Local Mixture Models in Hilbert Space

by
Zhiyue Huang

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2010

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, we study local mixture models with a Hilbert space structure. First, we consider the fibre bundle structure of local mixture models in a Hilbert space. Next, the spectral decomposition is introduced in order to construct local mixture models. We analyze the approximation error asymptotically in the Hilbert space. After that, we will discuss the convexity structure of local mixture models. There are two forms of convexity conditions to consider, first due to positivity in the -1-affine structure and the second by points having to lie inside the convex hull of a parametric family. It is shown that the set of mixture densities is located inside the intersection of the sets defined by these two convexities. Finally, we discuss the impact of the approximation error in the Hilbert space when the domain of mixing variable changes.

Acknowledgements

First, I thank my supervisor Professor Paul Marriott, for his support in the Master program. He is always there to listen and to give advice. He teaches me how to think questions carefully and in detail. I always feel lucky to have him as my supervisor. I also want to say 'thank you' to my second readers, Professor Don McLeish and Professor Christopher Small for their insightful comments and editing my thesis. Let me thank Jiheng Wang and Zhenhao Li in University of Waterloo, for the friendship and support. Last, but not least, I want thank my parents, Shouyi Huang, and Guyin Lou, for giving me life and educating me.

Contents

List of Figures

Chapter 1

Introduction

Suppose that a random variable X takes values in a sample space $\mathcal{S}$ and that its distribution can be represented by a probability density function of the form

$$f(x|\theta, Q) = \int f(x|\eta, \theta) dQ(\eta), \tag{1.1}$$

where we call η the *latent random variables* from the probability measure Q. Such distribution Q is called *mixing distribution* or *latent distribution*. Mixture distributions are applied widely, for example capture-recapture models ([Cha87], [XM07] and [BDKS05]), measurement error model ([Lin95] and [Mar03]), cluster analysis ([HT96], [FR02] and [TJP04]).

In this thesis, we introduce a Hilbert space structure to the mixture distributions, discuss the fibre bundle and convexity structure of local mixture models in the Hilbert space, and show the effect on inference on the parameter θ if the domain of η changes.

Let us get a first taste of mixture distributions by considering measurement error models.

Example 1. *(Measurement Error Models [Lin95] (Page 14)) Consider a simple linear regression model*

$$Y = \alpha + \beta X^* + \epsilon,$$

which obeys the usual Gauss-Markov conditions, with $\epsilon \sim N(0, \sigma^2)$*. However, suppose that* X^* *is measured with error, so that only*

$$X = X^* + \eta$$

is observed, where η *considered a latent variable with distribution* Q*. Furthermore,* η *is assumed independent of both* X^* *and* ϵ *with mean zero. The density of observed variables is*

$$f(x|\alpha, \beta, \sigma^2, Q) = \int f(x|\eta, \alpha, \beta, \sigma^2) dQ(\eta)$$

Mixture distributions, in this case, allow inference on the regression parameters β in the case where there is measurement error. Local mixture models, which are introduced later, further allow such inference when all that is known about the mixing distribution Q is that it has a relatively small variance. Our task is to make inference about the two population variables Y and X using measurement error models. There are two different statistical inferential problems in this example. One is to fit the regression model, i.e. to estimate the α and β in the model, here comes the first inferential problem of mixture distributions, *parameter estimation*. The other is the *estimation of the mixing distribution* Q. Different geometrical structures of mixture distribution have been discussed for each purpose.

1.1 Statistics Inference and Geometrical Structure of Mixture Distributions

1.1.1 Parameter Estimation

For parameter estimation, the geometrical structures such as *jet space* (See Appendix A.1.2) and *fibre bundle* (See Appendix A.1.2) in an affine space (See Appendix A.1.2) are chosen by Marriott ([Mar02] and [Mar07b]) and Anaya-Izquirerdo andMarriott [AIM07]. In 2002, Marriott [Mar02] defines

the sets X_{Mix} and V_{Mix} for a support set $\mathcal{S}$, by

$$X_{Mix} = \left\{ f(x) | f \in C^{\infty}(S, \mathbb{R}), f \in L^2(v), \int f(x)dv = 1 \right\};$$
$$V_{Mix} = \left\{ g(x) | g \in C^{\infty}(\mathcal{S}, \mathbb{R}), g \in L^2(v), \int g(x)dv = 0 \right\}.$$

It is shown that $(V_{Mix}, +)$ is a vector space and $(X_{Mix}, V_{Mix}, +)$ is an *affine space* (See Appendix A.1.1) with the natural addition operation $+$.

In [Mar02], we consider the vector space

$$T^2 M_{\theta_0} := span \left\{ \frac{\partial}{\partial \theta} f(x|\theta_0), \frac{\partial^2}{\partial \theta^2} f(x|\theta_0) \right\} \subset V_{Mix}$$

and attach it to a point $f(x|\theta_0) \in X_{Mix}$. Such structure is described by jet space. For a higher order jet space,

$$T^K M_{\theta_0} := span \left\{ \frac{\partial}{\partial \theta} f(x|\theta_0), \frac{\partial^2}{\partial \theta^2} f(x|\theta_0), \cdots, \frac{\partial^K}{\partial \theta^K} f(x|\theta_0) \right\}, \quad K = 1, 2, \cdots$$

the local mixture model is defined

$$f(x|\theta_0, \alpha) = f(x|\theta_0) + \sum_{i=1}^{K} \alpha_i \frac{\partial^i}{\partial \theta^i} f(x|\theta_0). \tag{1.2}$$

The local mixture model can be viewed as a Laplace expansion of the mixture density function $f(x|\theta, Q)$, see for example [Mar02] and [AI06]. Other than the equation above, we also need two boundaries for approximation, defined as follows.

Definition 1. *The hard boundary is defined by the condition that* $f(x|\theta, \alpha) \geq 0$ *for all* $x \in \mathcal{S}$.

Definition 2. *The soft boundary is defined by* $f(x|\theta, \alpha)$ *lying in convex hull of curve* $f(x|\theta)$ *in the mixture affine geometry.*

The hard boundary offers us the positivity condition which ensures that the approximation of mixture distribution is a density function, while the

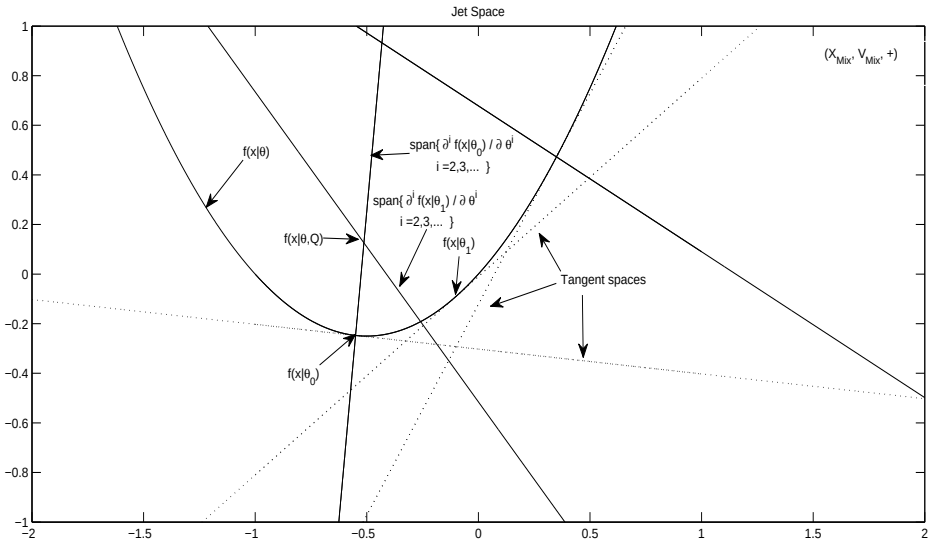

Figure 1.1: Jet Space Strucutre of the Mixture Distributions

soft boundary ensures that the resultant approximation can be realized by an exact mixture model, mentioned in [AIM07] and [Mar07b]. Then, we use such model to approximate the real distribution $f(x|\theta, Q)$.

The definition of a local mixture model as given on the previous page gives rise to a potential identification problem. There could be different approaches to approximate the mixture density functions, such as

$$f(x|\theta_0, \alpha) = f(x|\theta_0) + \sum_{i=1}^{K} \alpha_i \frac{\partial^i}{\partial \theta^i} f(x|\theta_0)$$

or

$$f(x|\theta_1, \alpha') = f(x|\theta_1) + \sum_{i=1}^{K} \alpha'_i \frac{\partial^i}{\partial \theta^i} f(x|\theta_1),$$

where $\theta_1 \neq \theta_0$. Both of the two models can be used to approximate $f(x|\theta, Q)$. Geometrically, the space of $f(x|\theta, Q)$ can be represented by a vector space

$span\left\{\frac{\partial^i}{\partial\theta^i}f(x|\theta_0), i=1,2,\cdots\right\}$ attached at point $f(x|\theta_0)$ or $span\left\{\frac{\partial^i}{\partial\theta^i}f(x|\theta_1), i=1,2,\cdots\right\}$ at point $f(x|\theta_1)$ as shown in Figure 1.1. The identification problem is solved later in the work of Anaya-Izquierdo and Marriott in [AIM07]. If $f(x|\theta,\eta)$ is in a regular exponential family, then we may set θ equal the mean parameter and with loss of generality, assume that the first moment of the mixing distribution is zero. The local mixture model is given by,

$$f(x|\theta,\alpha) = f(x|\theta) + \sum_{i=2}^{K} \alpha_i \frac{\partial^i}{\partial\theta^i} f(x|\theta), \tag{1.3}$$

where α_i are coefficients. We can see the main difference between local mixture model 1.2 and 1.3 is the drop of $\alpha_1 \partial f(x|\theta)/\partial\theta$. It is shown in [AI06] and [AIM07] that there is no loss in generality when interpreting local mixtures in terms of asymptotic expansion of mixture models.

In the later works of Marriott [Mar07b], a definition of the global extension of the local mixture model can be given as follows.

Definition 3. *The global extension of the local mixture model of a regular family $f(x|\theta) \in X_{Mix}$ is defined as*

$$f(x|\theta,\alpha) = f(x|\theta) + \sum_{i=1}^{K} \alpha_i g_i(x|\theta), \tag{1.4}$$

where α_i are coefficients and $g_i \in V_{Mix}$.

$f(x|\theta,\alpha)$ is an element in X_{Mix}, because

$$\int_{\mathcal{S}} f(x|\theta,\alpha)dx = \int_{\mathcal{S}} f(x|\theta) + \sum_{i=1}^{K} \alpha_i g_i(x|\theta)dx = 1.$$

Hence, the local mixture model has a structure in the affine space $(X_{Mix}, V_{Mix}, +)$. The fibre bundle (Appendix A.1.2) in the affine space $(X_{Mix}, V_{Mix}, +)$ can describe the geometrical structure of local mixture model. We consider a larger parametric family of density functions $f(x|\theta,\eta)$ instead of $f(x|\theta)$. Such structure satisfies the following

1. $f(x|\theta, 0)$ equals $f(x|\theta)$ for all θ.

2. For each fixed θ_0, the family $f(x|\theta, \eta)$ is Fisher orthogonal (See Appendix A.2) to $f(x|\theta)$ at θ_0.

3. For each fixed θ_0, the family $f(x|\theta, \eta)$ has zero -1-curvature with respect to η (See Appendix A.1.2) either at $(\theta_0, 0)$ or over all the support of η.

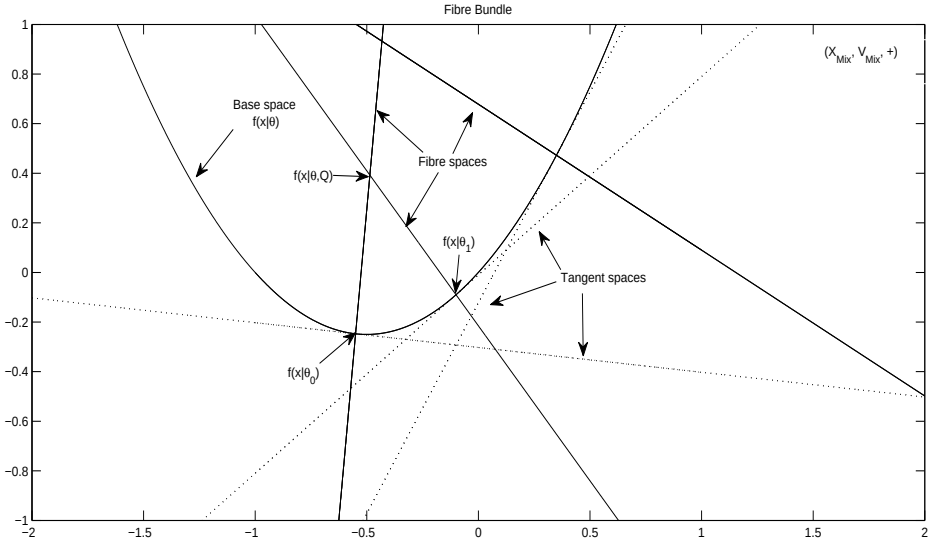

Figure 1.2: Fibre Bundle Structure of the Mixture Distributions

As shown in Figure 1.2, $\theta \in \mathcal{C}$ in the affine space $(X_{Mix}, V_{Mix}, +)$ forms the base space of a fibre bundle, while the fibre space spanned by $\{g_i(x|\theta), i = 1, 2, \cdots\}$ which is Fisher orthogonal to $f(x|\theta)$ at θ_0. To understand the Fisher orthogonal condition, we should go back to [Ama85] (Page 253) and [ABNK$^+$87] (Page 59). Amari suggests how to construct a Fisher orthogonal parametrization for the models. In other words, the coordinates of parameters are Fisher orthogonal to each other. In the fibre bundle of mixture

models, the base space is the curve $f(x|\theta)$, related to θ, the parameters that we are interested in, while the fibre space spanned by $\{g_i, i = 1, \cdots, n\}$ are the space of nuisance parameters. Note here, with the positivity condition, the vector space we use for local mixture models is a convex subspace of the fibre at point $f(x|\theta_0)$ by [Mar07b]. In Amari's construction, the parameter of interest is orthogonal to the nuisance parameters. Amari also shows in [Ama85] that the dimension of the fibre of the normal bundle can grow with the sample size without losing the efficiency of inference for the parameters of interest. Analogous results are discussed in [CR87] by Cox and Reid and discussion thereto. Again, the two boundaries are necessary here to keep the positivity condition and dealing with exact mixture distributions.

1.1.2 Estimation of Mixing Distribution

The other main statistical inferential problem relating to mixture distribution was the estimation the mixing distribution. It was first raised by Robbins [Rob56] in 1956. Since we do not observe $\eta_1, \cdots, \eta_N$, but $X_1, X_2, \cdots, X_N$, the estimation of mixing distribution are based on the value of X_i. For example in measurement error models, we want to estimate the distribution of Q. Different approaches are proposed for the problem, including Bayesian estimators ([Rob64], [Rol68] and [Mee72]) and maximum likelihood estimators ([Lai78], [Lin81] and [Ler92]). For estimating mixing distributions, Lindsay ([Lin83] and [Lin95]) uses *convex* (See Appendix A.1.2) geometry to study statistics in a finite dimensional space and Wood [Woo99] applies the *cyclic polytope* (See Appendix A.1.2) structure, especially for mixture of binomial distributions.

In [Lin83] and [Lin95], for the i.i.d observations $X_1, X_2, \cdots, X_N$, the likelihood function is defined as the function from the parameter space $\mathcal{C}$ to $\mathbb{R}$. In the parameter space $\mathcal{C}$, also an affine space defined in Appendix A.1.1, the image of the curve, $\{f(x|\theta), \theta \in \mathcal{C}\}$, is the set of all possible fitted values of the likelihood vector. Then, the mixture model $f(x|\theta, Q)$ is located inside the convex hull of this curve. Therefore, it can be written in the form

of convex combination of the elements in the image $\{f(x|\theta), \theta \in \mathcal{C}\}$. With the above fact and the results in convex geometry, Lindsay shows that the loglikelihood $\ell(Q) = \sum_{i=1}^{N} \log f(x|Q)$ has a unique maximum over the space of all distribution functions Q. Furthermore, the maximizer $\hat{Q}$ is a discrete distribution with no more than D distinct points of support, where D is the number of distinct points in $(x_1, x_2, \cdots, x_n)$.

In 1999, Wood [Woo99] builds on earlier work of Lindsay and elucidates the geometrical structure of the following question: given a mixture of binomial distributions, how do we estimate the unknown mixing distribution Q? For binomial distribution $Bin(n, p)$, let n be fixed. As p changes from 0 to 1, we obtain a binomial curve B_n in the simplex,

$$T = \left\{ x = (q_0, \cdots, q_n); \sum_i q_i = 1, q_i \geq 0, \forall i \right\}$$

where q_i stands the probability that random variable is equal to i. Wood used the fact that the convex set of mixtures of binomial distribution is affinely isomorphic to the cyclic polytopes. He uses a smoothing estimator $\hat{Q}$ to produce a 'nearest point' estimator $\hat{Q}_k$ in the sense of Kullback-Leibler distance on a face of the convex hull of $B_n(k)$, the k-segment piecewise linear approximation to binomial curve B_n. The estimator $\hat{Q}_k$ has a unique realization as a convex combination of the set of vertices $\{q_0, q_2, \cdots, q_n\}$ of the face. Such vertices and their weights offer us an estimator of the mixing distribution, i.e.

$$\hat{Q} = \sum_{i=0}^{n} p_i q_i.$$

where for all $i \in [1, n]$,

$$\sum_{i=0}^{n} p_i = 1, \quad p_i \geq 0.$$

1.2 Purpose and Outline

The thesis will follow Marriott's work more than the approach of Lindsay. The focus is on estimating parameter of interest θ in a large dimensional family of mixture distributions

$$f(x|\theta, Q) = \int_{\mathcal{Q}} f(x|\theta, \eta) dP(Q).$$

We have an infinite dimensional nuisance parameter $Q \in \mathcal{Q}$. [Mar07b] gives an infinite dimensional nuisance parameter problem in the Bayesian view. Given a set of observations, $\{x_1, x_2, \cdots, x_N\}$, we calculate the marginal posterior of parameter θ as

$$\int_{\mathcal{Q}} \left\{ \prod_{i=1}^{N} \int f(x_i|\eta, \theta) dQ(\eta) \right\} dP(Q),$$

for the prior measure dP on $\mathcal{Q}$ which will be some subset of the space of distribution. This results in the problem that we need to integrate over an infinite dimensional parameter $Q \in \mathcal{Q}$. The (global extension) local mixture model avoids such infinite dimensional nuisance parameters. By (global extension) local mixture model, we can approximate the marginal posterior by

$$\int \left\{ \prod_{i=1}^{N} \int f(x_i|\theta, \alpha) \right\} dP(\alpha),$$

where α is a vector of $\alpha_1, \alpha_2, \cdots, \alpha_K$. Rather than considering in the infinitely (high) dimensional space, we can consider a model in the finite (low) dimensional space. Furthermore, it is shown that there is little changes in the inference of θ in [Mar02], [AI06], [AIM07], [Mar07b], [AICMV09] and this thesis.

Different techniques for low dimensional reduction has been used in mixture models, such as Laplace asymptotic expansion ([Mar02], [Mar03], [AI06] and [AIM07]) and Principle Component Analysis ([MV04] and [Mar07b]). All

previous work are considering the affine space $(X_{Mix}, V_{Mix}, +)$. The purpose of this thesis is to consider the geometrical structure of the local mixture model with a Hilbert structure.

In Chapter 2, we will discuss the spectral decomposition of mixture density functions. Then, we will talk about the fibre bundle structure and convexity structure of mixture models in the Hilbert space in Definition 4. Then, we discuss the effect of the approximation error with the choice of compact region $\mathcal{C}$ of the latent random variables η. The part is mainly based on the previous work of Marriott ([Mar02], [MV04] and [Mar07b]) and K. A. Anaya-Izquierdo [AIM07].

We conclude in Chapter 3 with a final discussion and directions for future research.

Chapter 2

Linear Structure in Hilbert Space

In this chapter, the regularity conditions are given firstly. Then, We consider the space of distributions in the framework introduced in [AICMV09] and describe the fibre bundle structure introduced by Marriott [Mar07b] in a Hilbert space.

Definition 4. *For a density function $f(x)$ define a Hilbert space with -1-affine geometry by*

$$\mathcal{H}(f(x)) = \left\{ g(x) | \int f(x)^{-1} g(x)^2 dx < \infty \right\}$$

with the inner product

$$\langle g_i, g_j \rangle_{\mathcal{H}(f)} = \int_{\mathcal{S}} f^{-1} g_i g_j dx < \infty, \tag{2.1}$$

and the corresponding norm $\| \cdot \|_{\mathcal{H}(f)}$. The orthogonal condition in $\mathcal{H}(f(x))$,

$$\int_{\mathcal{S}} f^{-1} g_i g_j dx = 0, \quad i \neq j$$

also indicates that g_i and g_j are Fisher orthogonal (See Appendix A.2).

For +1-affine geometry, the inner product of the Hilbert space can be defined as

$$\langle g_i, g_j \rangle_{\mathcal{H}(f)} = \int_{\mathcal{S}} f g_i g_j dx < \infty,$$

Next, we decompose the mixture density functions in the Hilbert space. The convexity structure of the mixture distributions and an approximation error by the extending local mixture model 1.4 are also discussed. In [AIM07], Anaya-Izquierdo and Marriott explore the phenomena that the spectrum of the eigenvalues change with the change of the domain of the mixing parameter η. We will give the proof of the phenomena in Appendix A.3. The part is mainly based on the previous work of Marriott ([Mar02], [MV04] and [Mar07b]) and Anaya-Izquirerdo and Marriott [AIM07].

Before we start, the regularity conditions are given.

2.1 Regularity Conditions

Consider the mixture density function as follows,

$$f(x|\theta, Q) = \int_{\mathcal{C}} f(x|\theta, \eta) q(\eta) d\eta$$

where $f(x|\theta)$ is a family of probability density or mass functions and the mixing distribution is given by $q(\eta)$ with mixing variable η. Denote $\mathcal{S}$ be the sample space of x, $\mathcal{C}$ be the domain of η and Ω be the parameter space of θ. For simplification, we will restrict to models where mixing is of the form

$$f(x|\theta, Q) = \int_{\mathcal{C}} f(x|\theta + \eta) q(\eta) d\eta.$$

We give the regularity condition for $f(x|\theta, \eta)$ as follows.

Regularity Condition 1.

$f(x|\theta, \eta)$ is continuously differentiable with respect to η over a compact set $\mathcal{C}$.

Regularity Condition 2.

$f(x|\theta, Q)$ and $f(x|\theta)$ have common support $\mathcal{S}$ in x. For any $x \in \mathcal{S}$, $f(x|\theta, Q)$ is strictly positive.

Regularity Condition 3.

If $f(x|\theta, Q)$ is probability density function of x, then $f(x|\theta, \eta)$ and $\partial f(x|\theta, \eta)/\partial\theta$ are continuous over the region $\mathcal{S} \times \mathcal{C} \times \Omega$. By Leibniz's rule for differentiation under the integral sign (Appendix A.1.1), we have

$$\int_{\mathcal{S}} \frac{\partial}{\partial\theta} f(x|\theta, \eta) dx = \frac{\partial}{\partial\theta} \int_{\mathcal{S}} f(x|\theta, \eta) dx.$$

While $f(x|\theta, Q)$ is probability mass function of x, we need

$$\sum_{x_i \in \mathcal{S}} \frac{\partial}{\partial\theta} f(x_i|\theta, \eta) = \frac{\partial}{\partial\theta} \sum_{x_i \in \mathcal{S}} f(x_i|\theta, \eta).$$

Regularity Condition 4.

We assume

$$\int_{\mathcal{S}\times\mathcal{C}} |\mathcal{F}(x|\theta, \eta)| \, d(x, \eta) < \infty.$$

where $\mathcal{F}(x|\theta, \eta)$ is defined as Equation 2.7. By Fubini's theorem (Appendix A.1.1),

$$\int_{\mathcal{S}} \int_{\mathcal{C}} \mathcal{F}(x|\theta, \eta) d\eta dx = \int_{\mathcal{C}} \int_{\mathcal{S}} \mathcal{F}(x|\theta, \eta) dx d\eta.$$

Two more regularity conditions are given in next section in order to introduce the Hilbert structure.

2.2 Fibre Bundles in a Hilbert Space

A manifold structure as the space of distributions with a common support is well developed in [Ama85], [ABNK$^+$87] and [MR93]. In general however the

geometry of a manifold is not a good way to think of spaces of distributions. For example, consider the space of all distributions on three categories, B_0, B_1 and B_2. The space of all such distribution is determined by the triple of probabilities (π_0, π_1, π_2) with the constraints,

$$\pi_i \geq 0, \text{ and } \sum_{i=0}^{2} \pi_i = 1.$$

In the standard definition of a parameterization of an open subset of a manifold requires a diffeomorphism to an open set of Euclidean space. However, the distributions we are considering do not have a manifold structure but that of a simplex.

This idea can be generalised to a much more complex space of distributions. For example consider approximating any distribution on the real line by an infinite dimensional extended multinomial based on discretising the line into bins, whose probability mass function is defined as

$$f(x_1, x_2 \cdots | n, p_1, p_2, \cdots) = \begin{cases} \frac{n!}{x_1! x_2! \cdots} p_1^{x_1} p_2^{x_2} \cdots & while \sum_{i=1}^{\infty} x_i = n \\ 0 & otherwise \end{cases}.$$

We can approximate the continuous sample space $\mathcal{S}$ of a continuous distribution with density function $f(x)$, by an infinite set of bins

$$\{[n\epsilon, (n+1)\epsilon) | n \in \mathbb{Z}\},$$

for fixed $\epsilon > 0$. The probability on the bins $[n\epsilon, (n+1)\epsilon)$ is defined by

$$\pi_i := \mathbb{P}(B = [i\epsilon, (i+1)\epsilon)) = \int_{[i\epsilon,(i+1)\epsilon)} f(x)dx \geq 0. \tag{2.2}$$

So we have

$$\sum_i \pi_i = 1$$

Furthermore,

$$\Delta^{\infty} = \left\{ \pi \in \mathbb{R}^{\infty} | \sum_{i=1}^{\infty} \pi_i = 1, \pi_i \geq 0 \right\} \tag{2.3}$$

is called the standard infinite dimensional simplex.

Example 2. *Consider the probability density function of Normal distribution* $\mathcal{N}(\mu, 1)$ *given by*

$$\phi(x|\mu, 1) = \frac{1}{\sqrt{2\pi}} \exp\{(x-\mu)^2/2\}, \quad x \in (-\infty, +\infty).$$

We discretize the sample space into a set of bins with the form

$$B_i = [(i-1)\epsilon, i\epsilon), \quad i \in \mathbb{Z}$$

where $\epsilon > 0$*. The bin probabilities are given by*

$$\pi_i = \int_{B_i} \phi(x|\mu, 1)dx, \quad i \in \mathbb{Z}.$$

Then, we have an infinite number of finite width bins. π *is in the infinite dimensional simplex defined in Equation 2.3.*

Let the infinite dimensional simplex Δ^∞ be equipped with -1-affine structure introduced in [AICMV09]. In the infinite dimensional simplex Δ^∞ ,we can construct the affine space, $\langle X_{Mix}, V_{Mix}, + \rangle$, where

$$X_{Mix} = \left\{ x \in \mathbb{R}^\infty | \sum_{i=1}^{\infty} x_i = 1 \right\}, \quad V_{Mix} = \left\{ v \in \mathbb{R}^\infty | \sum_{i=1}^{\infty} v_i = 0 \right\}.$$

The operator $+$ is the normal addition operator. For generality, we consider the continuous case in the thesis, in which the affine space can be written as

$$X_{Mix} = \left\{ f | \int_{\mathcal{S}} f(x)dx = 1 \right\}, \quad V_{Mix} = \left\{ g | \int_{\mathcal{S}} g(x)dx = 0 \right\}.$$

With the positivity conditions, i.e. $\forall x_i \in \mathcal{S}$, $f(x_i|\theta, Q) > 0$, the elements in X_{Mix} are densities, while V_{Mix} forms a vector space. For all element $f \in X_{Mix}$, $\Pi(f)$ is a subspace of V_{Mix} defined as

$$\Pi^-(f) := \{ g | \exists \alpha > 0, \text{such that} f \pm \alpha g > 0, f \in X_{Mix}, g \in V_{Mix} \}. \tag{2.4}$$

Note that there is no guarantee that the Fisher information of $g \in \Pi^-(f)$ always exists. Let us look an example of infinite Fisher information given in [LCM09].

Example 3. *Let $X_1, \cdots, X_N$ be a random sample from the mixture exponential $(1-\alpha)Exp(1) + \alpha Exp(\theta)$, where $Exp(\theta)$ denotes the exponential distribution with mean θ. The score value for α at $\alpha = 0$ and given θ is*

$$S(\theta) = \sum_{i=1}^{N} \left\{ \frac{\theta^{-1} \exp(-\theta^{-1} X_i)}{\exp(-X_i)} - 1 \right\}.$$

However, under the homogeneous model where $\alpha = 0$, we find

$$\mathbb{E}[S^2(\theta)] = \begin{cases} \{n(1-\theta)^2/[\theta(2-\theta)]\} & \theta < 2 \\ \infty & \theta \geq 2 \end{cases}$$

Hence, for $\theta \geq 2$, we have an infinite Fisher information.

The fibre bundle structure of mixture model is introduced by Marriott in [Mar07b], following the idea of Amari [Ama85]. A distribution $f(x|\theta)$ with different values of θ forms a curve $c(\mathcal{C})$ in the infinite dimensional simplex Δ^∞. The infinite-dimensional simplex Δ^∞ has countable basis. Therefore, it is separable and the convex hull $\mathbb{K}$ of the set of distribution $f(x|\theta)$ can be introduced by Proposition 6. It forms a set for mixture distribution $f(x|\theta, Q)$. According to [AICMV09], we need the space of mixture distribution be a subspace of Δ^∞, in which all elements share same moment structure and support.

The following two regularity conditions are necessary.

Regularity Condition 5. *For all $\eta \in \mathcal{C}$, $f(x|\theta, \eta)$ share the same moment structure. In other words, for any $\eta_1, \eta_2 \in \mathcal{C}$ $\ln f(x|\theta, \eta_1) - \ln f(x|\theta, \eta_2)$ is an element of the set $\Pi^+(f(x|\theta, Q))$*

$$\Pi^+(f(x|\theta, Q)) := \left\{ g(x) | \exists \alpha > 0, \textit{such that} \int f(x|\theta, Q) \exp(\pm \alpha g(x)) dx < \infty \right\}. \tag{2.5}$$

Regularity Condition 6. *For any $\eta_1, \eta_2 \in \mathcal{C}$, $f(x|\theta, \eta_1) - f(x|\theta, \eta_2)$ is an element of $\Pi^-(f(x|\theta, Q))$.*

Denote $\mathcal{D}$ be the subset of Δ^∞ which satisfies Regular Condition 5 and 6. By the result in [AICMV09], $\mathcal{D}$ will form the Hilbert space defined in Definition 4. We denote the Hilbert space by Π.

Recall the mixture distribution can be given as

$$f(x|\theta, Q) = f(x|\theta_0) + \sum_{j=1}^{\infty} \alpha_j g_j,$$

where $g_j \in \Pi$ and $j = 1, 2, \cdots$. Let g_j, $j = 1, 2, \cdots$ be orthogonal to the score vector at $f(x|\theta_0)$. Second, the family $f(x|\theta, \eta)$ has zero -1 curvature either at θ_0 or globally. The base space is the set of $f(x|\theta)$ with different value of θ, while g_j, $i = 1, 2, \cdots$ span the fibre. We have a fibre bundle with Hilbert structure in its fibre.

2.3 Spectral Expansion of Mixture Density Functions

In the local mixture model with basis $\partial^i f(x|\theta_0)/\partial\theta^i$ in Equation 1.2, a large number of basis vectors is needed while the domain $\mathcal{C}$ of η is large. To solve this problem, Marriott introduce the Principle Component Analysis to find the basis span V_{Mix}. It is shown that Principle Component Analysis can keep the number of components low without great change of inference, even when η has a large domain $\mathcal{C}$. It is also applied to the likelihood function in [MV04].

Let $(\mathcal{C}, Q)$ be a measurable space, η is a random variable in the space with distribution Q over the compact set $\mathcal{C}$ by Regularity Condition 1. We have a Hilbert space Θ on it as follows

$$\Theta = \{f(\eta) | f \in C(\mathcal{C}, \mathbb{R})\},$$

The inner product is defined as

$$\langle g, h \rangle_\Theta = \int_{\mathcal{C}} g(\eta) h(\eta) d\eta. \tag{2.6}$$

The inner product of Θ exist because of the compactness of $\mathcal{C}$. Let θ_0 be a point in $\mathcal{C}$. To expand the vector $f(x|\theta,\eta) - f(x|\theta_0)$ in the subspace which is orthogonal to the term $\partial f(x|\theta)/\partial\theta|_{\theta_0}$ in Θ. For any $f(x|\theta,\eta), f(x|\theta_0) \in \mathcal{D}$, we choose $\mathcal{F}(x|\theta,\eta) \in \Theta$ as follows,

$$\mathcal{F}(x|\theta,\eta) := f(x|\theta,\eta) - f(x|\theta_0) - \langle f(x|\theta,\eta) - f(x|\theta_0), s_0(x)\rangle_\Pi \, s_0(x), \quad (2.7)$$

where

$$s_0(x) = \frac{1}{\|\frac{\partial}{\partial\theta} f(x|\theta_0)\|_\Pi} \frac{\partial}{\partial\theta} f(x|\theta_0),$$

where

$$\left\| \frac{\partial}{\partial\theta} f(x|\theta_0) \right\|_\Pi^2 = \int_{\mathcal{S}} f(x|\theta,Q)^{-1} \left[\frac{\partial}{\partial\theta} f(x|\theta_0) \right]^2 dx.$$

A kernel function $k(\eta_1,\eta_2)$ can be introduced as

$$k(\eta_1,\eta_2) = \int_{\mathcal{S}} f(x|\theta,Q)^{-1} \mathcal{F}(x|\theta,\eta_1)\mathcal{F}(x|\theta,\eta_2) dx. \quad (2.8)$$

Note here, the kernel depends on both θ and Q. By Regularity Condition 6, the kernel $k(\eta,\eta) < \infty$ for all $\eta \in \mathcal{C}$. For all $\eta_1,\eta_2 \in \mathcal{C}$, the kernel $k(\eta_1,\eta_2)$ is in $L_2(\mathcal{C}\times\mathcal{C})$. It is given as Lemma 2 in Appendix A.3. As mentioned in Appendix A.1.1, each kernel is associated with a reproducing kernel Hilbert space. So we have a reproducing kernel Hilbert space $(\Theta, k(\cdot,\cdot))$. Consider an integral operator $A(\cdot)$ on Θ, for $g \in \Theta$,

$$\begin{aligned}(Ag)(\eta_2) &= \int_{\mathcal{C}} g(\eta_1) k(\eta_1,\eta_2) d\eta_1 \\ &= \int_{\mathcal{C}} g(\eta_1) \int_{\mathcal{S}} f(x|\theta,Q)^{-1} \mathcal{F}(x|\theta,\eta_1)\mathcal{F}(x|\theta,\eta_2) dx d\eta_1. \end{aligned} \quad (2.9)$$

It has good properties which are proved in Appendix A.3.

Lemma 1. *The integral operator $A(\cdot)$ on $\mathcal{C}\times\mathcal{C}$ is compact, self-adjoint and positive. Furthermore, the operator $A(\cdot)$ is trace-class, i.e. the sum of all eigenvalues is finite.*

According to the book of L.Debnath, P.Minkusinski (Page 188 - 190 [DM05]), we know

Proposition 1. *(Spectral Theorem for Self-Adjoint Compact Operators) Let $A(\cdot)$ be a self-adjoint, compact operator on an infinite dimensional Hilbert space $\mathcal{H}$. Then, there exists in $\mathcal{H}$ a complete orthonormal system $\{\phi_1, \phi_2, \cdots\}$ consisting of eigenvectors of $A(\cdot)$. Moreover, for every $g \in \mathcal{H}$,*

$$g = \sum_{i=1}^{\infty} \langle g, \phi_i \rangle \, \phi_i,$$

where λ_n is the eigenvalue corresponding to ϕ_i. Furthermore, if A has infinitely many distinct eigenvalues $\lambda_1, \lambda_2, \cdots$, then $\lambda_n \to 0$ as $n \to \infty$.

Consider the eigenfunction equation,

$$\int_{\mathcal{C}} e_i(\eta_1) \int_{\mathcal{S}} f(y|\theta, Q)^{-1} \mathcal{F}(y|\theta, \eta_1) \mathcal{F}(y|\theta, \eta_2) dy d\eta_1 = \lambda e_i(\eta_2). \tag{2.10}$$

Applying Proposition 1 to the expansion of $\mathcal{F}(x|\theta, \eta)$, we have the expansion

$$\mathcal{F}(x|\theta, \eta) = \sum_{i=1}^{\infty} s_i(x) e_i(\eta),$$

where

$$\begin{aligned} s_i(x) &= \langle \mathcal{F}(x|\theta, \eta), e_i(\eta) \rangle_{\Theta} \\ &= \int_{\mathcal{C}} \mathcal{F}(x|\theta, \eta) e_i(\eta) d\eta. \end{aligned} \tag{2.11}$$

Then the spaced spanned by the vectors $s_i(x)$ has the following properties. The proofs are given in Appendix A.3.

Theorem 1. *The space spanned by $s_i(x)$, $i = 1, 2, \cdots$ is a subset of the vector space V_{Mix}. The set of $s_i(x)$, $i = 1, 2, \cdots$ is a complete orthogonal system of the Hilbert space Π, and the norm of s_i is λ_i.*

Theorem 2 is given as follows and the proof can be found in Appendix A.3.

Theorem 2. *The mixture density function $f(x|\theta, Q)$ can be expanded as*

$$\int f(x|\theta,\eta)dQ(\eta) = f(x|\theta_0) + \sum_{i=0}^{\infty} \alpha_i s_i(x),$$

where

$$\alpha_0 = -\int_{\mathcal{S}} f(x|\theta,Q)^{-1} f(x|\theta_0) s_0(x) dx.$$

and

$$\alpha_i = \int_{\mathcal{C}} e_i(\eta) q(\eta) d\eta, \quad i = 1, 2, \cdots.$$

In [Mar07b], Marriott suggests using a form of PCA to approximate the mixture density function. In other words, the eigenfunctions corresponding to the K-largest eigenvalues will be kept. The mixture density function is approximated by

$$\begin{aligned} f(x|\theta, Q) &\approx f(x|\theta, \alpha) \\ &= f(x|\theta) + \sum_{i=0}^{K} \alpha_i s_i(x). \end{aligned}$$

One critical note is that we are approximating the probability density function. Therefore, the positivity condition should be added

$$f(x|\theta, \alpha) > 0, \forall x \in \mathcal{S}.$$

Definition 5. *Consider all subspaces $\tilde{\Pi} \subseteq \Pi$, such that*

$$\sum_{i=1}^{K} \lambda_{(i)} \geq \alpha \sum_{i=1}^{\infty} \lambda_{(i)}, \quad 0 \leq \alpha \leq 1,$$

where $\lambda_{(i)}$ are eigenvalues ordered in descent. Among these affine spaces $\tilde{\Pi}$, the one with smallest K is called best α-space, denoted by Π_α, which is spanned by the $\{s_i(x), i = 1, 2, \cdots, K\}$. $s_i(x), i = 1, 2, \cdots, K$ correspond to the first K largest eigenvalues.

Note that here $\sum_{i=1}^{\infty}\lambda_i$ is finite because $A(\cdot)$ is a trace class operator as proved in Lemma 1.

Definition 6. *Consider a K-dimensional subspaces of Π spanned by the set of vectors $\{s_i(x), i = 0, 1, \cdots, K\}$, where $s_i(x), i = 1, 2, \cdots, K$ correspond to the K largest eigenvalues. Then, we call such subspace best K-space, denoted by Π_K.*

In fact, $f(x|\theta, \alpha)$ is the projection of $f(x|\theta, Q)$ onto the K-dimension space Π_K. It is obvious that the distance from $f(x|\theta, Q) \in \Pi$ to its projection on Π_K is $\sum_{i=K+1}^{\infty}\lambda_i$. We use the notation U_{Π_K} to describe the projection from Π to Π_K. Projection from Π to Π_α is similar.

Example 4. *Suppose X follows a a Binomial distribution $\mathcal{B}(10, 0.5 + \eta)$, while the η has a uniform distribution $\mathcal{U}(-0.4, 0.4)$. We know the probability density function of X is given by, for $x = 0, 1, 2, \cdots, 10$,*

$$\begin{aligned} f(x) &= \frac{\binom{10}{x}}{0.8}\int_{-0.4}^{0.4}(0.5+\eta)^x(0.5-\eta)^{10-x}d\eta \\ &= \frac{\binom{10}{x}}{0.8}\int_{0.1}^{0.9}\eta^x(1-\eta)^{10-x}d\eta \\ &= \frac{\binom{10}{x}}{0.8}\left(B(0.9; x+1, 11-x) - B(0.1; x+1, 11-x)\right), \end{aligned}$$

where $B(x; a, b)$ is the incomplete beta function defined as

$$B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt.$$

The density function is shown in Figure 2.1.

Let us expand $f(x|\eta)$ at point $f(x|\eta = 0)$. We obtain

$$\begin{aligned} \mathcal{F}(x|\eta) &= f(x|\eta) - f(x|\eta = 0) \\ &= \binom{10}{x}(0.5+\eta)^x(0.5-\eta)^{10-x} - \binom{10}{x}0.5^{10}. \end{aligned}$$

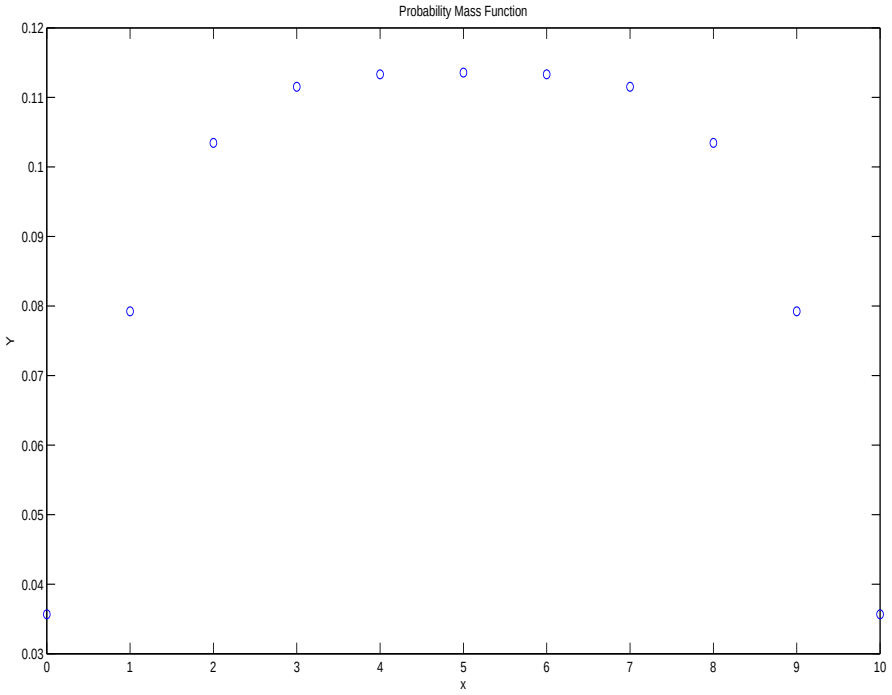

Figure 2.1: Probability Mass Function of $f(x)$

Since we know the value of θ*, therefore the last term of* $\mathcal{F}$ *in Equation 2.7 is* 0*. The kernel* $k(\eta_1, \eta_2)$ *is given by*

$$k(\eta_1, \eta_2) = \sum_{x=0}^{10} f(x_i)^{-1} \mathcal{F}(x_i|\eta_1)\mathcal{F}(x_i|\eta_2).$$

The eigenfunction equation can be written as

$$\int_{-0.4}^{0.4} e_i(\eta_1)k(\eta_1, \eta_2)d\eta_1 = \lambda_i e_i(\eta_2).$$

We select a uniform grid of 1000 *points from* $[-0.4, 0.4]$*. Then we can have a eigenvector equation to approximate the eigenfunction equation. The behavior of such approximation is discussed in Appendix A.5. Hence, we have*

$$Je_i = \lambda_i e_i, \tag{2.12}$$

where J is a 1000×1000 matrix and e_i is a 1000×1 vector. The eigenvalues and vectors are shown in Figure 2.2. According to the left panel of Figure 2.2, we seethe eigenvalues λ_i converge to zero quickly. Numerically, the seven largest nonzero eigenvalues are $\lambda_1 = 911.7859$, $\lambda_2 = 737.0085$, $\lambda_3 = 131.9741$, $\lambda_4 = 53.4189$, $\lambda_5 = 7.5543$, $\lambda_6 = 1.6140$ and $\lambda_7 = 0.1496$. Their eigenvectors are plotted in the right panel of Figure 2.2.

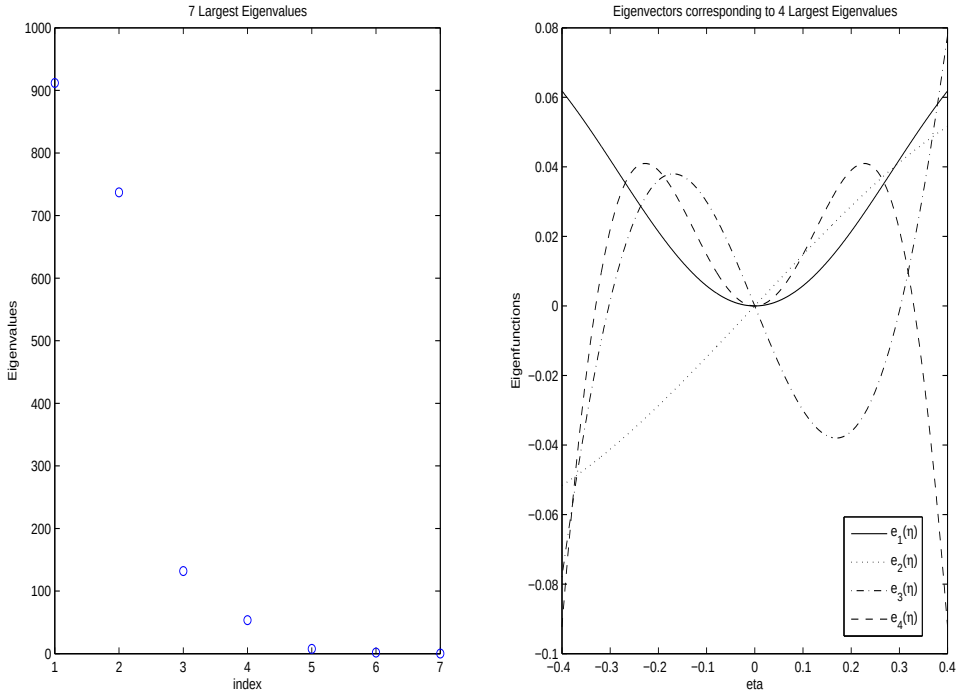

Figure 2.2: Eigenvalues and Vectors of Eigenvector Equation 2.12

Since $s(x) = \langle f(x|\eta), e(\eta)\rangle_\Theta$, and denote $\mathbf{f}(x) = (f(x|\eta_1), f(x|\eta_2), \cdots, f(x|\eta_{1000}))^T$, we have $s_i(x) = \mathbf{f}(x)^T e_i$. They are shown in Figure 2.3. We use $s_i(x)$, $i = 1, 2, 3, 4$ to approximate the mixture density functions. We calculate the values of α_i using

$$\alpha_i = \frac{1}{0.4} \int_{-0.2}^{0.2} e_i(\eta) d\eta.$$

We obtain $\alpha_1 = 0.0247$, $\alpha_2 = 0.0000$, $\alpha_3 = 0.0000$ *and* $\alpha_4 = 0.0109$.

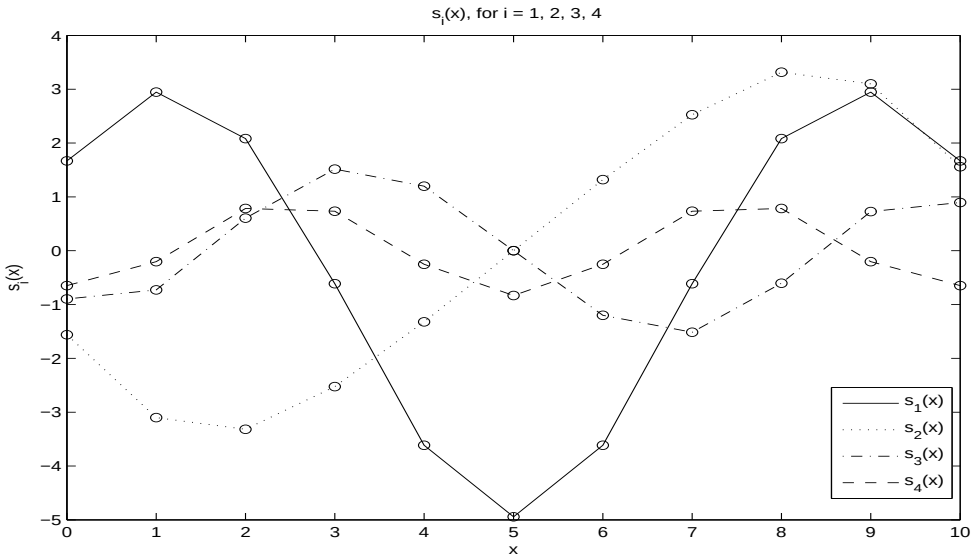

Figure 2.3: $s_i(x)$ for $i = 1, 2, 3, 4$ for $f(x|\theta, \alpha)$

Therefore, we approximate the mixture density function by

$$f_1(x_j) = \binom{10}{x} 0.5^{10} + 0.0247 s_1(x_j) + 0.0109 s_4(x_j), \quad x_j = 0, 1, 2, \cdots, 10. \tag{2.13}$$

The relative error, defined for each x_j, $j = 0, 1, 2, \cdots, 10$,

$$RE = f(x_j)^{-1}(f_1(x_j) - f(x_j))^2,$$

is shown in Figure 2.4.

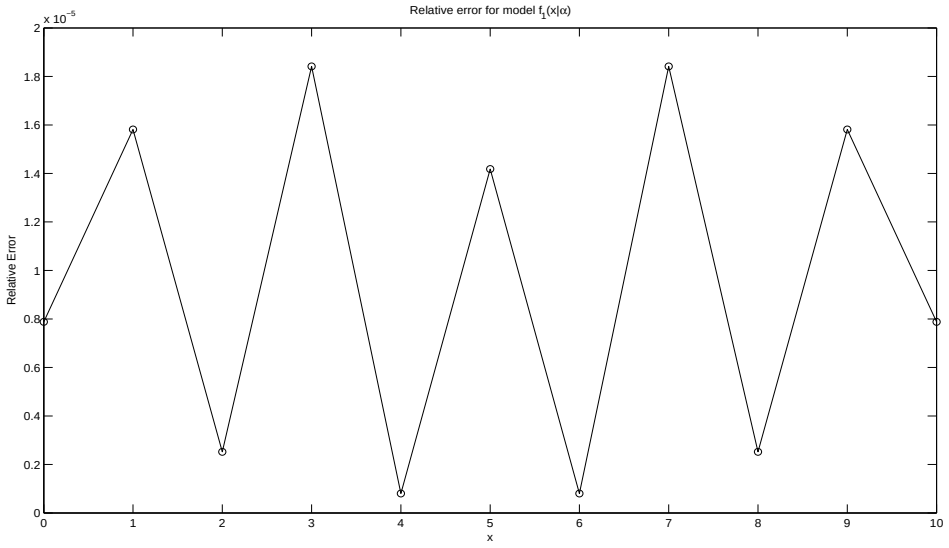

Figure 2.4: Relative Error for Model $f_1(x|\alpha)$

2.4 Convexity Structure of Mixture Model

The Hilbert space Π, whose inner product defined by Equation 2.1, with basis $\{\lambda_i^{-1/2} s_i(x), i = 0, 1, \cdots\}$ is separable by Proposition 6. The dimension of Π could be an infinite or finite over the field $\mathbb{R}$ of real numbers. Here let us treat it as infinite space for generality.

Consider the subspace $\Pi^{\mathcal{F}}$ of Π defined by, for fixed θ,

$$\Pi^{\mathcal{F}} := \{\mathcal{F}(x|\theta, \eta_i) | \eta_i \in \mathcal{C}, \mathcal{F}(x|\theta, \eta_i) \in \Pi\}.$$

The convex hull $co\Pi^{\mathcal{F}}$ of $\Pi^{\mathcal{F}}$ is located inside the Hilbert space Π.

$$co\Pi^{\mathcal{F}} = \left\{ g = \sum_{i=1}^{n} \beta_i \mathcal{F}_i | \beta_i \geq 0, \sum_{i=1}^{n} \beta_i = 1, \mathcal{F}_i \in \Pi^{\mathcal{F}}, n \in \mathbb{N} \right\}.$$

Define three sets,

$$\tilde{\mathbb{K}}^{\mathcal{F}} = \left\{ g = \sum_{j=0}^{\infty} \alpha_j s_j(x) | f(x_i|\theta_0) + g(x_i) > 0, \forall x_i \in \mathcal{S}, f(x_i|\theta) \in X_{Mix}, \alpha_j \in \mathbb{R} \right\},$$

$$\mathbb{K}^{\mathcal{F}} = \left\{ g = \sum_{j=0}^{\infty} \gamma_j s_j(x) | \gamma_j \in \left[\min_{\eta \in \mathcal{C}} e_j(\eta), \max_{\eta \in \mathcal{C}} e_j(\eta) \right] \right\},$$

$$\hat{\mathbb{K}}^{\mathcal{F}} = \left\{ g = \sum_{j=0}^{\infty} \beta_j s_j(x) | \beta_j \in \left[\min_{\eta \in \mathcal{C}} e_j(\eta), \max_{\eta \in \mathcal{C}} e_j(\eta) \right], g(x_i) + f(x_i|\theta_0) > 0, x_i \in \mathcal{S} \right\}.$$

We can show the following theorem and the proof is given in Appendix A.4.

Theorem 3. *The closed convex hull* $\overline{co\Pi^{\mathcal{F}}}$ *of* $\Pi^{\mathcal{F}}$ *is compact and closed in the closure of* $\hat{\mathbb{K}}^{\mathcal{F}}$.

Let the images of $\mathbb{K}^{\mathcal{F}}$, $\tilde{\mathbb{K}}^{\mathcal{F}}$ and $\hat{\mathbb{K}}^{\mathcal{F}}$ from Π to Π_K are

$$\tilde{\mathbb{K}}_K^{\mathcal{F}} = \left\{ g = \sum_{j=0}^{K} \alpha_j s_j(x) | g(x_i) + f(x_i|\theta_0) > 0, \forall x_i \in \mathcal{S} \right\},$$

$$\mathbb{K}_K^{\mathcal{F}} = \left\{ g = \sum_{j=0}^{K} \gamma_j s_j(x) | \gamma_j \in \left[\min_{\eta \in \mathcal{C}} e_j(\eta), \max_{\eta \in \mathcal{C}} e_j(\eta) \right] \right\},$$

$$\hat{\mathbb{K}}_K^{\mathcal{F}} = \left\{ g = \sum_{j=0}^{K} \beta_j s_j(x) | \beta_j \in \left[\min_{\eta \in \mathcal{C}} e_j(\eta), \max_{\eta \in \mathcal{C}} e_j(\eta) \right], g(x_i) + f(x_i|\theta_0) > 0, x_i \in \mathcal{S} \right\}.$$

When we use $f(x|\theta_0)+g$, $g \in \tilde{\mathbb{K}}_K^{\mathcal{F}}$ as the local mixture model, then $f(x|\theta_0)+g$ is a density function. However, it is possible that the coefficient α_j is not in the range of $[\min_{\eta \in \mathcal{C}} e_j(\eta), \max_{\eta \in \mathcal{C}} e_j(\eta)]$. When we use $f(x|\theta_0) + g$, $g \in \mathbb{K}_K^{\mathcal{F}}$, every coefficient $f(x|\theta_0)+g$ is in the range, but may not be a density function. The best approximation will be that $f(x|\theta_0) + g$, $g \in \hat{\mathbb{K}}_K^{\mathcal{F}}$, such that the approximation is a density function while all coefficients are in their own range.

Example 4. *(continued)*

The mixture model is given by

$$f_2(x_j|\alpha) = \binom{10}{x_j} 0.5^{10} + \sum_{i=1}^{4} \alpha_i s_i(x_j), \quad x_j = 0, 1, 2, \cdots, 10.$$

In matrix form, the linear equation system can be written as

$$\mathbf{f} = \vec{\alpha}\mathbf{X},$$

where $\mathbf{f} = (f_2(x_0|\alpha) - \binom{10}{x_0}0.5^{10}, f_2(x_1|\alpha) - \binom{10}{x_1}0.5^{10}, \cdots, f_2(x_{10}|\alpha) - \binom{10}{x_{10}}0.5^{10})^T$, $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$, *and* $\mathbf{X}$ *is a* 4×11 *matrix, whose element is* $X_{ij} = s_i(x_j)$. *We want to solve the optimization problem*

$$\min(\mathbf{f} - \vec{\alpha}\mathbf{X})^T \mathbf{A} (\mathbf{f} - \vec{\alpha}\mathbf{X}),$$

where $\mathbf{A}$ *is a diagonal matrix and* $A_{ii} = f(x_i)^{-1}$. *We estimate the coefficients by linear regression*

$$\vec{\alpha} = (\mathbf{X}\mathbf{A}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{A}\vec{f},$$

and have $\hat{\alpha}_1 = 0.0247$, $\hat{\alpha}_2 = 0$, $\hat{\alpha}_3 = 0$ *and* $\hat{\alpha}_4 = 0.0110$. *The relative error is shown in Figure 2.5*

For all $x_j = 0, 1, 2, \cdots, 10$, *we have* $f_2(x_j|\hat{\alpha}) > 0$. *Therefore, we know* $f_2(x_j|\alpha) - f(x|\eta = 0)$ *is located in* $\tilde{\mathbb{K}}_4$. *On the other hand,* $9.5418 \times 10^8 \leq e_1(\eta) \leq 0.0618$, $-0.0516 \leq e_2(\eta) \leq 0.0516$, $-0.0779 \leq e_3(\eta) \leq 0.0779$ *and* $-0.0923 \leq e_4(\eta) \leq 0.0410$ *for* $\eta \in [-0.4, 0.4]$. *For each* i, α_i *lies in the range of* $e_i(\eta)$ *over* $\eta \in [-0.4, 0.4]$. *Hence, we know* $f_2(x|\hat{\alpha}) - f(x|\eta = 0)$ *also are in* $\mathbb{K}_4$. *Therefore,* $f_2(x|\hat{\alpha}) - f(x|\eta = 0)$ *is an element in* $\hat{\mathbb{K}}_4$.

Example 5. *Consider the* X *from a Binomial distribution* $\mathcal{B}(10, 0.0001 + \eta)$, *where the* η *has a uniform distribution* $\mathcal{U}(-0.00009, 0.0001)$. *We expand the distribution at point* $f(x|\eta = 0)$. *The eigenvalues and eigenvectors are given in Figure 2.6.*

We use the eigenvectors corresponding to ten largest eigenvalues to approximate the mixture density functions.

$$f_3(x_j|\alpha) = \binom{10}{x_j} (0.0001)^{x_j} (0.9999)^{10-x_j} + \sum_{i=1}^{7} \alpha_i s_i(x_j), \quad x_j = 0, 1, 2, \cdots, 10.$$

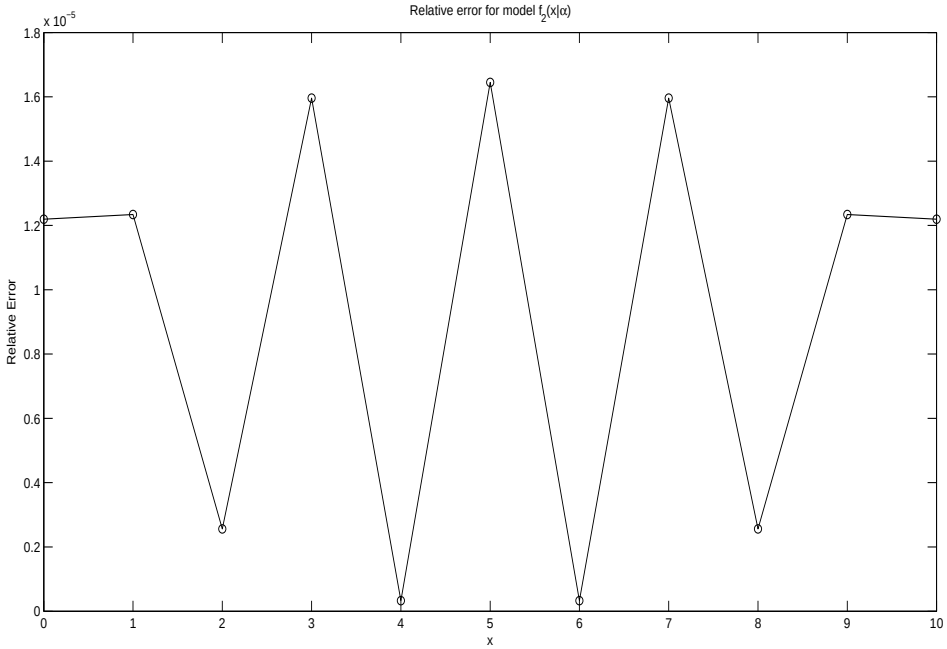

Figure 2.5: Relative Errors of $f_2(x|\hat{\alpha})$

By linear regression mentioned in last example, we have $\hat{\alpha}_1 = 0.0029$, $\hat{\alpha}_2 = 0.0233$, $\hat{\alpha}_3 = -0.0029$, $\hat{\alpha}_4 = 0.0124$, $\hat{\alpha}_5 = -0.4070$, $\hat{\alpha}_6 = 0.3210$ *and* $\hat{\alpha}_7 = 1.2060$.

For all $x_j = 0, 1, 2, \cdots, 10$, *we have* $f_3(x|\hat{\alpha}) > 0$. *Therefore, we know* $f_3(x|\hat{\alpha}) - f(x|\eta = 0)$ *is located in* $\tilde{\mathbb{K}}_7$. *On the other hand,* $-0.2759 \leq e_7(\eta) \leq 0.3412$ *for* $\eta \in [-0.00009, 0.0001]$, *while* $f_3(x|\hat{\alpha}) - f(x|\eta = 0)$ *are not in* $\mathbb{K}_7$, *because* $\alpha_7 = 1.2060 > 0.3421$. *In summary,* $f_3(x|\hat{\alpha})$ *is not in* $\mathbb{K}_7$ *but* $\tilde{\mathbb{K}}_7$.

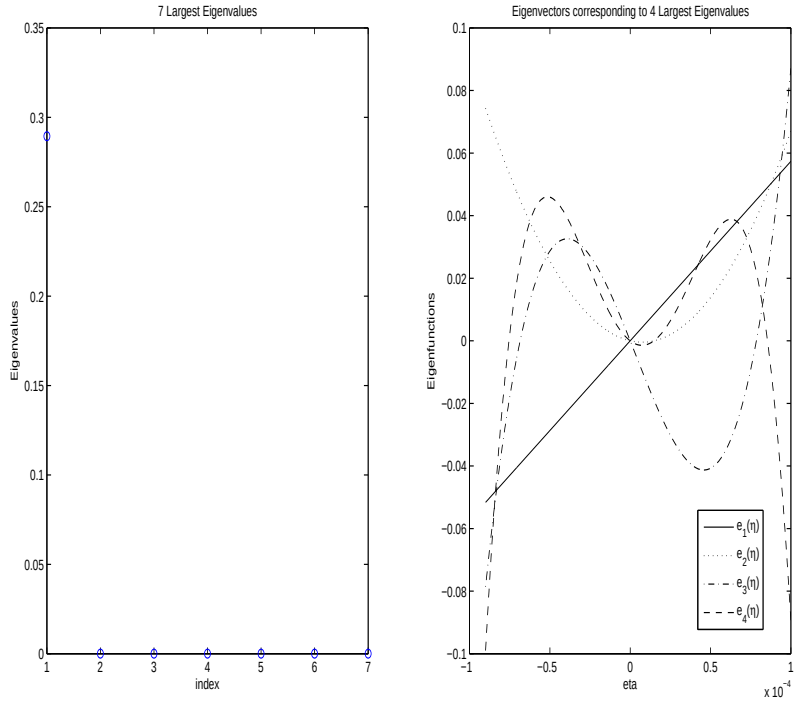

Figure 2.6: Eigenvalues and Eigenvectors for Example 5

If $U_{\Pi_K}\mathcal{F}(x|\theta,Q) \in \hat{\mathbb{K}}_K^{\mathcal{F}}$, the norm from $\mathcal{F}(x|\theta,Q)$ to $U_{\Pi_K}\mathcal{F}(x|\theta,Q)$ is

$$\begin{aligned}\|\mathcal{F}(x|\theta,Q) - U_{\Pi_K}\mathcal{F}(x|\theta,Q)\|_{\Pi}^2 &= \left\| \sum_{i=K+1}^{\infty} \alpha_i^2 s_i^2(x) \right\|_{\Pi}^2 \\ &\leq \sum_{i=K+1} \int_{\mathcal{S}} f(x|\theta,Q)^{-1} s_i^2(x) dx \\ &= \sum_{i=K+1}^{\infty} \lambda_i.\end{aligned}$$

To describe the error in approximation, some information on the decay of the eigenvalues is needed. Some work has been already done in the area. Considering the operators on a bounded interval, in Weyl's work [Wey12], for a general kernel $k(x,y) \in L_2(\mathcal{C} \times \mathcal{C})$, which is continuously differentiable

in $\mathcal{C} \times \mathcal{C}$, then $\lambda_K = o(K^{-3/2})$. In the case that $k(x, y)$ is a positive definite kernel, as shown by Reade [Rea83] and [Rea84], $\lambda_K = o(K^{-2})$. As described by Reade [Rea83] and [Rea84], $\lambda_K = o(K^{-2})$. It is also mentioned that

$$\sum_{K+1}^{\infty} \lambda_i = o(K^{-1}),$$

as $K \to \infty$. Therefore, we can give following theorem. Its proof is in Appendix A.4.

Theorem 4. *In the Hilbert space* Π*, if* $U_{\Pi_K}\mathcal{F}(x; Q) \in \hat{\mathbb{K}}_K^{\mathcal{F}}$*, the norm of the vector from the mixture density* $f(x|\theta, Q)$ *with a compact* $\mathcal{C}$ *and mixture models* $f(x|\theta, \alpha)$ *by PCA has the order* $o(K^{-1})$*.*

2.5 Dependence on Choice of Compact Region

The dependence on the choice of compact region of mixture models has been discussed by Anaya-Izquierdo and Marriott [AIM07]. Assume that we want keep the α-percentage of the sum of all eigenvalues of the eigenfunction Equation 2.10, we will find that the number of eigenfunctions needed to reconstruct the mixture density function changes, while the domain of η changes. In geometric view, it can be thought that the dimension of the parameter space changes. Such phenomena also indicates that a manifold can not describe the structure precisely. Furthermore, according to Lemma 4, we can see that the eigenvalues depends on the domain of η closely.

Lemma 3. *Let*

$$v_i(x) = f(x|\theta, Q)^{-1/2} s_i(x),$$

then

$$\int_{\mathcal{C}} e_i(\eta_1) \int_{\mathcal{S}} f(y|\theta, Q)^{-1} \mathcal{F}(y|\theta, \eta_1)\mathcal{F}(y|\theta, \eta_2) dy d\eta_1 = \lambda e_i(\eta_2)$$

is equivalent to,

$$\int_{\mathcal{S}} v_i(x) \int_{\mathcal{C}} \frac{\mathcal{F}(x|\theta,\eta)}{\sqrt{f(x|\theta,Q)}} \frac{\mathcal{F}(y|\theta,\eta)}{\sqrt{f(y|\theta,Q)}} d\eta dx = \lambda v_i(y). \tag{2.14}$$

The proof is given in Appendix A.5. We approximate the mixture density function with

$$f(x|\theta,Q) = f(x|\theta) + \sum_{i=0}^{K} \alpha_i s_i(x).$$

With the expansion of $\mathcal{C}$, the space Π should be expanded to contain the convex hull of curve $c(\mathcal{C})$. Such expansion of Π can be described in following two theorems. Both of the proofs are shown in Appendix A.5. To prove Theorem 6, discretization of the eigenvalue equation 2.14 is needed. See more detail in Appendix A.5.

Theorem 5. *With the expansion of region $\mathcal{C}$, the eigenvalues $\lambda_i, i \geq 1$ for the expansion increase.*

Theorem 6. *The larger the domain of η is, the more number of $s_i(x)$ are needed to contribute Π_α.*

Example 4. *(continued)*
Assume that the η has a uniform distribution $\eta \sim \mathcal{U}(-0.2, 0.2)$. Numerically, the nonzero eigenvalues are $\lambda_1 = 365.1704$, $\lambda_2 = 131.0270$, $\lambda_3 = 5.5456$, $\lambda_4 = 0.5690$, $\lambda_5 = 0.0175$, $\lambda_6 = 0.0008$ and $\lambda_7 = 0$. Comparing them with the case that $\mathcal{U}(-0.3, 0.3)$, whose $\lambda'_1 = 586.3474$, $\lambda'_2 = 438.4387$, $\lambda'_3 = 38.3676$, $\lambda'_4 = 8.6468$, $\lambda'_5 = 0.6241$, $\lambda'_6 = 0.0684$ and $\lambda'_7 = 0.0031$. We see for all $i = 1, 2, \cdots$, $\lambda'_i \geq \lambda_i$.

Let $\alpha = 98\%$, in the case of $\mathcal{U}(-0.2, 0.2)$, we need two $s_i(x)$ for approximation. On the other hand, in the case of $\mathcal{U}(-0.3, 0.3)$, we only need three $s_i(x)$ for approximation to reach the same rate of α.

We also consider the case of $\mathcal{U}(-0.4, 0.4)$. In Figure 2.7, we can see it clearly, that with the region of η expansion, the eigenvalues $\lambda_i, i \geq 1$ for the expansion increase. Furthermore, when $\eta \sim \mathcal{U}(-0.4, 0.4)$, we need four $s_i(x)$ to contribute $\Pi_{0.98}$.

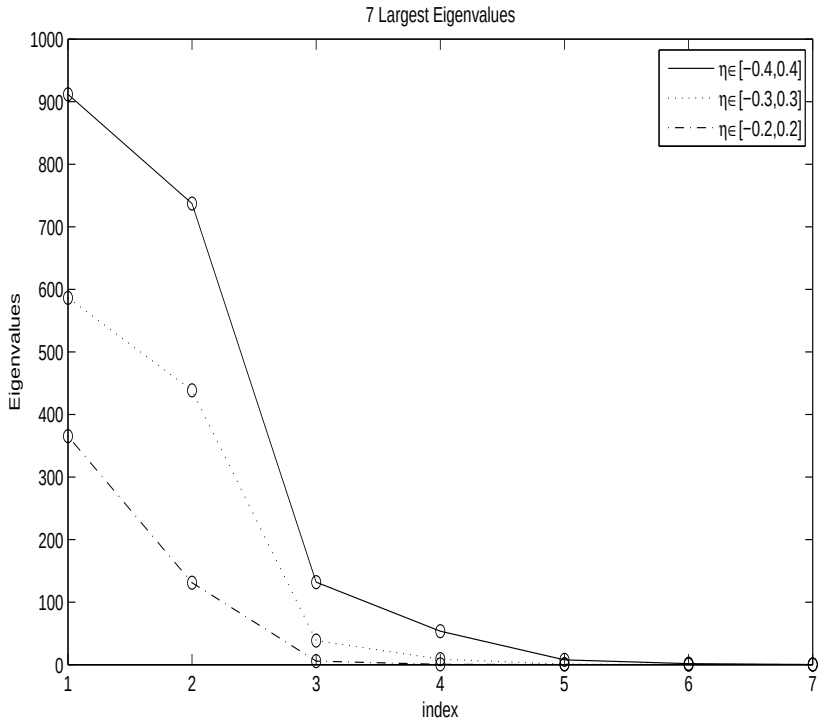

Figure 2.7: Asymptotic behavior of eigenvalues while $\eta \in [-0.4, 0.4]$, $\eta \in [-0.3, 0.3]$ and $\eta \in [-0.2, 0.2]$

Chapter 3

Conclusions and Future Work

Our contribution in this thesis have mainly included: thinking about the set of mixture distributions with a single Hilbert structure, whose inner product is the -1-expectation of the product of two vectors, introducing the fibre bundle structure to the set of mixture distributions, decomposing the mixture density functions spectrally in the Hilbert space, discussing the convexity structure of extending local mixture models in the Hilbert structure, showing the asymptotic behaviors of relative error in the approximation of mixture models and giving the proof of the effect of domain $\mathcal{C}$ to the relative errors in the approximation.

There are some possible future directions for this research.

Relaxation of the Regularity Conditions

One of the future work in mixture models is to relax the regularity conditions.

For Regularity Condition 1, we can relax the condition that η has a compact support $\mathcal{C}$ to $(-\infty, +\infty)$. In [Bue04], [BP06] and [BP07], integral operators with unbounded intervals are considered. In this work, the authors discuss the asymptotic behavior of eigenvalues of the integral operators with unbounded intervals. We can obtain a more general result in analyzing the asymptotic behavior of the relative errors of the approximation.

One of the challenges in the future is the case when the Fisher information is infinite. According to [AICMV09], the infinite dimensional simplex can be decomposed into a bunch of Hilbert spaces. How can we analyze two models in different Hilbert space? It is an interesting question for future work.

Nonlinear Approximation of Mixture Model

There is one concern in the PCA approximation when considering the Kolmogrov n-width (Appendix A.1.2) of the approximation,

$$d_n = \inf_{\Pi_K} \sup_{x \in \mathcal{S}} \inf_{\mathcal{F}(x|\theta,\alpha) \in \Pi_K} \frac{(f(x|\theta, Q) - f(x|\theta, \alpha))^2}{f(x|\theta, Q)},$$

where $\mathcal{F}(x|\theta, \alpha) = f(x|\theta, \alpha) - f(x|\theta_0)$. Unlike the norm we considered in the spectral decomposition, the Kolmogrov n-width considers the supremum of the error over all $x \in \mathcal{S}$. It approximates locally, while the norm defined in a Hilbert space Π global approximation. When we achieve a good approximation globally, this does not guarantee that we have as good a local approximation. This is one of the challenges in the PCA approach.

Example 6. *Consider the X from a Binomial distribution $\mathcal{B}(10, 0.06 + \eta)$, while the η has a uniform distribution $\mathcal{U}(-0.053, 0.1)$.*

$$f_5(x|\alpha) = \binom{10}{x} 0.06^x 0.94^{10-x} + \sum_{i=1}^{1000} \alpha_i s_i(x), \quad x_j = 0, 1, 2, \cdots, 10,$$

where

$$\alpha_i = \frac{1}{0.153} \int_{-0.053}^{0.1} e_i(\eta) d\eta.$$

and $s_i(x)$ is obtained by spectral decomposition.

For all $x = 0, 1, 2 \cdots, 10$, $f_5(x|\alpha) \geq 0$. However, the relative error of such approximation is given by Figure 3.1.

According to the Figure 3.1, we see the relative error at point $x = 1, 2, 3, 4$ are really big comparing the others. The Kolmogrov n-width is 5.5233×10^{-9}

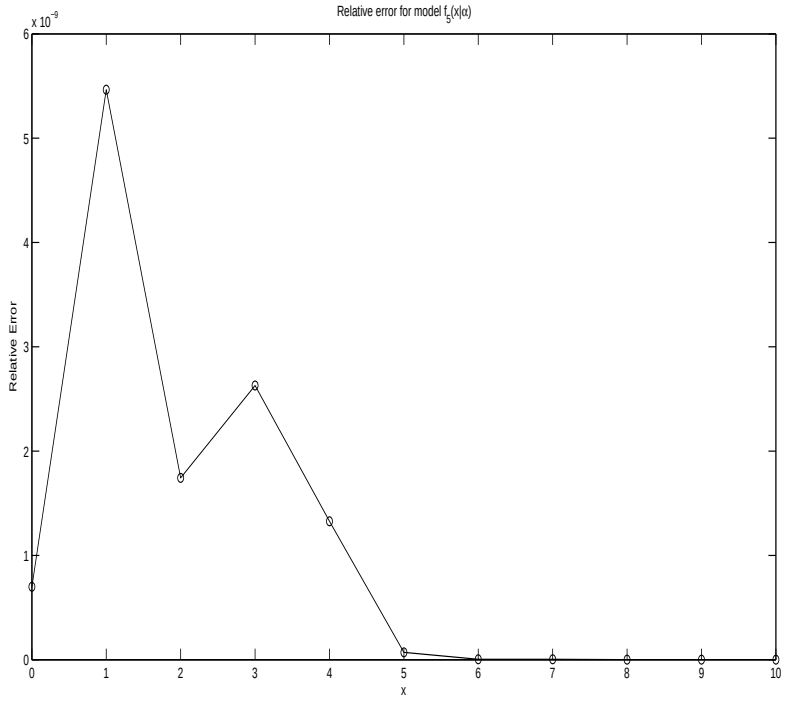

Figure 3.1: Relative Error for Model $f_5(x|\alpha)$

at $x_j = 1$*, while the other relative errors are below* 3×10^{-9}*. The relative errors of approximation are not in same level. In such sense, it is not a good approximation.*

One way to solve the problem is using nonlinear approximation, such as the sparse model. Let K still be the number of terms we want in the approximation. For all $x_j \in \mathcal{S}$, an approximation operator $\mathcal{A}_K$ on Π is

$$\left(\mathcal{A}_K\mathcal{F}\right)(x_j|\theta, Q) = \sum_{i\in I_K} \alpha_i s_i(x_j),$$

where $I_K := I_K(x_j)$ represents the set of indices corresponding to the K largest $s_i(x_j)$. Note here, for different $x_j \in \mathcal{S}$, we have different approximation $\left(\mathcal{A}_K\mathcal{F}\right)(x_j|\theta, Q)$. Such approximation is called best K-term approx-

imation. Because the vector of coefficients in $f(x_j|\theta, \alpha)$ is sparse, it is also called a sparse model. The correlation between sparse model and Kolmogrov n-width has been discussed in [CDD09]. It is shown that sparse model can offer an approximation with bounded Komogrov n-width.

There is a big challenge in the sparse model. The computing cost is high, because finding the K largest coefficients for each is an NP hard problem. A recent development in approximation is the compressed sensing algorithm [Don06] and [CDD09], which changes the NP hard problem into a convex optimization problem.

Furthermore, the basis from spectral decomposition are the optimal basis for linear approximation techniques. In [CD97], Cohen and D'ales show that the optimality is lost in nonlinear approximation. Therefore, the basis should be changed in wavelets or trigonometric system in nonlinear approximation of mixture models.

Bibliography

[ABNK+87] S. Amari, OE Barndo-Nieldon, RE Kass, SL Lauritzen, and CR Rao. Differential Geometry in Statistical Inference, volume 10 of IMS Lecture Notes Monograph. *Inst. Math. Stat., Hayward, CA*, 1987.

[AI06] K. Anaya-Izquierdo. *Statistical and geometric analysis of local mixture models and a proposal of some new tests of fit for censored data*. PhD thesis, Ph. D. thesis, Universidad Nacional Autónoma de México, 2006.

[AICMV09] K. Anaya-Izquierdo, F. Critchley, P. Marriott, and P. Vos. Towards information geometry on the space of all distributions. *Submitted to the Annals of Statistics*, 2009.

[AIM07] K. Anaya-Izquierdo and P. Marriott. Local mixture models of exponential families. *Bernoulli*, 13(3):623–640, 2007.

[Ama85] S. Amari. *Differential-geometrical methods in statistics*. Springer, 1985.

[Bar95] R.G. Bartle. *The elements of integration and Lebesgue measure*. Wiley, 1995.

[BBZ07] G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294, 2007.

[BDKS05] D. Böhning, E. Dietz, R. Kuhnert, and D. Schön. Mixture models for capture-recapture count data. *Statistical Methods and Applications*, 14(1):29–43, 2005.

[BNBC+92] O.E. Barndorff-Nielsen, P. Blæsild, A.L. Carey, PE Jupp, M. Mora, and MK Murray. Finite-dimensional algebraic representations of the infinite phylon group. *Acta Applicandae Mathematicae: An International Survey Journal on Applying Mathematics and Mathematical Applications*, 28(3):219–252, 1992.

[BP06] J. Buescu and A.C. Paixão. Eigenvalues of positive definite integral operators on unbounded intervals. *Positivity*, 10(4):627–646, 2006.

[BP07] J. Buescu and AC Paixão. Eigenvalue distribution of Mercer-like kernels. *Mathematische Nachrichten*, 280(9):984–995, 2007.

[Bue04] J. Buescu. Positive integral operators in unbounded domains. *Journal of Mathematical Analysis and Applications*, 296(1):244–255, 2004.

[CD97] A. Cohen and J.P. D'Ales. Nonlinear approximation of random functions. *SIAM Journal on Applied Mathematics*, 57(2):518–540, 1997.

[CDD09] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *American Mathematical Society*, 22(1):211–231, 2009.

[Cha87] A. Chao. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791, 1987.

[CR87] D.R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49:1–39, 1987.

[DM05] L. Debnath and P. Mikusiński. *Hilbert spaces with applications*. Academic Press, 2005.

[Don06] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[FHHP01] M.J. Fabian, P. Habala, P. Hájek, and J. Pelant. *Functional analysis and infinite-dimensional geometry*. Springer Verlag, 2001.

[FR02] C. Fraley and A.E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–632, 2002.

[GBSS05] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. Measuring statistical dependence with Hilbert-Schmidt norms. *Lecture Notes in Computer Science*, 3734:63–77, 2005.

[GM95] R.J. Gardner and P. McMullen. *Geometric tomography*. Cambridge University Press, 1995.

[GWZ93] P.M. Gruber, J.M. Wills, and GM Ziegler. *Handbook of convex geometry*. North-Holland Amsterdam, 1993.

[HB04] M. Hein and O. Bousquet. Kernels, associated structures and generalizations. *Max-Planck-Institut fuer biologische Kybernetik, Technical Report*, 2004.

[HT96] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):155–176, 1996.

[KSM99] I. Kolar, J. Slovak, and P.W. Michor. Natural operations in differential geometry. 1999.

[Lai78] N. Laird. Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

[LCM09] P. Li, J. Chen, and P. Marriott. Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96(2):411, 2009.

[Ler92] B.G. Leroux. CONSISTENT ESTIMATION OF A MIXING DISTRIBUTION. *The Annals of Statistics*, 20(3):1350–1360, 1992.

[Lin81] B.G. Lindsay. Properties of the maximum likelihood estimator of a mixing distribution. *Statistical Distributions in Scientific Work*, 5:95–109, 1981.

[Lin83] B.G. Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, 11(1):86–94, 1983.

[Lin95] B.G. Lindsay. Mixture models: theory, geometry, and applications. Ims, 1995.

[Mar02] P. Marriott. On the local geometry of mixture models. *Biometrika*, 89(1):77–93, 2002.

[Mar03] P. Marriott. On the geometry of measurement error models. *Biometrika*, 90(3):567, 2003.

[Mar07a] M.G. Marmorino. Comment: Improvement of Weyls Inequality. *Journal of Mathematical Chemistry*, 41(3):327–327, 2007.

[Mar07b] P. Marriott. Extending local mixture models. *Annals of the Institute of Statistical Mathematics*, 59(1):95–110, 2007.

[Mee72] G. Meeden. Bayes estimation of the mixing distribution, the discrete case. *The Annals of Mathematical Statistics*, 43(6):1993–1999, 1972.

[Mer09] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909.

[MR93] M.K. Murray and J.W. Rice. *Differential geometry and statistics*. Chapman & Hall/CRC, 1993.

[MV04] P. Marriott and P. Vos. On the global geometry of parametric models and information recovery. *Bernoulli*, 10(4):639–650, 2004.

[Pin86] A. Pinkus. n-widths and Optimal Recovery in Approximation Theory. In *Proceeding of Symposia in Applied Mathematics*, volume 36, 1986.

[Rea83] J. Reade. Eigenvalues of positive definite kernels. *SIAM J. Math. Anal*, 14(1):152–157, 1983.

[Rea84] J.B. Reade. Eigenvalues of positive definite kernels II. *SIAM Journal on Mathematical Analysis*, 15:137, 1984.

[Rob56] H. Robbins. An empirical Bayes approach to statistics. In *Proc. Third Berkeley Symp. Math. Statist. Probab*, volume 1, pages 157–163, 1956.

[Rob64] H. Robbins. The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35:1–20, 1964.

[Rol68] J.E. Rolph. Bayesian estimation of mixing distributions. *The Annals of Mathematical Statistics*, 39(4):1289–1302, 1968.

[Sau89] D.J. Saunders. *The geometry of jet bundles*. Cambridge Univ Pr, 1989.

[SZ08] S. Smale and D.X. Zhou. Geometry on probability spaces. *preprint*, 2008.

[TJP04] A. Topchy, A.K. Jain, and W. Punch. A mixture model for clustering ensembles. In *Proc. SIAM Intl. Conf. on Data Mining*, pages 379–390. Citeseer, 2004.

[Wei76] H.V. Weizsäcker. A note on infinite dimensional convex sets. *Mathematica Scandinavica*, 38:321, 1976.

[Wey12] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

[Woo99] G.R. Wood. Binomial mixtures: geometric estimation of the mixing distribution. *The Annals of Statistics*, 27(5):1706–1721, 1999.

[XM07] C. Xuan Mao. Estimating population sizes for capture–recapture sampling with binomial mixtures. *Computational Statistics and Data Analysis*, 51(11):5211–5219, 2007.

Appendix A

A.1 Mathematical Preliminaries

A.1.1 Background of Functional Analysis

The mathematical preliminaries of this part mainly serve the spectral decomposition. More details can be found in [DM05] and [HB04].

Leibniz's Rule and Fubini Theorem (Page 46, 111 [Bar95])

Proposition 2. *(Leibniz's Rule) Suppose that for some $t_0 \in [a, b]$, the function $x \to f(x, t_0)$ is integrable on X, that $\partial f/\partial t$ exists on $X \times [a, b]$, and that there exist an integrable function on X such that*

$$\left| \frac{\partial}{\partial t} f(x, t) \right| \leq g(x).$$

Then, we have

$$\frac{\partial}{\partial t} \int f(x, t) d\mu(x) = \int \frac{\partial}{\partial t} f(x, t) d\mu(x).$$

Proposition 3. *(Fubini Theorem) Suppose A and B are complete measure spaces. Suppose $f(x, y)$ is $A \times B$ measurable. If*

$$\int_{A \times B} |f(x, y)| d(x, y) < \infty,$$

where the integral is taken with respect to a product measure on the space over $A \times B$, then

$$\int_A \int_B f(x,y)dydx = \int_B \int_A f(x,y)dxdy = \int_{A\times B} f(x,y)d(x,y).$$

ℓ_p Norm and L_p Norm

For a vector $x \in \mathbb{R}^N$, the ℓ_p norm is defined as

$$\|x\|_{\ell_p} = \left(\sum_i^N |x_i|^p\right)^{1/p}.$$

For $1 \leq p < \infty$ and a measure space $(\mathcal{S}, \Sigma, \mu)$, consider the set of all measurable functions from $\mathcal{S}$ to $\mathbb{R}$, the L_p norm is defined as

$$\|f\|_{L_p} = \left(\int_{\mathcal{S}} |f|^p d\mu\right)^{1/p}.$$

Affine Space [AI06]

Definition 7. *An affine space is either the empty set or a triplet $(X, V, +)$ consisting of a nonempty set X of points, a real vector space V of translations and a action $+ : X \otimes V \to X$ satisfying the following conditions:*

- *Let $\vec{0}$ be the zero vector in V. For all $x \in X$*

$$x \oplus \vec{0} = x.$$

- *For all $\vec{u}, \vec{v} \in V$ and all $x \in X$,*

$$(x \oplus \vec{u}) \oplus \vec{v} = x \oplus (\vec{u} + \vec{v}).$$

- *For any two points $x, y \in X$, there is a unique $\vec{u} \in V$ such that*

$$x \oplus \vec{u} = y.$$

Hilbert Space (Page 87, 126 in [DM05])

Let $\mathcal{H}$ be a vector space. A mapping $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is called an *inner product* in $\mathcal{H}$ if for any $x, y, z \in \mathcal{H}$, and $\alpha, \beta \in \mathbb{R}$, the following conditions are satisfied:

- $\langle x, y \rangle = \overline{\langle y, x \rangle}$ (the bar denotes the complex conjugate);
- $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$;
- $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ implies $x = 0$.

Such vector space with an inner product is called a *inner product space*.

Definition 8. *A complete inner product space is called a Hilbert space.*

One of important theorems in Hilbert space is Riesz representation theorem given by following on page 126 in [DM05].

Proposition 4. *(Riesz Representation Theorem) Let f be a bounded linear functional on a Hilbert space $\mathcal{H}$. There exists exactly one $x_0 \in \mathcal{H}$ such that $f(x) = \langle x, x_0 \rangle$ for all $x \in \mathcal{H}$. Moreover, we have $\|f\| = \|x_0\|$.*

Trace Class and Mercer's Theorem [Mer09]

Definition 9. *A bounded linear operator A over a separable Hilbert space $\mathcal{H}$ is said to be in the trace class if for some orthonormal bases $\{\phi_i\}_i$ of $\mathcal{H}$ the sum of positive terms*

$$\sum_{i=1}^{\infty} \langle A\phi_i, \phi_i \rangle < \infty.$$

Proposition 5. *Let A be a positive, integral operator on $L_2[a, b]$ with continuous kernel $K(s, t) = K(t, s)$ on $[a, b]^2$ $(|a|, |b|) < \infty$. Then the kernel $K(s, t)$ can be represented by the bilinear series*

$$K(s, t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s) \phi_i(t)$$

absolutely and uniformly convergent on $[a,b]^2$*, where* $\lambda_i \geq 0$*,* $i = 1, 2, \cdots$ *are the eigenvalues of operator* A *and* ϕ_i*,* $i = 1, 2, 3, \cdots$ *are corresponding orthonormal eigenfunctions.*

Reproducing Kernel Hilbert Space [GBSS05]

Let $\mathcal{S}$ be the sample space of the random variables X and $\mathcal{H}$ a Hilbert space of real-valued functions on $\mathcal{S}$. We say $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) if every linear map of the form

$$L_x : f \mapsto f(x)$$

from $\mathcal{S}$ to $\mathbb{R}$ is continuous for any $x \in \mathcal{S}$. The Riesz representation theorem states that for every $x \in \mathcal{S}$ there exists a unique element $k(x, \cdot)$ of $\mathcal{H}$ with the property that:

$$f(x) = \langle f, k(x, \cdot) \rangle, \quad \forall f \in \mathcal{H}, \forall x \in \mathcal{S},$$

and

$$\langle k(x, \cdot), k(y, \cdot) \rangle = k(x, y), \quad \forall x, y \in \mathcal{S}.$$

The space $\mathcal{S}$ can be mapped into $\mathcal{H}$ via the feature mapping $x \in \mathcal{S} \mapsto \Phi(x) = k(x, \cdot) \in \mathcal{H}$. Therefore, $\langle \Phi(x), \Phi(y) \rangle = k(x, y)$.

Separable Hilbert Space (Page 127 in [DM05])

Definition 10. *A Hilbert space is called separable if it contains a complete orthonormal sequence.*

We also have a important theorem related to the separable Hilbert space.

Proposition 6. *A Hilbert space is separable if and only if it has a countable orthonormal basis.*

Hilbert-Schmidt Norm and Operators [BBZ07]

A linear operator L from $\mathcal{H}$ to $\mathcal{H}$ is called *Hilbert-Schmidt operator*, if $\sum_{i\geq 1}\|Le_i\|^2_{\mathcal{H}} < \infty$, where $\{e_i, i = 1, 2, \cdots\}$ is the orthonormal basis of $\mathcal{H}$. The set of all Hilbert-Schmidt operators on Π^* is denoted by $HS(\mathcal{H})$ with the inner product

$$\langle A, B\rangle_{HS(\mathcal{H})} = \sum_{i\geq 1}\langle Ae_i, Be_i\rangle .$$

Orthogonal Projectors [BBZ07]

An *orthogonal projector* in $\mathcal{H}$ is a linear operator U such that

$$U^2 = U = U^T.$$

A.1.2 Background of Geometry

The geometry background includes differential geometry and convex geometry. Structure such as convexities and jet space, are used to describe mixture models structure.

Convex Body (Page 45 in [GWZ93])

Definition 11. *A subset* $\mathbb{K}$ *is convex if* $(1-\lambda)x+\lambda y \in \mathbb{K}$ *for all* $x, y \in \mathbb{K}$ *and* $0 < \lambda < 1$. *If the convex subset* $\mathbb{K}$ *is compact and with nonempty interior, they are called convex bodies.*

Convex Hull and Polytope (Page 7 in [GM95] and Page 487 in [GWZ93])

The *convex hull* of a set $\mathbb{K}$ is the smallest convex set containing it. The following theorem is given to define a polytope in [GWZ93].

Proposition 7. $P \subset \mathbb{R}^n$ *is a polytope if and only if it is the convex hull of a finite set of points in* $\mathbb{R}^n$.

Generalization of Milman's Theorem [Wei76]

Proposition 8. *Let $\mathbb{K}$ be a convex locally compact subset of a Hilbert space $\mathcal{H}$. Then for any compact subset $\mathbb{L}$ of $\mathbb{K}$ the closed convex hull $\bar{co}\mathbb{L}$ of $\mathbb{L}$ in $\mathcal{H}$ is compact and contained in $\mathbb{K}$.*

This proposition also hold in a locally convex linear space. The fact is that every Banach space is a locally convex linear space. So, therefore, is a Hilbert space. See Page 107 in [FHHP01] for details.

Cyclic Polytope (Page 493 in [GWZ93])

Given integer $n \geq 2$ and $k \geq n+1$, take the convex hull of any n distinct points on the moment curve $(x, x^2, \cdots, x^n)$. The combinatorial structure of the resulting simplicial n-polytope is independent of the actual choice of points, and such polytope is called *cyclic n-polytope with k vertices*.

Infinite Dimensional Simplex [AICMV09]

A distribution on R^∞ is a point on the infinite-dimensional simplex

$$\Delta^\infty = \left\{ f(x_i) \in \mathbb{R}^\infty | \sum f(x_i) = 1, f(x_i) \geq 0 \right\}.$$

Manifold (Page 4 in [KSM99])

A *topological manifold* is a separable Hausdorff space $\mathbb{M}$ which is locally homeomorphic to $\mathbb{R}^n$.

Fibre and Bundle (Page 6 in [Sau89])

Definition 12. *A fibred manifold is a triple $(\mathcal{E}, \pi, \mathbb{M})$ where $\mathcal{E}$ and $\mathbb{M}$ are manifolds and $\pi : \mathcal{E} \to \mathbb{M}$ is a surjective submersion. $\mathcal{E}$ is the total space, π the projection, and $\mathbb{M}$ the base space.*

Definition 13. *If $(\mathcal{E}, \pi, \mathbb{M})$ is a fibred manifold and $p \in \mathbb{M}$ then a local trivialisation of π around p is a triple (W_p, F_p, t_p) where W_p is a neighborhood*

of p, F_p is a manifold and $t_p : \pi^{-1}(W_p) \to W_p \times F_p$ is a diffeomorphism satisfying the condition

$$pr_1 \circ t_p = \pi|_{\pi^{-1}(W_p)}.$$

A fibred manifold which has at least one local trivialistation around each point of its base space is known as a bundle.

Amari use such structure in [Ama85] for statistical inference. The structure is explained explicitly by Marriott in [MV04].

Jet Space (Page 161 in [Sau89])

Definition 14. *Let $(E, \pi, \mathbb{M})$ be a bundle, and let $p \in \mathbb{M}$. Define the local section $\phi, \psi \in \Gamma_p(\pi)$ to be $2-$equivalent at p if $\phi(p) = \psi(p)$ and if, in some adapted coordinate system (x^i, μ^α) around $\phi(p)$,*

$$\left.\frac{\partial \phi^\alpha}{\partial x^i}\right|_p = \left.\frac{\partial \psi^\alpha}{\partial x^i}\right|_p \quad \text{and} \quad \left.\frac{\partial^2 \phi^\alpha}{\partial x^i \partial x^j}\right|_p = \left.\frac{\partial^2 \psi^\alpha}{\partial x^i \partial x^j}\right|_p$$

for $1 \le i, j \le m$ and $1 \le \alpha \le n$. The equivalence class containing ϕ is called the $2-$ject of ϕ at p.

Such structure applied for general statistical propose can be found in Page 243 in [MR93] and [BNBC+92].

Gel'fand n-width and Kolmogorov n-width [Pin86]

Definition 15. *The Gel'fand n-width of X with respect to the ℓ_2^m norm is defined as*

$$d^n(X; \ell_2^m) = \inf_{V_n} \sup\{\|x\|_{\ell_2} : x \in V_n^\perp \cap X\},$$

where the infimum is over n-dimensional linear subspace of $\mathbb{R}^m$, and $V_n^\perp$ denotes the orthogonal complement of V_n with respect to the standard Eucildean inner product.

Definition 16. *Let $X \subset \mathbb{R}^m$ be a bounded set. The Kolmogorov n-width of X with respect the ℓ_2^m norm is defined as*

$$d_n(X; \ell_2^m) = \inf_{V_n} \sup_{x \in X} \inf_{y \in V_n} \|x - y\|_{\ell_2},$$

where the infimum is over n-dimensional linear subspaces of $\mathbb{R}^m$.

These two width are equivalent. Comparing with Gel'fand n-width, Kolomogrov n-width is more widely used to evaluate the approximation.

A.2 Fisher Orthogonal and -1 Representation

In Armari's book [Ama85], the α-representation of density functions has been discussed in detail. Let $\ell_\alpha(x|\theta)$ be a one parameter family of functions defined by

$$\ell_\alpha(x|\theta) = \begin{cases} \frac{2}{1-\alpha} f(x|\theta)^{(1-\alpha)/2} & \alpha \neq 1 \\ \log f(x|\theta) & \alpha = 1 \end{cases}.$$

The -1-representation of $f(x|\theta)$ is given by

$$\ell_{-1}(x,\theta) = f(x|\theta).$$

Furthermore, the α-expectation is also introduced in [Ama85],

$$\mathbb{E}_\alpha\left[g(x)\right] = \int g(x) f(x)^\alpha dx.$$

The Fisher information can be expressed as

$$\begin{aligned} i(\theta) &= \mathbb{E}_\alpha\left[\frac{\partial}{\partial\theta}\ell_\alpha(x|\theta)^2\right] \\ &= \mathbb{E}_{-1}\left[\frac{\partial}{\partial\theta}\ell_{-1}(x|\theta)^2\right] \\ &= \int_{\mathcal{S}} f(x|\theta)^{-1}\left[\frac{\partial}{\partial\theta}f(x|\theta)\right]^2 dx. \end{aligned}$$

Correspondingly, we define two vectors g_i, g_j to be Fisher orthogonal in -1 representation if

$$\int_{\mathcal{S}} f(x|\theta)^{-1} g_i g_j dx = 0, \quad i \neq j.$$

A.3 Proof for Properties Spectral Decomposition

Proposition 9. *A continuous image of a compact space is compact.*

Lemma 1. *The integral operator $A(\cdot)$ on $\mathcal{C} \times \mathcal{C}$ is compact, self-adjoint and positive. Furthermore, the operator $A(\cdot)$ is trace-class, i.e. the sum of all eigenvalues is finite.*

Proof. For any $f, g \in \Theta$, we have

$$\begin{aligned}\langle f, Ag \rangle &= \int_{\mathcal{C}} \int_{\mathcal{C}} g(\eta_1) k(\eta_1, \eta_2) d\eta_1 f(\eta_2) d\eta_2 \\ &= \int_{\mathcal{C}} \int_{\mathcal{C}} f(\eta_2) k(\eta_1, \eta_2) d\eta_2 g(\eta_1) d\eta_1 \\ &= \langle Af, g \rangle .\end{aligned}$$

It is self-adjoint.

Furthermore,

$$\begin{aligned}\langle Ag, g \rangle &= \int_{\mathcal{C}} g(\eta_1) \int_{\mathcal{C}} \int_{\mathcal{S}} f(x|\theta, Q)^{-1} \mathcal{F}(x|\theta, \eta_1) \mathcal{F}(x|\theta, \eta_2) d\eta_1 dx g(\eta_2) d\eta_2 \\ &= \int_{\mathcal{S}} f(x|\theta, Q)^{-1} \int_{\mathcal{C}} \mathcal{F}(x|\theta, \eta_1) g(\eta_1) d\eta_1 \int_{\mathcal{C}} \mathcal{F}(x|\theta, \eta_2) g(\eta_2) d\eta_2 dx \\ &= \int_{\mathcal{S}} f(x|\theta, Q)^{-1} \left(\int_{\mathcal{C}} \mathcal{F}(x|\theta, \eta_1) g(\eta_1) d\eta_1 \right)^2 dx \\ &\geq 0\end{aligned}$$

Hence, A is a positive operator.

If η_2 have a compact support $\mathcal{C}$ and $\int_{\mathcal{C}} g(\eta_1) k(\eta_1, \eta_2) d\eta_1$ is continuous with respect to η_2, therefore $A(\cdot)$ is a compact operator by Proposition 9.

Now, we want show operator A is trace class. Since the integral operator $A(\cdot)$ is positive on a compact support $\mathcal{C}$ with a continuous $k(\eta_1, \eta_2)$ on $\mathcal{C} \times \mathcal{C}$ (proved later in Lemma 2), apply Proposition 5 to it, we have the trace of

the operator is absolutely and uniformly convergent, i.e.

$$k(\eta_1, \eta_2) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\eta_1) \phi_i(\eta_2) < \infty.$$

Set $\eta_1 = \eta_2$, we have

$$\begin{aligned} \int_{\mathcal{C}} k(\eta, \eta) d\eta &= \int_{\mathcal{C}} \sum_{i=1}^{\infty} \lambda_i \phi_i(\eta) \phi_i(\eta) d\eta \\ &= \sum_{i=1}^{\infty} \lambda_i. \end{aligned}$$

Because $k(\eta, \eta) < \infty$ and continuous in compact $\mathcal{C}$, $\int_{\mathcal{C}} k(\eta, \eta) d\eta = \sum_{i=1}^{\infty} \lambda_i < \infty$.

On the other hand, by Regularity Condition 4, we have

$$\begin{aligned} \sum_{i=1}^{\infty} \langle A\phi_i, \phi_i \rangle &= \sum_{i=1}^{\infty} \int_{\mathcal{C}} \int_{\mathcal{C}} \phi_i(\eta_2) k(\eta_1, \eta_2) d\eta_2 \phi_i(\eta_1) d\eta_1 \\ &= \sum_{i=1}^{\infty} \int_{\mathcal{C}} \int_{\mathcal{C}} \phi_i(\eta_2) \left\{ \sum_{j=1}^{\infty} \lambda_j \phi_j(\eta_1) \phi_j(\eta_2) \right\} d\eta_2 \phi_i(\eta_1) d\eta_1 \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i \int_{\mathcal{C}} \phi_j(\eta_2) \phi_i(\eta_2) d\eta_2 \int_{\mathcal{C}} \phi_j(\eta_1) \phi_i(\eta_1) d\eta_1 \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i \delta_{ij} \\ &= \sum_{i=1}^{\infty} \lambda_i \end{aligned}$$

where ϕ_i are the eigenfunctions and δ_{ij} is the Delta function. According to Definition 9, $\sum_{i=1}^{\infty} \langle A\phi_i, \phi_i \rangle < \infty$, $A(\cdot)$ is trace operator.

□

Theorem 1. *The space spanned by $s_i(x)$, $i = 1, 2, \cdots$ is a subset of the vector space V_{Mix}. The set of $s_i(x)$, $i = 1, 2, \cdots$ is a complete orthogonal system of the Hilbert space Π, and the norm of s_i is λ_i.*

Proof. First, we want show that the L_1 integral of $s_i(x)$ defined in Equation 2.11 is equal to zero over the sample space $\mathcal{S}$.

$$\begin{aligned}\int_{\mathcal{S}} s_i(x)dx &= \int_{\mathcal{S}}\int_{\mathcal{C}} \mathcal{F}(x|\theta,\eta)e_i(\eta)d\eta dx \\ &= \int_{\mathcal{C}}\int_{\mathcal{S}} \mathcal{F}(x|\theta,\eta)dx e_i(\eta)d\eta \\ &= 0.\end{aligned}$$

Next, for any vector $v(x) \in span\{s_i(x), i = 1, 2, \cdots\}$, we can write in the form $v = \sum_i \alpha_i s_i(x)$, where α_i are coefficients.

$$\int_{\mathcal{S}} v(x)dx = \sum_i \alpha_i \int_{\mathcal{S}} s_i(x)dx = 0.$$

Hence, we know $v(x) \in V_{Mix}$. Hence, the space spanned by $s_i(x)$, $i = 1, 2, \cdots$ is a subset of the vector space V_{Mix}.

For the Hilbert space Θ, we have a complete orthogonal basis $\{e_i(\eta)\}$, by which, for all $f(x|\theta,\eta) - f(x|\theta_0) \in \Theta$,

$$f(x|\theta,\eta) - f(x|\theta_0) = \langle f(x|\theta,\eta) - f(x|\theta_0), s_0(x)\rangle_{\Pi} s_0(x) + \sum_{i=1}^{\infty} s_i(x)e_i(\eta).$$

On the other hand, for all $f(x|\theta,\eta) - f(x|\theta_0)$ also locates in the Hilbert space Π, then $s_i(x)$, $i = 0, 1, 2, \cdots$, become a complete basis of Π.

For $s_i(x)$ and $s_j(x)$ in the Hilbert space Π, we have

$$\begin{aligned}\langle s_i(x), s_j(x)\rangle_{\Pi} &= \int_{\mathcal{S}} s_i(x)s_j(x)f(x|\theta,Q)^{-1}dx \\ &= \int_{\mathcal{S}}\int_{\mathcal{C}} e_i(\eta_1)\mathcal{F}(x|\theta,\eta_1)d\eta_1 \int_{\mathcal{C}} e_j(\eta_2)\mathcal{F}(x|\theta,\eta_2)d\eta_2 f(x|\theta,Q)^{-1}dx \\ &= \int_{\mathcal{C}} e_i(\eta_1)\int_{\mathcal{C}} e_j(\eta_2)\int_{\mathcal{S}} \mathcal{F}(x|\theta,\eta_1)\mathcal{F}(x|\theta,\eta_2)f(x|\theta,Q)^{-1}dx d\eta_2 d\eta_1 \\ &= \int_{\mathcal{C}} e_i(\eta_1)\lambda_j e_j(\eta_1)d\eta_1 \\ &= \lambda_i\delta_{ij},\end{aligned}$$

where δ_{ij} is the Delta function.

□

Theorem 2. *The mixture density function $f(x|\theta, Q)$ can be expanded as*

$$\int f(x|\theta,\eta)dQ(\eta) = f(x|\theta_0) + \sum_{i=0}^{\infty} \alpha_i s_i(x),$$

where

$$\alpha_0 = -\int_{\mathcal{S}} f(x|\theta, Q)^{-1} f(x|\theta_0) s_0(x) dx.$$

and

$$\alpha_i = \int_{\mathcal{C}} e_i(\eta) q(\eta) d\eta, \quad i = 1, 2, \cdots.$$

Proof. By spectral decomposition, we have the expansion

$$f(x|\theta, Q) = f(x|\theta_0) + \sum_{i=0}^{\infty} \alpha_i s_i(x),$$

where

$$\begin{aligned}
\alpha_0 &= \int \langle f(x|\theta,\eta) - f(x|\theta_0), s_0(x)\rangle_{\Pi}\, dQ(\eta) \\
&= \int_{\mathcal{C}}\int_{\mathcal{S}} f(x|\theta,Q)^{-1} f(x|\theta,\eta) s_0(x) dx q(\eta) d\eta - \int_{\mathcal{C}}\int_{\mathcal{S}} f(x|\theta,Q)^{-1} f(x|\theta_0) s_0(x) dx q(\eta) d\eta \\
&= \int_{\mathcal{S}} f(x|\theta,Q)^{-1} \int_{\mathcal{C}} f(x|\theta,\eta) q(\eta) d\eta s_0(x) dx - \int_{\mathcal{S}} f(x|\theta,Q)^{-1} f(x|\theta_0) s_0(x) dx \\
&= \int_{\mathcal{S}} s_0(x) dx - \int_{\mathcal{S}} f(x|\theta,Q)^{-1} f(x|\theta_0) s_0(x) dx \\
&= -\int_{\mathcal{S}} f(x|\theta,Q)^{-1} f(x|\theta_0) s_0(x) dx
\end{aligned}$$

For α_i, $i = 1, 2, \cdots$, according to Cauchy-Schwarz inequality, we have

that

$$\begin{aligned}\alpha_i^2 &= \left(\int_{\mathcal{C}} e_i(\eta)q(\eta)d\eta\right)^2 \\ &\leq \int_{\mathcal{C}} e_i(\eta)^2 d\eta \int_{\mathcal{C}} q(\eta)^2 d\eta \\ &\leq 1.\end{aligned}$$

□

Lemma 2. *The kernel $k(\eta_1, \eta_2)$ is in $L_2(\mathcal{C} \times \mathcal{C})$.*

Proof.

$$\begin{aligned}\int_{\mathcal{C}}\int_{\mathcal{C}} k(\eta_1,\eta_2)^2 d\eta_1 d\eta_2 &= \int_{\mathcal{C}}\int_{\mathcal{C}} \left[\int_{\mathcal{S}} f(x|\theta,Q)^{-1}\mathcal{F}(x|\theta,\eta_1)\mathcal{F}(x|\theta,\eta_2)dx\right]^2 d\eta_1 d\eta_2 \\ &\leq \int_{\mathcal{C}}\int_{\mathcal{C}}\int_{\mathcal{S}} \frac{\mathcal{F}(x|\theta,\eta_1)^2}{f(x|\theta,Q)}dx \int_{\mathcal{S}} \frac{\mathcal{F}(x|\theta,\eta_2)^2}{f(x|\theta,Q)}dx d\eta_1 d\eta_2 \\ &= \left[\int_{\mathcal{C}}\int_{\mathcal{S}} \frac{\mathcal{F}(x|\theta,\eta)^2}{f(x|\theta,Q)}dx d\eta\right]^2.\end{aligned}$$

The inequality holds because of Cauchy-Schwarz inequality. According to spectral decomposition, we have

$$\begin{aligned}\mathcal{F}(x|\theta,\eta)^2 &= \left[\sum_{i=1}^{\infty} s_i(x)e_i(\eta)\right]^2 \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} s_i(x)s_j(x)e_i(\eta)e_j(\eta),\end{aligned}$$

where $e_i(\eta)$ and $e_j(\eta)$ are orthogonal to each other, i.e. $\int_{\mathcal{C}} e_i(\eta)e_j(\eta)d\eta = 0$. Furthermore, $\|e_i(\eta)\|_{\Theta}^2 = 1$ Therefore, we have

$$\int_{\mathcal{C}}\int_{\mathcal{C}} k(\eta_1,\eta_2)^2 d\eta_1 d\eta_2 \leq \left[\int_{\mathcal{S}} \frac{s_i(x)^2}{f(x|\theta,Q)}dx\right]^2$$

By Theorem 1, we obtain

$$\int_{\mathcal{C}}\int_{\mathcal{C}} k(\eta_1,\eta_2)^2 d\eta_1 d\eta_2 \leq \left[\sum_{i=1}^{\infty}\lambda_i\right]^2 < \infty.$$

Therefore, we have $k(\eta_1, \eta_2) \in L_2(\mathcal{C} \times \mathcal{C})$. □

A.4 Proof for Convexity Structure of Mixture Models

Lemma 3. *The convex hull* $co\Pi^{\mathcal{F}}$ *of* $\Pi^{\mathcal{F}}$ *is a subset of* $\hat{\mathbb{K}}^{\mathcal{F}}$.

Proof. Since for any $g \in co\Pi^{\mathcal{F}}$, we have

$$\begin{aligned}\sum_{i=1}^{n} \beta_i f(x|\theta, \eta_i) &= f(x|\theta_0) + \sum_{i=1}^{n} \beta_i \mathcal{F}_i \\ &= f(x|\theta_0) + g,\end{aligned}$$

where $\beta_i \geq 0$, $\sum_{i=1}^{n} \beta_i = 1$ and $n \in \mathbb{N}$. Therefore, $f(x|\theta_0) + g$ is the mixture of the set of distribution $f(x|\theta, \eta)$. For all $x_i \in \mathcal{S}$, $f(x_i|\theta, \eta) > 0$. Therefore, we can have a set $co\Pi^{\mathcal{F}} \subseteq \tilde{\mathbb{K}}^{\mathcal{F}}$.

On the other hand, since for any $\mathcal{F}_i \in \Pi$ can be expanded as

$$\mathcal{F}_i = \sum_{j=0}^{\infty} e_j(\eta_i) s_j(x),$$

the convex hull $co\Pi^{\mathcal{F}}$ also has the form

$$co\Pi^{\mathcal{F}} = \left\{ g = \sum_{j=0}^{\infty} \gamma_j s_j(x) | \gamma_j = \sum_{i=1}^{n} \beta_i e_j(\eta_i), \beta_i \geq 0, \sum_{i=1}^{n} \beta_i = 1, \eta_i \in \mathcal{C}, n \in \mathbb{N} \right\}.$$

Because $\mathcal{C}$ is compact and $e_i(\eta)$ is a continuous real function of η, we know that $e_i(\eta)$ is bounded. It means

$$\inf_{\eta \in \mathcal{C}} e_j(\eta) \sum_{j=1}^{n} \beta_j \leq \sum_{j=1}^{n} \beta_j e_j(\eta) \leq \sup_{\eta \in \mathcal{C}} e_j(\eta) \sum_{j=1}^{n} \beta_j$$

i.e.

$$\gamma_j \in \left[\inf_{\eta \in \mathcal{C}} e_j(\eta), \sup_{\eta \in \mathcal{C}} e_j(\eta) \right]. \tag{A-1}$$

Therefore, we know $co\Pi^{\mathcal{F}} \subseteq \mathbb{K}^{\mathcal{F}}$. We know that $\hat{\mathbb{K}}^{\mathcal{F}} = \mathbb{K}^{\mathcal{F}} \cap \tilde{\mathbb{K}}^{\mathcal{F}}$, therefore $co\Pi^{\mathcal{F}} \subseteq \hat{\mathbb{K}}^{\mathcal{F}}$. □

Theorem 3. *The closed convex hull* $\overline{co\Pi^{\mathcal{F}}}$ *of* $\Pi^{\mathcal{F}}$ *is compact and closed in the closure of* $\hat{\mathbb{K}}^{\mathcal{F}}$.

Proof. The closure of $\hat{\mathbb{K}}^{\mathcal{F}}$ is given by

$$\overline{co\hat{\mathbb{K}}^{\mathcal{F}}} = \left\{ g = \sum_{j=0}^{\infty} \beta_j s_j(x) | \beta_j \in \left[\min_{\eta \in \mathcal{C}} e_j(\eta), \max_{\eta \in \mathcal{C}} e_j(\eta) \right], g(x_i) + f(x_i|\theta_0) \geq 0, x_i \in \mathcal{S} \right\}.$$

The map from $\mathcal{C}$ to $\overline{co\hat{\mathbb{K}}^{\mathcal{F}}}$ is continuous, by Proposition 9, we have $\overline{co\hat{\mathbb{K}}^{\mathcal{F}}}$ is compact. It is a convex locally compact subset of a Hilbert space.

On the other hand, $\overline{co\Pi^{\mathcal{F}}}$ is the closed convex hull of $\Pi^{\mathcal{F}} \subseteq \overline{co\hat{\mathbb{K}}^{\mathcal{F}}}$, then by Proposition 8, we have that the closed convex hull $\overline{co\Pi^{\mathcal{F}}}$ of $\Pi^{\mathcal{F}}$ is compact and closed in the closure of $\hat{\mathbb{K}}^{\mathcal{F}}$.

□

Theorem 4. *In the Hilbert space* Π, *if* $U_{\Pi_K}\mathcal{F}(x;Q) \in \hat{\mathbb{K}}_K^{\mathcal{F}}$, *the norm of the vector from the mixture density* $f(x|\theta, Q)$ *with a compact* $\mathcal{C}$ *and mixture models* $f(x|\theta, \alpha)$ *by PCA has the order* $o(K^{-1})$.

Proof. First, by Lemma 2, we know that the kernel $k(\eta_1, \eta_2)$ is in $L_2(\mathcal{C} \times \mathcal{C})$.

Next, we want show that $k(\eta_1, \eta_2)$ is continuously differentiable. We know that

$$\mathcal{F}(x|\theta, \eta) = f(x|\theta, \eta) - f(x|\theta_0) - \frac{\partial}{\partial \theta} f(x|\theta).$$

It is continuously differentiable with respect to η because of Regularity Condition 1. Take derivative of $k(\eta_1, \eta_2)$ with respect to η_1

$$\begin{aligned} \frac{\partial}{\partial \eta_1} k(\eta_1, \eta_2) &= \frac{\partial}{\partial \eta_1} \int_{\mathcal{S}} f(x|\theta, Q)^{-1} \mathcal{F}(x|\theta, \eta_1) \mathcal{F}(x|\theta, \eta_2) dx \\ &= \int_{\mathcal{S}} f(x|\theta, Q)^{-1} \mathcal{F}(x|\theta, \eta_2) \frac{\partial}{\partial \eta_1} \mathcal{F}(x|\theta, \eta_1) dx. \end{aligned}$$

Because $\frac{\partial}{\partial \eta_1}\mathcal{F}(x|\theta, \eta_1)$ is continuous, $\frac{\partial}{\partial \eta_1} k(\eta_1, \eta_2)$ is continuous with respect to η_1. Similarly, we know $\frac{\partial}{\partial \eta_2} k(\eta_1, \eta_2)$ is continuous with respect to η_2.

Then we know $k(\eta_1, \eta_2) \in L_2(\mathcal{C} \times \mathcal{C})$ is continuously differentiable. Furthermore, it is also a positive definite kernel. Then by [Rea83],

$$\sum_{i=K+1}^{\infty} \lambda_i = o(K^{-1}).$$

□

A.5 Proof for Properties of Dependence on Choice of Compact Region

Lemma 4. *Let*

$$v_i(x) = f(x|\theta, Q)^{-1/2} s_i(x),$$

then

$$\int_{\mathcal{C}} e_i(\eta_1) \int_{\mathcal{S}} f(y|\theta, Q)^{-1} \mathcal{F}(y|\theta, \eta_1) \mathcal{F}(y|\theta, \eta_2) dy d\eta_1 = \lambda e_i(\eta_2)$$

is equivalent to,

$$\int_{\mathcal{S}} v_i(x) \int_{\mathcal{C}} \frac{\mathcal{F}(x|\theta, \eta)}{\sqrt{f(x|\theta, Q)}} \frac{\mathcal{F}(y|\theta, \eta)}{\sqrt{f(y|\theta, Q)}} d\eta dx = \lambda v_i(y).$$

Proof. Rewrite Equation (2.10)

$$\int_{\mathcal{C}} e_i(\eta_1) \int_{\mathcal{S}} \frac{\mathcal{F}(y|\theta, \eta_1)}{\sqrt{f(y|\theta, Q)}} \frac{\mathcal{F}(y|\theta, \eta_2)}{\sqrt{f(y|\theta, Q)}} dy d\eta_1 = \lambda e_i(\eta_2)$$

By Regularity Condition 4, we can change the order of integration, and get

$$\int_{\mathcal{S}} \frac{\mathcal{F}(y|\theta, \eta_2)}{\sqrt{f(y|\theta, Q)}} \int_{\mathcal{C}} e_i(\eta_1) \frac{\mathcal{F}(y|\theta, \eta_1)}{\sqrt{f(y|\theta, Q)}} d\eta_1 dy = \lambda e_i(\eta_2)$$

Note that

$$v_i(y) = \int_{\mathcal{C}} e_i(\eta_1) \frac{\mathcal{F}(y|\theta, \eta_1)}{\sqrt{f(y|\theta, Q)}} d\eta_1,$$

so,

$$\int_{\mathcal{S}} \frac{\mathcal{F}(y|\theta, \eta_2)}{\sqrt{f(y|\theta, Q)}} v_i(y) dy = \lambda e_i(\eta_2).$$

Combine two equations above, we obtain

$$\int_{\mathcal{C}} \int_{\mathcal{S}} \frac{\mathcal{F}(x|\theta, \eta_1)}{\sqrt{f(x|\theta, Q)}} v_i(x) dx \frac{\mathcal{F}(y|\theta, \eta_1)}{\sqrt{f(y|\theta, Q)}} d\eta_1 = \lambda v_i(y),$$

Based on Regularity Condition 4, it can be simplified as

$$\int_{\mathcal{S}} v_i(x) \int_{\mathcal{C}} \frac{\mathcal{F}(x|\theta,\eta_1)}{\sqrt{f(x|\theta,Q)}} \frac{\mathcal{F}(y|\theta,\eta_1)}{\sqrt{f(y|\theta,Q)}} d\eta_1 dx = \lambda v_i(y).$$

□

Proposition 10. *Weyl's inequality: Define $\lambda_n^{(C)}$ to be the nth eigenvalue of the operator $C(\cdot)$ ordered from lowest to highest. Let S_n be an n-dimensional subspace of a Hilbert space $\mathcal{H}$ and L_n be a subspace of dimension less than or equal to n. Assume that operator $C(\cdot)$ is the sum of operators $A(\cdot)$ and $B(\cdot)$. we have*

$$\lambda_{a+b-1}^{(C)} \geq \lambda_a^{(A)} + \lambda_b^{(B)}.$$

Details of the proof can be found in [Mar07a]. Proposition A-4 is needed to prove Theorem 5. In [DM05], it is known that

Proposition 11. *All eigenvalues of a positive operator are non-negative.*

Consider the integral operator $C(\cdot)$ on Θ associated with the kernel $k^*(x,y)$ is defined by

$$(Cf)(x) = \int_{\mathcal{S}} f(y) k^*(x,y) dy.$$

where

$$k^*(x,y) = \int_{\mathcal{C}} \frac{\mathcal{F}(y|\theta,\eta)}{\sqrt{f(y|\theta)}} \frac{\mathcal{F}(x|\theta,\eta)}{\sqrt{f(x|\theta)}} d\eta.$$

Now, we can prove Theorem 1.

Theorem 5. *With the region of η expansion, the eigenvalues $\lambda_i, i \geq 1$ for the expansion increase.*

Proof. Assume that the regions of η is $\mathcal{C}$, and $\Delta\mathcal{C}$ is the part of expansion. To the positive, compact operator

$$\begin{aligned}(C + \Delta C)g &= \int_{\mathcal{S}} g \int_{\mathcal{C}+\Delta\mathcal{C}} \frac{\mathcal{F}(y|\theta,\eta)}{\sqrt{f(y|\theta)}} \frac{\mathcal{F}(x|\theta,\eta)}{\sqrt{f(x|\theta)}} d\eta dy \\ &= Cg + \Delta Cg,\end{aligned}$$

where

$$Cg = \int_{\mathcal{S}} g \int_{\mathcal{C}} \frac{\mathcal{F}(y|\theta,\eta)}{\sqrt{f(y|\theta)}} \frac{\mathcal{F}(x|\theta,\eta)}{\sqrt{f(x|\theta)}} d\eta dy$$
$$\Delta Cg = \int_{\mathcal{S}} g \int_{\Delta\mathcal{C}} \frac{\mathcal{F}(y|\theta,\eta)}{\sqrt{f(y|\theta)}} \frac{\mathcal{F}(x|\theta,\eta)}{\sqrt{f(x|\theta)}} d\eta dy.$$

It is easy to show that ΔC is positive operator. According to Proposition 6, all eigenvalues of B is non-negative, i.e. $\lambda_1^{(\Delta C)} \geq 0$. Apply Weyl's inequality, we show that to any $i \geq 1$,

$$\lambda_i^{(\mathcal{C}+\Delta\mathcal{C})} \geq \lambda_i^{(\mathcal{C})}.$$

□

All we have done above are theoretical, things are much harder in practical. The main reason is that we can not always obtain a closed form for the eigenfunction equation Equation 2.10. In such case, we need to approximate the eigenfunction equation by a eigenvector equation. The discretization of the operators and the behavior of approximated eigenvalues is discussed in [SZ08]. Assume that $X = (x_1, x_2, \cdots, x_m)^T$ is a sample independently drawn according to an uniform distribution over the support $\mathcal{C}$. We introduce a sampling operator $R_X : (\Pi, k^*(\cdot,\cdot)) \to \ell_2$ such that

$$R_X(g) = (g(x_1), g(x_2), \cdots, g(x_m))^T.$$

The adjoint of the sampling operator, $R_X^T : \ell_2 \to (\Pi, k^*(\cdot,\cdot))$ is given by

$$R_X^T v = \sum_{i=1}^{m} v_i k^*(x_i, \cdot), \quad v \in \ell^2.$$

In [SZ08], it is pointed out that the operator $\frac{1}{m} R_X^T R_X$, denoted by J, converges to the integral operator $C(\cdot)$, when the the number of pieces m tends to infinity.

Proposition 12. *Assume*

$$\kappa := \sqrt{\sup_{x \in \mathcal{S}} k(x,x)} < \infty.$$

Let X be a sample independently drawn from a $f(x|\theta, Q)$ distribution of $\mathcal{S}$. With confidence $1-\delta$, we have

$$\|J - A\|_{HS} \leq \frac{4\kappa^2 \log(2/\delta)}{\sqrt{m}},$$

where $\|\cdot\|_{HS}$ is the Hilbert-Schmidt norm defined in **Appendix A**.

Furthermore, it is also give a proposition for approximation of the eigenvalues and eigenvectors from two operators.

Proposition 13. *Let $A(\cdot)$ and $\hat{A}(\cdot)$ be two compact positive definite operators on a Hilbert space $\mathcal{H}$, with nondecreasing eigenvalues $\{\lambda_i\}$ and $\{\hat{\lambda}_i\}$ with multiplicity. Then, there holds*

$$\max_{j \geq 1} \|\lambda_j - \hat{\lambda}_j\| \leq \|A - \hat{A}\|_{HS}.$$

According to these propositions, we can replace the eigenfunction equation

$$(Cf)(x) = \lambda f(x)$$

by

$$JS = \lambda S,$$

where

$$S = (s(x_1), s(x_2), \cdots, s(x_m))^T,$$

while m is large enough. Denote the sum of m largest eigenvalues of matrix J by σ_m,

$$\sigma_m = \sum_{i=1}^{m} \lambda_i,$$

where λ_i are ordered by descent.

In [HB04], let $\mathcal{M}_{n,m}$ stands for the set of $n \times m$ real matrices. It is mentioned that

Proposition 14. *Let J be a $n \times n$ real symmetric matrix,*

$$\sigma_m(J) = \max\{tr(JY) : Y \in R_m\},$$

with

$$R_m := \{XX^T : X \in \mathcal{M}_{n,m}(\mathbb{R}), X^TX = I_m\};$$

and the convex hull of R_m, Ω_m is

$$\begin{aligned}\Omega_m &:= co\{XX^T : X \in \mathcal{M}_{n,m}(\mathbb{R}), X^TX = I_m\} \\ &= \{C \geq 0 : trC = m, \lambda_1(C) \leq 1\}.\end{aligned}$$

By the proposition, we can prove Lemma 5, which is used in the proof of Theorem 6.

Lemma 5. *Whenever $k \leq l$, we have*

$$\frac{\sigma_l}{l} \leq \frac{\sigma_k}{k}.$$

Proof. The convex hull Ω_m is

$$\Omega_m = \{C \geq 0 : trC = m, \lambda_1(C) \leq 1\}.$$

Then, we have

$$\frac{\Omega_m}{m} = \left\{C \geq 0 : trC = 1, \lambda_1(C) \leq \frac{1}{m}\right\}.$$

By the form above, we know, for any $k \leq l$,

$$\frac{\Omega_l}{l} \leq \frac{\Omega_k}{k}. \tag{A-2}$$

On the other hand, let J is the $n \times n$ real symmetric matrix,

$$R_m(J) := \{tr(JY) : Y \in R_m\}, \quad 1 \leq m \leq n,$$

are m-numerical ranges of J. Geometrically they represent the shadow of R_m along the line directed by J. Clearly, the bounds of $R_m(A)$ and those of

$$\Omega_m(J) = \{tr(JC) : C \in \Omega_m\}$$

are the same. In other words,

$$\max R_m(A) = \max \Omega_m(A), \quad \min R_m(A) = \min \Omega_m(A).$$

Combine the fact with equation (A-2), we have

$$\frac{\sigma_l}{l} \leq \frac{\sigma_k}{k}.$$

□

Theorem 6. *The larger the domain of η is, the more number of $s_i(x)$ are needed to contribute Π_α.*

Proof. In other words, we want to prove that

$$f(J(\mathcal{C})) = \frac{\sigma_m(J(\mathcal{C}))}{\sigma_N(J(\mathcal{C}))}, \tag{A-3}$$

where $J(\mathcal{C})$ is a symmetric matrix and $\mathcal{C}$ is the domain of η, decreases while $\mathcal{C}$ is enlarged with direction H, which is a positive-definite symmetric matrix. Differential (A-3) with respect to matrix J,

$$\frac{df}{dJ} = \frac{\sigma'_m(J,H)\sigma_N(J) - \sigma'_N(J,H)\sigma_m(J)}{\sigma_N^2(J)}.$$

Since $\sigma_N(J,H) > 0$, our aim is to show

$$\alpha \leq \frac{\sigma'_m(J,H)}{\sigma_m(J)} \leq \frac{\sigma'_N(J,H)}{\sigma_N(J)}$$

In [HB04], we know that

$$\sigma'_m(J,H) = \max\{tr(CH), C \in \partial\sigma_m(J)\}$$

where

$$\partial_m(J) = \{C \geq 0, tr(C) = m, \lambda_1(C) \leq 1, tr(JC) = \sigma_m(J)\},$$

and

$$\frac{\sigma'_m(J,H)}{\sigma_m(J)} = \max(tr(DH)), D \in \Omega_1,$$

where

$$\begin{aligned}\Omega_1 &= \frac{1}{\sigma_m}\partial\sigma_m(J) \\ &= \{D_1 \geq 0, tr(D_1) = \frac{m}{\sigma_m(J)}, \lambda_1(D_1) \leq \frac{1}{\sigma_m(J)}, tr(JD_1) = 1\}.\end{aligned}$$

Similarly, we have

$$\frac{\sigma'_{m+1}(J,H)}{\sigma_{m+1}(J)} = \max(tr(DH)), D \in \Omega_2,$$

where

$$\begin{aligned}\Omega_2 &= \frac{1}{\sigma_{m+1}}\partial\sigma_{m+1}(J) \\ &= \{D_2 \geq 0, tr(D_2) = \frac{m+1}{\sigma_{m+1}(J)}, \lambda_1(D_2) \leq \frac{1}{\sigma_{m+1}(J)}, tr(JD_2) = 1\}.\end{aligned}$$

$\exists D_1 \in \Omega_1$,

$$\frac{\sigma'_m(J,H)}{\sigma_m} = tr(D_1H);$$

$\exists D_2 \in \Omega_2$,

$$\frac{\sigma'_{m+1}(J,H)}{\sigma_{m+1}} = tr(D_2H).$$

The inequality could be proved as follows:

$$\begin{aligned}
\frac{\sigma'_m(J,H)}{\sigma_m(J)} - \frac{\sigma'_{m+1}(J,H)}{\sigma_{m+1}(J)} &= tr(D_1 H) - tr(D_2 H) \\
&= tr\left(D_1 H - D_2 H\right) \\
&= tr\left((D_1 - D_2)H\right) \\
&\leq tr\left(D_1 - D_2\right) tr\left(H\right) \\
&= (tr(D_1) - tr(D_2))tr(H) \\
&= \left(\frac{m}{\sigma_m} - \frac{m+1}{\sigma_{m+1}}\right) tr(H) \qquad \text{(A-4)} \\
&\leq 0 \qquad \text{(A-5)}
\end{aligned}$$

(A-4) to (A-5) is because of Lemma 5. □