note that all the previous results extend for groups of size N. We now show that if $\mathbb{E}_P \|q - q_{\hat{\alpha},\hat{\pi}}\|_{L^1(\mathbb{R}^2)} \to 0$ almost surely, the underlying joint densities converge in $L^1(\mathcal{X} \times \mathbb{R})$. Appealing to Theorem 5 of [RVS20], we will show the mixing weights are consistently estimated.

Theorem 8 (Consistent Recovery)**.** *Suppose distribution Q is 2-identifiable. Further assume that Q_X is absolutely continuous with respect to the Lesbesgue measure. Under the conditions of Corollary 5, $\hat{\pi}_m \overset{a.s.}{\to} \pi_m^*$ and $\|f_m^* - f_m^{\hat{\alpha}}\|_{L^1(Q_X)} \overset{a.s.}{\to} 0$, up to a permutation.*

3.3.4 Identifiability of NMR

Any finite mixture model with M mixture components is identifiable from groups of size $2M - 1$, but mixture models with certain types of structure are identifiable from groups of a much smaller size [VS19]. In this section we provide a necessary and sufficient condition for the mixture of regressions Q to be identifiable from groups of size 2 and 3. As will be seen, nearly any problem of practical interest will be 3-identifiable. Furthermore an interesting subset of problems are 2-identifiable, including nearly every problem of practical interest where $M = 2$.

A finite mixture is identifiable from groups of size 2 if the mixture component distributions satisfy a notion known as *joint irreducibility* [VS19]. A collection of distributions is jointly irreducible if a valid probability distribution can be obtained through their linear combination only when coefficients are non-negative [BS14]. It is satisfied, for example, each distribution has a region of support not shared by the others. The result below shows that the mixture of regressions Q is 2-identifiable if and only if, for each underlying mean function f_m^*, there is a set of nonzero probability where f_m^* is is strictly greater than (or strictly less than) all other mean functions.

Lemma 5. *The collection of joint densities $\{p_m(x, y)\}_{m=1}^M$, where $p_m(x, y) = \phi(y; f_m^*(x), \sigma^2)p(x)$, is jointly irreducible if and only if for each mean function in the collection $\{f_m^*\}_{m=1}^M$, there exists a subset $S_m \subseteq \mathcal{X}$ with $Q_X(S_m) > 0$ such that, for all $j \in [M]$ with $j \neq m$, either $f_m^*(x) > f_j^*(x)$ for all $x \in S_m$, or $f_m^*(x) < f_j^*(x)$ for all $x \in S_m$.*

Remark 2. *(The case $M = 2$.) When there are exactly two mixture components, we note that Lemma 5 applies when the two mean functions differ on a set of positive probability. In other words, paired observations are sufficient for identifiability provided the two components correspond to* different regression problems.

Remark 3. *(The case $M = 3$.) When there are exactly 3 mixture components, Lemma 5 applies when the ordering of the values of the mean functions changes on a set of positive probability.*

If the conditions of Lemma 5 are not satisfied, we can make use of the fact that a finite mixture with M is identifiable from groups of size 3 if the mixture component distributions are linearly independent [VS19]. As it turns out, linear independence of the mixture components is satisfied under very general conditions. The result below shows that the mixture of regressions is 3-identifiable when, for each underlying mean function f_m^*, there is a set of positive probability where f_m^* differs from all other mean functions.

Lemma 6. *The collection of joint densities $\{q_m(x,y)\}_{m=1}^M$, where $q_m(x,y) = \phi(y; f_m^*(x), \sigma^2)q_X(x)$, is linearly independent if $Q_X(\{x \mid f_m^*(x) \neq f_j^*(x)\ \forall j \neq m\}) > 0$ for all $m \in [M]$.*

3.4 Conclusion

In this work we introduced a novel variant of the kernel density estimator that yields consistent estimates of any identifiable nonparametric mixture of regressions from grouped observations. We established a consistency result for estimators of the form $\widehat{q}_{\hat{\pi},\hat{\alpha}}$. Namely, consistent estimation of q implies consistent estimates of the underlying components when the mixture of regressions is identifiable. Furthermore, we establish necessary and sufficient conditions for the mixture of regressions model to be 2-identifiable. Finally, we derive an EM algorithm which is extensible to various forms of mean function estimator and arbitrary group size.

3.5 Supplement

3.5.1 Proof of Theorem 5

Theorem 5. *If $(\widehat{\pi}, \widehat{f})$ is a global minimizer of 3.8, then for each m, there exist $\alpha_m, \alpha'_m \in \mathbb{R}^n$ such that*

$$\widehat{f}_m(x) = \sum_{i=1}^n \alpha_{m,i} k(x, X_i) + \sum_{i=1}^n \alpha'_{m,i} k(x, X'_i).$$

Proof. Let

$$\mathcal{S} = \operatorname{span}\{k(\cdot, X_1), \ldots, k(\cdot, X_n), k(\cdot, X'_1), \ldots, k(\cdot, X'_n)\} \subseteq \mathcal{H}.$$

For each m, decompose

$$\widehat{f}_m = \widehat{f}_{m,\parallel} + \widehat{f}_{m,\perp},$$

where

$$\widehat{f}_{m,\|} \in \mathcal{S}, \qquad \widehat{f}_{m,\perp} \in \mathcal{S}^{\perp}.$$

By the reproducing property, for every $i = 1, \ldots, n$,

$$\widehat{f}_{m,\perp}(X_i) = \langle \widehat{f}_{m,\perp}, k(\cdot, X_i) \rangle_{\mathcal{H}} = 0,$$

and similarly,

$$\widehat{f}_{m,\perp}(X_i') = \langle \widehat{f}_{m,\perp}, k(\cdot, X_i') \rangle_{\mathcal{H}} = 0.$$

Therefore

$$\widehat{f}_m(X_i) = \widehat{f}_{m,\|}(X_i), \qquad \widehat{f}_m(X_i') = \widehat{f}_{m,\|}(X_i')$$

for all i and all m.

Hence the empirical risk depends only on the parallel components:

$$\widehat{J}(\widehat{\pi}, \widehat{f}) = \widehat{J}(\widehat{\pi}, \widehat{f}_{\|}),$$

where

$$\widehat{f}_{\|} = (\widehat{f}_{1,\|}, \ldots, \widehat{f}_{M,\|}).$$

On the other hand, by orthogonality,

$$\|\widehat{f}_m\|_{\mathcal{H}}^2 = \|\widehat{f}_{m,\|}\|_{\mathcal{H}}^2 + \|\widehat{f}_{m,\perp}\|_{\mathcal{H}}^2 \geq \|\widehat{f}_{m,\|}\|_{\mathcal{H}}^2.$$

Therefore,

$$\widehat{J}(\widehat{\pi}, \widehat{f}_{\|}) + \lambda \sum_{m=1}^{M} \|\widehat{f}_{m,\|}\|_{\mathcal{H}}^2 \leq \widehat{J}(\widehat{\pi}, \widehat{f}) + \lambda \sum_{m=1}^{M} \|\widehat{f}_m\|_{\mathcal{H}}^2.$$

Since $(\widehat{\pi}, \widehat{f})$ is a minimizer, equality must hold. Because $\lambda > 0$, equality in the norm inequality implies

$$\|\widehat{f}_{m,\perp}\|_{\mathcal{H}}^2 = 0$$

for every m. Hence

$$\widehat{f}_m = \widehat{f}_{m,\|} \in \mathcal{S}.$$

Thus each $\widehat{f}_m$ admits an expansion of the form

$$\widehat{f}_m(x) = \sum_{i=1}^{n} \alpha_{m,i} k(x, X_i) + \sum_{i=1}^{n} \alpha'_{m,i} k(x, X_i'),$$

as claimed. □

3.5.2 Proof of Theorem 6

Lemma 7 (Gaussian tail bounds)**.** *Let $W \sim N(0,1)$. Then, for every $u > 0$,*

$$\Pr(|W| > u) \le 2e^{-u^2/2}.$$

Moreover, for every $u \ge 1$,

$$\mathbb{E}\big[W^2 \mathbf{1}_{\{|W|>u\}}\big] \le 2ue^{-u^2/2}.$$

Proof. For $\lambda > 0$, Chernoff's bound gives

$$\Pr(W > u) = \Pr(e^{\lambda W} > e^{\lambda u}) \le e^{-\lambda u}\mathbb{E}e^{\lambda W} = e^{-\lambda u + \lambda^2/2}.$$

Choosing $\lambda = u$ yields

$$\Pr(W > u) \le e^{-u^2/2}.$$

By symmetry,

$$\Pr(|W| > u) \le 2e^{-u^2/2}.$$

For the second bound,

$$\mathbb{E}\big[W^2 \mathbf{1}_{\{|W|>u\}}\big] = 2\int_u^\infty w^2 \phi(w)\, dw.$$

Since

$$\phi'(w) = -w\phi(w),$$

integration by parts gives

$$\int_u^\infty w^2\phi(w)\, dw = u\phi(u) + \int_u^\infty \phi(w)\, dw.$$

Also,

$$\int_u^\infty \phi(w)\, dw \le \frac{1}{u}\int_u^\infty w\phi(w)\, dw = \frac{\phi(u)}{u}.$$

Therefore, for $u \ge 1$,

$$\mathbb{E}\big[W^2 \mathbf{1}_{\{|W|>u\}}\big] \le 2\left(u + \frac{1}{u}\right)\phi(u) \le 4u\phi(u) \le 2ue^{-u^2/2}.$$

□

Lemma 8 (Properties of the truncated loss)**.** *Assume* $\pi, \rho \in \Delta_\nu^{M-1}$, $|y| \le T$, $|y'| \le T$, *and*

$$|t_m|, |s_m|, |t'_m|, |s'_m| \le A$$

for all $m = 1, \dots, M$. *Then*

$$\log(2\pi\sigma^2) \le \ell(y, y', \pi, t, t') \le -\log \nu + \log(2\pi\sigma^2) + \frac{(T+A)^2}{\sigma^2}.$$

Moreover,

$$|\ell(y, y', \pi, t, t') - \ell(y, y', \rho, t, t')| \le \frac{1}{\nu}\|\pi - \rho\|_1,$$

and, for every fixed $\rho \in \Delta_\nu^{M-1}$,

$$|\ell(y, y', \rho, t, t') - \ell(y, y', \rho, s, s')| \le \frac{T+A}{\sigma^2}\left(\|t - s\|_1 + \|t' - s'\|_1\right).$$

Proof. First, since $\pi_m \ge \nu$,

$$\sum_{m=1}^{M} \pi_m \phi(y; t_m, \sigma^2)\phi(y'; t'_m, \sigma^2) \ge \nu\phi(y; t_1, \sigma^2)\phi(y'; t'_1, \sigma^2).$$

The choice of index 1 here is arbitrary. Therefore,

$$\ell(y, y', \pi, t, t') \le -\log \nu + \log(2\pi\sigma^2) + \frac{(y - t_1)^2 + (y' - t'_1)^2}{2\sigma^2}.$$

Using $|y|, |y'| \le T$ and $|t_1|, |t'_1| \le A$ gives the stated upper bound. For the lower bound, each Gaussian density is at most $(2\pi\sigma^2)^{-1/2}$, so

$$\sum_{m=1}^{M} \pi_m \phi(y; t_m, \sigma^2)\phi(y'; t'_m, \sigma^2) \le \frac{1}{2\pi\sigma^2}.$$

Thus $\ell(y, y', \pi, t, t') \ge \log(2\pi\sigma^2)$.

Next, for fixed y, y', t, t', define $H_m = \phi(y; t_m, \sigma^2)\phi(y'; t'_m, \sigma^2)$. Using

$$|\log a - \log b| \le \frac{|a - b|}{\min\{a, b\}}, \qquad a, b > 0,$$

we get

$$|\ell(y, y', \pi, t, t') - \ell(y, y', \rho, t, t')| \le \frac{|\sum_m (\pi_m - \rho_m) H_m|}{\min\{\sum_m \pi_m H_m, \sum_m \rho_m H_m\}}.$$

Since $\sum_m \pi_m H_m \geq \nu \sum_m H_m$, $\sum_m \rho_m H_m \geq \nu \sum_m H_m$, and

$$\left| \sum_m (\pi_m - \rho_m) H_m \right| \leq \|\pi - \rho\|_1 \sum_m H_m,$$

the π-Lipschitz bound follows.

Finally, fix ρ. Differentiating gives

$$\frac{\partial}{\partial t_j} \ell(y, y', \rho, t, t') = -\frac{\rho_j H_j}{\sum_m \rho_m H_m} \frac{y - t_j}{\sigma^2}.$$

Because $0 \leq \rho_j H_j / \sum_m \rho_m H_m \leq 1$, we have

$$\left| \frac{\partial}{\partial t_j} \ell(y, y', \rho, t, t') \right| \leq \frac{|y| + |t_j|}{\sigma^2} \leq \frac{T + A}{\sigma^2}.$$

The same bound holds for the derivative with respect to t'_j. The final Lipschitz bound follows by integrating these coordinatewise derivative bounds along the line segment from (t, t') to (s, s'). $\square$

Lemma 9 (ℓ_1-covering number of the simplex)**.** *Let*

$$\Delta^{M-1} = \left\{ \pi \in \mathbb{R}^M : \pi_m \geq 0,\ \sum_{m=1}^{M} \pi_m = 1 \right\}.$$

For every $\varepsilon > 0$, there exists an ε-net $\Pi_\varepsilon \subseteq \Delta^{M-1}$ with respect to $\|\cdot\|_1$ such that

$$|\Pi_\varepsilon| \leq \left(\frac{2M}{\varepsilon} + 1 \right)^{M-1}.$$

Consequently, for every $n \geq 1$, there exists a $1/n$-net $\Pi_n \subseteq \Delta^{M-1}$ satisfying

$$|\Pi_n| \leq (2Mn + 1)^{M-1}.$$

The same bound holds for $\Delta_\nu^{M-1} \subseteq \Delta^{M-1}$.

Proof. Let $\varepsilon > 0$. Define Π_ε to be the set of all $\rho \in \Delta^{M-1}$ such that

$$\rho_m \in \left\{ \frac{k\varepsilon}{2M} : k \in \mathbb{Z}_{\geq 0} \right\}, \qquad m = 1, \ldots, M-1.$$

Since $\rho \in \Delta^{M-1}$, the final coordinate is determined by

$$\rho_M = 1 - \sum_{m=1}^{M-1} \rho_m.$$

For each of the first $M-1$ coordinates, there are at most

$$\frac{2M}{\varepsilon} + 1$$

possible grid values in the interval $[0, 1]$. Therefore,

$$|\Pi_\varepsilon| \leq \left(\frac{2M}{\varepsilon} + 1\right)^{M-1}.$$

Now fix $\pi \in \Delta^{M-1}$. For each $m = 1, \ldots, M-1$, choose ρ_m so that

$$\rho_m \in \left\{\frac{k\varepsilon}{2M} : k \in \mathbb{Z}_{\geq 0}\right\}$$

and

$$|\pi_m - \rho_m| \leq \frac{\varepsilon}{2M}.$$

Define

$$\rho_M = 1 - \sum_{m=1}^{M-1} \rho_m.$$

Then $\rho \in \Delta^{M-1}$, and

$$\sum_{m=1}^{M-1} |\pi_m - \rho_m| \leq (M-1)\frac{\varepsilon}{2M}.$$

Moreover,

$$\begin{aligned} |\pi_M - \rho_M| &= \left|\sum_{m=1}^{M-1} (\rho_m - \pi_m)\right| \\ &\leq \sum_{m=1}^{M-1} |\rho_m - \pi_m| \\ &\leq (M-1)\frac{\varepsilon}{2M}. \end{aligned}$$

Hence,

$$\begin{aligned}\|\pi-\rho\|_1 &= \sum_{m=1}^{M-1}|\pi_m-\rho_m| + |\pi_M-\rho_M| \\ &\le 2(M-1)\frac{\varepsilon}{2M} \\ &\le \varepsilon.\end{aligned}$$

Thus Π_ε is an ε-net of Δ^{M-1}.

Taking $\varepsilon = 1/n$ gives

$$|\Pi_n| \le (2Mn+1)^{M-1}.$$

Finally, since

$$\Delta_\nu^{M-1} \subseteq \Delta^{M-1},$$

the same net also covers Δ_ν^{M-1}. □

Lemma 10 (Maurer's vector contraction inequality [Mau16])**.** *Let $\mathcal{Z}$ be an arbitrary set, and let $w_1,\dots,w_n \in \mathcal{W}$ be fixed points. Let $\mathcal{G}$ be a class of functions on $\mathcal{W}$ taking values in $\mathbb{R}^d$. For each $i=1,\dots,n$, let $\psi_i : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz. Let $\epsilon_1,\dots,\epsilon_n$ be independent Rademacher random variables. Then*

$$\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \psi_i(g(w_i)) \le \sqrt{2}\,L\,\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^d \epsilon_{ij} g_j(w_i).$$

Theorem 9 (Uniform deviation bound)**.** *Assume that $\mathcal{F}$ satisfies assumption* **(BR)***, and that*

$$\max_{1\le m\le M}\sup_{x\in\mathcal{X}} |f_m^*(x)| \le C_* < \infty.$$

For $n \ge 2$ and $\delta \in (0,1)$, define

$$u_{n,\delta} = \sqrt{2\log\frac{8n}{\delta}},$$

$$L_{n,\delta} = \frac{A + C_* + \sigma u_{n,\delta}}{\sigma^2},$$

$$C_{n,\delta} = -\log\nu + \max\{0, \log(2\pi\sigma^2)\} + \frac{(A+C_*+\sigma u_{n,\delta})^2}{\sigma^2}.$$

Also define

$$C_0 = -\log\nu + |\log(2\pi\sigma^2)| + \frac{2A^2}{\sigma^2} + \frac{4C_*^2}{\sigma^2}.$$

Then, with probability at least $1-\delta$*,*

$$\sup_{\pi\in\Delta_\nu^{M-1},\, f\in\mathcal{F}^M} |J(\pi,f)-\widehat{J}(\pi,f)| \le r_n(\delta),$$

where

$$r_n(\delta) = \frac{8M^{3/2}L_{n,\delta}B}{\sqrt{n}} + C_{n,\delta}\sqrt{\frac{(M-1)\log(2Mn+1)+\log(4/\delta)}{2n}} + \frac{2}{\nu n} + \frac{\delta}{n}\left(\frac{C_0}{2}+1+\sqrt{2\log\frac{8n}{\delta}}\right).$$

Consequently, if

$$\delta_\epsilon(n) = \Pr\left(\sup_{\pi\in\Delta_\nu^{M-1},\, f\in\mathcal{F}^M} |J(\pi,f)-\widehat{J}(\pi,f)| > \epsilon\right),$$

then for every fixed $\epsilon>0$*,*

$$\sum_{n=1}^{\infty}\delta_\epsilon(n) < \infty.$$

Proof. Write

$$Ph = \mathbb{E}[h(X,Y,X',Y')], \qquad P_nh = \frac{1}{n}\sum_{i=1}^{n} h(X_i,Y_i,X_i',Y_i').$$

Also define

$$\ell_{\pi,f}(X,Y,X',Y') = \ell(Y,Y',\pi,f(X),f(X')).$$

Then

$$J(\pi,f) = P\ell_{\pi,f}, \qquad \widehat{J}(\pi,f) = P_n\ell_{\pi,f}.$$

Step 1: Gaussian tail decomposition. Fix $u\ge 1$, set $T_u = C_* + \sigma u$, and define

$$\Omega_u = \{|Y|\le T_u,\ |Y'|\le T_u\}.$$

and

$$\Omega_{i,u} = \{|Y_i|\le T_u,\ |Y_i'|\le T_u\}.$$

Let $S\in\{1,\dots,M\}$ denote the latent mixture component, with

$$\Pr(S=m) = \pi_m^*.$$

Conditional on S, X, X',

$$Y = f_S^*(X) + \sigma W, \qquad Y' = f_S^*(X') + \sigma W',$$

where $W, W' \sim N(0,1)$.

By Lemma 7, $\Pr(\Omega_u^c) \le 4e^{-u^2/2}$, and therefore

$$\Pr\left(\cup_i \Omega_{i,u}^c\right) \le 4ne^{-u^2/2}.$$

Taking

$$u = u_{n,\delta} = \sqrt{2\log\frac{8n}{\delta}},$$

gives

$$\Pr\left(\cup_i \Omega_{i,u}^c\right) \le \frac{\delta}{2}.$$

Now decompose

$$\begin{aligned} J(\pi, f) - \widehat{J}(\pi, f) &= \mathbb{E}[\ell_{\pi,f}(X, X', Y, Y')(\mathbf{1}_{\Omega_u} + \mathbf{1}_{\Omega_u^c})] \\ &\quad - \frac{1}{n}\sum_{i=1}^n \ell_{\pi,f}(X_i, X_i', Y_i, Y_i')(\mathbf{1}_{\Omega_{i,u}} + \mathbf{1}_{\Omega_{i,u}^c}) \\ &= (P - P_n)(\ell_{\pi,f}\mathbf{1}_{\Omega_u}) + (P - P_n)(\ell_{\pi,f}\mathbf{1}_{\Omega_u^c}) \end{aligned}$$

On the event that $\cap_i \Omega_{i,u}$,

$$P_n(\ell_{\pi,f}\mathbf{1}_{\Omega_u^c}) = \frac{1}{n}\sum_{i=1}^n \ell_{\pi,f}(X_i, X_i', Y_i, Y_i')\mathbf{1}_{\Omega_{i,u}^c} = 0$$

for every π and f. We address the remaining terms in steps 2 and 3 below.

Step 2: Tail contribution.

Since $\pi_m \ge \nu$,

$$\ell(y, y', \pi, f(x), f(x')) \le -\log\nu + \log(2\pi\sigma^2) + \frac{(y - f_1(x))^2 + (y' - f_1(x'))^2}{2\sigma^2}.$$

Using

$$(a-b)^2 \le 2a^2 + 2b^2,$$

together with $|f_1(x)| \le A$ and $|f_1(x')| \le A$, we obtain

$$(y - f_1(x))^2 \le 2y^2 + 2A^2, \qquad (y' - f_1(x'))^2 \le 2(y')^2 + 2A^2.$$

Therefore

$$|\ell(y, y', \pi, f(x), f(x'))| \leq -\log \nu + |\log(2\pi\sigma^2)| + \frac{2A^2}{\sigma^2} + \frac{y^2 + (y')^2}{\sigma^2}.$$

Recall that

$$Y = f_S^*(X) + \sigma W, \qquad Y' = f_S^*(X') + \sigma W',$$

where $W, W' \sim N(0, 1)$, and

$$|f_S^*(X)|, |f_S^*(X')| \leq C_*.$$

Hence

$$Y^2 \leq 2C_*^2 + 2\sigma^2 W^2, \qquad (Y')^2 \leq 2C_*^2 + 2\sigma^2 (W')^2.$$

Substituting into the previous inequality yields

$$|\ell_{\pi,f}| \leq C_0 + 2W^2 + 2(W')^2,$$

where

$$C_0 = -\log \nu + |\log(2\pi\sigma^2)| + \frac{2A^2}{\sigma^2} + \frac{4C_*^2}{\sigma^2}.$$

Now,

$$\Omega_u^c = \{|Y| > C_* + \sigma u\} \cup \{|Y'| > C_* + \sigma u\}.$$

Since

$$|Y| = |f_S^*(X) + \sigma W| \leq C_* + \sigma |W|,$$

the event $|Y| > C_* + \sigma u$ implies $|W| > u$. Similarly, $|Y'| > C_* + \sigma u \implies |W'| > u$. Therefore,

$$\Omega_u^c \subseteq \{|W| > u\} \cup \{|W'| > u\}.$$

Consequently,

$$\begin{aligned}
\sup_{\pi,f} \left| P(\ell_{\pi,f} \mathbf{1}_{\Omega_u^c}) \right| &\leq \mathbb{E}\left[\left(C_0 + 2W^2 + 2(W')^2 \right) \mathbf{1}_{\{|W|>u\} \cup \{|W'|>u\}} \right] \\
&\leq \mathbb{E}\left[\left(C_0 + 2W^2 + 2(W')^2 \right) \left(\mathbf{1}_{\{|W|>u\}} + \mathbf{1}_{\{|W'|>u\}} \right) \right].
\end{aligned}$$

Expanding the expectation and using independence gives

$$\sup_{\pi,f} \left| P(\ell_{\pi,f} \mathbf{1}_{\Omega_u^c}) \right| \leq 2C_0 \Pr(|W| > u) + 4\,\mathbb{E}\left[W^2 \mathbf{1}_{\{|W|>u\}} \right] + 4\,\mathbb{E}[W^2] \Pr(|W| > u).$$

Since $\mathbb{E}[W^2] = 1$,

$$\sup_{\pi,f} \left| P(\ell_{\pi,f} \mathbf{1}_{\Omega_u^c}) \right| \leq (2C_0 + 4) \Pr(|W| > u) + 4\, \mathbb{E}\left[W^2 \mathbf{1}_{\{|W|>u\}} \right].$$

Applying the Gaussian tail lemma,

$$\Pr(|W| > u) \leq 2e^{-u^2/2}.$$

Since $n \geq 2$ and $\delta \in (0,1)$, we have

$$u_{n,\delta} = \sqrt{2 \log \frac{8n}{\delta}} \geq \sqrt{2 \log(8n)} \geq \sqrt{2 \log 16} > 1.$$

Thus the second Gaussian tail inequality applies with $u = u_{n,\delta}$, giving

$$\mathbb{E}\left[W^2 \mathbf{1}_{\{|W|>u\}} \right] \leq 2ue^{-u^2/2}.$$

In sum, taking $u = u_{n,\delta}$

$$\begin{aligned}
\sup_{\pi,f} \left| P(\ell_{\pi,f} \mathbf{1}_{\Omega_u^c}) \right| &\leq (2C_0 + 4) \cdot 2e^{-u_{n,\delta}^2/2} + 4 \cdot 2u_{n,\delta} e^{-u_{n,\delta}^2/2} \\
&= (4C_0 + 8 + 8u_{n,\delta}) e^{-u_{n,\delta}^2/2} \\
&= \left(4C_0 + 8 + 8\sqrt{2 \log \frac{8n}{\delta}} \right) \frac{\delta}{8n} \\
&= \frac{\delta}{n} \left(\frac{C_0}{2} + 1 + \sqrt{2 \log \frac{8n}{\delta}} \right).
\end{aligned}$$

Step 3: Truncated empirical process.

Define

$$g_{\pi,f,u}(x, y, x', y') = \ell(y, y', \pi, f(x), f(x')) \mathbf{1}_{\{|y| \leq T_u,\ |y'| \leq T_u\}}.$$

Step 3a: Simplex covering number. Let Π_n be a $1/n$-net of Δ_ν^{M-1} in ℓ_1, with cardinality bounded by Lemma 9 as

$$|\Pi_n| \leq (2Mn + 1)^{M-1}.$$

For every $\pi \in \Delta_\nu^{M-1}$, choose $\rho(\pi) \in \Pi_n$ such that

$$\|\pi - \rho(\pi)\|_1 \leq \frac{1}{n}.$$

Using the Lipschitz bound in π of Lemma 8,

$$|\ell(y, y', \pi, t, t') - \ell(y, y', \rho, t, t')| \le \frac{1}{\nu}\|\pi - \rho\|_1,$$

we obtain

$$|g_{\pi,f,u}(z) - g_{\rho(\pi),f,u}(z)| \le \frac{1}{\nu n}.$$

Therefore

$$\sup_{\pi,f} |Pg_{\pi,f,u} - P_n g_{\pi,f,u}| \le \sup_{\rho\in\Pi_n,f} |Pg_{\rho,f,u} - P_n g_{\rho,f,u}| + \frac{2}{\nu n}.$$

Step 3b: Symmetrization and vector contraction. Fix $\rho \in \Pi_n$. Define

$$V_\rho^+ = \sup_{f\in\mathcal{F}^M} \{Pg_{\rho,f,u} - P_n g_{\rho,f,u}\},$$

and

$$V_\rho^- = \sup_{f\in\mathcal{F}^M} \{P_n g_{\rho,f,u} - Pg_{\rho,f,u}\}.$$

Introduce the notation $S = (Z_1, \ldots, Z_n)$ where $Z_i = (X_i, X_i', Y_i, Y_i')$. By one-sided symmetrization (the ghost-sample argument),

$$\mathbb{E}_S V_\rho^\pm \le 2\mathbb{E}_S\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}^M} \frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\rho,f,u}(Z_i).$$

For fixed data $Z_i = (X_i, Y_i, X_i', Y_i')$, by Lemma 8, the map

$$(t, t') \mapsto \ell(Y_i, Y_i', \rho, t, t')\mathbf{1}_{\{|Y_i|\le T_u,\ |Y_i'|\le T_u\}}$$

is L_u-Lipschitz with respect to

$$\|t - s\|_1 + \|t' - s'\|_1,$$

where

$$L_u = \frac{A + C_* + \sigma u}{\sigma^2}.$$

We now apply Lemma 10. Let

$$w_i = (X_i, X_i') \in \mathcal{X} \times \mathcal{X}.$$

Set

$$\mathcal{W} = \mathcal{X} \times \mathcal{X}, \qquad d = 2M.$$

Define

$$\mathcal{G} = \left\{ g_f : \mathcal{W} \to \mathbb{R}^{2M} : f = (f_1, \ldots, f_M) \in \mathcal{F}^M \right\},$$

where

$$g_f(x, x') = \big(f_1(x), \ldots, f_M(x), f_1(x'), \ldots, f_M(x')\big).$$

For each $i = 1, \ldots, n$, define

$$\psi_i(t_1, \ldots, t_M, t'_1, \ldots, t'_M) = \ell(Y_i, Y'_i, \rho, t, t')\mathbf{1}_{\{|Y_i| \le T_u,\ |Y'_i| \le T_u\}},$$

where $t = (t_1, \ldots, t_M)$ and $t' = (t'_1, \ldots, t'_M)$. Then

$$g_{\rho,f,u}(Z_i) = \psi_i(g_f(w_i)).$$

The truncated-loss Lipschitz bound gives

$$|\psi_i(t, t') - \psi_i(s, s')| \le L_u\big(\|t - s\|_1 + \|t' - s'\|_1\big).$$

Since

$$\|t - s\|_1 + \|t' - s'\|_1 = \|(t, t') - (s, s')\|_1 \le \sqrt{2M}\, \|(t, t') - (s, s')\|_2,$$

each ψ_i is $\sqrt{2M}L_u$-Lipschitz with respect to the Euclidean norm. Therefore Maurer's vector contraction inequality yields

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}^M} \frac{1}{n} \sum_{i=1}^n \epsilon_i g_{\rho,f,u}(Z_i) &= \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \psi_i(g(w_i)) \\
&\le 2\sqrt{M}\, L_u\, \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{2M} \epsilon_{ij} g_j(w_i).
\end{aligned}$$

By the definition of $\mathcal{G}$,

$$\begin{aligned}
\sup_{g \in \mathcal{G}} \sum_{i=1}^n \sum_{j=1}^{2M} \epsilon_{ij} g_j(w_i) &= \sup_{f_1, \ldots, f_M \in \mathcal{F}} \sum_{m=1}^M \left[\sum_{i=1}^n \epsilon_{i,m} f_m(X_i) + \sum_{i=1}^n \epsilon_{i,M+m} f_m(X'_i) \right] \\
&= \sum_{m=1}^M \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^n \epsilon_{i,m} f(X_i) + \sum_{i=1}^n \epsilon_{i,M+m} f(X'_i) \right].
\end{aligned}$$

Using the elementary inequality $\sup_f(A_f + B_f) \le \sup_f A_f + \sup_f B_f$,

$$\sup_{f\in\mathcal{F}}\left[\sum_{i=1}^n \epsilon_{i,m} f(X_i) + \sum_{i=1}^n \epsilon_{i,M+m} f(X_i')\right] \le \sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_{i,m} f(X_i) + \sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_{i,M+m} f(X_i').$$

Therefore

$$\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{2M} \epsilon_{ij} g_j(w_i) \le \sum_{m=1}^M \left\{ \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_{i,m} f(X_i) + \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_{i,M+m} f(X_i') \right\}.$$

Using the Rademacher complexity assumption,

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_{i,m} f(X_i) \le \frac{B}{\sqrt{n}}, \qquad \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_{i,M+m} f(X_i') \le \frac{B}{\sqrt{n}}.$$

Hence

$$\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^{2M} \epsilon_{ij} g_j(w_i) \le \frac{2MB}{\sqrt{n}}.$$

Combining the preceding displays yields

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}^M} \frac{1}{n}\sum_{i=1}^n \epsilon_i g_{\rho,f,u}(Z_i) \le \frac{4M^{3/2}L_u B}{\sqrt{n}}.$$

Consequently,

$$\mathbb{E}V_\rho^{\pm} \le \frac{8M^{3/2}L_u B}{\sqrt{n}}.$$

Step 3c: Bounded difference inequality. By Lemma 8,

$$\min\{0, \log(2\pi\sigma^2)\} \le \ell(y, y', \pi, f(x), f(x')) \le -\log\nu + \log(2\pi\sigma^2) + \frac{(A + C_* + \sigma u)^2}{\sigma^2}.$$

Hence, every $g_{\pi,f,u}$ takes values in an interval of length

$$C_u = -\log\nu + \max\{0, \log(2\pi\sigma^2)\} + \frac{(A + C_* + \sigma u)^2}{\sigma^2}.$$

Changing one observation changes $V_\rho^{\pm}$ by at most C_u/n. To see this, recall $S = (Z_1, \ldots, Z_n)$ and consider

$$V_\rho^+ = V_\rho^+(S) = \sup_{f\in\mathcal{F}^M} \{P g_{\rho,f,u} - P_n g_{\rho,f,u}\}.$$

Let $\tilde{S}_i$ be S but with the i-term changed to $\tilde{Z}_i = (\tilde{X}_i, \tilde{X}_i', \tilde{Y}_i, \tilde{Y}_i')$. Also let $\tilde{P}_n^i g_{\rho,f,u}$ be the empirical

average based on $\tilde{S}_i$. Then

$$\begin{aligned} V_\rho^+(S) - V_\rho^+(\tilde{S}_i) &= \sup_{f \in \mathcal{F}^M} \{P g_{\rho,f,u} - P_n g_{\rho,f,u}\} - \sup_{f \in \mathcal{F}^M} \{P g_{\rho,f,u} - \tilde{P}_n^i g_{\rho,f,u}\} \\ &\le \sup_{f \in \mathcal{F}^M} \{P_n g_{\rho,f,u} - \tilde{P}_n^i g_{\rho,f,u}\} \\ &= \sup_{f \in \mathcal{F}^M} \frac{1}{n}(g_{\rho,f,u}(X_i, Y_i, X_i', Y_i') - g_{\rho,f,u}(\tilde{X}_i, \tilde{X}_i', \tilde{Y}_i, \tilde{Y}_i') \le \frac{C_u}{n}. \end{aligned}$$

The argument for V_ρ^- is similar. Therefore, McDiarmid's inequality [McD89] gives

$$\Pr\left(V_\rho^\pm - \mathbb{E}V_\rho^\pm \ge t\right) \le \exp\left(-\frac{2nt^2}{C_u^2}\right).$$

Applying a union bound over both signs and over $\rho \in \Pi_n$, we obtain with probability at least $1 - \delta/2$,

$$\sup_{\rho \in \Pi_n, f} |P g_{\rho,f,u} - P_n g_{\rho,f,u}| \le \frac{8M^{3/2} L_u B}{\sqrt{n}} + C_u \sqrt{\frac{(M-1)\log(2Mn+1) + \log(4/\delta)}{2n}}.$$

Combining this with the net approximation yields

$$\sup_{\pi,f} |P g_{\pi,f,u} - P_n g_{\pi,f,u}| \le \frac{8M^{3/2} L_u B}{\sqrt{n}} + C_u \sqrt{\frac{(M-1)\log(2Mn+1) + \log(4/\delta)}{2n}} + \frac{2}{\nu n}.$$

Now take $u = u_{n,\delta}$. With probability at least $1 - \delta/2$, all samples lie in their good events. With probability at least $1 - \delta/2$, the empirical-process bound above holds.

Therefore, by the union bound, both events hold with probability at least $1 - \delta$. On this event,

$$\begin{aligned} \sup_{\pi,f} |J(\pi, f) - \widehat{J}(\pi, f)| &\le \sup_{\pi,f} |P g_{\pi,f,u} - P_n g_{\pi,f,u}| + \sup_{\pi,f} \left|P(\ell_{\pi,f} \mathbf{1}_{\Omega_u^c})\right| \\ &\le \frac{8M^{3/2} L_{n,\delta} B}{\sqrt{n}} + C_{n,\delta} \sqrt{\frac{(M-1)\log(2Mn+1) + \log(4/\delta)}{2n}} \\ &\quad + \frac{2}{\nu n} + \frac{\delta}{n}\left(\frac{C_0}{2} + 1 + \sqrt{2\log\frac{8n}{\delta}}\right) \\ &= r_n(\delta). \end{aligned}$$

This proves the stated high-probability bound.

It remains to prove summability. Fix $\epsilon > 0$, and set

$$\delta_n = n^{-2}.$$

Then

$$r_n(\delta_n) = O\left(\frac{\log^{3/2}(n)}{\sqrt{n}}\right).$$

Thus, for all sufficiently large n, $r_n(\delta_n) < \epsilon$. For those n, $\delta_\epsilon(n) \leq \delta_n = n^{-2}$, and therefore

$$\sum_{n=1}^{\infty} \delta_\epsilon(n) < \infty.$$

□

3.5.3 Proof of Theorem 7

Lemma 11. *Assume $\mathcal{X}$ is compact and Q_X is a Borel probability measure on $\mathcal{X}$. Let k be a universal kernel on $\mathcal{X}$, with RKHS $\mathcal{H}$. Suppose*

$$f_m^* \in L^2(Q_X), \qquad m = 1, \ldots, M.$$

Fix $\pi^* \in \Delta^{M-1}$ and $\sigma^2 > 0$. Then for every $\varepsilon_1 > 0$, there exists

$$g = (g_1, \ldots, g_M) \in \mathcal{H}^M$$

such that

$$J(\pi^*, g) \leq J(\pi^*, f^*) + \varepsilon_1.$$

Proof. We first recall the log-sum inequality. If

$$a_1, \ldots, a_M \geq 0$$

and

$$b_1, \ldots, b_M > 0,$$

then

$$\left(\sum_{m=1}^{M} a_m\right) \log \frac{\sum_{m=1}^{M} a_m}{\sum_{m=1}^{M} b_m} \leq \sum_{m=1}^{M} a_m \log \frac{a_m}{b_m}.$$

For fixed $x, x' \in \mathcal{X}$, define

$$p_{f^*}(y, y' \mid x, x') = \sum_{m=1}^{M} \pi_m^* \phi(y; f_m^*(x), \sigma^2)\phi(y'; f_m^*(x'), \sigma^2),$$

and similarly define

$$p_g(y, y' \mid x, x') = \sum_{m=1}^{M} \pi_m^* \phi(y; g_m(x), \sigma^2)\phi(y'; g_m(x'), \sigma^2).$$

For each m, write

$$p_{m,f^*}(y, y' \mid x, x') = \phi(y; f_m^*(x), \sigma^2)\phi(y'; f_m^*(x'), \sigma^2),$$

and

$$p_{m,g}(y, y' \mid x, x') = \phi(y; g_m(x), \sigma^2)\phi(y'; g_m(x'), \sigma^2).$$

Apply the log-sum inequality with

$$a_m = \pi_m^* p_{m,f^*}(y, y' \mid x, x'), \qquad b_m = \pi_m^* p_{m,g}(y, y' \mid x, x').$$

This gives

$$p_{f^*}(y, y' \mid x, x') \log \frac{p_{f^*}(y, y' \mid x, x')}{p_g(y, y' \mid x, x')} \le \sum_{m=1}^{M} \pi_m^* p_{m,f^*}(y, y' \mid x, x') \log \frac{p_{m,f^*}(y, y' \mid x, x')}{p_{m,g}(y, y' \mid x, x')}.$$

Integrating both sides with respect to y, y' gives

$$\mathrm{KL}(p_{f^*}(\cdot, \cdot \mid x, x') \,\|\, p_g(\cdot, \cdot \mid x, x')) \le \sum_{m=1}^{M} \pi_m^* \mathrm{KL}(p_{m,f^*}(\cdot, \cdot \mid x, x') \,\|\, p_{m,g}(\cdot, \cdot \mid x, x')) \,.$$

Since p_{m,f^*} and $p_{m,g}$ are products of two Gaussian densities with common variance σ^2,

$$\mathrm{KL}(p_{m,f^*}(\cdot, \cdot \mid x, x') \,\|\, p_{m,g}(\cdot, \cdot \mid x, x')) = \frac{(f_m^*(x) - g_m(x))^2 + (f_m^*(x') - g_m(x'))^2}{2\sigma^2}.$$

Therefore,

$$\mathrm{KL}(p_{f^*}(\cdot, \cdot \mid x, x') \,\|\, p_g(\cdot, \cdot \mid x, x')) \le \frac{1}{2\sigma^2} \sum_{m=1}^{M} \pi_m^* \left[(f_m^*(x) - g_m(x))^2 + (f_m^*(x') - g_m(x'))^2\right] .$$

Taking expectation over X, X' gives

$$J(\pi^*, g) - J(\pi^*, f^*) \le \frac{1}{2\sigma^2} \sum_{m=1}^{M} \pi_m^* \mathbb{E}\left[(f_m^*(X) - g_m(X))^2 + (f_m^*(X') - g_m(X'))^2\right] .$$

Since X and X' have the same distribution Q_X,

$$\mathbb{E}(f_m^*(X) - g_m(X))^2 = \mathbb{E}(f_m^*(X') - g_m(X'))^2 = \|f_m^* - g_m\|_{L^2(Q_X)}^2.$$

Hence

$$J(\pi^*, g) - J(\pi^*, f^*) \leq \frac{1}{\sigma^2} \sum_{m=1}^{M} \pi_m^* \|f_m^* - g_m\|_{L^2(Q_X)}^2.$$

It remains to approximate each f_m^* in $L^2(Q_X)$ by an element of $\mathcal{H}$.

Since continuous functions are dense in $L^2(Q_X)$, for every $\epsilon_2 > 0$ and every m, there exists a continuous function h_m such that

$$\|f_m^* - h_m\|_{L^2(Q_X)} < \epsilon_2.$$

Since k is universal, for every m there exists $g_m \in \mathcal{H}$ such that

$$\sup_{x \in \mathcal{X}} |h_m(x) - g_m(x)| < \epsilon_2.$$

Because Q_X is a probability measure,

$$\|h_m - g_m\|_{L^2(Q_X)}^2 = \int_{\mathcal{X}} |h_m(x) - g_m(x)|^2 \, dQ_X(x) \leq \int_{\mathcal{X}} \epsilon_2^2 \, dQ_X(x) = \epsilon_2^2.$$

Thus

$$\|h_m - g_m\|_{L^2(Q_X)} \leq \epsilon_2.$$

Therefore, by the triangle inequality,

$$\|f_m^* - g_m\|_{L^2(Q_X)} \leq \|f_m^* - h_m\|_{L^2(Q_X)} + \|h_m - g_m\|_{L^2(Q_X)} < 2\epsilon_2.$$

Substituting into the risk bound,

$$J(\pi^*, g) - J(\pi^*, f^*) \leq \frac{1}{\sigma^2} \sum_{m=1}^{M} \pi_m^* (2\epsilon_2)^2 = \frac{4\epsilon_2^2}{\sigma^2} \sum_{m=1}^{M} \pi_m^* = \frac{4\epsilon_2^2}{\sigma^2}.$$

Choose $\epsilon_2 > 0$ small enough that

$$\frac{4\epsilon_2^2}{\sigma^2} < \epsilon_1.$$

Then

$$J(\pi^*, g) - J(\pi^*, f^*) < \epsilon_1,$$

and hence

$$J(\pi^*, g) \leq J(\pi^*, f^*) + \epsilon_1.$$

□

Theorem 10 (Consistency of the kernel method)**.** *Assume that $\mathcal{X}$ is compact, Q_X is a Borel probability measure on $\mathcal{X}$, and k is a bounded universal kernel on $\mathcal{X}$, with RKHS $\mathcal{H}$. Assume*

$$f_m^* \in L^2(Q_X), \qquad m = 1, \ldots, M,$$

and assume also that there exists $C_* < \infty$ such that

$$\max_{1 \leq m \leq M} \sup_{x \in \mathcal{X}} |f_m^*(x)| \leq C_*.$$

Assume

$$\pi_m^* > 0, \qquad m = 1, \ldots, M.$$

Let $\nu_n \in (0, 1/M)$ and $\lambda_n > 0$ satisfy

$$\nu_n \to 0, \qquad \lambda_n \to 0,$$

and

$$\frac{1}{\nu_n n} \to 0, \qquad \frac{1}{\lambda_n} \sqrt{\frac{\log n}{n}} \to 0.$$

For each n, let

$$(\widehat{\pi}_n, \widehat{f}_n) \in \operatorname*{arg\,min}_{(\pi, f) \in \Delta_{\nu_n}^{M-1} \times \mathcal{H}^M} \left\{ \widehat{J}(\pi, f) + \lambda_n \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}}^2 \right\}.$$

Then

$$J(\widehat{\pi}_n, \widehat{f}_n) \to J(\pi^*, f^*) \qquad \textit{almost surely.}$$

Proof. Choose a deterministic sequence $a_n \to \infty$ so slowly that

$$\frac{a_n^2}{\lambda_n} \sqrt{\frac{\log n}{n}} \to 0.$$

This is possible because

$$\frac{1}{\lambda_n} \sqrt{\frac{\log n}{n}} \to 0.$$

Define

$$R_n = \frac{a_n}{\sqrt{\lambda_n}}.$$

Fix $\varepsilon > 0$. By the RKHS approximation lemma, choose

$$g = (g_1, \dots, g_M) \in \mathcal{H}^M$$

such that

$$J(\pi^*, g) \le J(\pi^*, f^*) + \varepsilon.$$

Let

$$C_g = \sum_{m=1}^{M} \|g_m\|_{\mathcal{H}}^2.$$

Since $\nu_n \to 0$ and $\pi_m^* > 0$ for all m, there exists N_π such that

$$\pi^* \in \Delta_{\nu_n}^{M-1} \qquad \text{for all } n \ge N_\pi.$$

Let Ω_1 be the event on which

$$\widehat{J}(\pi^*, g) \to J(\pi^*, g).$$

By the strong law of large numbers, $\Pr(\Omega_1) = 1$.

Now fix $\omega \in \Omega_1$. Since

$$\widehat{J}(\pi^*, g)(\omega) \to J(\pi^*, g) < \infty$$

and $\lambda_n C_g \to 0$, there exist $D(\omega) < \infty$ and $N_g(\omega) < \infty$ such that, for all $n \ge N_g(\omega)$,

$$\widehat{J}(\pi^*, g)(\omega) - \log(2\pi\sigma^2) + \lambda_n C_g \le D(\omega).$$

For $n \ge \max\{N_\pi, N_g(\omega)\}$, optimality gives

$$\widehat{J}(\widehat{\pi}_n, \widehat{f}_n) + \lambda_n \sum_{m=1}^{M} \|\widehat{f}_{n,m}\|_{\mathcal{H}}^2 \le \widehat{J}(\pi^*, g) + \lambda_n C_g.$$

Since

$$\ell(y, y', \pi, t, t') \ge \log(2\pi\sigma^2),$$

we obtain

$$\lambda_n \sum_{m=1}^{M} \|\widehat{f}_{n,m}(\omega)\|_{\mathcal{H}}^2 \le D(\omega).$$

Thus

$$\sum_{m=1}^{M} \|\widehat{f}_{n,m}(\omega)\|_{\mathcal{H}}^2 \le \frac{D(\omega)}{\lambda_n}.$$

Because $a_n \to \infty$, there exists $N_R(\omega) < \infty$ such that

$$D(\omega) \le a_n^2 \qquad \text{for all } n \ge N_R(\omega).$$

Therefore, for all sufficiently large n,

$$\widehat{f}_n(\omega) \in \mathcal{H}^M(R_n),$$

where

$$\mathcal{H}(R_n) = \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \le R_n\}.$$

Now apply the uniform deviation theorem to the deterministic class $\mathcal{H}(R_n)$ with $\nu = \nu_n$ and $\delta_n = n^{-2}$. Let

$$E_n = \left\{ \sup_{\pi \in \Delta_{\nu_n}^{M-1},\, f \in \mathcal{H}^M(R_n)} |J(\pi, f) - \widehat{J}(\pi, f)| \le r_n(\delta_n; \nu_n, R_n) \right\}.$$

The uniform deviation theorem gives

$$\Pr(E_n^c) \le \delta_n = n^{-2}.$$

Therefore

$$\sum_{n=1}^{\infty} \Pr(E_n^c) \le \sum_{n=1}^{\infty} n^{-2} < \infty.$$

By the Borel–Cantelli lemma,

$$\Pr(E_n^c \text{ infinitely often}) = 0.$$

Equivalently, there exists an event Ω_2 with $\Pr(\Omega_2) = 1$ such that, for every $\omega \in \Omega_2$, there exists $N_U(\omega) < \infty$ with

$$E_n(\omega) \text{ holding for all } n \ge N_U(\omega).$$

It remains to check that

$$r_n(\delta_n; \nu_n, R_n) \to 0.$$

Since k is bounded, if

$$K = \sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty,$$

then the boundedness and Rademacher constants for $\mathcal{H}(R_n)$ satisfy

$$A_n = B_n = KR_n.$$

With $\delta_n = n^{-2}$,

$$u_{n,\delta_n} = O(\sqrt{\log n}).$$

The leading RKHS-radius contribution is

$$R_n^2\sqrt{\frac{\log n}{n}} = \frac{a_n^2}{\lambda_n}\sqrt{\frac{\log n}{n}} \to 0.$$

The net-approximation term satisfies

$$\frac{2}{\nu_n n} \to 0.$$

Finally,

$$\frac{1}{\nu_n n} \to 0$$

implies eventually

$$\log\left(\frac{1}{\nu_n}\right) \le \log n,$$

and therefore

$$\log\left(\frac{1}{\nu_n}\right)\sqrt{\frac{\log n}{n}} \le \frac{(\log n)^{3/2}}{\sqrt{n}} \to 0.$$

Hence

$$r_n(\delta_n; \nu_n, R_n) \to 0.$$

Now fix $\omega \in \Omega_1 \cap \Omega_2$. This event has probability one. For all sufficiently large n, we have

$$\widehat{f}_n(\omega) \in \mathcal{H}^M(R_n), \qquad \widehat{\pi}_n(\omega) \in \Delta_{\nu_n}^{M-1},$$

and $E_n(\omega)$ holds. Hence

$$J(\widehat{\pi}_n, \widehat{f}_n) \le \widehat{J}(\widehat{\pi}_n, \widehat{f}_n) + r_n(\delta_n; \nu_n, R_n).$$

By optimality,

$$\widehat{J}(\widehat{\pi}_n, \widehat{f}_n) + \lambda_n \sum_{m=1}^{M} \|\widehat{f}_{n,m}\|_{\mathcal{H}}^2 \le \widehat{J}(\pi^*, g) + \lambda_n C_g.$$

Dropping the nonnegative penalty term gives

$$\widehat{J}(\widehat{\pi}_n, \widehat{f}_n) \le \widehat{J}(\pi^*, g) + \lambda_n C_g.$$

Therefore, along this fixed sample path,

$$J(\widehat{\pi}_n, \widehat{f}_n) \le \widehat{J}(\pi^*, g) + \lambda_n C_g + r_n(\delta_n; \nu_n, R_n).$$

Taking $\limsup$ and using

$$\widehat{J}(\pi^*, g)(\omega) \to J(\pi^*, g), \qquad \lambda_n \to 0, \qquad r_n(\delta_n; \nu_n, R_n) \to 0,$$

gives

$$\limsup_{n\to\infty} J(\widehat{\pi}_n, \widehat{f}_n) \le J(\pi^*, g) \le J(\pi^*, f^*) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary,

$$\limsup_{n\to\infty} J(\widehat{\pi}_n, \widehat{f}_n) \le J(\pi^*, f^*).$$

On the other hand, population optimality gives

$$J(\widehat{\pi}_n, \widehat{f}_n) \ge J(\pi^*, f^*) \qquad \text{for every } n.$$

Therefore,

$$J(\widehat{\pi}_n, \widehat{f}_n) \to J(\pi^*, f^*)$$

on $\Omega_1 \cap \Omega_2$. Since $\Pr(\Omega_1 \cap \Omega_2) = 1$, the convergence holds almost surely. □

3.5.4 Proof of Theorem 5

Corollary 6 (L^1-consistency)**.** *Under the conditions of Theorem 7,*

$$\mathbb{E}_P \left[\| q(\cdot, \cdot \mid x, x') - q_{\hat{\alpha},\hat{\pi}}(\cdot, \cdot \mid x, x') \|_{L^1(\mathbb{R}^2)} \right] \overset{a.s.}{\to} 0.$$

Proof. By Theorem 7, we have $J(\widehat{\pi}, \widehat{\alpha}) = J(\widehat{\pi}, \widehat{f}) \overset{\text{a.s.}}{\to} J(pi^*, f^*)$. Applying Pinsker's inequality followed by Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}_P \left[\| q(\cdot, \cdot \mid x, x') - q_{\hat{\alpha},\hat{\pi}}(\cdot, \cdot \mid x, x') \|_{L^1(\mathbb{R}^2)} \right] &\le \mathbb{E}_P \sqrt{2 D_{\mathrm{KL}}(q(\cdot, \cdot \mid x, x') \,\|\, q_{\hat{\alpha},\hat{\pi}}(\cdot, \cdot \mid x, x'))} \\
&\le \sqrt{2 \mathbb{E}_P D_{\mathrm{KL}}(q(\cdot, \cdot \mid x, x') \,\|\, q_{\hat{\alpha},\hat{\pi}}(\cdot, \cdot \mid x, x'))} \\
&= \sqrt{2 \left(J(\hat{\pi}, \hat{\alpha}) - J(\pi^*, f^*) \right)} \overset{\text{a.s.}}{\to} 0.
\end{aligned}$$

□

3.5.5 Proof of Theorem 8

Theorem 11 (Consistent Recovery of π^* and f^*)**.** *Suppose distribution Q is 2-identifiable. Further assume that Q_X is absolutely continuous with respect to the Lesbesgue measure. Under the conditions of Theorem 5, $\hat{\pi}_m \xrightarrow{a.s.} \pi^*_m$ and $\|f^*_m - f^{\hat{\alpha}}_m\|_{L^1(Q_X)} \xrightarrow{a.s.} 0$, up to a permutation.*

Proof. From Theorem 5, we have

$$\mathbb{E}_P\left[\|q(\cdot,\cdot \mid x,x') - q_{\hat{\alpha},\hat{\pi}}(\cdot,\cdot \mid x,x')\|_{L^1(\mathbb{R}^2)}\right] \xrightarrow{\text{a.s.}} 0.$$

Recall that the marginal distribution Q_X is absolutely continuous w.r.t. the Lebesgue measure, so it has a valid density $p(x)$. The marginal distribution of pairs (X, X') is the product measure $P = Q_X^{\otimes 2}$ and therefore has density $p(x)p(x')$. It is easy to show that the LHS is the L^1 distance of two joint densities on $(\mathcal{X} \times \mathbb{R})^2$

$$\begin{aligned}
&\mathbb{E}_P\left[\|q(\cdot,\cdot \mid x,x') - q_{\hat{\alpha},\hat{\pi}}(\cdot,\cdot \mid x,x')\|_{L^1(\mathbb{R}^2)}\right] \\
&\quad= \int_{\mathcal{X}^2}\int_{\mathbb{R}^2} |q(y,y' \mid x,x') - q_{\hat{\alpha},\hat{\pi}}(y,y' \mid x,x')| d(y,y') dP(x,x') \\
&\quad= \int_{\mathcal{X}^2} p(x)p(x') \int_{\mathbb{R}^2} |q(y,y' \mid x,x') - q_{\hat{\alpha},\hat{\pi}}(y,y' \mid x,x')| d(y,y') d(x,x') \\
&\quad= \int_{\mathcal{X}^2}\int_{\mathbb{R}^2} |p(x)p(x')q(y,y' \mid x,x') - p(x)p(x')q_{\hat{\alpha},\hat{\pi}}(y,y' \mid x,x')| d(y,y') d(x,x') \\
&\quad= \int_{\mathcal{X}^2\times\mathbb{R}^2} |p(x)p(x')q(y,y' \mid x,x') - p(x)p(x')q_{\hat{\alpha},\hat{\pi}}(y,y' \mid x,x')| d(y,y',x,x') \\
&\quad= \int_{(\mathcal{X}\times\mathbb{R})^2} |p(x)p(x')q(y,y' \mid x,x') - p(x)p(x')q_{\hat{\alpha},\hat{\pi}}(y,y' \mid x,x')| d((x,y),(x',y')) \\
&\quad= \|r((x,y),(x',y')) - \hat{r}((x,y),(x',y'))\|_{L^1((\mathcal{X}\times\mathbb{R})^2)},
\end{aligned}$$

where $r((x,y),(x',y')) \coloneqq p(x)p(x')q(y,y' \mid x,x')$ and $\hat{r}((x,y),(x',y')) \coloneqq p(x)p(x')q_{\hat{\alpha},\hat{\pi}}(y,y' \mid x,x')$ are densities on $(\mathcal{X} \times \mathbb{R})^2$. By definition of q and $q_{\alpha,\pi}$, we can write r and $\hat{r}$ as

$$\begin{aligned}
r((x,y),(x',y')) &= \sum_{m=1}^{M} \pi^*_m \left(p(x)\phi(y; f^*_m(x), \sigma^2)\right)\left(p(x')\phi(y'; f^*_m(x'), \sigma^2)\right), \\
\hat{r}((x,y),(x',y')) &= \sum_{m=1}^{M} \hat{\pi}_m \left(p(x)\phi(y; f^{\hat{\alpha}}_m(x), \sigma^2)\right)\left(p(x')\phi(y'; f^{\hat{\alpha}}_m(x'), \sigma^2)\right),
\end{aligned}$$

where $p(x)\phi(y; f^*_m(x), \sigma^2)$ and $p(x)\phi(y; f^*_m(x), \sigma^2)$ are densities in their own right. Let the refer-

ence measure ν be the Lebesgue measure, and define for $m \in [M]$ the lifted measures $\mu_m, \hat{\mu}_m$ on $\mathcal{X} \times \mathbb{R}$ with corresponding Radon-Nikodym derivatives

$$\frac{d\mu_m}{d\nu} = p(x)\phi(y; f_m^*(x), \sigma^2),$$
$$\frac{d\hat{\mu}_m}{d\nu} = p(x)\phi(y; f_m^{\hat{\alpha}}(x), \sigma^2).$$

Then, then r and $\hat{r}$ are well-defined densities of the mixtures of product measures

$$\mu = \sum_{m=1}^{M} \pi_m^* \mu_m^{\otimes 2},$$
$$\hat{\mu} = \sum_{m=1}^{M} \hat{\pi}_m \hat{\mu}_m^{\otimes 2},$$

and the total variation distance between them, in terms of the densities r and $\hat{r}$, is

$$\begin{aligned}\delta(\mu, \hat{\mu}) &= \frac{1}{2}\|r((x,y),(x',y')) - \hat{r}((x,y),(x',y'))\|_{L^1((\mathcal{X}\times\mathbb{R})^2)} \\ &= \frac{1}{2}\mathbb{E}_P\left[\|q(\cdot,\cdot \mid x,x') - q_{\hat{\alpha},\hat{\pi}}(\cdot,\cdot \mid x,x')\|_{L^1(\mathbb{R}^2)}\right] \overset{\text{a.s.}}{\to} 0.\end{aligned}$$

This satisfies the conditions of Theorem 5 of [RVS20], so $\hat{\pi}_m \to \pi_m^*$ and $\delta(\mu_m, \hat{\mu}_m) \to 0$ for all $m \in [M]$ up to a permutation, giving us the first statement. It remains to show that this also implies the second statement. The following argument holds for all $m \in [M]$, so fix m. We quickly see that

$$\begin{aligned}\delta(\mu_m, \hat{\mu}_m) &= \int_{\mathcal{X}\times\mathbb{R}} |p(x)\phi(y; f_m^*(x), \sigma^2) - p(x)\phi(y; f_m^{\hat{\alpha}}(x), \sigma^2)| d(x,y) \\ &= \int_{\mathcal{X}} p(x) \int_{\mathbb{R}} |\phi(y; f_m^*(x), \sigma^2) - \phi(y; f_m^{\hat{\alpha}}(x), \sigma^2)| dy dx \\ &= \mathbb{E}_{Q_X}\left[\|\phi(\cdot; f_m^*(x), \sigma^2) - \phi(\cdot; f_m^{\hat{\alpha}}(x), \sigma^2)\|_{L^1(\mathbb{R})}\right] =: \Psi_n(x),\end{aligned}$$

Clearly $\Psi_n(x) \overset{\text{a.s.}}{\to} 0$. The above says $\Psi_n(x)$ converges to zero in Q_X-expectation almost surely over the data. Applying Markov's inequality gives, for all $\epsilon > 0$,

$$Q_X\left(\{x \in X : \Psi_n(x) > \epsilon\}\right) \leq \frac{\mathbb{E}_{Q_X}[\Psi_n(X)]}{\epsilon} \overset{\text{a.s.}}{\to} 0.$$

So $\Psi_n(x)$ converges to zero in Q_X-probability almost surely over the data, which implies there is a subsequence $\{n_j\}$ such that $\Psi_{n_j}(x) \overset{\text{a.s.}}{\to} 0$ for Q_X-a.e. x. Let $E \subset \mathcal{X}$ with $Q_X(E) = 1$ be the set such that $\Psi_{n_j}(x) \overset{\text{a.s.}}{\to} 0$ along the subsequence. Since, by assumption 1, $|f_m^{\alpha}(x)| \leq C_\alpha C_k$

for all $x \in E$ for all α, by Balzano-Weierstrass there exists a sub-subsequence $\{n_{j_s}\}$ such that $f^{\hat{\alpha}}_{m,n_{j_s}}(x) \to \tau(x)$ for some value $|\tau(x)| \leq C_\alpha C_k$ for all $x \in E$.

We now apply the dominated convergence theorem. Since the map $\mu \mapsto \phi(\cdot; \mu, \sigma^2)$ from $\mathbb{R}$ to $L^1(\mathbb{R})$ is continuous in μ, $\phi(y; f^{\hat{\alpha}}_{m,n_{j_s}}(x), \sigma^2) \to \phi(y; \tau(x), \sigma^2)$ pointwise in y. Any Gaussian density on the bounded set $\{\mu : \ \mu \leq C_\alpha C_k\}$ can be dominated by a suitably scaled Gaussian density of higher variance, for instance:

$$\phi(y; \mu, \sigma^2) \leq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-y^2}{4\sigma^2}\right) \exp\left(\frac{C_\alpha^2 C_k^2}{2\sigma^2}\right).$$

Therefore, $\phi(y; f^{\hat{\alpha}}_{m,n_{j_s}}(x), \sigma^2) \to \phi(y; \tau(x), \sigma^2)$ in $L^1(\mathbb{R})$ for all x in E. Because we already know $\phi(y; f^{\hat{\alpha}}_{m,n_{j_s}}(x), \sigma^2) \to \phi(y; f^*_m(x), \sigma^2)$ in $L^1(\mathbb{R})$ for all $x \in E$, the triangle inequality gives $\phi(y; \tau(x), \sigma^2) = \phi(y; f^*_m(x), \sigma^2)$ in $L^1(\mathbb{R})$ for all $x \in E$. Since the map $\mu \mapsto \phi(\cdot; \mu, \sigma^2)$ is continuous, and therefore an injection from $\mathbb{R}$ to $L^1(\mathbb{R})$, we have $f_{m,n_{j_s}} = f^*_m(x)$ in $L^1(\mathbb{R})$ for all $x \in E$, and since $Q_X(E) = 1$ we conclude that $\|f_{m,n_{j_s}} - f^*_m(x)\|_{L^1(Q_X)} \overset{\text{a.s.}}{\to} 0$.

To conclude that the full sequence converges, we apply the subsequence principle. For any sequence of real numbers $\{a_n\}$, if every subsequence $\{a_{n_k}\}$ has a further subsequence $\{a_{n_{j_s}}\}$ converging to a common limit a, then $a_n \to a$. In this case, when we defined $\Psi_n(x)$, we could have restricted ourselves to any subsequence $\{n'\} \subset \mathbb{N}$ and since $\Psi_{n'}(x) \overset{\text{a.s.}}{\to} 0$, we must conclude $\|f_{m,n'_{j'}} - f^*_m(x)\|_{L^1(Q_X)} \overset{\text{a.s.}}{\to} 0$ for some suitable subsequence $\{n'_{j'}\}$. The result follows. □

3.5.6 Proof of Identifiability Results

Lemma 12. *The collection of joint densities $\{q_m(x,y)\}_{m=1}^M$, where $q_m(x,y) = \phi(y; f^*_m(x), \sigma^2) q_X(x)$, is jointly irreducible if and only if, for each mean function in the collection $\{f^*_m\}_{m=1}^M$, there exists a subset $S_m \subseteq \mathcal{X}$ with $Q_X(S_m) > 0$ such that, for all $j \in [M]$ with $j \neq m$, either $f^*_m(x) > f^*_j(x)$ for all $x \in S_m$, or $f^*_m(x) < f^*_j(x)$ for all $x \in S_m$.*

Proof. We start by proving sufficiency.

Assume that the collection of joint densities is not jointly irreducible. Then there exists $\{\gamma_m\}_{m=1}^M$ and index i with $\gamma_i < 0$ and $\gamma_j \geq 0$ for all $j \neq i$, such that $\sum_{m=1}^M \gamma_m q_m(x,y)$ is a probability density. Note that we can always obtain this condition on the γ_m, since if there were multiple negative coefficients we could set all but one to zero and rescale the coefficients accordingly to obtain a density.

For all $x \in S_i$, supposing the case $f^*_i(x) > f^*_j(x)$,

$$\lim_{y\to\infty} \frac{|\gamma_i| q_i(x,y)}{\sum_{m\neq i} \gamma_m q_m(x,y)} = \lim_{y\to\infty} \frac{|\gamma_i| \phi(y; f^*_i(x), \sigma^2)}{\sum_{m\neq i} \gamma_m \phi(y; f^*_m(x), \sigma^2)} = \infty.$$

This implies there exists a finite y_x, depending on x, such that

$$\sum_{m=1}^{M} \gamma_m q(x, y) < 0, \quad \forall y > y_x.$$

Note this is true for all x in S_i, which has positive Q_X-measure and Q_X is absolutely continuous with respect to the Lebesgue measure. In addition, note that $\int_{y>y_x} \phi(y; f_m^*(x), \sigma^2)\ dy > 0$ for all $m \in [M]$, and any Gaussian measure is absolutely continuous with respect to the Lebesgue measure. It follows that

$$\sum_{m=1}^{M} \gamma_m q(x, y) < 0$$

on a set of positive Lebesgue measure on the space $\mathcal{X} \times \mathbb{R}$ and is not a valid probability density.

The above argument holds for all $i \in [M]$. Therefore, $\sum_m \gamma_m q_m(x, y)$ is not a valid density unless $\gamma_m \geq 0$ for all $m \in [M]$, a contradiction. We conclude the condition in the statement of the Lemma is sufficient for the collection of joint densities to be jointly irreducible.

We now prove necessity.

Suppose for some $m \in [M]$ that $Q_X(S_m) = 0$. Then there exist indices $i_x, j_x \in [M]$, possibly depending on x, such that

$$f_{i_x}^*(x) \leq f_m^*(x) \leq f_{j_x}^*(x)$$

for Q_X-a.e. x. Since the conditional Gaussian densities share a common variance, we can select constants $\gamma_m < 0$ and, for all $i, j \neq m$, $0 < \gamma_i, \gamma_j < 1$, with γ_m sufficiently small in magnitude relative to γ_i, γ_j, such that

$$|\gamma_m| q_m(x, y) \leq \gamma_{i_x} q_{i_x}(x, y) + \gamma_{j_x} q_{j_x}(x, y), \quad \forall y, \text{ for } Q_X\text{-a.e. } x.$$

Therefore,

$$\sum_{k=1}^{M} \gamma_k q_k(x, y) \geq 0, \quad \forall y, \text{ for } Q_X\text{-a.e. } x,$$

and therefore on a set of positive Lebesgue measure on the space $\mathcal{X} \times \mathbb{R}$. Suitably scaling the coefficients γ_k by a positive constant yields a valid probability density in $\mathcal{X} \times \mathbb{R}$. Since the scaled γ_m is still negative, we conclude the assumption given in the statement of the Lemma is necessary for the collection of joint densities to be jointly irreducible. □

To prove the corresponding result for linear independence, we will first need the following two Lemmas.

Lemma 13. *Let $b_1, b_2, \ldots, b_M \in \mathbb{R}$ be distinct constants. The sequence of exponential functions $\{\phi_i(y) = \exp(b_i y)\}_{i=1}^M$ is linearly independent.*

Proof. Let $c_i \in \mathbb{R}$ be such that

$$\sum_{i=1}^{M} c_i \phi_i(y) = \sum_{i=1}^{M} c_i \exp(b_i y) = 0 \qquad \forall y \in \mathbb{R}. \tag{3.12}$$

We can construct a system of equations by noticing that a sum of exponential functions is infinitely differentiable, and the derivatives of all orders k w.r.t. y must satisfy

$$\frac{d^k}{dy^k} \sum_{i=1}^{M} c_i \exp(b_i y) = \sum_{i=1}^{M} c_i b_i^k \exp(b_i y) = 0 \qquad \forall k.$$

Considering derivatives of orders up to $M-1$ and evaluating at $y = 0$ gives the system of equations

$$\sum_{i=1}^{M} c_i b_i^k = 0, \quad k = 0, 1, \ldots, M-1.$$

This is the linear system $V^\top c = 0$, $V \in \mathbb{R}^{M \times M}$ is the Vandermonde matrix $V_{i,k} = b_i^k$. If V is full rank, then this system only has the solution $c = 0$. The determinant of the Vandermonde matrix V is given by

$$\det V = \prod_{1 \le i < j \le M} (b_j - b_i),$$

which is nonzero since the b_i are distinct. Therefore, V is full rank and only has the solution $c_i = 0$ for all $i \in [M]$. Therefore, $\phi_1, \ldots, \phi_M$ are linearly independent. □

Lemma 14. *Fix $\sigma^2 > 0$ and let $\mu_1, \mu_2, \ldots, \mu_M \in \mathbb{R}$ be distinct. The Gaussian densities*

$$\phi_i(y; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu_i)^2}{2\sigma^2}\right), \qquad i = 1, \ldots, M,$$

are linearly independent.

Proof. By expanding the exponent, for all $i \in [M]$, we have

$$\phi_i(y; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \exp\left(\frac{\mu_i y}{\sigma^2}\right) \exp\left(-\frac{\mu_i^2}{2\sigma^2}\right),$$

where $\exp(\frac{\mu_i y}{\sigma^2})$ is the only term that is both non-constant with respect to y and dependent on the index i. Let $c_i \in \mathbb{R}$ be such that

$$\sum_{i=1}^{M} c_i \phi_i(y; \mu_i, \sigma^2) = 0 \qquad \forall y \in \mathbb{R}. \tag{3.13}$$

Dividing by $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) > 0$ and defining constants $a_i = \exp\left(-\frac{\mu_i^2}{2\sigma^2}\right) > 0$ and $b_i = \frac{\mu_i}{\sigma^2}$, the above is equivalent to

$$\sum_{i=1}^{M} (c_i a_i) \exp(b_i y) = 0.$$

Since the μ_i are distinct, so are the b_i. Therefore, $c_i a_i = 0$ for all $i \in [M]$ by lemma 13. Since $a_i > 0$, it follows that $c_i = 0$ for all $i \in [M]$. We conclude the densities $\phi_1, \ldots, \phi_M$ are linearly independent. □

We are now ready to prove the linear independence result given in the main document.

Lemma 15. *The collection of joint densities* $\{q_m(x,y)\}_{m=1}^M$*, where* $q_m(x,y) = \phi(y; f_m^*(x), \sigma^2) q_X(x)$*, is linearly independent if* $Q_X(\{x \mid f_m^*(x) \neq f_j^*(x)\ \forall j \neq m\}) > 0$ *for all* $m \in [M]$.

Proof. Write $A_m := \{x : f_m^*(x) \neq f_j^*(x)\ \forall j \neq m\}$, so the condition becomes $Q_X(A_m) > 0$ for all $m \in [M]$. Let λ^d be the Lebesgue measure on $\mathbb{R}^d$ for any integer $d \geq 1$, where the superscript is omitted in the case $d = 1$. Given $c \in \mathbb{R}^M$, suppose that

$$\sum_{m=1}^{M} c_m \, p_m(x,y) = 0, \quad (Q_X \otimes \lambda) - \text{a.e.}$$

We will show $c = 0$.

By Fubini there is a $\lambda^{\dim(\mathcal{X})}$-null (and hence Q_X-null) set N such that for every $x \notin N$, $q_X(x) \sum_m c_m \phi(y; f_m^*(x), \sigma^2) = 0$ for a.e. y. For $x \in \mathrm{Supp}(Q_X) \setminus N$, we can divide by $q_X(x)$ and define the set

$$E := \Big\{ x : \sum_{m=1}^{M} c_m \, \phi(y; f_m^*(x), \sigma^2) = 0 \ \ \forall y \in \mathbb{R} \Big\},$$

noting that $E \supseteq \mathrm{Supp}(Q_X) \setminus N$. It follows that $Q_X(E) = 1$.

Now fix $x \in E$, let $\mu_1, \ldots, \mu_{K_x}$ be the distinct values among $\{f_m^*(x)\}_{m=1}^M$, and set $I_k = \{m :$

$f_m^*(x) = \mu_k\}$. Then

$$0 = \sum_{m=1}^{M} c_m \, \phi(y; f_m^*(x), \sigma^2) = \sum_{k=1}^{K_x} \Big(\sum_{m \in I_k} c_m \Big) \phi(y; \mu_k, \sigma^2) \qquad \forall y,$$

and the μ_k are distinct, so Lemma 14 forces

$$\sum_{m \in I_k} c_m = 0 \qquad \forall k \in [K_x].$$

Fix $m \in [M]$. Since $Q_X(E^c) = 0$,

$$Q_X(A_m \cap E) \geq Q_X(A_m) - Q_X(E^c) = Q_X(A_m) > 0,$$

so $A_m \cap E \neq \varnothing$. For all $x \in A_m \cap E$, $f_m^*(x) \neq f_j^*(x)$ for all $j \neq m$. Without loss of generality, let $\mu_1 = f_m^*(x)$ and note the index set $I_k = \{m\}$ is a singleton. From the previous argument we get $c_m = 0$. As $m \in [M]$ was arbitrary, we must have $c = 0$. We conclude the collection $\{q_m(x, y)\}_{m=1}^M$ is linearly independent in $L^1(\lambda^{\dim(\mathcal{X})} \otimes \lambda)$.

□

CHAPTER 4

Supervised PCA: A Multiobjective Approach

4.1 Introduction

Supervised principal component analysis (SPCA) is, as its name suggests, the problem of learning a low dimensional data representation in the spirit of PCA, while ensuring that the learned representation is also useful for supervised learning tasks. SPCA has received considerable interest outside of machine learning [Liu20, SWW$^+$20, RM06, CWSZ08, CWH$^+$10, VBT$^+$21], owing to the broad appeal of PCA and a desire to perform supervised dimension reduction. We view SPCA as a fundamental and important problem, and in this work aim to advance the state of the art in SPCA.

Toward that end, our prior work [RSB$^+$19] introduced the following formulation of the SPCA problem:

$$\begin{aligned} &\min_{L\in\mathbb{R}^{p\times r},\beta\in\mathbb{R}^{r\times q}} \sum_{i=1}^{n} g(\boldsymbol{y}_i, \boldsymbol{x}_i, L, \beta) + \lambda\|X - XLL^\top\|_F^2 \\ &s.t.\ L^\top L = I_r, \end{aligned} \tag{4.1}$$

where g is a supervised loss function of interest. This straightforward approach optimizes a criterion that explicitly balances the empirical risk associated to a supervised learning problem with the variance explained by the learned representation.

Compared to other work on SPCA [BHPT06, YYT$^+$06, BGAJ11, LYNS16, PV18, GK24, PY25], problem (4.1) has several advantages. First, many prior works are specific to regression or

This chapter contains joint work with Clayton Scott, Laura Balzano, Daniel Kessler, and Chandra S. Sripada. A preliminary version of this work was published as *Supervised Principal Component Analysis via Manifold Optimization* at IEEE Data Science Workshop 2019. This chapter is an elaboration of that work as well as our pre-print [RBK$^+$20] which shares a title with this chapter, but from which the current work has diverged substantially. This work is currently in preparation for journal submission, jointly with the same authors.

classification, while this approach accommodates arbitrary loss functions. Second, several existing approaches operate in two stages, first learning the representation by one criterion, and subsequently inferring a prediction model by another. These approaches typically use correlation of the learned representation with the response variables as a proxy for the criterion of ultimate interest, e.g., classification accuracy. Third, many existing approaches do not have a means of specifying a trade-off between prediction error (PE) and variation explained (VE), which can lead to poor performance. In problem (4.1), this trade-off is governed by a tuning parameter. In particular, this approach has been found empirically [RSB$^+$19, XVC$^+$21, PY25] to outperform existing SPCA methods in terms of both PE and VE, suggesting VE serves as a form of regularization for the supervised learning problem.

Compared to our prior workshop paper [RSB$^+$19], which introduced problem (4.1) and an alternating optimization algorithm without theoretical guarantees, this paper (i) recasts (4.1) as a joint optimization problem on the product manifold $\mathrm{St}(p,r) \times \mathbb{R}^{r\times q}$, (ii) establishes an $O(1/\epsilon^2)$ convergence rate for Riemannian gradient descent to a stationary point, (iii) provides a sufficient condition for local minima under general smooth supervised losses, (iv) gives a simplified characterization of local minima and a PCA subspace recovery result for the squared-error case, and (v) evaluates the performance of the proposed method on real and simulated data, supporting the claims mentioned above.

4.1.1 Notation

Let $X \in \mathbb{R}^{n\times p}$ be a data matrix which has p-dimensional data observations as rows, denoting the ith row by column vector $\mathbf{x}_i$. Similarly, let $Y \in \mathbb{R}^{n\times q}$ be an associated matrix of q-dimensional responses with the ith row denoted by column vector $\boldsymbol{y}_i$. To simplify notation, X is assumed to have been centered, meaning the columns have zero mean. Y is assumed to have been centered when its entries are realizations of continuous random variables. Denote by $\mathrm{St}(p,r)$ the Stiefel manifold of $p \times r$ real-valued matrices with orthonormal columns. Denote by I_r the $r \times r$ identity matrix. Column vectors are written as bold lowercase letters, with matrices in plain capital letters. For a real-valued matrix A, let $A^\top$ denote the transpose, A^+ the pseudoinverse, $\|A\|_F$ the Frobenius norm, and $\sigma_{min}(A)$ the minimal singular value of A with $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{min}(A)$.

4.2 Background and Related Work

The goal of dimensionality reduction (DR) is to find an r-dimensional representation of the input data, $r < p$. If Y is used to find this representation, the problem is referred to as supervised dimensionality reduction (SDR). In this section we review PCA as it relates to our contribution. We also review prior work on SPCA and other forms of SDR.

4.2.1 PCA

PCA was first formulated by Karl Pearson in 1901 [Pea01] and later reinvented by Harold Hotelling [Hot33]. Geoemtrically, it can be thought of as the problem of finding an affine subspace of best fit to a collection of points in the squared error sense. As an optimization problem, PCA can be written

$$\min_{L \in \mathbb{R}^{p \times r}} \ \|X - XLL^\top\|_F^2 \ \ s.t. \ \ L^\top L = I_r, \tag{4.2}$$

where X is assumed to be centered, I_r is the $r \times r$ identity matrix, and $\|A\|_F$ is the Frobenius norm of a matrix A. Projection of X to the subspace spanned by columns of the optimal L gives the best rank-r approximation of X in terms of squared reconstruction error. Equivalently, this projection has the statistical interpretation of capturing the largest possible variance in the data among all rank-r projections. That is, we maximize VE, which is given by

$$\text{VE} = \frac{\|XL\|_F^2}{\|X\|_F^2} \ \in [0, 1]. \tag{4.3}$$

Note that this formulation of VE makes sense for any L with orthonormal columns.

The process of performing PCA prior to a regression task is referred to as principal component regression (PCR), a nice discussion of which is given by Jolliffe [Jol82]. To the authors' knowledge, no such name exists for the analogous approach for classification. This work will refer to that method as principal component classification (PCC).

4.2.2 Supervised Dimension Reduction Methods

PCA has enjoyed immense popularity in statistical analysis for the past century or so. It remains a useful tool for dimension reduction (DR) due to its effectiveness and ease of computation. However, PCA does not make use of any supervisory information, and therefore DR via PCA may not be useful for subsequent classification or regression tasks. This stems from the fact that in most problems of interest, there is a tradeoff between directions that explain variation in X, and those that are predictive of Y. To overcome this limitation, several approaches to SDR have been proposed. We first describe some fundamental SDR methods and highlight their connections to PCA, and then proceed to review existing approaches to SPCA.

Fisher's linear discriminant, or Fisher discriminant analysis (FDA) is arguably the canonical example of supervised dimension reduction in the classification setting. FDA finds a dimension reduced representation of X such that interclass variation is maximized while intraclass variation is minimized. Though it may seem that FDA is generally preferable to PCA for classification, this has been shown not always to be the case, especially when the number of training samples is

small [MK01]. A number of extensions of FDA have been proposed. For example, local Fisher discriminant analysis (LFDA) [Sug07] modifies FDA by approximately preserving local distances between same-class points.

Partial least squares (PLS) regression finds projections of the input data that account for a high amount of variation, but are also highly correlated with projections of the dependent variables. It is somewhat different from other methods presented here, in that both X and Y are projected to a new space to determine their relationship. Without means of specifying the trade-off between correlation and variation, PLS tends to put preference on directions that account for high variation rather than high correlation, causing it to behave similarly to PCR [FHT01].

Reduced rank regression (RRR) [A$^+$51, Ize75] attempts to minimize regression error under the constraint that the coefficient matrix be low rank. Such models arise in econometrics and other settings where the underlying relationship between predictor and response is believed to be low rank. This model is intimately related to PCA [Ize75, VR13]. Yee and Hastie [YH03] extend RRR to encompass categorical response variables through what they call reduced rank vector generalized linear models (RRVGLMs). Their work primarily explores the case of reduced rank logistic regression.

The earliest of the SPCA approaches [BHPT06], which we call Bair's method, is a simple two stage procedure. First, feature selection based on univariate regression coefficients is performed. Second, PCA is performed on the data matrix consisting only of the selected features. This approach may not be optimal, especially in the case where features are jointly predictive but not individually predictive. Furthermore, the method is only applicable to univariate regression and binary classification. On the other hand, this approach has some rigorous theory including a consistency result under an assumption of perfect variable selection with high probability. Recently, the method of iterative supervised principal components (ISPCA) [PV18] has extended Bair's method to multiclass classification and reduced computational complexity via an iterative deflationary scheme.

A method herein referred to as Barshan's method [BGAJ11] approaches SPCA by means of the Hilbert-Schmidt Independence Criterion (HSIC). In a universal reproducing kernel Hilbert space (RKHS), two random variables are independent if and only if their HSIC is zero. Barshan's method maximizes an empirical measure of the HSIC, which has the form of a trace maximization problem similar to PCA. This method has also been extended to sparse SPCA [SGCE17] as well as the federated setting [BYTM24].

The supervised singular value decomposition [LYNS16] (SSVD), takes a somewhat different approach. They propose an inverse regression model in which Y is a factor in a low rank generative process for X. Specifically, the SSVD model has the form $X = UL^\top + E$, where $U = YB + F$ where E and F are error matrices, U is a low-rank score matrix, and B is a coefficient matrix. This

method has only been developed for regression.

In the classification setting the Supervised Linear Centroid-Encoder (SLCE) [GK24] finds an orthogonal projection minimizing the reconstruction error of class centroids rather than individual samples. The resulting optimization admits a closed-form solution via eigendecomposition of the symmetric matrix $X\tilde{C}^T + \tilde{C}X^T - XX^T$, where $\tilde{C}$ contains the class centroids.

Papazoglou et al. [PY25] propose Covariance Supervised PCA (CSPCA), which maximizes a weighted combination of the covariance between projected data and responses and the variance of the projected data, controlled by a regularization parameter κ. The problem reduces to eigendecomposition of $C = X^T YY^T X + \kappa X^T X$, yielding a closed-form solution. The authors extend the method to high-dimensional settings via Nyström approximation and to classification via a delta kernel.

The approach most similar to our work, and the only SPCA method to model PE directly, is supervised probabilistic principal component analysis [YYT+06] (SPPCA). SPPCA extends the probabilistic principal component analysis (PPCA) [TB99] framework. As with PPCA, the likelihood model of SPPCA allows for statistical testing and Bayesian inference. The method uses an EM algorithm, which can be slow to converge. In addition, this approach places the same amount of emphasis on the dependent and independent variables, and is sensitive to the relative dimensions of X and Y. However, SPPCA provides a convenient and straightforward extension to the semi-supervised setting.

Finally, we mention some related lines of work. The supervised sparse PCA problem has been explored from a regularization perspective [KFTS15, KFTS18, Kaw20]. Recent work has also explored the SPCA problem in the dynamic subspace setting [OWHZ24].

4.2.3 Theory

This work considers an iterative algorithm for solving a non-convex matrix factorization problem that is identifiable only up to an orthogonal transformation. As such, there are a number of highly related lines of work in principal subspace estimation and non-convex matrix factorization.

The convergence properties of gradient-based algorithms for principal component analysis have been studied extensively in both online and batch settings. In the online setting, the widely studied Oja's rule [Oja82], which applies a simple projected gradient update, has been shown to converge under general conditions [YYLT05, BDF13, JJK+16]. The GROUSE algorithm [BNR10], which is similar to the current work in that it performs online subspace tracking directly on the Grassmannian via incremental gradient steps, has been shown to be equivalent to Oja's rule, inheriting existing convergence results [Bal22]. Beyond PCA, there have several results showing that many nonconvex matrix factorization problems have benign optimization landscapes, including

the Burer-Monteiro approach for smooth semidefinite programs [BVB16, BVB20] among others. This line of work has been extended to the supervised matrix factorization setting by lifting the problem to a higher dimensional space, providing a globally convergent algorithm [LLY23], and landscape analysis [LLY24].

4.3 Method

In this section we recast the supervised PCA problem (4.1) as an unconstrained optimization problem over the product manifold $\mathcal{M} \coloneqq \mathrm{St}(p,r) \times \mathbb{R}^{r\times q}$ and introduce the Riemannian gradient descent algorithm used to solve it. Throughout, $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$ denote the row-stacked covariate and response matrices.

4.3.1 Problem Reformulation

We first observe that, on the Stiefel manifold, the PCA term in (4.1) admits a simpler equivalent form. For any $L \in \mathrm{St}(p,r)$ we have $L^\top L = I_r$, so $LL^\top$ is the orthogonal projector onto $\mathrm{range}(L)$ and

$$\|X - XLL^\top\|_F^2 \;=\; \|X\|_F^2 \;-\; \|XL\|_F^2.$$

Since $\|X\|_F^2$ is independent of the optimization variables, minimizing the reconstruction term on $\mathrm{St}(p,r)$ is equivalent to maximizing $\|XL\|_F^2$. Utilizing this fact and reparameterizing the trade-off between the supervised and unsupervised terms as a convex combination, we recast (4.1). Defining $\theta \coloneqq (L,\beta) \in \mathcal{M}$ and

$$h(\theta) \coloneqq \|XL\|_F^2, \tag{4.4}$$

$$G(\theta) \coloneqq \sum_{i=1}^{n} g(\boldsymbol{y}_i, \boldsymbol{x}_i, \theta), \tag{4.5}$$

$$f(\theta,\lambda) \coloneqq (1-\lambda)G(\theta) \;-\; \lambda\, h(\theta), \tag{4.6}$$

problem (4.1) becomes

$$\min_{\theta\in\mathcal{M}} \; f(\theta,\lambda), \tag{4.7}$$

with $\lambda \in [0,1]$. This reparameterization confines λ to an closed interval that interpolates between the purely supervised solution ($\lambda = 0$) and the unsupervised PCA solution ($\lambda = 1$).

For concreteness we explicitly consider the squared error and logistic losses given by

$$g_{\text{LS}}(\boldsymbol{y}, \boldsymbol{x}, L, \beta) = \|\boldsymbol{y} - \beta^\top L^\top \boldsymbol{x}\|_2^2 \quad \text{and}$$
$$g_{\text{LR}}(\boldsymbol{y}, \boldsymbol{x}, L, \beta) = \log \frac{\exp(\boldsymbol{x}^\top L \beta_{\boldsymbol{y}})}{\sum_{j^\top=1}^{q} \exp(\boldsymbol{x}^\top L \beta_{j^\top})},$$

respectively, where $\beta_{\boldsymbol{y}}$ is the column of β corresponding to the class given by one-hot vector $\boldsymbol{y}$. These two methods will be referred to as least squares PCA (LSPCA) and logistic regression PCA (LRPCA).

Remark 4 (From alternating to joint optimization)**.** *Our prior work [RSB+19] solved* (4.1) *by alternating optimization over L and β, with the L-update performed on the Grassmannian. The joint formulation* (4.7) *eliminates the inner subproblem, admits a single Lipschitz-gradient analysis on the product manifold, and yields the cleaner convergence analysis developed in Section 4.4.*

Remark 5 (Stiefel vs. Grassmann)**.** *We work on $\mathrm{St}(p, r)$ rather than $\mathrm{Gr}(p, r)$ because it clarifies the relationship between iterates of L and iterates of β, simplifying the convergence analysis. Note the objective $f(\theta, \lambda)$ is invariant under the reparameterization $(L, \beta) \mapsto (LQ, Q^\top \beta)$ for unitary Q, hence the results of Section-4.4 characterizing local minima descend cleanly to the Grassmannian.*

4.3.2 Manifold Geometry Necessities

In this section we introduce quantities and notation that will be relevant for the results in Section-4.4.

We endow $\mathrm{St}(p, r)$ with the Euclidean metric inherited from $\mathbb{R}^{p\times r}$, and endow $\mathbb{R}^{r\times q}$ with its standard inner product, making $\mathcal{M}$ a Riemannian product manifold [Bou23, Prop.-3.20]. The linearization of a manifold around any point is referred to as its *tangent space*. The tangent space at $L \in \mathrm{St}(p, r)$ is

$$T_L \,\mathrm{St}(p, r) \;=\; \{U \in \mathbb{R}^{p\times r} \;:\; L^\top U + U^\top L = 0\}, \tag{4.8}$$

[Bou23, Ch.-3] and the orthogonal projection onto it is

$$P_L(Z) \;=\; Z - L\,\mathrm{sym}(L^\top Z), \qquad \mathrm{sym}(A) \coloneqq \tfrac{1}{2}(A + A^\top) \tag{4.9}$$

[Bou23, Sec.7.3]. The tangent space at $\theta = (L, \beta) \in \mathcal{M}$ is the product $T_\theta\mathcal{M} = T_L \,\mathrm{St}(p, r) \times \mathbb{R}^{r\times q}$ [Bou23, Prop.-3.20].

Given a smooth extension $\bar{\mathrm{f}} : \mathbb{R}^{p\times r} \times \mathbb{R}^{r\times q} \to \mathbb{R}$ of f to a neighborhood of $\mathcal{M}$ in the ambient space, the Riemannian gradient of f at θ decomposes blockwise as

$$\mathrm{grad}\, f(\theta) \;=\; \left(P_L\left(\nabla_L \bar{\mathrm{f}}\right),\; \nabla_\beta \bar{\mathrm{f}}\right), \tag{4.10}$$

Algorithm 2 Riemannian gradient descent for Problem (4.7)

Require: Data (X, Y); rank r; weight $\lambda \in [0, 1]$; step size $\eta > 0$; tolerance $\varepsilon > 0$; max iterations T.

Initialize $L_0 \in \mathrm{St}(p, r)$ at the top-r right singular vectors of X.
Initialize $\beta_0 = \min_\beta \sum_{i=1}^n g(\mathbf{x}_i, \boldsymbol{y}_i, (L_0, \beta))$.
for $t = 0, 1, \ldots, T-1$ **do**
    Compute Euclidean partials $\partial_L \bar{\mathrm{f}}(\theta_t, \lambda)$ and $\partial_\beta \bar{\mathrm{f}}(\theta_t, \lambda)$.
    $\xi_t \leftarrow \partial_L \bar{\mathrm{f}}(\theta_t, \lambda) - L_t \,\mathrm{sym}\big(L_t^\top \partial_L \bar{\mathrm{f}}(\theta_t, \lambda)\big)$.
    $L_{t+1} \leftarrow \mathrm{qf}\big(L_t - \eta\, \xi_t\big)$.
    $\beta_{t+1} \leftarrow \beta_t - \eta\, \partial_\beta \bar{\mathrm{f}}(\theta_t, \lambda)$.
    if $\| \operatorname{grad} f(\theta_{t+1}, \lambda)\| \le \varepsilon$ **then break**
return $\theta_T = (L_T, \beta_T)$.

where the β-component coincides with the Euclidean gradient [Bou23, Ex.-3.67]. We refer to a point $\theta^* = (L^*, \beta^*)$ where $\operatorname{grad} f(\theta) = 0$ as a *Riemannian critical point* of f.

We retract tangent steps back to $\mathcal{M}$ using the QR retraction for L, while β needs no retraction. Given $\xi = (U, W) \in T_\theta \mathcal{M}$, the retraction is given by

$$R_\theta(\xi) = (R_L(U), R_\beta(W)) \;=\; (\mathrm{qf}(L + U),\; \beta + W), \tag{4.11}$$

where $\mathrm{qf}(\cdot)$ returns the Q-factor of the thin QR decomposition. This retraction is first-order and costs $O(pr^2)$ per call. The polar retraction $R_L^{\mathrm{pol}}(\xi) = (L + \xi)\big((L + \xi)^\top (L + \xi)\big)^{-1/2}$ is a second-order alternative and may be substituted without affecting the asymptotic guarantees of Section 4.4. We refer the reader to [AMS09, Bou23] for a complete treatment of Riemannian optimization on Stiefel and Grassmann manifolds.

4.3.3 Riemannian Gradient Descent

We solve (4.7) by Riemannian gradient descent (RGD) onthe product manifold $\mathcal{M}$. At each iteration we evaluate the Riemannian gradient via (4.10), take a step in the negative-gradient direction with fixed step size η, and retract back to the manifold. The procedure is summarized in Algorithm 2.

For a general smooth supervised loss g, evaluating $\partial_L \bar{\mathrm{f}}$ and $\partial_\beta \bar{\mathrm{f}}$ is dominated by the cost of forming $XL \in \mathbb{R}^{n \times r}$ and applying $X^\top$ once, i.e., $O(npr)$ flops. The Stiefel projection and the QR retraction each contribute $O(pr^2)$. Hence the overall per-iteration cost is $O(npr + pr^2)$, which under the typical regime $r \ll \min(n, p)$ is dominated by the data-product term.

4.3.4 Kernel SPCA

In this section we extend all proposed methods to perform kernel SDR. Further details are provided in the suplementary material.

Kernel PCA (kPCA) [SSM97] is a means of performing non-linear unsupervised dimension reduction by performing PCA in a high-dimensional feature space associated to a symmetric positive definite kernel. Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a symmetric positive definite kernel function. Associated to k is a high dimensional feature space $\mathcal{F}$ and mapping Φ such that $\Phi : \mathbb{R}^p \to \mathcal{F}$. The kernel matrix associated to k is $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $k(\boldsymbol{y}, \boldsymbol{z}) = \langle \Phi(\boldsymbol{y}), \Phi(\boldsymbol{z}) \rangle_{\mathcal{F}}$ $\forall \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^p$. Let X_Φ be the matrix with n rows where each row is the representation in $\mathcal{F}$ of the corresponding row of X. Note that kPCA finds the projection of X_Φ onto its top r principal components, rather than the principal components themselves. Computing the principal components is usually impractical or intractable as Φ may be unknown and/or $\mathcal{F}$ may be of arbitrarily high or even infinite dimension. Computationally, all that is required is to find the eigenvectors corresponding to the r largest eigenvalues of the centered kernel matrix $\widetilde{K} = K - \frac{1}{n}\mathbf{1}\mathbf{1}'K - \frac{1}{n}K\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'K\mathbf{1}\mathbf{1}'$. This amounts to solving

$$\hat{L} = \min_{L} \|\widetilde{K} - \widetilde{K}LL'\|_F^2 \;\; s.t. \;\; L'L = I_r, \tag{4.12}$$

where now $p = n$, and so L is $n \times r$. Noting (4.12) has the same form as (4.2), and that the projection of X_Φ onto its top r principal components is given by $\widetilde{K}\hat{L}$, we can kerneliize LSPCA and LRPCA (which we call kLSPCA and kLRPCA, respectively) by simply substituting $\widetilde{K}$ for X in (4.7), i.e.,

$$\min_{(L,\beta)\in\mathcal{M}} (1-\lambda) G\left((L, \beta)\right) - \lambda \|\widetilde{K}L\| \tag{4.13}$$

Further details are given in section 4.8.2.

4.4 Theoretical Results

In this section we establish three results characterizing the behavior of Algorithm 2 and the geometry of the objective $f(\cdot, \lambda)$. We first show that Algorithm 2 converges to a stationary point at the standard $O(1/\varepsilon^2)$ rate (Lemma 17). We then give a second-order sufficient condition for a stationary point to be a local minimum, valid for any smooth supervised loss g (Theorem 12). Finally, we specialize to the LSPCA setting and obtain (i) a complete characterization of local minima in terms of the smallest singular value of β^* (Corollary 7), and (ii) a guarantee that, when $q < r \leq n$, the top $r - q$ principal directions of X are contained in the range of the globally optimal L^* (Theorem 13).

The following mild assumptions are used in the results of this section.

Assumption 6. $\lambda \in [0, 1]$.

Assumption 7. *For each i, $g(\boldsymbol{y}_i, \boldsymbol{x}_i, \cdot)$ admits a C^2 extension $\bar{\mathrm{g}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \cdot)$ to an open neighborhood of $\mathcal{M}$ in $\mathbb{R}^{p\times r} \times \mathbb{R}^{r\times q}$.*

Assumption 8. *The Euclidean gradient of $\bar{\mathrm{g}}(\theta)$ is Lipschitz with constant γ_g on an open neighborhood of $\mathcal{M}$.*

Assumption 9. *The supervised loss function $g(\cdot, \cdot, \theta)$ is such that the sublevel sets*

$$\Omega_{\theta_0,\lambda} = \{\theta \in \mathcal{M} : \ f(\theta, \lambda) \leq f(\theta_0, \lambda)\},$$

are bounded, and hence compact, where $\theta_0 \in \mathcal{M}$ denotes the initializer of Algorithm-2.

Assumption 7 is satisfied by every loss considered in our experiments (squared error, logistic). Assumption 8 follows from Assumption 7 together with the compactness of $\mathrm{St}(p, r)$; we state it separately because the Lipschitz constant γ_f appears explicitly in the step-size choice for Algorithm 2. Specifically, we have the following lemma regarding the Lipschitz constant of the gradient of $\bar{\mathrm{f}}$.

Lemma 16. *Under assumption 8, $\nabla \bar{\mathrm{f}}(\theta)$ is Lipschitz continuous on an open neighborhood of $\mathcal{M}$ with constant $\gamma_f = (1 - \lambda)n\gamma_g + 2\lambda\|X\|_2^2$.*

4.4.1 Convergence to a Stationary Point

Lemma 17. *Let Assumptions 6-9 hold. Let $F := \sup_{\theta\in\Omega} \|\operatorname{grad} f(\theta, \lambda)\| < \infty$. For fixed step size $\eta \in (0, \min(c/F, 1/\gamma_f))$ and any $\epsilon > 0$, Algorithm-2 produces an iterate $\theta_T = (L_T, \beta_T)$ with $f(\theta_T, \lambda) \leq f(\theta_0, \lambda)$ and*

$$\|\operatorname{grad} f(\theta_T, \lambda)\|_F \leq \epsilon,$$

where T is upper bounded by

$$T \leq \left\lceil \frac{2(f(\theta_0, \lambda) - f(\theta_T, \lambda))}{\eta\epsilon^2} \right\rceil.$$

The lemma is a direct instantiation of the standard $O(1/\varepsilon^2)$ rate for Riemannian gradient descent with a Lipschitz-gradient objective on a compact submanifold [Bou23, Ch. 4], applied to the product manifold $\mathcal{M}$. Compactness of $\mathrm{St}(p, r)$ ensures $f_{\min}$ is attained and that the gradient of f pulled back through the retraction is well-defined and bounded; the unconstrained factor $\mathbb{R}^{r\times q}$ contributes only the standard Euclidean Lipschitz bound on $\partial_\beta \bar{\mathrm{f}}$, which is finite by Assumption 8.

4.4.2 Characterization of Critical Points

Due to the coupling of L and β through the prediction term, the landscape of the SPCA objective is considerably more complex than that of PCA alone, where all local minima are global and all other critical points are strict saddles [LLA+19]. In the most general case, we can identify local minima via a generalized inequality involving the Hessian of the supervised loss function $g(\cdot)$. The condition holds for any smooth supervised loss.

Theorem 12 (Second-order sufficient condition)**.** *Let Assumption 7 hold and let $\theta^* = (L^*, \beta^*) \in \mathcal{M}$ be a Riemannian critical point of $f(\cdot, \lambda)$. Define the Lagrange multiplier matrix*

$$\Lambda^* \;\coloneqq\; \tfrac{1}{2}(1-\lambda)\, L^{*\top} \nabla_L \bar{\mathrm{G}}(\theta^*) \;-\; \lambda\, L^{*\top} X^\top X L^* \;\in\; \mathbb{R}^{r\times r}, \tag{4.14}$$

where $\bar{\mathrm{G}}(\theta) \coloneqq \sum_{i=1}^n \bar{\mathrm{g}}(\boldsymbol{y}_i, \mathbf{x}_i, \theta)$. Then θ^* is a strict local minimum of $f(\cdot, \lambda)$ if, for all nonzero $\xi = (U, W) \in T_{\theta^*}\mathcal{M}$,

$$(1-\lambda)\, \nabla^2 \bar{\mathrm{G}}(\theta^*)[\xi, \xi] \;>\; 2\lambda \|XU\|_F^2 \;+\; 2\, \mathrm{tr}\big(U^\top U\, \Lambda^*\big). \tag{4.15}$$

Remark 6 (Interpretation)**.** *The inequality in Theorem-(12) balances the curvature of the supervised loss against the curvature of the unsupervised PCA term on the tangent space at θ^*. As $\lambda \to 0^+$ the condition reduces to positive definiteness of $\sum_i \nabla^2 \bar{\mathrm{g}}$, the usual second-order condition for the supervised problem alone. As $\lambda \to 1^-$ the unsupervised curvature dominates, and the local minima of f concentrate at L^* spanning the leading principal directions of X. Intermediate values of λ produce a continuous trade-off.*

4.4.3 Results for Squared-Error

When g is the squared error loss, we show that the sufficient condition for local minima can be written strictly in terms of the tangent vector $U \in \mathrm{St}(p, r)$.

Corollary 7 (Local Minima Conditions for LSPCA)**.** *Let g be the squared error and let $\theta^* = (L^*, \beta^*)$ be a Riemannian critical point of $f(\cdot, \lambda)$. Assume $L^{*\top} X^\top X L^*$ is invertible and β^* has*

full column rank. Then θ^* is a strict local minimum of $f(\cdot,\lambda)$ if for all nonzero $U \in T_{L^*}\,\mathrm{St}(p,r)$,

$$\begin{aligned}(1-\lambda)\left(\|XU\beta^*\|_F^2 + \|XL^*\Sigma_L^{-1}b(U)\|_F^2\right) + \lambda\|XL^*U^\top\|_F^2 \\ > \lambda\|XU\|_F^2 + 2(1-\lambda)\|\Sigma_L^{-1/2}b(U)\|_F^2,\end{aligned} \tag{4.16}$$

where $\Sigma_L = L^{*\top}X^\top XL$, and $b(U) = L^{*\top}X^\top XU\beta^* - U^\top X^\top E^*$.

The next result shows for the squared error loss, when $r > q$, that the global minimizer $\theta^*_{gl} = (L^*_{gl}, \beta^*_{gl})$ of $f(\theta,\lambda)$ contains the $(r-q)$-dimensional principal subspace of X in $\mathrm{range}(L^*_{gl})$, regardless of $\lambda \in (0,1]$.

Theorem 13. *Let Assumption 6 hold. Let g be the squared error and assume $r > q$ and $\sigma_{r-q}(X) > \sigma_{r-q+1}(X)$. Let $X = U_X\Sigma_X V_X^\top$ be the SVD of X with $V_{r-q} \in \mathbb{R}^{p\times(r-q)}$. Denote the global minimizer $\theta^*_{gl} = (L^*_{gl}, \beta^*_{gl})$ of $f(\theta,\lambda)$. Then*

$$\mathrm{span}(V_{r-q}) \;\subseteq\; \mathrm{span}(L^*_{gl}). \tag{4.17}$$

Remark 7 (Interpretation)**.** *If $r > q$, $\beta \in \mathbb{R}^{r\times q}$ has rank at most q. So $\mathrm{range}(L\beta)$ is a subspace of $\mathbb{R}^p$ with $\dim(\mathrm{range}(L\beta)) \le q$, while $\mathrm{range}(L)$ is a subspace of $\mathbb{R}^p$ with $\dim(\mathrm{range}(L)) = p > q$ by definition. The remaining $r-q$ directions are therefore uncoupled from the supervised loss and are pinned by the unsupervised term to the top $r-q$ principal directions of X. Theorem 13 has a useful practical consequence: when $r > q$ one can initialize Algorithm 2 by fixing the first $r-q$ columns of L_0 to V_{r-q} and optimizing only the remaining $p \times q$ submatrix. Additionally, the top $(r-q)\times q$ submatrix of β is determined from the first $r-q$ singular values and left singular vectors of X, meaning only the bottom $q \times q$ submatrix of β must be optimized if Algorithm-2 is initialized from the truncated SVD of X.*

Conjecture 1. *Assume $r > q$, $\sigma_{r-q}(X) > \sigma_{r-q+1}(X)$, and let g be the squared error loss. Let $U_{r-q}\Sigma_{r-q}V_{r-q}^T$ be the SVD of X truncated to the first $r-q$ singular triplets. Let $\theta_0 = (L_0, \beta_0)$ be the initialization of Algorithm-2 given by*

$$L_0 = \begin{bmatrix} V_{r-q}, & \tilde{L}_0 \end{bmatrix},$$
$$\beta_0 = \begin{bmatrix} \Sigma_{r-q}^{-1}U_{r-q}Y \\ \tilde{\beta}_0 \end{bmatrix},$$

where $\tilde{L}_0 \in \mathrm{St}(p,q)$ satisfies $\mathrm{range}(\tilde{L}_0) \cap \mathrm{range}(V_{r-q}) = \{0\}$ and $\tilde{\beta}_0 \in \mathbb{R}^{q\times q}$. Then the iterates $\theta_k = (L_k, \beta_k)$ of Algorithm-2 satisfy $\mathrm{range}(V_{r-q}) \subsetneq \mathrm{range}(L_k)$ for all $k \ge 0$.

4.5 Experiments

In this section we evaluate the proposed method against a suite of existing approaches. Among SPCA methods, we compare against Barshan's method, SPPCA, SSVD, SLCE, CSPCA, and ISPCA (subsuming Bair's method). We include PCR/PCC as an unsupervised baseline, and provide broader context through comparison with general SDR methods PLS and RRR in the regression setting, and FDA and RRLR in the classification setting. For the proposed method, we report results for both the joint Riemannian gradient descent algorithm (Algorithm 2) and the block coordinate descent (BCD) variant from our prior work [RSB+19]. In the kernel setting, we compare kLSPCA and kLRPCA against kPCR/kPCC and kBarshan, the only kernelized competitor.

The datasets used are described in Table 4.1, and span both the $n < p$ and $n > p$ regimes. Most are drawn from the UCI machine learning repository and the Arizona State feature selection repository; dataset-specific links are provided in Appendix 4.8.3. Datasets were selected on three grounds: (i) the HCP connectomics dataset represents a high-dimensional scientific application and was the primary motivation for the proposed methods; (ii) MNIST and FMNIST provide a controlled setting for visualizing the learned features; and (iii) several smaller datasets (Ionosphere, Barshan A, etc.) were included for direct comparability with prior SPCA work [BGAJ11, PV18]. In section 4.5.1, results are presented for comparison on the prediction task, as other SPCA works only consider this metric. As such, in section 4.5.1 CV is performed to minimize PE. In section 4.5.4 methods are evaluated on the basis of Pareto optimality.

For each experiment, 20% of the dataset was uniformly selected at random as an independent test set. For methods requiring parameter tuning, the remaining 80% was used in a 5-fold CV procedure; all methods were then retrained on the full 80% with the best parameters before evaluation on the held-out test set. This process, including test set selection, was repeated 10 times to produce the results in Table 4.2. In Table 4.2, the best overall linear and kernel methods are shown in red and blue, respectively, while the best SPCA method in each setting is marked with an asterisk ($*$).

4.5.1 Prediction Performance

We evaluate all methods at a fixed subspace dimension $r = 2$ and with r chosen by 10-fold CV. Results are averaged over 10 random train/test splits and reported in Table 4.2. We deliberately include the $r = 2$ setting because it corresponds to common practice in visualization, and serves as a stress test of each method's performance under severe dimensionality constraints.

In the regression setting the proposed methods are consistently competitive across all three datasets and both values of r. The two LSPCA variants attain the best result on Residential in

https://archive.ics.uci.edu/ml/datasets.php
https://jundongl.github.io/scikit-feature/datasets.html

Table 4.1: Datasets used in the experiments. For classification, q denotes the number of classes; for regression, the dimension of the response. The Music dataset was subsampled to $n = 100$ to obtain an $n<p$ setting (full dataset size in parentheses).

Name	Type	q	n	p	Source
Ionosphere	class.	2	354	34	UCI
Colon	class.	2	62	2000	ASU
Arcene	class.	2	200	10000	ASU
Residential	regr.	2	372	103	UCI
Music	regr.	2	100 (1059)	116	UCI
Barshan A	regr.	1	100	4	[BGAJ11]
MNIST	class.	10	60,000	784	[LBBH98]
FMNIST	class.	10	60,000	784	[XRV17]
HCP	regr.	-	863	34716	[VdG08, SAR+19]

both columns, tied with each other and statistically tied with RRR, whose nominal $r = 2$ error is marginally lower. LSPCA (BCD) is the best method on Music in both columns. On Barshan A, both LSPCA variants are statistically tied for the best linear result in both columns, alongside the closed-form Barshan, CSPCA, and ISPCA estimates. Among kernel methods, kLSPCA (Joint) achieves the best result on every regression dataset and column. SPPCA and SSVD essentially match the unsupervised PCR baseline on Residential and Music, with standard error identical to PCR on Music, suggesting their subspace estimates are only weakly supervised; both improve markedly over PCR on Barshan A, where they are competitive with the other supervised methods.

In the classification setting the picture is more nuanced, though the proposed methods fare well throughout. SLCE and CSPCA, both of which admit closed-form solutions, are strong competitors: SLCE is among the best on all three datasets, and CSPCA is strong under CV, tied for the best rate on Ionosphere and among the best on Colon. The proposed LRPCA methods are at least as strong: both variants tie for the lowest error on Arcene , and LRPCA (Joint) matches or beats SLCE in every column, tied at $r = 2$ on Ionosphere and lower under CV, and lower under CV on Colon. On Colon at $r = 2$ the lowest error is shared by RRLR, Barshan, and CSPCA, with SLCE and both LRPCA variants a slightly behind. In the kernel setting, kLRPCA (Joint) attains the best rate on every classification dataset and column. We note, however, that SLCE does not have a tunable trade-off between PE and VE; we return to this point in §4.5.4.

Performance of Algorithm 2 is nearly identical to BCD on five of six tasks, with the exception of the Music dataset.

Table 4.2: Comparison of mean squared error (regression) or error rate (classification) of competing methods, with standard error. Subspace dimension ($r = 2$) was held fixed for results in the first column of each dataset. For results in the second column, subspace dimension was chosen by 10-fold CV. SPCA methods have names shown in **bold**. For each experiment, the best linear method is shown in ***red***, the best kernel method is shown in ***blue***

	Regression					
	Residential		Barshan A		Music	
	$r = 2$	**CV**	$r = 2$	**CV**	$r = 2$	**CV**
PCR	1.052 ± 0.372	0.442 ± 0.140	0.712 ± 0.346	0.401 ± 0.259	1.953 ± 0.305	1.879 ± 0.302
PLS	0.458 ± 0.151	0.161 ± 0.037	$\mathbf{0.238 \pm 0.090}$	$\mathbf{0.238 \pm 0.090}$	1.919 ± 0.350	1.885 ± 0.324
RRR	$\mathbf{0.070 \pm 0.029}$	-	$\mathbf{0.238 \pm 0.091}$	-	30.6 ± 4.8	-
ISPCA	0.262 ± 0.209	0.209 ± 0.065	$\mathbf{0.230 \pm 0.059}$	$\mathbf{0.210 \pm 0.059}$	1.953 ± 0.305	1.879 ± 0.302
SPPCA	1.051 ± 0.370	0.427 ± 0.133	0.264 ± 0.064	0.246 ± 0.067	1.953 ± 0.305	1.879 ± 0.302
Barshan	0.614 ± 0.220	0.285 ± 0.087	$\mathbf{0.228 \pm 0.084}$	$\mathbf{0.232 \pm 0.084}$	1.949 ± 0.349	1.847 ± 0.326
SSVD	1.051 ± 0.371	0.425 ± 0.129	0.228 ± 0.083	0.232 ± 0.084	1.953 ± 0.305	1.879 ± 0.302
CSPCA	0.612 ± 0.213	0.284 ± 0.087	$\mathbf{0.229 \pm 0.083}$	$\mathbf{0.232 \pm 0.082}$	1.984 ± 0.320	1.864 ± 0.317
LSPCA (BCD)	$\mathbf{0.072 \pm 0.031}$	$\mathbf{0.074 \pm 0.030}$	$\mathbf{0.233 \pm 0.083}$	$\mathbf{0.233 \pm 0.081}$	$\mathbf{1.632 \pm 0.156}$	$\mathbf{1.667 \pm 0.133}$
LSPCA (Joint)	$\mathbf{0.072 \pm 0.031}$	$\mathbf{0.074 \pm 0.030}$	$\mathbf{0.233 \pm 0.085}$	$\mathbf{0.231 \pm 0.082}$	1.953 ± 0.305	1.880 ± 0.303
kPCR	1.053 ± 0.372	0.443 ± 0.146	0.405 ± 0.100	0.208 ± 0.053	1.955 ± 0.307	1.883 ± 0.300
kBarshan	0.899 ± 0.212	0.761 ± 0.166	0.276 ± 0.099	0.269 ± 0.099	2.054 ± 1.070	2.054 ± 1.077
kLSPCA (BCD)	0.287 ± 0.121	0.138 ± 0.097	0.163 ± 0.068	0.162 ± 0.065	2.061 ± 1.067	2.042 ± 1.069
kLSPCA (Joint)	$\mathbf{0.209 \pm 0.052}$	$\mathbf{0.205 \pm 0.048}$	$\mathbf{0.086 \pm 0.011}$	$\mathbf{0.069 \pm 0.014}$	$\mathbf{1.921 \pm 0.365}$	$\mathbf{1.882 \pm 0.301}$
	Classification					
	Ionosphere		Colon		Arcene	
	$r = 2$	**CV**	$r = 2$	**CV**	$r = 2$	**CV**
PCC	0.405 ± 0.037	0.133 ± 0.013	0.278 ± 0.073	0.194 ± 0.028	0.317 ± 0.022	0.250 ± 0.014
FDA	0.129 ± 0.016	-	0.242 ± 0.107	-	0.228 ± 0.084	-
RRLR	0.133 ± 0.017	-	$\mathbf{0.139 \pm 0.028}$	-	$\mathbf{0.100 \pm 0.025}$	-
ISPCA	0.129 ± 0.008	0.124 ± 0.005	0.217 ± 0.137	0.258 ± 0.133	0.313 ± 0.058	0.269 ± 0.077
SPPCA	0.367 ± 0.039	0.129 ± 0.008	0.278 ± 0.073	$\mathbf{0.167 \pm 0.048}$	0.325 ± 0.014	0.242 ± 0.022
Barshan	0.124 ± 0.005	$\mathbf{0.095 \pm 0.005}$	$\mathbf{0.139 \pm 0.028}$	$\mathbf{0.167 \pm 0.048}$	0.275 ± 0.014	0.142 ± 0.022
CSPCA	0.124 ± 0.028	$\mathbf{0.095 \pm 0.005}$	$\mathbf{0.139 \pm 0.028}$	$\mathbf{0.167 \pm 0.048}$	0.275 ± 0.014	0.150 ± 0.029
SLCE	$\mathbf{0.110 \pm 0.010}$	0.114 ± 0.008	$\mathbf{0.167 \pm 0.023}$	$\mathbf{0.167 \pm 0.023}$	$\mathbf{0.100 \pm 0.014}$	$\mathbf{0.100 \pm 0.014}$
LRPCA (BCD)	$\mathbf{0.114 \pm 0.008}$	0.110 ± 0.010	$\mathbf{0.167 \pm 0.048}$	$\mathbf{0.139 \pm 0.028}$	$\mathbf{0.100 \pm 0.014}$	$\mathbf{0.100 \pm 0.014}$
LRPCA (Joint)	$\mathbf{0.110 \pm 0.013}$	$\mathbf{0.100 \pm 0.014}$	$\mathbf{0.167 \pm 0.048}$	$\mathbf{0.139 \pm 0.028}$	$\mathbf{0.100 \pm 0.014}$	$\mathbf{0.100 \pm 0.014}$
kPCC	0.381 ± 0.017	0.081 ± 0.005	0.250 ± 0.083	0.250 ± 0.083	0.333 ± 0.008	0.283 ± 0.008
kBarshan	0.300 ± 0.045	0.327 ± 0.124	0.333 ± 0.162	0.333 ± 0.162	0.359 ± 0.048	0.379 ± 0.081
kLRPCA (BCD)	0.071 ± 0.040	0.056 ± 0.030	0.225 ± 0.111	0.225 ± 0.111	0.231 ± 0.073	0.215 ± 0.047
kLRPCA (Joint)	$\mathbf{0.047 \pm 0.010}$	$\mathbf{0.043 \pm 0.014}$	$\mathbf{0.194 \pm 0.028}$	$\mathbf{0.194 \pm 0.073}$	$\mathbf{0.108 \pm 0.022}$	$\mathbf{0.108 \pm 0.022}$

4.5.2 Visualizing the Learned Features

A key motivation for the proposed method is that maximizing VE alongside PE encourages learned features to reflect meaningful structure in the data, not merely class boundaries. To investigate this, we compare two-dimensional embeddings learned by PCA, LSPCA, LFDA, LDA, and PLS on binary classification tasks from MNIST and FMNIST (Figure 4.1). We include LFDA as a representative of methods that optimize purely for discriminability, without a variance objective, and LDA as an additional reference, noting that its embedding is one-dimensional (artificially spread vertically for visualization). All other methods use $r = 2$.

For MNIST, the task is binary classification of ones and sevens. LSPCA achieves the highest test accuracy (0.990) while producing a spatially coherent embedding: moving up and to the right along the direction of maximum variation, the vertical section of either digit rotates clockwise; moving

up and to the left, the horizontal crossbar that distinguishes sevens from ones grows longer. Intra-class structure is also visible; ones vary primarily in rotation while sevens exhibit variation along both axes and points near the class boundary naturally have short vertical sections, consistent with visual ambiguity between the two digits. The PCA and PLS embeddings capture similar structure with slightly lower accuracy. The LFDA embedding preserves the boundary property (ambiguous digits appear near the decision boundary) but otherwise lacks continuous spatial variation: there is no apparent attribute of the digits that changes along the vertical axis, and intra-class variation is difficult to interpret. LDA achieves 0.981 accuracy from a single dimension but the learned feature captures only a weak notion of vertical versus horizontal orientation.

For FMNIST, the task is binary classification of shirts and dresses. LSPCA again achieves the best two-dimensional accuracy (0.920). Moving up and to the right, the length-to-width ratio of the garment decreases, an obvious discriminating feature between dresses and shirts; moving up and to the left, brightness decreases. While brightness is not class-discriminative, it constitutes a major source of intra-class variation, and its capture illustrates the benefit of the variance objective. PCA learns a similar embedding. LFDA produces the lowest accuracy of all five methods on this task (0.832) and the embedding, while showing some length-to-width organization, discards most of the intra-class structure. LDA achieves the highest accuracy overall (0.960) from a single feature that cleanly represents length-to-width ratio but by design ignores variation orthogonal to the class boundary, which may be informative in other problems. PLS performs comparably to PCA in both accuracy and interpretability on this task.

4.5.3 HCP

We consider the application of LSPCA to connectomic data from the HCP [VdG08]. The data are constructed from functional magnetic resonance imaging (FMRI) of subjects brains, which are time-series, by a number of processing steps, the full details of which are given in [SAR+19]. Intuitively, FMRI measures activity patterns in the brain over time. Looking at the correlation between brain regions of activity over time gives a rough idea of which different parts of the brain interact, presumably working in conjunction to accomplish certain tasks or govern thought and behavior. The covariance structure of activity in different regions of an individuals brain has been found to be useful in explaining certain aspects of mood, behavior, or cognitive performance [SAR+19]. Finding the brain structures that account for the most variation in brain activity is a task for PCA. This section explores the use of LSPCA to find similar structures that have more explanatory power for a certain trait or performance on a task.

First, for each subject, voxels are collected into a coarse partition consisting of 264 functional areas, known as the Power parcellation [P+11]. Next, voxel-wise behavior is spatially averaged

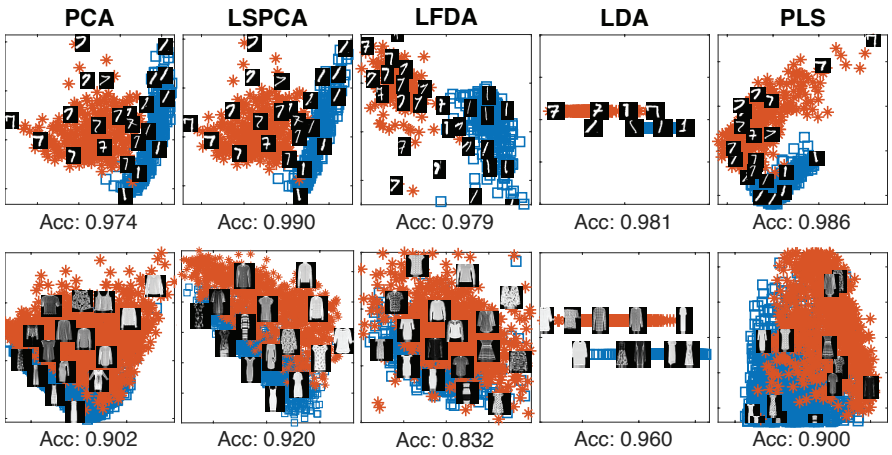
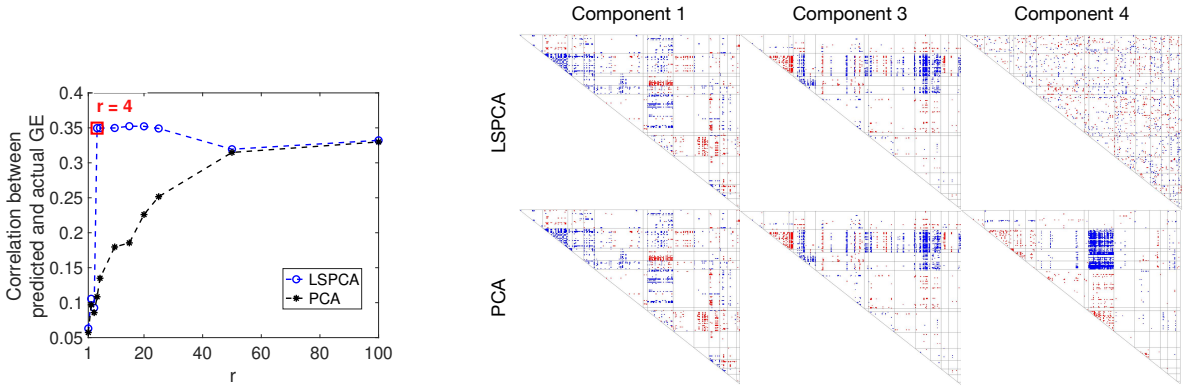


Figure 4.1: Comparison of features learned by (from left to right) PCA, LRPCA, and LFDA, LDA, and PLS on the task of classifying (top) ones and sevens from MNIST and (bottom) dresses and shirts from FMNIST. Shown below each embedding is the resulting classification accuracy on test data.

within each functional area, resulting in 264 time-series from which correlation matrices are constructed. We then construct our data by vectorizing the upper-triangular portions of the subjects' correlation matrices. As in Sripada et al. [SAR+19], our task is to identify patterns of correlated brain activity that predict certain response variables, called phenotypes, associated to each of the subjects, e.g., extroversion, processing speed. Figure 4.2a shows correlation of actual and predicted General Executive (GE) phenotype [SAR+19] on HCP as a function of subspace dimension. The experimental procedure used for this experiment is identical to that described in section 4.5.1. One approach employed in this task is brain basis set (BBS) modeling [SAR+19], which learns a subspace of small dimension using PCA. LSPCA using $r = 4$ is able to achieve equivalent or better predictive performance to BBS with $r = 100$, and substantially outperforms BBS with $r = 4$. Figure 4.2b shows three of the first four components produced by PCA and LSPCA

(PCs and LSPCs, respectively), reorganized according to the intrinsic connectivity network (ICN) assignments of Power [P$^+$11]. Components $1-3$ are substantially similar between LSPCA and PCA, but the respective fourth components bear little similarity. Inclusion of the fourth LSPCA component increases average test set correlation of the predicted phenotype from 0.11 to 0.33, while inclusion of the fourth PC increases average correlation from 0.11 to 0.13. The ICN assignments are determined strictly by intra-individual phenomenon, while the PCs and LSPCs are determined by inter-individual variation. It is therefore remarkable that there should be such alignment between PCs and ICN structure (visible in several components depicted in Figure 4.2b where weights concentrate in regions demarcated by overlaid gridlines that reflect ICN structure), a matter which is discussed further by Sripada et al. [SAR$^+$19]. However, the fourth LSPC, which is the most predictive component, does not demonstrate substantial ICN structure. This suggests that the bulk of the predictive connectivity for GE is not aligned with the Power ICN, and thus perhaps the structure of the Power ICNs alone are insufficient to fully understand inter-individual differences in GE.



(a) Correlation vs. subpsace dimension.

(b) Visualization of components one, three, and four.

Figure 4.2: (4.2a) Correlation between predicted and actual GE as a function of subspace dimension. (4.2b) Comparison of components produced by PCA and LSPCA on HCP data, with the components reshaped to reflect ICN assignments of Power [P$^+$11]. Blue (red) denote component entries that are two standard deviations above (below) the component mean.

4.5.4 Evaluating Pareto Optimality

In this section we compare the proposed approach to competitors through the lens of multiobjective optimization. The plots shown in Figure 4.3 correspond to the tests in Table 4.2 on the residential dataset, where the dimension $r = 2$ is fixed so the comparisons between methods can be direct and meaningful. Solutions that don't generalize to unseen data are of little practical use. We

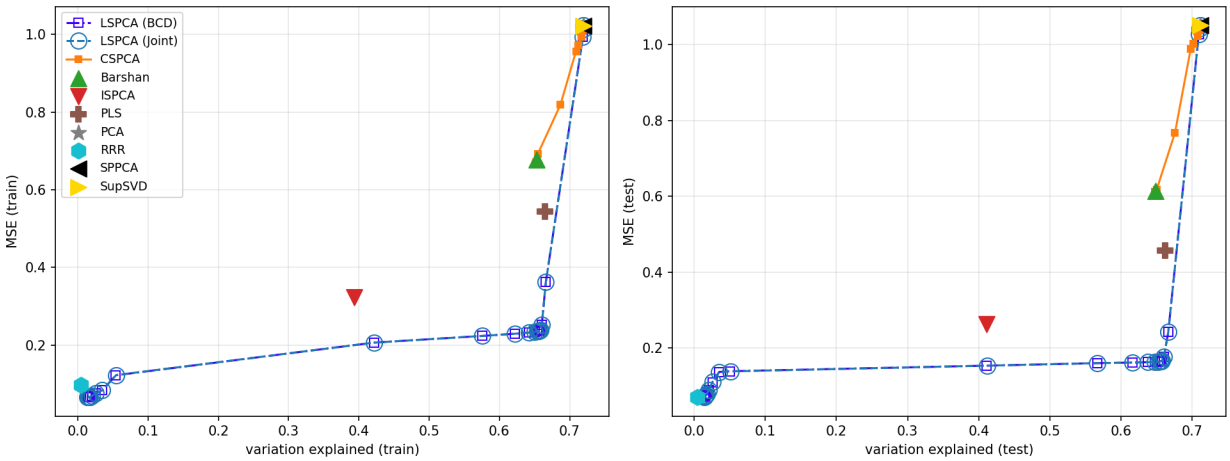


Figure 4.3: Comparison of Pareto optimality of competing methods on the residential dataset in terms of prediction error and variation explained, with subspace dimension $r = 2$.

therefore show plots corresponding to training and test data. The curves shown for LSPCA are parameterized by λ, while the curves shown for CSPCA are parameterized by κ. All plots were generated according to the same procedure described in section 4.5.1.

We find that the proposed methods dominate all SPCA competitors in the Pareto sense. In cases where one performance criterion is close, there appears to exist a value of λ for which our method performs significantly better in the other criterion. Moreover, the proposed methods appear able to decrease PE substantially while losing little VE until a point of diminishing returns is reached. After this point, PE can be decreased only marginally for the price of substantial VE. This behavior suggests that our methods are finding points on or close to the Pareto frontier, while most competing methods seem to correspond approximately to individual points on the Pareto frontier. This supports evidence from section 4.5.1 that these methods are able to perform well when $r > 2$ is allowed, but suffer when subspace dimension is severely restricted.

4.6 Conclusion

We recast SPCA as a joint optimization problem over a product manifold, and characterize conditions for local minimizers. For the proposed algorithm, we provide guarantees for convergence to a stationary point. The resulting method generalizes PCA, RRR, and other reduced rank prediction problems, and extends to the kernel setting. We demonstrated that the proposed approach dominates existing SPCA methods and is competitive with other SDR methods in terms of prediction while outperforming them in VE.

Finally, the use of PCA is ubiquitous in experimental research of the hard sciences, as well as the social sciences. Applying our approach with the proper link functions could yield meaningful

insight into important problems. For example, given the prevalence of PCA and ordinal regression tasks in neuroscience [DAZ+13], biology [DLB+17], and other fields, applying our method with the ordered logit response models could be a promising new approach.

4.7 Supplement

4.7.1 Theorem Proofs

We first restate the assumptions needed for our results. Then we restate each result and follow it with the proof.

Assumption 6. $\lambda \in [0, 1]$.

Assumption 7. *For each i, $g(\boldsymbol{y}_i, \boldsymbol{x}_i, \cdot)$ admits a C^2 extension $\bar{\mathrm{g}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \cdot)$ to an open neighborhood of $\mathcal{M}$ in $\mathbb{R}^{p\times r} \times \mathbb{R}^{r\times q}$.*

Assumption 8. *The Euclidean gradient of $\bar{\mathrm{g}}(\theta)$ is Lipschitz with constant γ_g on an open neighborhood of $\mathcal{M}$.*

Assumption 9. *The supervised loss function $g(\cdot, \cdot, \theta)$ is such that the sublevel sets*

$$\Omega_{\theta_0,\lambda} = \{\theta \in \mathcal{M} : \ f(\theta, \lambda) \leq f(\theta_0, \lambda)\},$$

are bounded, and hence compact, where $\theta_0 \in \mathcal{M}$ denotes the initializer of Algorithm-2.

Since $\mathrm{St}(p, r)$ is compact, Assumption 9 reduces to boundedness of the β-component of the sublevel set. It ensures that the iterates of Algorithm 2 remain in a fixed compact subset of $\mathcal{M}$, on which f is bounded below and $\mathrm{grad}\, f$ has a uniform Lipschitz constant; the rate in Lemma 17 then follows from the standard descent-lemma argument [Bou23, Prop. 4.13].

Lemma 18. *Under assumption 8, $\nabla \bar{\mathrm{f}}(\theta)$ is Lipschitz continuous on an open neighborhood of $\mathcal{M}$ with constant $\gamma_f = (1 - \lambda)n\gamma_g + 2\lambda\|X\|_2^2$.*

Proof. First we observe that $\bar{\mathrm{h}}$ is not a function of β, so the Lipschitz constant of $\nabla_\theta \bar{\mathrm{h}}$ must only depend on the gradient with respect to L. We have $\nabla \bar{\mathrm{h}}_L(L) = 2X^\top XL$, so

$$\begin{aligned}\|\nabla_L \bar{\mathrm{h}}(L_1) - \nabla_L \bar{\mathrm{h}}(L_2)\|_F &= \|2X^\top X(L_1 - L_2)\|_F \\ &\leq 2\|X^\top X\|_2\|L_1 - L_2\|_F \\ &= 2\|X\|_2^2\|L_1 - L_2\|_F,\end{aligned}$$

where $\|X\|_2$ is the spectral norm of X. Letting γ_g, be the Lipschitz constant for $\nabla \bar{\mathrm{g}}$ By the triangle inequality,

$$\begin{aligned}\|\nabla_\theta \bar{\mathrm{f}}(\theta_1) - \nabla_\theta \bar{\mathrm{f}}(\theta_2)\|_F &\leq (1-\lambda)\|\nabla_\theta \bar{\mathrm{G}}(\theta_1) - \nabla_\beta \bar{\mathrm{G}}(\theta_2)\|_F + \lambda\|\nabla_L \bar{\mathrm{h}}(L_1) - \nabla_L \bar{\mathrm{h}}(L_2)\|_F \\ &= (1-\lambda)n\gamma_g\|\theta_1 - \theta_2\|_F + 2\lambda\|X\|_2^2\|L_1 - L_2\|_F \\ &= \left((1-\lambda)n\gamma_g + 2\lambda\|X\|_2^2\right)\|\theta_1 - \theta_2\|_F\end{aligned}$$

gives an admissible Lipschitz constant for $\nabla \bar{\mathrm{f}}$. □

4.7.1.1 Convergence Proof

Lemma 17. *Let Assumptions 6-9 hold. Let $F \coloneqq \sup_{\theta\in\Omega} \|\operatorname{grad} f(\theta,\lambda)\| < \infty$. For fixed step size $\eta \in (0, \min(c/F, 1/\gamma_f))$ and any $\epsilon > 0$, Algorithm-2 produces an iterate $\theta_T = (L_T, \beta_T)$ with $f(\theta_T,\lambda) \leq f(\theta_0,\lambda)$ and*

$$\|\operatorname{grad} f(\theta_T,\lambda)\|_F \leq \epsilon,$$

where T is upper bounded by

$$T \leq \left\lceil \frac{2(f(\theta_0,\lambda) - f(\theta_T,\lambda))}{\eta\epsilon^2} \right\rceil.$$

Proof of Lemma 17. Let $\bar{\mathrm{f}}$ denote the C^2 extension of f to a neighborhood of $\mathcal{M}$ in the ambient space (Assumption 7), and let γ_f be the Lipschitz constant of $\nabla \bar{\mathrm{f}}$ on a neighborhood of the sublevel set $\Omega \coloneqq \{\theta \in \mathcal{M} : f(\theta,\lambda) \leq f(\theta_0,\lambda)\}$, which is compact by Assumption 9. For every retraction, by the Taylor expansion we have $R_\theta(\xi) = (L + U + O(\|U\|^2), \beta + W)$. From compactness of Ω and the $O(\|\xi\|^2)$ error of the QR retraction, there exists a constant $\alpha > 0$ such that for all $\theta \in \Omega$ and $\xi \in T_\theta\mathcal{M}$,

$$\|R_\theta(\xi) - (\theta + \xi)\| \leq \alpha\|\xi\|^2,$$

provided $\|\xi\| \leq c$ for some $c > 0$ depending on Ω [Bou23, Lemma 4.4]. Combined with the Lipschitz bound on $\nabla \bar{\mathrm{f}}$, this yields the retraction-smoothness inequality [Bou23, Lemma 4.7]: there exists $\gamma > 0$ (depending on γ_f, α, and $\sup_{\theta\in\Omega} \|\nabla \bar{\mathrm{f}}(\theta)\|$) such that for all $\theta \in \Omega$ and $\xi \in T_\theta\mathcal{M}$ with $\|\xi\| \leq c$,

$$f(R_\theta(\xi),\lambda) \;\leq\; f(\theta,\lambda) + \langle \operatorname{grad} f(\theta,\lambda), \xi\rangle + \frac{\gamma}{2}\|\xi\|^2. \tag{4.18}$$

Choose step size $\eta \in \left(0,\ \min(c\,/\,F,\ 1/\gamma)\right]$, where $F \coloneqq \sup_{\theta\in\Omega} \|\operatorname{grad} f(\theta,\lambda)\| < \infty$ (finite by compactness of Ω). Inductively, suppose $\theta_t \in \Omega$. Then $\| - \eta \operatorname{grad} f(\theta_t,\lambda)\| \leq \eta F \leq c$, so (4.18)

applies to $\theta_{t+1} = R_{\theta_t}(-\eta \operatorname{grad} f(\theta_t, \lambda))$:

$$f(\theta_{t+1}, \lambda) \le f(\theta_t, \lambda) - \eta \,\|\operatorname{grad} f(\theta_t, \lambda)\|^2 + \frac{\gamma\eta^2}{2}\|\operatorname{grad} f(\theta_t, \lambda)\|^2.$$

With $\eta \le 1/\gamma$, the coefficient $1 - \gamma\eta/2 \ge 1/2$, giving

$$f(\theta_{t+1}, \lambda) \;\le\; f(\theta_t, \lambda) - \tfrac{\eta}{2}\,\|\operatorname{grad} f(\theta_t, \lambda)\|^2. \tag{4.19}$$

In particular $f(\theta_{t+1}, \lambda) \le f(\theta_t, \lambda)$, so $\theta_{t+1} \in \Omega$ and the induction proceeds.

Let $f_{\min} \coloneqq \inf_{\theta\in\Omega} f(\theta, \lambda) > -\infty$ (continuity of f on compact Ω). Summing (4.19) from $t = 0$ to $t = T - 1$:

$$\sum_{t=0}^{T-1} \tfrac{\eta}{2}\,\|\operatorname{grad} f(\theta_t, \lambda)\|^2 \le f(\theta_0, \lambda) - f(\theta_T, \lambda) \le f(\theta_0, \lambda) - f_{\min}.$$

Hence

$$\begin{aligned}\min_{0\le t<T} \|\operatorname{grad} f(\theta_t, \lambda)\|^2 &\le \frac{1}{T}\sum_{t=0}^{T-1} \|\operatorname{grad} f(\theta_t, \lambda)\|^2 \\ &\le \frac{2\big(f(\theta_0, \lambda) - f_{\min}\big)}{\eta T}.\end{aligned}$$

For any $\varepsilon > 0$, choosing $T \ge 2\big(f(\theta_0, \lambda) - f_{\min}\big)/(\eta\varepsilon^2)$ guarantees some iterate $\theta^* \in \{\theta_0, \dots, \theta_{T-1}\}$ with $\|\operatorname{grad} f(\theta^*, \lambda)\| \le \varepsilon$. The descent inequality (4.19) also yields $f(\theta^*, \lambda) \le f(\theta_0, \lambda)$, completing the proof. □

4.7.1.2 The General Case

Theorem 12 (Second-order sufficient condition)**.** *Let Assumption 7 hold and let $\theta^* = (L^*, \beta^*) \in \mathcal{M}$ be a Riemannian critical point of $f(\cdot, \lambda)$. Define the Lagrange multiplier matrix*

$$\Lambda^* \;\coloneqq\; \tfrac{1}{2}(1-\lambda)\, L^{*\top}\nabla_L\, \bar{\mathrm{G}}(\theta^*) \;-\; \lambda\, L^{*\top} X^\top X L^* \;\in\; \mathbb{R}^{r\times r}, \tag{4.14}$$

where $\bar{\mathrm{G}}(\theta) \coloneqq \sum_{i=1}^n \bar{\mathrm{g}}(\boldsymbol{y}_i, \mathbf{x}_i, \theta)$. Then θ^* is a strict local minimum of $f(\cdot, \lambda)$ if, for all nonzero $\xi = (U, W) \in T_{\theta^*}\mathcal{M}$,

$$(1-\lambda)\,\nabla^2 \bar{\mathrm{G}}(\theta^*)[\xi, \xi] \;>\; 2\lambda\|XU\|_F^2 \;+\; 2\,\operatorname{tr}\big(U^\top U\,\Lambda^*\big). \tag{4.15}$$

Proof. The proof below works, but can be greatly simplified by appealing to EQ.-7.29 of [Bou23].

just a few lines.

The proof has three steps: (i) write the Lagrangian for the Stiefel constraint and identify the multiplier Λ^*; (ii) compute the Riemannian Hessian of f on $\mathcal{M}$ at θ^* as the Hessian of the Lagrangian restricted to the tangent space; (iii) apply the standard second-order sufficient condition.

The product manifold $\mathcal{M} = \mathrm{St}(p, r) \times \mathbb{R}^{r\times q}$ carries only the Stiefel constraint $\tau(L) \coloneqq L^\top L - I_r = 0$, with the β-factor unconstrained. With a symmetric multiplier $\Lambda \in \mathbb{R}^{r\times r}$, the Lagrangian is

$$\mathcal{L}(\theta, \Lambda) \;=\; f(\theta, \lambda) \;-\; \mathrm{tr}\big(\Lambda\, \tau(L)\big).$$

The first-order condition $\nabla_L \mathcal{L} = 0$ yields $\nabla_L f(\theta^*) = 2L^*\Lambda^*$. Left-multiplying by $L^{*\top}$ and using $L^{*\top}L^* = I_r$,

$$\Lambda^* \;=\; L^{*\top}\nabla_L f(\theta^*).$$

$$\nabla_L f(\theta^*, \lambda) = (1-\lambda)X^T(Y - XL^*\beta^*)\beta^{*T} - \lambda X^T X L^* = 2L^*\Lambda^*$$

Substituting $\nabla_L f(\theta) = (1-\lambda)\nabla_L \bar{\mathrm{G}}(\theta) - 2\lambda X^\top X L$ yields (4.14). The β-block of the gradient is unconstrained, so the criticality of θ^* also gives $(1-\lambda)\nabla_\beta \bar{\mathrm{G}}(\theta^*) = 0$, which also holds if $\lambda = 1$ without assumptions on the criticality of β^*.

For an embedded submanifold of Euclidean space with the induced metric, the Riemannian Hessian of f at a critical point equals the Hessian of the Lagrangian at (θ^*, Λ^*), restricted as a bilinear form to the tangent space $T_{\theta^*}\mathcal{M}$ [AMS09, Prop. 5.5.4]. For $\eta = (U, W) \in T_{\theta^*}\mathcal{M}$,

$$\mathrm{Hess}\, f(\theta^*)[\eta, \eta] \;=\; \nabla^2 f(\theta^*)[\eta, \eta] \;-\; \mathrm{tr}\big(\Lambda^*\, \nabla^2 c(L^*)[U, U]\big).$$

The map $L \mapsto L^\top L$ is quadratic with constant Hessian $\nabla^2\tau(L)[U, U] = 2\,U^\top U$, so the correction term equals $-2\ \mathrm{tr}(U^\top U\Lambda^*)$.

The ambient Hessian decomposes by linearity as $\nabla^2 f \;=\; (1-\lambda)\nabla^2\,\bar{\mathrm{G}} - \lambda\nabla^2\,\bar{\mathrm{h}}$, where $\bar{\mathrm{h}}(\theta) = \|XL\|_F^2 = \mathrm{tr}(L^\top X^\top X L)$ is quadratic in L and independent of β, giving $\nabla^2\,\bar{\mathrm{h}}(\theta^*)[\eta, \eta] = 2\|XU\|_F^2$. Combining,

$$\begin{aligned}\mathrm{Hess}\, f(\theta^*)[\eta, \eta] = (1-\lambda)\nabla^2\,\bar{\mathrm{G}}(\theta^*)[\eta, \eta] - 2\lambda\|XU\|_F^2 \\ - 2\,\mathrm{tr}\big(U^\top U\,\Lambda^*\big).\end{aligned}$$

By the standard second-order sufficient condition for Riemannian optimization on embedded submanifolds [AMS09, Prop. 5.5.7], if $\mathrm{Hess}\, f(\theta^*)[\eta, \eta] > 0$ for all nonzero $\eta \in T_{\theta^*}\mathcal{M}$, then θ^* is a strict local minimum of $f(\cdot, \lambda)$. Rearranging for strict positivity gives (4.15). □

4.7.1.3 The Squared Error Case

We now specialize Theorem 12 to the squared-error loss $g(\boldsymbol{y}_i, \mathbf{x}_i, \theta) = \|\boldsymbol{y}_i - \beta^\top L^\top \mathbf{x}_i\|_2^2$, for which the multiplier Λ^* admits a particularly clean form.

Lemma 19 (Multiplier simplification at an LSPCA critical point)**.** *Let g be the squared error and let $\theta^* = (L^*, \beta^*)$ be a critical point of $f(\cdot, \lambda)$ with $\lambda \in (0, 1)$. Define the residual $E^* := Y - XL^*\beta^*$. Then*

$$L^{*\top} X^\top E^* \;=\; 0, \qquad \textit{and} \qquad \Lambda^* \;=\; -\lambda\, L^{*\top} X^\top X L^*. \tag{4.20}$$

Proof. For the squared error, $\bar{G}(\theta) = \|Y - XL\beta\|_F^2$ has Euclidean gradients

$$\begin{aligned}\nabla_L \bar{G}(\theta) &= -2X^\top (Y - XL\beta)\,\beta^\top \\ \nabla_\beta \bar{G}(\theta) &= -2\, L^\top X^\top (Y - XL\beta).\end{aligned}$$

The β-block of the criticality condition for f, $\nabla_\beta \bar{G}(\theta^*) = 0$, gives $L^{*\top} X^\top E^* = 0$. Substituting into the multiplier expression (4.14),

$$\begin{aligned}\Lambda^* &= (1-\lambda)\, L^{*\top}\big(-2X^\top E^* \beta^{*\top}\big) - 2\lambda\, L^{*\top} X^\top X L^* \\ &= -2(1-\lambda)\, \underbrace{(L^{*\top} X^\top E^*)}_{=0}\, \beta^{*\top} - 2\lambda\, L^{*\top} X^\top X L^* \\ &= -2\lambda L^{*\top} X^\top X L^*.\end{aligned}$$

□

Corollary 7 (Local Minima Conditions for LSPCA)**.** *Let g be the squared error and let $\theta^* = (L^*, \beta^*)$ be a Riemannian critical point of $f(\cdot, \lambda)$. Assume $L^{*\top} X^\top X L^*$ is invertible and β^* has full column rank. Then θ^* is a strict local minimum of $f(\cdot, \lambda)$ if for all nonzero $U \in T_{L^*} \mathrm{St}(p, r)$,*

$$\begin{aligned}(1-\lambda)\left(\|XU\beta^*\|_F^2 + \|XL^*\Sigma_L^{-1} b(U)\|_F^2\right) + \lambda\|XL^*U^\top\|_F^2 \\ > \lambda\|XU\|_F^2 + 2(1-\lambda)\|\Sigma_L^{-1/2} b(U)\|_F^2,\end{aligned} \tag{4.16}$$

where $\Sigma_L = L^{*\top} X^\top X L$, and $b(U) = L^{*\top} X^\top X U \beta^* - U^\top X^\top E^*$.

Proof. We apply Theorem 12 with Λ^* given by Lemma 19. The Euclidean Hessian of $\bar{G}$ at θ^*

along $\eta = (U, W) \in T_{\theta^*}\mathcal{M}$ is, by a direct computation,

$$\begin{aligned}\nabla^2 \bar{G}(\theta^*)[\eta, \eta] = \|XU\beta^* + XL^*W\|_F^2 \\ - 2\ \mathrm{tr}\big(W^\top U^\top X^\top E^*\big).\end{aligned} \tag{4.21}$$

The first term is the contribution of the bilinear part of the supervised gradient (the directional derivative of $\nabla_L \bar{G} = -2X^\top E\beta^\top$ and $\nabla_\beta \bar{G} = -2L^\top X^\top E$ in direction (U, W), contracted against (U, W)). The second term arises from differentiating the residual E in the direction (U, W).

Substituting $\Lambda^* = -2\lambda L^{*\top} X^\top X L^*$ from Lemma 19 into the second-order inequality (4.15) of Theorem 12, we denote the smaller side of the inequality

$$\begin{aligned}L(U) &\coloneqq \lambda\Big[\|XU\|_F^2 - \mathrm{tr}\big(U^\top U \cdot L^{*\top} X^\top X L^*\big)\Big] \\ &= \lambda\Big[\|XU\|_F^2 - \|XL^*U^\top\|_F^2\Big],\end{aligned}$$

and the larger side of the inequality

$$\begin{aligned}R(U, W) &= (1-\lambda)\,\|XU\beta^* + XL^*W\|_F^2 \\ &\quad - 2(1-\lambda)\ \mathrm{tr}(W^\top U^\top X^\top E^*) \\ &= (1-\lambda)\left(\|XU\beta^*\|_F^W + \|XL^*W\|_F^2 + 2\,\mathrm{tr}(W^\top b(U)\right)\end{aligned}$$

for all nonzero $(U, W) \in T_{\theta^*}\mathcal{M}$, where $b(U) \coloneqq L^{*\top} X^\top X U\beta^* - U^\top X^\top E^*$, and $R(U, W)$ refers to the RHS of the inequality, and $L(U)$ to the LHS. We wish to discern the conditions under which

$$L(U) < R(U, W). \tag{4.22}$$

For each fixed U, the RHS $R(U, W)$ is a strictly convex quadratic in W (the leading $W^\top \Sigma_L W$ term has $\Sigma_L \coloneqq L^{*\top} X^\top X L^* \succ 0$ by hypothesis). Setting the gradient in W to zero and solving yields the minimizing W^*

$$\begin{aligned}W^* &= \Sigma_L^{-1}\big[U^\top X^\top E^* - L^{*\top} X^\top X U\beta^*\big] \\ &= -\Sigma_L^{-1} b(U).\end{aligned}$$

The minimum value of $R^*(U) \coloneqq R(U, W^*)$, is a quadratic form in U alone. Denoting $\Sigma_{L^*} = L^{*\top} X^\top X L^*$, $\Sigma_\perp = L_\perp^{*\top} X^\top X L_\perp^*$, and $\Sigma_\times = L^{*\top} X^\top X L_\perp^*$, the condition (4.22) for strict local

minimality is then equivalent to

$$\begin{aligned} L(U) &< R^*(U) \\ &= (1-\lambda)\Big[\|XU\beta^*\|_F^2 + \|XL^*W^*\|_F^2 + 2\,\mathrm{tr}(W^{*\top} b(U))\Big] \end{aligned}$$

□

4.7.1.4 LSPCA Contains PCA Subspace

Theorem 13. *Let Assumption 6 hold. Let g be the squared error and assume $r > q$ and $\sigma_{r-q}(X) > \sigma_{r-q+1}(X)$. Let $X = U_X \Sigma_X V_X^\top$ be the SVD of X with $V_{r-q} \in \mathbb{R}^{p\times(r-q)}$. Denote the global minimizer $\theta_{gl}^* = (L_{gl}^*, \beta_{gl}^*)$ of $f(\theta, \lambda)$. Then*

$$\mathrm{span}(V_{r-q}) \subseteq \mathrm{span}(L_{gl}^*). \tag{4.17}$$

Proof. Let $(L^*, \beta^*) = (L_{gl}^*, \beta_{gl}^*)$. Assume $\mathrm{span}(V_{r-q}) \not\subset \mathrm{span}(L^*)$. Let $W = L^*(XL^*)^+ Y = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q] \in \mathbb{R}^{p\times q}$, and let $PDS = \mathrm{svd}((I - V_{r-q}V_{r-q}^\top)W)$ where $P = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_q]$. Now let $L = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{r-q}, P]$, so that $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q \in \mathrm{span}(L)$. It follows that

$$\begin{aligned} \min_\beta \|Y - XL\beta\|^2 &\le \|Y - XL[W^\top V_{r-q},\ S^\top D^\top]^\top\|^2 \\ &= \|Y - X(V_{r-q}V_{r-q}^\top W + (I - V_{r-q}V_{r-q}^\top)W)\|^2 \\ &= \|Y - XW\|^2 \\ &= \min_\beta \|Y - XL^*\beta\|^2. \end{aligned} \tag{4.23}$$

The above shows that L performs at least as well as L^* in terms of regression error. We now consider the performance of L in terms of variance explained. We have

$$\|XL\|_F^2 = \|X[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{r-q}, P]\|_F^2$$

By definition, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q \in \mathrm{span}(L^*)$, we can be sure that $\hat{\boldsymbol{w}} \in \mathrm{span}(L^*)$. Then there exists a matrix with orthonormal columns $\widetilde{L} = [\widetilde{\boldsymbol{\ell}}_1, \widetilde{\boldsymbol{\ell}}_2, \ldots, \widetilde{\boldsymbol{\ell}}_{r-1}, \hat{\boldsymbol{w}}]$ and $\mathrm{span}(\widetilde{L}) = \mathrm{span}(\hat{L})$. Define the

matrix $\widetilde{L}_{r-1} = [\widetilde{\boldsymbol{\ell}}_1, \widetilde{\boldsymbol{\ell}}_2, \ldots, \widetilde{\boldsymbol{\ell}}_{r-1}]$. It follows

$$
\begin{aligned}
&\|X - X\hat{L}\hat{L}^\top\|^2 - \|X - XLL^\top\|^2 \\
&= \|X - X\widetilde{L}\widetilde{L}^\top\|^2 - \|X - XLL^\top\|^2 \\
&= \operatorname{tr}\Big(X^\top X - 2X^\top X\widetilde{L}\widetilde{L}^\top + \widetilde{L}\widetilde{L}^\top X^\top X\widetilde{L}\widetilde{L}^\top \\
&\qquad - X^\top X + 2X^\top XLL^\top - LL^\top X^\top XLL^\top\Big) \\
&= \operatorname{tr}\Big(- 2X^\top X\widetilde{L}\widetilde{L}^\top + \widetilde{L}\widetilde{L}^\top X^\top X\widetilde{L}\widetilde{L}^\top \\
&\qquad + 2X^\top XLL^\top - LL^\top X^\top XLL^\top\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \operatorname{tr}\Big(2X^\top XLL^\top - X^\top XLL^\top LL^\top \\
&\qquad - 2X^\top X\widetilde{L}\widetilde{L}^\top + X^\top X\widetilde{L}\widetilde{L}^\top\widetilde{L}\widetilde{L}^\top\Big) && (4.24)\\
&= \operatorname{tr}\Big(X^\top XLL^\top - X^\top X\widetilde{L}\widetilde{L}^\top\Big) && (4.25)\\
&= \operatorname{tr}\Big(X^\top X(V_{r-1}V_{r-1}^\top + \hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top) \\
&\qquad - X^\top X(\widetilde{L}_{r-1}\widetilde{L}_{r-1}^\top + \hat{\boldsymbol{w}}\hat{\boldsymbol{w}}^\top)\Big) \\
&= \operatorname{tr}\Big(X^\top XV_{r-1}V_{r-1}^\top - X^\top X\widetilde{L}_{r-1}\widetilde{L}_{r-1}^\top\Big) \\
&= \|XV_{r-1}\|^2 - \|X\widetilde{L}_{r-1}\|^2 \\
&> 0. && (4.26)
\end{aligned}
$$

where (4.24) follows from the invariance of the trace to cyclic permutations, (4.25) is because L and $\widetilde{L}$ have orthonormal columns. The final inequality follows from the initial assumption that $\operatorname{span}(V_{r-1}) \not\subset \operatorname{span}(\hat{L})$ as longs as $\sigma_{r-1}(X) > \sigma_r(X)$. Together (4.23) and (4.26) imply that

$$
\begin{aligned}
(1-\lambda)\|Y - XL(XL)^+Y\|_F^2 + \lambda\|X - XLL^\top\|_F^2 < \\
(1-\lambda)\|Y - X\hat{L}\hat{\beta}\|_F^2 + \lambda\|X - X\hat{L}\hat{L}^\top\|_F^2
\end{aligned}
$$

for any $\lambda \in (0,1)$ which, in turn, implies $(\hat{L}, \hat{\beta})$ is not a global minimizer, a contradiction. □

4.8 Related Statistical Generative Model

4.8.1 Derivation of NLL

In this section we derive, in general terms, the NLL for the following model, which is a statistical generative model for our problem.

$$x \sim N(0, \sigma_x^2 I_p + \alpha LL'), \quad y|x \sim P_{y|x},$$

We put this in the functional form of the optimization formulation. Letting ξ represent any additional parameters related to $P_{y|x}$, we can write the NLL directly as

$$\begin{aligned} G(L, \beta, \alpha, \sigma_x^2, \xi) \triangleq & -\sum_{i=1}^{n} \ell_{y|x}(L, \beta, \xi; \boldsymbol{x}_i, \boldsymbol{y}_i) \\ & -\sum_{i=1}^{n} \ell_x(L, \sigma_x^2, \alpha; \boldsymbol{x}_i). \end{aligned}$$

We are interested in finding a simplified form for ℓ_x. In order to draw a connection to the optimization formulation, we make a few observations. First we can rewrite the covariance matrix of x as

$$\sigma_x^2 I_p + \alpha LL' = (\sigma_x I_p + \eta LL')^2,$$

where $\eta = \sqrt{\sigma_x^2 + \alpha} - \sigma_x$. Second, we can write the inverse of the covariance matrix

$$\begin{aligned} \left(\sigma_x^2 I_p + \alpha LL'\right)^{-1} &= (\sigma_x I_p + \eta LL')^{-2} \\ &= \frac{1}{\sigma_x^2}\left(I_p + \frac{\eta}{\sigma_x} LL'\right)^{-2} \\ &= \frac{1}{\sigma_x^2}\left(I_p - \frac{\eta}{\sigma_x} L(I_r + \frac{\eta}{\sigma_x} L'L)^{-1} L'\right)^2 \\ &= \frac{1}{\sigma_x^2}\left(I_p - \frac{\eta}{\sigma_x} L(\frac{\sigma_x + \eta}{\sigma_x} I_r)^{-1} L'\right)^2 \\ &= \frac{1}{\sigma_x^2}\left(I_p - \frac{\eta}{\sigma_x + \eta} LL'\right)^2 \end{aligned}$$

where the second step uses the matrix inversion lemma. Third, we simplify the determinant of the covariance matrix

$$\begin{aligned}\left|\sigma_x^2 I_p + \alpha LL'\right| &= \sigma_x^{2p}\left|I_p + \frac{\alpha}{\sigma_x^2}LL'\right| \\ &= \sigma_x^{2p}\left|I_r + \frac{\alpha}{\sigma_x^2}L'L\right| \\ &= \sigma_x^{2p}\left|\frac{\sigma_x^2+\alpha}{\sigma_x^2}I_r\right| \\ &= \sigma_x^{2p}\left(\frac{\sigma_x^2+\alpha}{\sigma_x^2}\right)^r,\end{aligned}$$

where the first step makes use of the Weinstein–Aronszajn identity [Poz14] (sometimes referred to as Sylvester's determinant theorem).

We now rewrite the second term of the NLL omitting additive constants as

$$\begin{aligned}-\sum_{i=1}^n \ell_x(L,\sigma_x^2,\alpha;\boldsymbol{x}_i) &= \frac{1}{2}\operatorname{Tr}\left(X(\sigma_x^2 I_p + \alpha LL')^{-1}X'\right) \\ &\quad + \frac{1}{2}n\log\left|\sigma_x^2 I_p + \alpha LL'\right| \\ &= \frac{1}{2\sigma_x^2}\operatorname{Tr}\left(X\left(I_p - \frac{\eta}{\sigma_x+\eta}LL'\right)^2 X'\right) \\ &\quad + \frac{1}{2}n\log\left(\sigma_x^{2p}\left(\frac{\sigma_x^2+\alpha}{\sigma_x^2}\right)^r\right) \\ &= \frac{1}{2\sigma_x^2}\|X - \frac{\eta}{\sigma_x+\eta}XLL'\|_F^2 \\ &\quad + \frac{1}{2}\left(n(p-r)\log(\sigma_x^2) + nr\log(\sigma_x^2+\alpha)\right).\end{aligned}$$

Resubstituting $\eta = \sqrt{\sigma_x^2+\alpha} - \sigma_x$ and simplifying yields

$$\begin{aligned}G(L,\beta,\alpha,\sigma_x^2,\xi) = &-\sum_{i=1}^n g(\boldsymbol{x}_i,\boldsymbol{y}_i,L,\beta) + \lambda\|X - \gamma XLL'\|_F^2 \\ &+ \frac{1}{2}\left(n(p-r)\log(\sigma_x^2) + nr\log(\sigma_x^2+\alpha)\right),\end{aligned}$$

where $\gamma = 1 - \left(\frac{\sigma_x^2}{\sigma_x^2+\alpha}\right)^{1/2}$, and $\lambda = \frac{1}{2\sigma_x^2}$ in the case of LSPCA and $\lambda = \frac{\sigma_y^2}{\sigma_x^2}$ in the case of LRPCA.

4.8.1.1 On Supervised Probabilistic PCA

SPPCA takes a latent variable approach similar to PPCA, extending PPCA to the supervised setting by modeling the conditional distribution of $y|z$. Furthermore, SPPCA assumes conditional independence of $y|z$ and $x|z$. The resulting model is

$$\begin{aligned} y|z &\sim N(W_y z, \sigma_y^2 I_q), \\ x|z &\sim N(W_x z, \sigma_x^2 I_p), \\ z &\sim N(0, \sigma_z^2 I_r). \end{aligned}$$

The conditional independence assumption may be overly strong, especially when the subspace dimension is misspecified, e.g., the subspace dimension is set too small to capture the full relationship between y and x.

Now consider a latent variable model for LSPCA, where all variables retain their previous definitions (L still has orthonormal columns):

$$y|x \sim N(\beta' L' x, \sigma_y^2 I_q),\ x|z \sim N(Lz, \sigma_x^2 I_p),\ z \sim N(0, \sigma_z^2 I_r).$$

Conditioning y on x alleviates the issue caused by the conditional independence assumption of SPPCA. Forming the joint distribution of x and y, ignoring log terms and additive constants, and integrating out the latent variable yields

$$\begin{aligned} f_{x,y}(x,y) =& f_{y|x}(y|x) \int_{-\infty}^{\infty} f_{x|z}(x) f_z(z) dz \\ &\propto \exp\big(-\frac{1}{2\sigma_y^2} \|y - \beta' L' x\|_2^2 \\ &\quad - \frac{\sigma_z^2}{2\sigma_x^2(\sigma_x^2+\sigma_z^2)} x' \big(\frac{\sigma_x^2+\sigma_z^2}{\sigma_z^2} I_p - LL'\big) x \big) \end{aligned}$$

where the $f_{(\cdot)}$ are the corresponding density functions, and logarithmic terms are ignored. If we write $\sigma_z^2 = \alpha \sigma_x^2$ with $\alpha = 2\eta\sigma_x + \eta^2$ (implying $\eta = \sqrt{\sigma_x^2 + \alpha} - \sigma_x$, as before) the negative log likelihood evaluated on data matrices X and Y reduces to

$$-G_{\text{LS}} \propto \|Y - XL\beta\|_F^2 + \frac{\sigma_y^2}{\sigma_x^2} \|X(I_p - \frac{\eta}{\sigma_x + \eta} LL')\|_F^2,$$

which matches the form of LSPCA.

Crucially, how should the conditioning of y on x rather than on z be interpreted? In the sug-

gested model

$$x = Lz + \zeta_x \implies y = \beta'(z + L'\zeta_x) + \zeta_y,$$

where $\zeta_x \sim N(0, \sigma_x^2 I_p)$ and $\zeta_y \sim N(0, \sigma_y^2 I_q)$. First note that $L'\zeta_x \sim N(0, \sigma_x^2 I_r)$, i.e., it is isotropic Gaussian noise in the latent subspace. With the substitution $\sigma_z^2 = \alpha\sigma_x^2$ the model becomes

$$\begin{aligned} x &= L(z + \frac{z'}{\sqrt{\alpha}}) + (I_p - LL')\zeta_x \\ &= L\widetilde{z} + (I_p - LL')\zeta_x, \\ y &= \beta'(z + \frac{z'}{\sqrt{\alpha}}) + \underbrace{\beta' L'(I_p - LL')\zeta_x}_{=0} + \zeta_y \\ &= \beta'\widetilde{z} + \zeta_y \end{aligned}$$

where $z, z' \overset{\text{i.i.d.}}{\sim} N(0, \alpha\sigma_x^2 I_r)$ and $\widetilde{z} \sim N(0, (1+\alpha)\sigma_x^2 I_r)$. We can now write the conditional distributions of y and x on $\widetilde{z}$

$$\begin{aligned} y|\widetilde{z} &\sim N(\beta'\widetilde{z}, \sigma_y^2 I_q), \\ x|\widetilde{z} &\sim N(L\widetilde{z}, \sigma_x^2(I_p - LL')), \\ \widetilde{z} &\sim N(0, (1+\alpha)\sigma_x^2 I_r), \\ \implies x &\sim N(0, \sigma_x^2(I_p + \alpha LL')). \end{aligned}$$

Reparameterizing such that $\alpha \leftarrow \sigma_x^2\alpha$ and integrating out the reparameterized latent variable $\widetilde{z}$ yields the LSPCA model.

4.8.1.2 MLEs of the Nuisance Parameters

In this section we derive the MLE updates of the nuisance parameters. For LRPCA, the nuisance parameters are σ_x^2 and α, while LSPCA adds an additional nuisance parameter σ_y^2. The partial derivatives w.r.t. σ_x^2 and α will be the same for G_{LS} and G_{LR}, implying the MLEs σ_x^2 and α will also be the same for both problems. Therefore, we derive the MLEs for LSPCA only.

Taking the partial derivative of G_{LS} with respect to σ_y^2 yields

$$\frac{\partial G_{\text{LS}}}{\partial \sigma_y^2} = \frac{1}{\sigma_y^2}\left(-\frac{1}{\sigma_y^2}\|Y - XL\beta\|_F^2 + nq\right)$$

which has a single zero at $\sigma_y^2 = \frac{1}{nq}\|Y - XL\beta\|_F^2$. The second partial derivative is positive at this point, making this the unique minimal σ_y^2.

Looking at the partial derivative with respect to α yields

$$\frac{\partial G_{\text{LS}}}{\partial \alpha} = -\frac{1}{(\sigma_x^2+\alpha)^2}\|XL\|_F^2 + \frac{nr}{\sigma_x^2+\alpha}$$

which has a single zero at $\alpha = \frac{1}{nr}\|XL\|_F^2 - \sigma_x^2$. Furthermore, G_{LS} is strictly increasing for $\alpha > \frac{1}{nr}\|XL\|_F^2 - \sigma_x^2$ and strictly decreasing for $\alpha < \frac{1}{nr}\|XL\|_F^2 - \sigma_x^2$, making this critical point a minimizer. Given the nonnegativity constraint on α, note this also implies that if $\frac{1}{nr}\|XL\|_F^2 - \sigma_x^2 < 0$, then the minimizer is $\alpha = 0$.

For the partial derivative with respect to σ_x^2 we have

$$\begin{aligned}\frac{\partial G_{\text{LS}}}{\partial \sigma_x^2} =& -\frac{1}{\sigma_x^4}\|X\|_F^2 + \frac{\alpha(2\sigma_x^2+\alpha)}{\sigma_x^4(\sigma_x^2+\alpha^2)}\|XL\|_F^2 \\ &+ \frac{n(p-r)}{\sigma_x^2} + \frac{nr}{\sigma_x^2+\alpha}.\end{aligned}$$

Evaluating the above at the optimal α, we find that if $\alpha = 0$, $\sigma_x^2 = \frac{1}{np}\|X\|_F^2$, which is exactly what is expected from the model. On the other hand, if $\alpha > 0$ we can note the following. The partial derivative is zero when

$$\begin{aligned}0 =& -(\sigma_x^2+\alpha)^2\|X\|_F^2 + \left(\alpha\sigma_x^2 + \alpha(\sigma_x^2+\alpha)\right)\|XL\|_F^2 \\ &+ (p-r)\sigma_x^2(\sigma_x^2+\alpha)^2 + nr\sigma_x^4(\sigma_x^2+\alpha).\end{aligned}$$

Plugging in $\alpha = \frac{1}{nr}\|XL\|_F^2 - \sigma_x^2$ yields the optimality condition

$$0 = \sigma_x^2\frac{p-r}{nr^2}\|XL\|_F^4 + \frac{1}{n^2r^2}\|XL\|_F^4(\|XL\|_F^2 - \|X\|_F^2),$$

which implies $\sigma_x^2 = \frac{1}{n(p-r)}(\|X\|_F^2 - \|XL\|_F^2)$. In either case G_{LS} is strictly decreasing as σ_x^2 approaches the critical point from the left, and strictly increasing as $\sigma_x^2 \to \infty$ from the right of the critical point, making the corresponding critical points minimizers.

In summary, given L and β, the maximum likelihood estimates of σ_y^2, σ_x^2, and α are

$$\hat{\alpha} = \max\left(\frac{1}{nr}\|XL\|_F^2 - \hat{\sigma}_x^2, 0\right) \tag{4.27}$$

$$\hat{\sigma}_x^2 = \begin{cases} \frac{1}{np}\|X\|_F^2 & \hat{\alpha} = 0 \\ \frac{1}{n(p-r)}\left(\|X\|_F^2 - \|XL\|_F^2\right) & \hat{\alpha} > 0 \end{cases} \tag{4.28}$$

$$\hat{\sigma}_y^2 = \frac{1}{nq}\|Y - XL\beta\|_F^2. \tag{4.29}$$

4.8.2 Kernel Supervised Dimension Reduction

In this section we extend all proposed methods to incorporate kernels. We begin by reviewing kernel PCA.

4.8.2.1 Kernel PCA

A function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is *symmetric* if $k(\boldsymbol{y}, \boldsymbol{z}) = k(\boldsymbol{z}, \boldsymbol{y})$ for all $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^p$. k is said to be *positive definite* if, for any $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$, the $m \times m$ matrix $[k(\boldsymbol{y}_i, \boldsymbol{y}_j)]_{i,j=1}^n$ is positive semi-definite. If k is symmetric and positive definite, it is called a *kernel*. Kernel PCA (kPCA) [SSM97] is a means of performing non-linear unsupervised dimension reduction by performing PCA in a high-dimensional feature space associated k. In particular, associated to k is a high dimensional feature space $\mathcal{F}$ and mapping Φ such that $\Phi : \mathbb{R}^p \to \mathcal{F}$. The kernel matrix associated to k is $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $k(\boldsymbol{y}, \boldsymbol{z}) = \langle \Phi(\boldsymbol{y}), \Phi(\boldsymbol{z}) \rangle_{\mathcal{F}}$ $\forall \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^p$. Let X_Φ be the matrix with n rows where each row is the representation in $\mathcal{F}$ of the corresponding row of X. Kernel PCA (kPCA) finds the projection of X_Φ onto its top r principal components, rather than the principal components themselves. Computing the principal components themselves is usually impractical or intractable as Φ may be unknown and/or $\mathcal{F}$ may be of extremely high dimension. Computationally, all that is required is to find the eigenvectors corresponding to the r largest eigenvalues of the centered kernel matrix $\widetilde{K} = K - \frac{1}{n}\mathbf{1}\mathbf{1}'K - \frac{1}{n}K\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'K\mathbf{1}\mathbf{1}'$. This amounts to solving

$$\begin{aligned} \hat{L} = &\min_L \|\widetilde{K} - \widetilde{K}LL'\|_F^2 \\ &s.t.\ \ L'L = I_r, \end{aligned} \tag{4.30}$$

where now $p = n$, and so L is $n \times r$. Let $\{\boldsymbol{v}_i\}_{i=1}^r$ be the top r principal components in the new feature space $\mathcal{F}$, and let $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r]$. The columns of $\hat{L}$ are such that

$$\boldsymbol{v}_i = \sum_{j=1}^n \hat{L}_{ji} \widetilde{\Phi}(\boldsymbol{x}_j),$$

where $\widetilde{\Phi}(\boldsymbol{x}_j) = \Phi(\boldsymbol{x}_j) - \frac{1}{n}\sum_{j'=1}^n \Phi(\boldsymbol{x}_{j'})$ is the centered representation of $\boldsymbol{x}_j$ in $\mathcal{F}$. To ensure the $\boldsymbol{v}_i$ are unit norm, the columns of $\hat{L}$ must be normalized to obtain $\bar{L}$ such that $\bar{L}'K\bar{L} = I_r$. The projection of a data point $\boldsymbol{x}$ onto the ith component is

$$\langle \boldsymbol{v}_i, \widetilde{\Phi}(\boldsymbol{x}) \rangle_{\mathcal{F}} = \sum_{j=1}^n \bar{L}_{ji} \widetilde{k}(\boldsymbol{x}, \boldsymbol{x}_j),$$

where

$$\begin{aligned}\widetilde{k}(\boldsymbol{y},\boldsymbol{z}) &= \langle\widetilde{\Phi}(\boldsymbol{y}),\widetilde{\Phi}(\boldsymbol{z})\rangle_{\mathcal{F}} \\ &= k(\boldsymbol{y},\boldsymbol{z}) - \frac{1}{n}\sum_{i=1}^{n}\left(k(\boldsymbol{y},\boldsymbol{x}_i)+k(\boldsymbol{x}_i,\boldsymbol{z})\right) \\ &+ \frac{1}{n^2}\sum_{j=1}^{n}\sum_{j'=1}^{n}k(\boldsymbol{x}_j,\boldsymbol{x}_{j'}).\end{aligned}$$

Most importantly for our purposes, the weights of the projection of the training data are given by

$$\Pi_{\{\boldsymbol{v}_i\}_{i=1}^{r}}(X)V = \widetilde{K}\bar{L} \in \mathbb{R}^{n\times r}, \tag{4.31}$$

i.e., $(\widetilde{K}\bar{L})_{ji} = \langle \boldsymbol{v}_i, \widetilde{\Phi}(\boldsymbol{x}_j)\rangle_{\mathcal{F}}$. We refer to using kPCA in procedures analogous to PCR and PCC as kPCR and kPCC, respectively.

4.8.3 Links to Datasets

The datasets used in this work that are directly available online are Ionosphere, Sonar, Colon, Arcene, Residential, and Music.

Ionosphere: https://archive.ics.uci.edu/ml/datasets/Ionosphere

Sonar: https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+%28Sonar%2C+Mines+vs.+Rocks%29

Colon: https://jundongl.github.io/scikit-feature/datasets.html

Arcene: https://jundongl.github.io/scikit-feature/datasets.html

Residential: https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set

Music: https://archive.ics.uci.edu/ml/datasets/Geographical+Original+of+Music

CHAPTER 5

Conclusions and Future Work

This dissertation has developed methods for recovering low-dimensional and latent structure from data, unified by a common aim: improving estimation by exploiting structure inherent in how the data are generated or constrained. This aim was pursued in two complementary settings: 1) the estimation of nonparametric mixture models, and 2) supervised dimension reduction, each developed with provable guarantees and practical algorithms.

Chapters 2 and 3 addressed the recovery of mixture components under minimal distributional assumptions. Chapter 2 introduced a variant of the kernel density estimator that consistently estimates any identifiable nonparametric mixture model from grouped observations, requiring only that the components be square integrable. The analysis rested on an oracle inequality for weighted kernel density estimators and a general principle linking consistent estimation of the group distribution to consistent recovery of the underlying components whenever the mixture is identifiable. For the practically important case of paired observations, an efficient algorithm was given and shown to be effective on tasks ranging from nuclear source detection to nonparametric continuous topic modeling of very short texts. Chapter 3 carried this program forward, extending the grouped-observation framework from the estimation of mixture densities to nonparametric mixtures of regressions, broadening the reach of identifiable nonparametric mixture estimation into a supervised, predictive setting.

Chapter 4 turned to supervised dimension reduction, studying supervised PCA as a single multiobjective problem that explicitly trades empirical risk against variance explained. By recasting the formulation as a joint optimization over the product manifold $\mathrm{St}(p, r) \times \mathbb{R}^{r \times q}$, the chapter established an $O(1/\epsilon^2)$ convergence rate for Riemannian gradient descent to a stationary point, a sufficient condition for local minimality under general smooth supervised losses, and, for the squared error, a simplified characterization of local minima together with a guarantee that the recovered representation retains the leading principal subspace. Empirically, the resulting method matched or exceeded existing SPCA approaches in both prediction error and variance explained, supporting the view that variance explained serves as a useful form of regularization for the supervised task.

Taken together, these chapters share more than an emphasis on theory paired with algorithms. Each begins with a problem that is ill-posed or non-convex in its natural form makes and progress by identifying the right structure to exploit: the grouping of observations in the mixture-model chapters, and the geometry of the Stiefel manifold together with a multiobjective formulation in the dimension-reduction chapter. In every case, identifiability and recovery are the central questions, and in every case the theoretical guarantees are matched to estimators and practical algorithms.

The broader contribution of this work is a set of tools for learning latent or low-dimensional structure from data while relying on as few assumptions as possible. The applications considered here, from particle detection to topic modeling to supervised representation learning, suggest that these tools are useful beyond the specific problems studied. However, a number of questions remain open.

For the nonparametric mixtures of regressions problem, extending the EM algorithm to other classes of function estimators, e.g., neural networks, is straightforward, as discussed in Chapter 3. However, it may also be of interest to expand the scope of the theory developed in Chapter 3 for other function estimators. Furthermore, the question of convergence of the proposed EM algorithm remains open. Known theoretical properties of the EM algorithm in the setting of mixtures of linear regressions [YCS14, KC20, BWY17, KHC21] suggest this topic is worth exploring. For example, a significant first step might involve characterizing a loose radius of convergence for the EM algorithm.

For the supervised PCA problem, experimental results suggest convergence to a global minimizer from a suitable initialization. However, the coupling of the supervised loss and PCA term create nontrivial differences in analysis from vanilla PCA. The fact that the global solution is known when $\lambda = 1$ (PCA) and, for some supervised losses of interest, when $\lambda = 0$, suggests homotopy methods [IWIT22, LYZZ23] and analysis of bifurcations as a promising approach. To this end, see Figure 5.1 which shows the LSPCA loss landscape stereographically projected from $\mathrm{St}(3,1)$ to the unit disc in $\mathbb{R}^2$. The landscape is shown across a range of λ values for a synthetic regression problem where the true optimal regression subspace is given by $0.25v_1 + 0.75v_2$ (a weighted average of the first two principal directions of X). From this figure we see empirically that the basin of attraction of the global minimizer appears to continuously deform between the solutions corresponding to $\lambda = 1$ and $\lambda = 0$, supporting the claim that continuation methods could recover the global solution for any value of λ. We note that certain degenerate cases may exist, e.g., if the optimal subspace for $\lambda = 0$ is closely aligned with the last principal direction of X. Furthermore, the long history of streaming PCA [Oja82, BCL18] and the successful application of manifold optimization to streaming PCA [BNR10, Bal22] suggests that an extension of the approach developed in Chapter 4 to the streaming setting could be both useful and feasible.

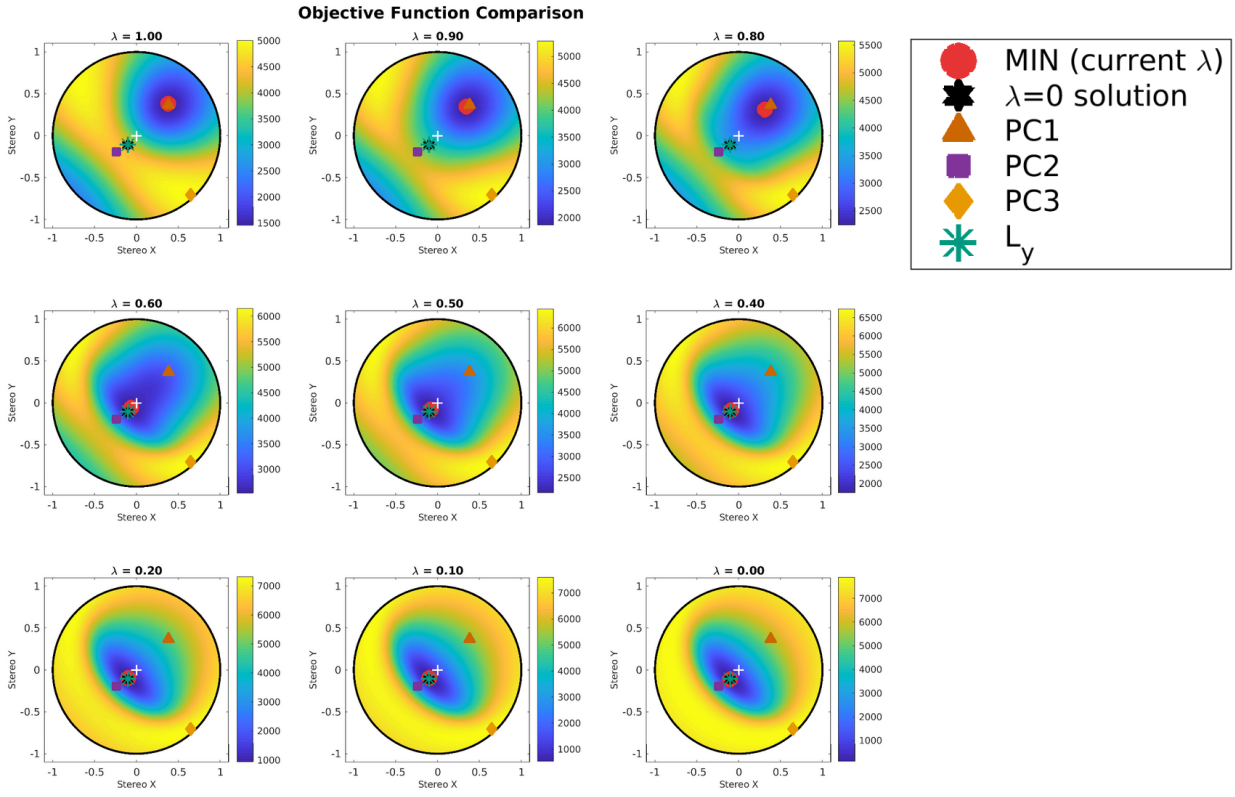


Figure 5.1: Stereographic projection of the LSPCA loss landscape for a planted regression model with $r = q = 1$, where the optimal regression subspace is $L_y = 0.25v_1 + 0.75v_2$ and v_i is the ith principal direction of the data matrix X.

Bibliography

[A+51] Theodore Wilbur Anderson et al. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951.

[ABG+14] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. In *Conference on Learning Theory*, pages 1135–1164, 2014.

[ADRX18] Bryon Aragam, Chen Dan, Pradeep Ravikumar, and Eric P Xing. Identifiability of nonparametric mixture models and Bayes optimal clustering. *arXiv preprint arXiv:1802.04397*, 2018.

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[AGKM12] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization – provably. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 145–162, New York, NY, USA, 2012. ACM.

[AMR09] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 12 2009.

[AMS09] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[Bal22] Laura Balzano. On the equivalence of oja's algorithm and grouse. In *International Conference on Artificial Intelligence and Statistics*, pages 7014–7030. PMLR, 2022.

[Ban32] Stefan Banach. *Théorie des opérations linéaires*. 1932.

[BCG+04] RK Bock, A Chilingarian, M Gaug, F Hakl, Th Hengstebeck, M Jiřina, J Klaschka, E Kotrč, P Savickỳ, and S Towers. Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 516(2-3):511–528, 2004.

[BCL18] Laura Balzano, Yuejie Chi, and Yue M Lu. Streaming pca and subspace tracking: The missing data case. *Proceedings of the IEEE*, 106(8):1293–1310, 2018.

[BDF13] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental pca. *Advances in neural information processing systems*, 26, 2013.

[BDW08] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained clustering: Advances in algorithms, theory, and applications*, chapter 3. CRC Press, 2008.

[BGAJ11] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011.

[BHPT06] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

[BNR10] Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *2010 48th Annual allerton conference on communication, control, and computing (Allerton)*, pages 704–711. IEEE, 2010.

[BNS18] Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *International Conference on Machine Learning*, pages 452–461, 2018.

[Bou23] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[BS14] Gilles Blanchard and Clayton Scott. Decontamination of mutually contaminated models. In *Artificial Intelligence and Statistics*, pages 1–9, 2014.

[BVB16] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *Advances in Neural Information Processing Systems*, 29, 2016.

[BVB20] Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. Deterministic guarantees for burer-monteiro factorizations of smooth semidefinite programs. *Communications on Pure and Applied Mathematics*, 73(3):581–608, 2020.

[BWY17] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[BYTM24] William Briguglio, Waleed A. Yousef, Issa Traoré, and Mohammad Mamun. Federated supervised principal component analysis. *IEEE Transactions on Information Forensics and Security*, 19:646–660, 2024.

[Coh84] Elizabeth A Cohen. Some effects of inharmonic partials on interval perception. *Music Perception*, 1(3):323–349, 1984.

[CWH+10] Xi Chen, Lily Wang, Bo Hu, Mingsheng Guo, John Barnard, and Xiaofeng Zhu. Pathway-based analysis for genome-wide association studies using supervised principal components. *Genetic epidemiology*, 34(7):716–724, 2010.

[CWSZ08] Xi Chen, Lily Wang, Jonathan D Smith, and Bing Zhang. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics*, 24(21):2474–2481, 2008.

[DAZ+13] Orla M Doyle, John Ashburner, FO Zelaya, Stephen CR Williams, Mitul A Mehta, and Andre F Marquand. Multivariate decoding of brain images using ordinal regression. *NeuroImage*, 81:347–357, 2013.

[DLB+17] Ben W Dulken, Dena S Leeman, Stéphane C Boutet, Katja Hebestreit, and Anne Brunet. Single-cell transcriptomic analysis defines heterogeneity and transcriptional dynamics in the adult neural stem cell lineage. *Cell reports*, 18(3):777–790, 2017.

[DV89] Richard D De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, 2001.

[Fol99] Gerald B. Folland. *Real analysis: modern techniques and their applications*. Pure and applied mathematics. Wiley, 1999.

[For65] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

[GCFL20] Pierre Gançarski, Bruno Crémilleux, Germain Forestier, and Thomas Lampert. Constrained clustering: Current and new trends. A Guided Tour of AI Research, In press. hal-02548212, 2020.

[GK24] Tomojit Ghosh and Michael Kirby. Linear centroid encoder for supervised principal component analysis. *Pattern Recognit.*, 155:110634, 2024.

[GM90] L. Gyorfi and E. Masry. The L_1 and L_2 strong consistency of recursive kernel density estimation from dependent samples. *IEEE Transactions on Information Theory*, 36(3):531–539, 1990.

[GPW+19] Wang Gao, Min Peng, Hua Wang, Yanchun Zhang, Qianqian Xie, and Gang Tian. Incorporating word embeddings into topic modeling of short text. *Knowledge and Information Systems*, 61(2):1123–1145, 2019.

[HA85] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[HLW13] Mian Huang, Runze Li, and Shaoli Wang. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108(503):929–941, 2013.

[Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[IWIT22] Hidenori Iwakiri, Yuhang Wang, Shinji Ito, and Akiko Takeda. Single loop gaussian homotopy method for non-convex optimization. *Advances in Neural Information Processing Systems*, 35:7065–7076, 2022.

[Ize75] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[JJ94] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[JJK+16] Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Conference on learning theory*, pages 1147–1164. PMLR, 2016.

[JM00] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[Jol82] Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982.

[Kaw20] Shuichi Kawano. Sparse principal component regression via singular value decomposition approach. arXiv preprint, February 2020. https://arxiv.org/abs/2002.09188.

[KC20] Jeongyeol Kwon and Constantine Caramanis. Em converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pages 1727–1736. PMLR, 2020.

[KFTS15] Shuichi Kawano, Hironori Fujisawa, Toyoyuki Takada, and Toshihiko Shiroishi. Sparse principal component regression with adaptive loading. *Computational Statistics & Data Analysis*, 89:192–203, 2015.

[KFTS18] Shuichi Kawano, Hironori Fujisawa, Toyoyuki Takada, and Toshihiko Shiroishi. Sparse principal component regression for generalized linear models. *Computational Statistics & Data Analysis*, 124:180–196, 2018.

[KHC21] Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2021.

[KR83] R.V. Kadison and J.R. Ringrose. *Fundamentals of the theory of operator algebras. V1: Elementary theory*. Pure and Applied Mathematics. Elsevier Science, 1983.

[KS12] JooSeuk Kim and Clayton D Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13(Sep):2529–2565, 2012.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lin95] Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

[Liu20] Zhenqiu Liu. Visualizing single-cell rna-seq data with semisupervised principal component analysis. *International journal of molecular sciences*, 21(16):5797, 2020.

[LLA+19] Xingguo Li, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, Zhaoran Wang, and Tuo Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019.

[Llo82] Stuart Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[LLY23] Joowon Lee, Hanbaek Lyu, and Weixin Yao. Exponentially convergent algorithms for supervised matrix factorization. volume 36, pages 76947–76959, 2023.

[LLY24] Joowon Lee, Hanbaek Lyu, and Weixin Yao. Supervised matrix factorization: Local landscape analysis and applications. In *Forty-first International Conference on Machine Learning*, 2024.

[Lux07] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[LW20] Darren L Linvill and Patrick L Warren. Troll factories: Manufacturing specialized disinformation on twitter. *Political Communication*, pages 1–21, 2020.

[LWZ+16] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174, 2016.

[LYNS16] Gen Li, Dan Yang, Andrew B Nobel, and Haipeng Shen. Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis*, 146:7–17, 2016.

[LYZZ23] Xi Lin, Zhiyuan Yang, Xiaoyuan Zhang, and Qingfu Zhang. Continuation path learning for homotopy optimization. In *International Conference on Machine Learning*, pages 21288–21311. PMLR, 2023.

[Mau16] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 3–17. Springer International Publishing, 2016.

[McD89] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[MJJ10] Pavan Kumar Mallapragada, Rong Jin, and Anil Jain. Non-parametric mixture models for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 334–343. Springer, 2010.

[MK01] Aleix M Martínez and Avinash C Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (2):228–233, 2001.

[NBDJ15] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313, 2015.

[NLGB10] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

[Oja82] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[OWHZ24] Wenbo Ouyang, Ruiyang Wu, Ning Hao, and Hao Helen Zhang. Dynamic supervised principal component analysis for classification. *Journal of Computational and Graphical Statistics*, 2024.

[P+11] Jonathan D Power et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

[Pea01] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[Poz14] Constantine Pozrikidis. *An introduction to grids, graphs, and networks*. Oxford University Press, 2014.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[PV18] Juho Piironen and Aki Vehtari. Iterative supervised principal components. In *International Conference on Artificial Intelligence and Statistics*, pages 106–114, 2018.

[PY25] Theodosios Papazoglou and Guosheng Yin. Covariance supervised principal component analysis. 2025.

[QCWW17] Jipeng Qiang, Ping Chen, Tong Wang, and Xindong Wu. Topic modeling over short texts by incorporating word embeddings. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 363–374. Springer, 2017.

[QZL+19] Jipeng Qiang, Qian Zhenyu, Yun Li, Yunhao Yuan, and Xindong Wu. Short text topic modeling techniques, applications, and performance: A survey. *arXiv preprint arXiv:1904.07695*, 2019.

[RBK+20] Alexander Ritchie, Laura Balzano, Daniel Kessler, Chandra S Sripada, and Clayton Scott. Supervised pca: A multiobjective approach. *arXiv preprint arXiv:2011.05309*, 2020.

[Rei89] R.D. Reiss. *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer series in statistics. Springer, 1989.

[RM06] Steven Roberts and Michael A Martin. Using supervised principal components analysis to assess multiple pollutant effects. *Environmental health perspectives*, 114(12):1877–1882, 2006.

[RSB+19] Alexander Ritchie, Clayton Scott, Laura Balzano, Daniel Kessler, and Chandra S Sripada. Supervised principal component analysis via manifold optimization. In *Proceedings of 2019 IEEE Data Science Workshop (DSW)*, 2019.

[RSS14] Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, pages 207–224, New York, NY, USA, 2014. ACM.

[RVS20] Alexander Ritchie, Robert A Vandermeulen, and Clayton Scott. Consistent estimation of identifiable nonparametric mixture models from grouped observations. *Advances in Neural Information Processing Systems*, 33:11676–11686, 2020.

[SADX14] Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric estimation of multi-view latent variable models. In *International Conference on Machine Learning*, pages 640–648, 2014.

[SAR+19] Chandra Sripada, Mike Angstadt, Saige Rutherford, Daniel Kessler, Yura Kim, Mike Yee, and Elizaveta Levina. Basic units of inter-individual variation in resting state connectomes. *Scientific reports*, 9(1):1900, 2019.

[Sco15] David W Scott. *Multivariate density estimation: theory, practice, and visualization*, page 164. John Wiley & Sons, 2015.

[SGCE17] Sara Sharifzadeh, Ali Ghodsi, Line H Clemmensen, and Bjarne K Ersbøll. Sparse supervised principal component analysis (sspca) for dimension reduction and variable selection. *Engineering Applications of Artificial Intelligence*, 65:168–177, 2017.

[SSM97] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[ST10] Jerry Scripps and Pang-Ning Tan. Constrained overlapping clusters: minimizing the negative effects of bridge-nodes. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(1):20–37, 2010.

[Sug07] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of machine learning research*, 8(May):1027–1061, 2007.

[SWW+20] Xin Yi See, Xuelan Wen, T Alexander Wheeler, Channing K Klein, Jason D Goodpaster, Benjamin R Reiner, and Ian A Tonks. Iterative supervised principal component analysis driven ligand design for regioselective ti-catalyzed pyrrole synthesis. *ACS Catalysis*, 10(22):13504–13517, 2020.

[SWY15] Geoffrey Schiebinger, Martin J Wainwright, and Bin Yu. The geometry of kernelized spectral clustering. *The Annals of Statistics*, 43(2):819–846, 2015.

[SZ20] Clayton Scott and Jianxin Zhang. Learning from label proportions: A mutual contamination framework. *Advances in neural information processing systems*, 33:22256–22267, 2020.

[TB99] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[VBT+21] Joshua T Vogelstein, Eric W Bridgeford, Minh Tang, Da Zheng, Christopher Douville, Randal Burns, and Mauro Maggioni. Supervised dimensionality reduction for big data. *Nature communications*, 12(1):1–9, 2021.

[VdG08] Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[VR13] Raja Velu and Gregory C Reinsel. *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media, 2013.

[VS19] Robert A. Vandermeulen and Clayton D. Scott. An operator theoretic approach to nonparametric mixture models. *The Annals of Statistics*, 47(5):2704–2733, October 2019.

[WBD06] Kiri L Wagstaff, Sugato Basu, and Ian Davidson. When is constrained clustering beneficial, and why? *Ionosphere*, 58(60.1):62–63, 2006.

[WC00] Kiri Wagstaff and Claire Cardie. Clustering with instance-level constraints. *AAAI/IAAI*, 1097:577–584, 2000.

[WCRS01] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pages 577–584, 2001.

[WQD14] Xiang Wang, Buyue Qian, and Ian Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28(1):1–30, 2014.

[XLZ19] Junyu Xuan, Jie Lu, and Guangquan Zhang. A survey on Bayesian nonparametric learning. *ACM Computing Surveys (CSUR)*, 52(1):1–36, 2019.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[XVC+21] Shaojie Xu, Joel Vaughan, Jie Chen, Agus Sudjianto, and Vijayan Nair. Supervised linear dimension-reduction methods: Review, extensions, and comparisons. *arXiv preprint arXiv:2109.04244*, 2021.

[XY18] Sijia Xiang and Weixin Yao. Semiparametric mixtures of nonparametric regressions. *Annals of the Institute of Statistical Mathematics*, 70(1):131–154, 2018.

[YCS14] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621. PMLR, 2014.

[YH03] Thomas W Yee and Trevor J Hastie. Reduced-rank vector generalized linear models. *Statistical modelling*, 3(1):15–41, 2003.

[YS68] Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 209–214, 1968.

[YYLT05] Zhang Yi, Mao Ye, Jian Cheng Lv, and Kok Kiong Tan. Convergence analysis of a deterministic discrete time system of oja's pca learning algorithm. *IEEE Transactions on Neural Networks*, 16(6):1318–1328, 2005.

[YYT+06] Shipeng Yu, Kai Yu, Volker Tresp, Hans-Peter Kriegel, and Mingrui Wu. Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 464–473. ACM, 2006.

[ZW19] Chaowen Zheng and Yichao Wu. Nonparametric estimation of multivariate mixtures. *Journal of the American Statistical Association*, pages 1–16, 2019.