Author:
Spiteri, Jake D

Title:
Nonparametric Estimation with Kernel Mean Embeddings

Nonparametric Estimation with Kernel Mean Embeddings

By

JAKE SPITERI

School of Mathematics
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Science.

SEPTEMBER 2023

Word count: 43556

Abstract

This thesis uses kernel mean embeddings to develop novel nonparametric methods for estimating hidden Markov models (HMMs) and state-space models (SSMs).

Our first result formalizes and generalizes a remark made in Song et al. [2014] to show that given an estimated embedding of a probability distribution in a reproducing kernel Hilbert space (RKHS), a quantity referred to as an estimated kernel mean embedding, the density of the embedded distribution can be estimated consistently at no additional cost. This is a substantial result towards the understanding of RKHS embeddings, and opens up the opportunity for end-to-end modelling using kernel mean embeddings. We show that the result can be used to estimate conditional densities using conditional mean embeddings, and demonstrate that empirically this outperforms existing kernel-based methods for conditional density estimation.

Our second contribution is the proposal of a nonparametric method for estimating hidden Markov models. The method we propose estimates the RKHS embeddings of the distributions that characterize the HMM, using spectral theory and the decomposition of several operators. In this setting we also propose a novel model-based kernel Bayes' rule which allows for inference in the HMM without estimating the underlying densities. We show empirically that our method outperforms a related method with an order of magnitude less data.

Our final contribution is in the proposal of a nonparametric method for estimating state-space models under minimal assumptions. The estimation of a nonparametric state-space model poses significant difficulty without strong assumptions on the underlying process, and there are few existing nonparametric methods in this setting. To the best of our knowledge our assumptions are the most general, and we show empirically that our proof of concept method captures the underlying dynamics of several synthetic models.

DEDICATION AND ACKNOWLEDGEMENTS

First and foremost, I extend my sincere gratitude to my supervisors, Dr Mathieu Gerber and Professor Anthony Lee. This thesis would not be the same without their thoughtful guidance. Research has brought its fair share of highs and lows; their persistent optimism has given me the courage to overcome obstacles, and their sage advice has guided me in moments of uncertainty. I am immensely grateful for their patience during my PhD journey and for continually urging me to consider the broader perspective. They taught me what it means to do research, and how to find joy in the process. I will always hold dear the fond memories of our time working together.

I am very grateful to my PhD examiners Dr Song Liu and Professor Pierre Alquier for taking the time to read this thesis, and for providing invaluable feedback that has undoubtedly improved the quality of this work.

My experience at the University of Bristol has been enriched by my peers and colleagues. Special thanks go to those whom I have had the opportunity to work closely with over the years: Daniel Williams, Dominic Owens, Michael Whitehouse, Mingxuan Yi, and Yuan Chen. Your friendship has been invaluable.

Finally, no acknowledgement would be complete without expressing my profound gratitude to my family. My father, Michael Spiteri, has been an unwavering pillar of support throughout my life. His sacrifices have shaped me into who I am today, and I would not be where I am without him. My mother, Josephine Spiteri, remains forever in my thoughts. I will always cherish the memories that we have shared and I hope to have made her proud. My siblings: Michael, Jessica, and Amy, deserve recognition for their constant support and advice. I owe an immeasurable debt of gratitude to my dear wife, Bingshin Tsai, whose love and encouragement have been my strength; I could not have persevered throughout this journey without her by my side. A special acknowledgement goes to Bingshin's family, particularly Minhua Tsai and Jingaw Tsai, for welcoming me as one of their own.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .... DATE: ....

Table of Contents

Page

List of Tables

Table **Page**

List of Figures

Figure **Page**

Chapter 1

Introduction

Kernel methods have seen widespread adoption in the statistics and machine learning literature since their inception, and have enabled the development of many non-linear algorithms for learning complex patterns [Schölkopf and Smola, 2001]. Kernels have enabled the development of non-linear methods for dimensionality reduction [Schölkopf et al., 1997], regression [Vovk, 2013], and classification [Cristianini and Shawe-Taylor, 2000], to name a few. In this classical setting, data points are mapped to a high-dimensional (and often infinite-dimensional) space in which intricate patterns within the data become evident and can be addressed using a linear method. Over the last 20 years this idea has been generalized: rather than embed data points into a reproducing kernel Hilbert space, we can embed probability distributions [Berlinet and Thomas-Agnan, 2011, Smola et al., 2007].

The embedding of probability distributions in reproducing kernel Hilbert spaces, often called kernel mean embeddings, provides a flexible framework for conducting nonparametric inference. Kernel mean embeddings can be estimated with a set of samples and manipulated via equivalents of the sum, product, and Bayes' rule to obtain embeddings of other distributions of interest without parametric assumptions [Song et al., 2009, Muandet et al., 2017]. The framework has seen applications in several areas, such as probabilistic modelling [Fukumizu et al., 2013], statistical inference [Gretton et al., 2006], and causal inference [Lopez-Paz et al., 2015]. Kernel mean embeddings allow us to compute expectations of functions which belong to the RKHS, however in general recovering information from embedded probability distributions is a non-trivial task, and precisely what information can be recovered from an embedding has remained an open question.

Hidden Markov models and state-space models are popular statistical models used for modelling time series. Since their inception [Baum and Petrie, 1966] they have seen extensive use in fields such as speech recognition [Rabiner, 1989], finance [Taylor, 1982], and bioinformatics

[Stanke et al., 2003]. Both HMMs and SSMs have the same underlying structure that consists of an unobservable latent process that is a Markov chain, and an observable process that depends on the latent process. The HMM assumes that the latent process takes values in a finite set, whilst the SSM relaxes this assumption such that the latent process takes continuous values. HMMs allow for elegant identifiability results even in the nonparametric setting [Gassiat et al., 2016], and their simplicity allows for analytic inference procedures [Rabiner, 1989]. On the other hand, the generalization to a continuous latent space poses significant difficulties: state-space models are not identifiable without strong parametric assumptions, and in many cases one must use computational methods such as particle filters in order to conduct inference [Doucet et al., 2001, Chopin et al., 2020].

This thesis presents several novel contributions.

1. We answer an open question in the kernel mean embedding literature regarding what information can be recovered from embeddings: we produce a density estimator that is uniformly consistent in probability which can be obtained at no additional computational cost, under reasonable assumptions on the kernel mean embedding. We apply the estimator to conditional mean embeddings to obtain a conditional density estimate, and show empirically that this estimator outperforms existing kernel-based approaches to conditional density estimation. The recovery of densities from estimated embeddings paves the way for a new generation of kernel-based algorithms, enabling downstream tasks and the possibility for end-to-end statistical modelling using kernel mean embeddings.

2. We propose a novel nonparametric method for estimating hidden Markov models using kernel mean embeddings and spectral theory, motivated by recent identifiability results for nonparametric HMMs. We use our density estimator to perform inference in the filtering task, and develop a novel alternative kernel Bayes' rule that allows for inference in the HMM without estimation of the densities. We also derive an estimator of the HMM order that is almost surely consistent. Our experiments show that our method outperforms a related method with an order of magnitude less data, and our order estimator is capable of correctly estimating the HMM order when the related method cannot. Performing inference using the alternative kernel Bayes' rule produces a significant improvement in performance, and we are not aware of any other nonparametric methods for HMMs that completely avoid estimating the underlying HMM densities.

3. We develop a novel nonparametric method for estimating a state-space model using kernel mean embeddings under minimal assumptions on the data-generating process. We prove that without additional assumptions the model is non-identifiable, and we use this to our advantage as the non-identifiability of the model allows us to sample latent states from a distribution of our choice. We derive a decomposition of a kernel mean embedding in terms of RKHS operators and use the decomposition to motivate an optimization problem.

There are very few existing methods in this setting, and to the best of our knowledge our assumptions on the data-generating process are the most general. We show in our experiments that this proof of concept approach to estimating nonparametric SSMs can effectively capture the dynamics of the underlying system.

1.1 Notation and assumptions

Throughout the thesis we use the following notation. Let $\mathscr{Y}$ and $\mathscr{X}$ denote the observation space and latent space respectively, then we define reproducing kernel Hilbert spaces $\mathscr{H}_{\mathscr{Y}}$ and $\mathscr{H}_{\mathscr{X}}$ on $\mathscr{Y}$ and $\mathscr{X}$ with associated kernel functions $k : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$ and $l : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$. We frequently reference the canonical feature maps induced by the kernel functions, which satisfy $\phi_Y(y) = k(y, \cdot)$ and $\varphi_X(x) = l(x, \cdot)$ for $y \in \mathscr{Y}$ and $x \in \mathscr{X}$. The canonical feature maps satisfy the reproducing property, and in particular

$$k(y, y') = \langle \phi_Y(y), \phi_Y(y') \rangle_{\mathscr{H}_{\mathscr{Y}}}, \quad \text{and} \quad l(x, x') = \langle \varphi_X(x), \varphi_X(x') \rangle_{\mathscr{H}_{\mathscr{X}}},$$

for $y, y' \in \mathscr{Y}$ and $x, x' \in \mathscr{X}$.

To ensure the existence of the RKHS embeddings, we require the following assumption.

Assumption 1.1. *The kernel functions on $\mathscr{Y}$ and $\mathscr{X}$ are measurable and bounded, and their associated reproducing kernel Hilbert spaces are separable.*

These assumptions on the kernel function are not restrictive: many popular kernel functions such as the Gaussian, Laplace, and Matérn kernels are bounded and measurable, and separability of their associated RKHSs follows when the underlying space is a separable Borel space or an analytic subset of a Polish space, such as $\mathbb{R}^d$, for $d \geq 1$ [Owhadi and Scovel, 2017].

1.2 Background

This section provides all of the background needed in the following chapters.

1.2.1 Hidden Markov and state-space models

Hidden Markov models and state-space models are statistical models used to model time series. The models consist of two processes: an observable process and an unobservable hidden/latent process (we use the phrases hidden and latent interchangeably). The hidden process is a Markov chain and the observed process is conditionally independent of all past values of the hidden process and all past and future values of the observable process, given the hidden process at the current time. An HMM differs from an SSM in that the hidden process of an HMM only take values from a finite set and thus one can think of a hidden Markov model as a special case of a state-space model. They are both characterized by the initial distribution of the hidden process, the transition distribution of the hidden process, and the distribution of the observable process given the hidden process. We refer to these quantities as the parameters of the model, and in statistical modelling we aim to estimate these parameters from a sequence of observations. A directed acyclic graph depicting HMMs and SSMs is given in Figure 1.1.

Hidden Markov models and state-space models are well-studied in the statistics literature, however a majority of the literature focuses on parametric modelling. Such models are specified by making distributional assumptions on the model, characterized by a finite-dimensional parameter which must be estimated via maximum likelihood estimation (MLE). This is often done using the Expectation-Maximization algorithm, which can suffer from slow convergence and convergence to sub-optimal local extrema. For HMMs, the MLE is consistent and asymptotically normal when the observation distribution, that is the distribution of the observable process given a hidden state, has finite support [Baum and Petrie, 1966], and these results have been extended to SSMs. See Douc et al. [2004] and references therein for details.

Nonparametric latent variable models have been become increasingly popular in applied sciences and have been applied to problems such as climate modelling [Lambert et al., 2003], genomics [Yau et al., 2011], animal movement [Langrock et al., 2015, 2018], speech recognition [Couvreur and Couvreur, 2000], facial expression recognition [Shang and Chan, 2009], and biology [Volant et al., 2014]. Their popularity stems from the fact that parametric modelling can often lead to poor performance. Specifying a parametric family of distributions that encompasses a wide-class of models whilst resulting in a computationally tractable model can be very difficult. When the underlying distributions are multi-modal, skewed, or heavy-tailed then specifying a parametric model becomes particularly difficult, and a poor choice of model assumptions may lead to poor performance in inference tasks. However, parametric models have their advantages: given domain expertise and reasonable assumptions one may produce a parametric model that adequately models data and captures meaningful details regarding the underlying system.

In many cases the hidden process is of primary interest. Hence, common inference tasks for the HMM involve inferring the hidden states given observations of the observable process. When the hidden process only takes finite values, such inference tasks can be performed analytically for an estimated model via the forward-backward algorithm. When the hidden process can take continuous values, inference procedures are only tractable under strong assumptions such as linearity and Gaussianity of the observation distribution and Markov transition, as finite sums become integrals in the continuous regime. When modelling a non-linear dynamical system one often uses approximate methods such as sequential Monte Carlo.

See Cappé et al. [2009] and Chopin et al. [2020] for a comprehensive introduction to hidden Markov models and state-space models.

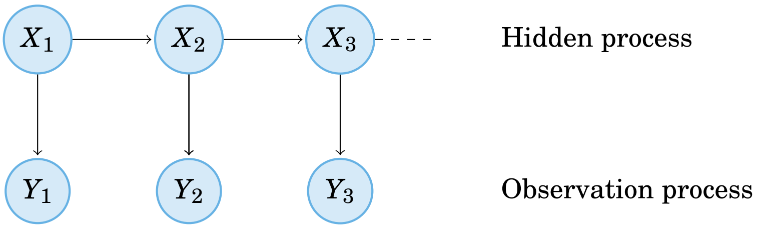


Figure 1.1: A directed acyclic graph describing hidden Markov models and state-space models, with observations (Y_1, Y_2, Y_3) and hidden states (X_1, X_2, X_3).

1.2.1.1 Hidden Markov models

A hidden Markov model consists of a pair of discrete-time stochastic processes $(X_t)_{t\geq 1}$ and $(Y_t)_{t\geq 1}$, referred to as the hidden process and observable process respectively. Let K and d be positive integers, and let $\mathscr{X}$ denote the set of hidden states $\{1,\dots,K\}$, and $\mathscr{Y} \subset \mathbb{R}^d$ the observation space. The hidden process $(X_t)_{t\geq 1}$ is a Markov chain on $\mathscr{X}$ with $K \times K$ transition matrix $\boldsymbol{Q}$, and initial distribution $\pi \in \Delta_K$, where Δ_K is the space of probability measures on $\mathscr{X}$ identified to the $(K-1)$-dimensional simplex. The observable process $(Y_t)_{t\geq 1}$ takes values in the observation space $\mathscr{Y}$, and we assume that for any $t > 1$ the observation Y_t is conditionally independent of all other observations $(Y_i)_{i\geq 1, i\neq t}$ and previous hidden states $(X_i)_{i=1}^{t-1}$ given X_t. Together, the pair of processes $(X_t, Y_t)_{t\geq 1}$ is a hidden Markov model.

We assume that the distribution of Y_t conditional on $X_t = k$, for any $k \in \mathscr{X}$, has density f_k with respect to the Lebesgue measure on $\mathscr{Y}$. We refer to f_k as an observation density and denote by $F = \{f_1, \dots, f_K\}$ the set of observation densities with respect to the Lebesgue measure on $\mathscr{Y}$.

If the initial distribution is the stationary distribution, then the hidden Markov model is uniquely defined by the stationary distribution π, the transition matrix $\boldsymbol{Q}$, and the set of observation densities F. For this reason, we state that the HMM is parametrized by $(F, \boldsymbol{Q}, \pi)$.

Figure 1.2 depicts a hidden-Markov model with three observed states as a three-view mixture model with hidden variable X_2. This interpretation provides intuition underlying Chapter 3: Y_1, Y_2, and Y_3 are conditionally independent given X_2, and expectations over $Y_1|X_2$ and $Y_3|X_2$ can be expressed in terms of expectations over $Y_2|X_2$ and the model parameters $\boldsymbol{Q}$ and π (see Lemma 3.1 and Proposition 8 of Anandkumar et al. [2012]).

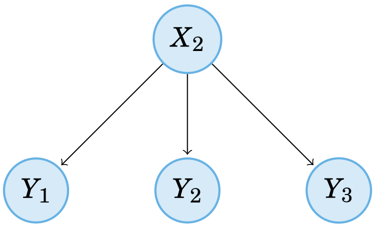


Figure 1.2: A multi-view representation of a hidden Markov model with three states.

1.2.1.2 State-space models

A state-space model consists of two processes: a latent process $(X_t)_{t\geq 1}$, and an observable process $(Y_t)_{t\geq 1}$ defined over discrete time. The latent process is Markovian, unobserved, and takes values in $\mathscr{X}$, and the observable process is observed and comprises of random variables that take values in $\mathscr{Y}$ which are conditionally independent given the latent process at time t. The conditional distribution of Y_t given $(X_1,\ldots,X_t,Y_1,\ldots,Y_{t-1},Y_{t+1},\ldots)$ depends only on X_t.

The structure and conditional dependencies of the state-space model can be visualized via the directed acyclic graph depicted in Figure 1.1. The model is determined by the conditional distributions of $Y_t|X_t$ and $X_{t+1}|X_t$, and the invariant distribution of $(X_t)_{t\geq 1}$. Typically $\mathscr{X}$ and $\mathscr{Y}$ are taken to be multidimensional Euclidean spaces, however other spaces can also be used.

Let $M_+^1(\mathscr{Z})$ denote the space of probability measures on a topological space $\mathscr{Z}$. We define $M:\mathscr{X}\to M_+^1(\mathscr{X})$ to be the Markov transition and $O:\mathscr{X}\to M_+^1(\mathscr{Y})$ the observation distribution such that $Y_t|X_t \sim O(X_t,dy_t)$ and $X_{t+1}|X_t \sim M(X_t,dx_{t+1})$. We assume that there exist probability density functions g and f such that $O(x_t,dy)=g(y|x_t)dy$ and $M(x_t,dx)=f(x|x_t)dx$ for $x_t\in\mathscr{X}$, and that the initial distribution of the Markov chain is the invariant distribution $\pi\in M_+^1(\mathscr{X})$. The collection (O,M,π) defines the state-space model, and we refer to this collection as the state-space representation.

A state-space model can also be defined by describing how the observed and latent space variables are related to sequences of noise, as seen below.

Example 1.1 (Linear Gaussian). *In a linear Gaussian state-space model the processes X_t and Y_t are linear functions of random variables X_{t-1} and X_t respectively, with independent Gaussian noise. That is,*

$$\begin{aligned} X_t &= AX_{t-1}+U_t,\\ Y_t &= BX_t+V_t, \end{aligned}$$

where A and B are $d_x\times d_x$ and $d_y\times d_x$ matrices, and U_t and V_t are vectors of i.i.d. Gaussian noise.

1.2.1.3 Inference

Several inference problems arise when working with sequential models such as hidden Markov models and state-space models. Many revolve around inferring the posterior of the hidden state X_i given a sequence of observations $y_{1:j}$, and the nature of the problem depends on the relationship between i and j:

- if $i=j+1$, we have the prediction problem wherein we must infer the next hidden state.
- if $i=j$, we have the filtering problem wherein we must infer the current hidden state.
- if $i<j$, we have the smoothing problem wherein we must infer the previous hidden states. Typically, given a sequence $y_{1:j}$ we infer all hidden states for $i<j$.

Additionally, beyond the hidden states, one might also be interested in predicting the next observation Y_{t+1} given a sequence of observations $y_{1:t}$, which can provide valuable insights into the evolution of the observable process.

1.2.2 Reproducing kernel Hilbert spaces

Definition 1.1 (Reproducing kernel Hilbert space [Aronszajn, 1950])**.** *Let $\mathscr{X}$ be a topological space, and let $\mathscr{H}$ be a Hilbert space of functions mapping from $\mathscr{X}$ to $\mathbb{R}$, equipped with inner product $\langle \cdot,\cdot \rangle_{\mathscr{H}}$. We call $\mathscr{H}$ a reproducing kernel Hilbert space if there exists a symmetric and positive-definite function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ which satisfies the following properties*

- $k(x,\cdot) \in \mathscr{H}$, $\forall x \in \mathscr{X}$,
- $\langle f, k(x,\cdot)\rangle_{\mathscr{H}} = f(x)$, $\forall x \in \mathscr{X}$, $\forall f \in \mathscr{H}$.

The function k described above is called a *reproducing kernel*, and the kernel k is uniquely defined by the RKHS. Conversely, for any symmetric and positive-definite kernel function there exists a unique RKHS for which it is the reproducing kernel [Aronszajn, 1950]. The second property is referred to as the reproducing property, and it follows that $\left\langle k(x,\cdot), k(x',\cdot)\right\rangle_{\mathscr{H}} = k(x,x')$ for $x, x' \in \mathscr{X}$. The notation $k(x,\cdot)$ is used to denote the kernel function with one argument fixed at $x \in \mathscr{X}$, and one free, and is often referred to as the canonical feature map $\phi(x) := k(x,\cdot) \in \mathscr{H}$.

Throughout this thesis we primarily use the Gaussian kernel function, however the proposed methods can be applied with a wide-variety of kernel functions. Several popular kernel functions are defined below.

Definition 1.2 (Gaussian, Laplace, and Matérn kernel functions)**.** *Let $x, x' \in \mathbb{R}^d$ for $d \geq 1$ and let $\|\cdot\|_2$ denote the L^2 norm. The Gaussian kernel function is defined as*

$$k(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\gamma^2}\right), \tag{1.1}$$

where $\gamma > 0$ is a kernel hyperparameter which determines the width of the kernel.

The Laplace kernel function is defined by

$$k(x,x') = \exp\left(-\frac{\|x-x'\|_2}{\gamma}\right), \tag{1.2}$$

where $\gamma > 0$ is a kernel hyperparameter which determines the width of the kernel.

The Matérn kernel function is defined by

$$k(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|x-x'\|_2}{\rho}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\|x-x'\|_2}{\rho}\right), \tag{1.3}$$

where $\nu > 0$ determines the smoothness, $\rho > 0$ determines the scale, Γ is the Gamma function, and K_ν is a modified Bessel function of the second kind and order ν.

See Williams and Rasmussen [2006] for an in-depth coverage of kernel functions and their usage, and Sriperumbudur et al. [2011] for a discussion of the richness and properties of their reproducing kernel Hilbert spaces.

1.2.2.1 Kernel mean embeddings

We consider random variables X and Y over topological spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively, and reproducing kernel Hilbert spaces $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ over $\mathcal{X}$ and $\mathcal{Y}$, with associated kernel functions $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and canonical feature mappings φ_X and ϕ_Y respectively.

Definition 1.3 (Kernel mean embedding [Smola et al., 2007, Berlinet and Thomas-Agnan, 2011]). *Let $\mathcal{X}$ be a topological space, and let $M^1_+(\mathcal{X})$ denote the space of probability measures on $\mathcal{X}$. The embedding of $\mathbb{P} \in M^1_+(\mathcal{X})$ in the reproducing kernel Hilbert space $\mathcal{H}_{\mathcal{X}}$ on $\mathcal{X}$ is defined by the mapping*

$$\mu_{\mathbb{P}} : M^1_+(\mathcal{X}) \to \mathcal{H}_{\mathcal{X}}, \qquad \mathbb{P} \mapsto \int l(x,\cdot) d\mathbb{P}(x), \tag{1.4}$$

where l is the kernel function associated with $\mathcal{H}_{\mathcal{X}}$, and the integral is a Bochner integral (we refer the reader to Diestel and Uhl [1977], Dinculeanu [2000] for the definition of the Bochner integral).

The following lemma provides a condition for the existence of the embedding.

Lemma 1.1 (Smola et al. [2007]). *If $\mathbb{E}_{X\sim\mathbb{P}}[\sqrt{l(X,X)}] < \infty$, then $\mu_{\mathbb{P}} \in \mathcal{H}_{\mathcal{X}}$ and expectations of functions belonging to $\mathcal{H}_{\mathcal{X}}$ can be computed as $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}_{\mathcal{X}}}$ for any $f \in \mathcal{H}_{\mathcal{X}}$.*

The kernel mean embedding defines a mapping from the probability space to a reproducing kernel Hilbert space, and kernel functions which ensure that the mapping is injective are termed *characteristic*. Characteristic kernel functions ensure that probability distributions are mapped to unique elements in the RKHS, and hence the embedding captures all of the information of the embedded distribution.

In practice we do not have direct access to the distribution $\mathbb{P}$, and we need to rely on an alternative method to compute the embedding. A simple approach is to approximate the expectation in Equation (1.4) with the sample mean. Suppose we observe n i.i.d. samples $x_1, \ldots, x_n$ from a distribution $\mathbb{P}$ on $\mathcal{X}$, then we can approximate its kernel mean embedding using the empirical measure $\mathbb{P}_n := \frac{1}{n}\delta_{x_i}$:

$$\hat{\mu}_{\mathbb{P}_n} = \int l(x,\cdot) d\mathbb{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} l(x_i,\cdot).$$

This is one of a number of ways to obtain the approximate embedding of a distribution. Other approaches may involve the use of an alternative estimator such as a shrinkage estimator [Muandet et al., 2016], the kernel sum, product or Bayes' rule as described below, or they can be obtained as the solution to a vector-valued kernel regression problem [Grünewälder et al., 2012].

Just as the mean element is used to define the embedding of a marginal distribution over $\mathcal{X}$, we may use a covariance element to define the embedding of the joint distribution of (X,Y), where X and Y are random variables on topological spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. This embedding is often referred to as the cross-covariance operator and it can be considered a mapping between RKHSs.

Definition 1.4 (Cross-covariance operator)**.** *Let (X,Y) be a random variable taking values on $\mathcal{X} \times \mathcal{Y}$, and define reproducing kernel Hilbert spaces $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let l and k be the measurable kernels associated with $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. Then the cross-covariance operator $\mathcal{C}_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ is defined as*

$$\mathcal{C}_{YX} := \mathbb{E}_{YX}[k(Y,\cdot) \otimes l(X,\cdot)] \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}},$$

where $\otimes$ denotes the tensor product. The operator is defined as the expectation of an element in the tensor product space $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$, and thus defines an element in $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$ and an operator from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}}$. Suppose $\mathbb{P}_{YX}$ is the joint distribution of (Y,X), then the cross-covariance operator is the kernel mean embedding of the joint distribution: $\mathcal{C}_{YX} = \mu_{\mathbb{P}_{YX}} \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$.

The following lemma provides a condition for the existence of the cross-covariance operator.

Lemma 1.2 (Baker [1973])**.** *If $\mathbb{E}_X[l(X,X)] < \infty$ and $\mathbb{E}_Y[k(Y,Y)] < \infty$, then $\mathcal{C}_{YX} \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$.*

The covariance operator can be estimated empirically given pairs of samples. For a set of samples $(y_i, x_i)_{i=1}^n$ from the joint distribution of (Y,X), we may estimate the cross-covariance operator as $\hat{\mathcal{C}}_{YX} := \frac{1}{n}\sum_{i=1}^n (k(y_i,\cdot) \otimes l(x_i,\cdot))$.

1.2.2.2 Conditional mean embedding

It is also possible to embed conditional distributions into reproducing kernel Hilbert spaces.

Definition 1.5 (Conditional mean embedding)**.** *Let $\mathcal{Y}$ and $\mathcal{X}$ be topological spaces. The conditional mean embedding of the conditional distribution $\mathbb{P}(Y|X=x)$, $x \in \mathcal{X}$, is defined as*

$$\mu_{Y|X=x} = \int_{\mathcal{Y}} k(y,\cdot) d\mathbb{P}(y|x) = \mathbb{E}[k(Y,\cdot)|X=x] \in \mathcal{H}_{\mathcal{Y}}. \tag{1.5}$$

The conditional mean embedding is a representer of expectation in the RKHS: for any $f \in \mathcal{H}_{\mathcal{Y}}$, $\mathbb{E}[f(Y)|X=x] = \langle f, \mu_{Y|X=x} \rangle_{\mathcal{H}_{\mathcal{Y}}}$. The following assumption is often used in the literature [Fukumizu et al., 2003, Song et al., 2009, Klebanov et al., 2020]

Assumption 1.2. *For all functions $g \in \mathcal{H}_{\mathcal{Y}}$ we have $\mathbb{E}[g(Y)|X=\cdot] \in \mathcal{H}_{\mathcal{X}}$.*

Under Assumption 1.2 there exists an operator $\mathcal{U}_{Y|X}$ mapping from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}}$ which satisfies $\mu_{Y|X=x} = \mathcal{U}_{Y|X} l(x,\cdot)$ for $x \in \mathcal{X}$. Additionally, the operator satisfies

$$\mu_{Y|X=x} = \left(\mathcal{C}_{XX}^{\dagger} \mathcal{C}_{YX}^{*}\right)^{*} l(x,\cdot), \quad x \in \mathcal{X}, \tag{1.6}$$

where A^* denotes the adjoint and $A^\dagger$ the Moore-Penrose pseudoinverse of an operator A, and the operator is also bounded [Klebanov et al., 2020]. One can think of $\mathscr{U}_{Y|X} := \left(\mathscr{C}_{XX}^\dagger \mathscr{C}_{YX}^*\right)^*$ as the kernel mean embedding of the distribution $\mathbb{P}(Y|X)$. The operator $\mathscr{U}_{Y|X}$ is referred to as the conditional mean operator, and $\mu_{Y|X=x}$ is referred to as a conditional mean embedding. Equation (1.6) shows that the operator traces out a path of conditional mean embeddings in the RKHS $\mathscr{H}_\mathscr{Y}$.

The conditional mean embedding can be approximated using Tikhonov regularization and a set of i.i.d. samples $(y_i, x_i)_{i=1}^n$ from the joint distribution of (Y, X). Let $\hat{\mathscr{C}}_{XX}$ and $\hat{\mathscr{C}}_{YX}$ denote empirical cross-covariance operators, then the empirical conditional mean embedding is

$$\hat{\mu}_{Y|X=x} = \hat{\mathscr{C}}_{YX}\left(\hat{\mathscr{C}}_{XX} + \lambda \mathscr{I}_{\mathscr{H}_\mathscr{X}^{\otimes 2}}\right)^{-1} l(x, \cdot),$$

where $\lambda > 0$ is a regularization parameter and $\mathscr{I}_{\mathscr{H}_\mathscr{X}^{\otimes 2}}$ denotes the identity operator in the tensor product RKHS $\mathscr{H}_\mathscr{X} \otimes \mathscr{H}_\mathscr{X}$. The empirical conditional mean embedding can be expressed as the weighted sum $\hat{\mu}_{Y|x} = \sum_{i=1}^n w_{n,i} k(y_i, \cdot)$, for $w_n = (K + n\lambda I_n)^{-1} l_x$, where $l_x = [l(x_1, x), \ldots, l(x_n, x)]$ and K is a kernel matrix with (i, j)-th entry $[K]_{i,j} = l(x_i, x_j)$ for $i, j \in \{1, \ldots, n\}$ [Song et al., 2009]. Hence, approximating the conditional mean embedding from a set of paired samples amounts to computing the inverse of a kernel matrix and a matrix-vector multiplication. This empirical estimator has been widely studied and is known to be consistent under certain assumptions [Grünewälder et al., 2012, Fukumizu, 2015, Park and Muandet, 2020, Li et al., 2022].

1.2.2.3 Kernel sum, product, and Bayes' rule

Embeddings can be obtained via operations such as the kernel sum, product, and Bayes' rule, and these operations provide a way to manipulate various embeddings to obtain other quantities of interest, as one would do in a purely probabilistic setting.

The following rules require that we commute the expectation and the conditional mean operator. This can be done using Proposition 1.1 when the conditional mean operator is bounded, the element being mapped is Bochner integrable, and the reproducing kernel Hilbert spaces are separable. The latter two conditions are satisfied under Assumption 1.1, and the operator is bounded under Assumption 1.2 [Klebanov et al., 2020].

Kernel sum rule. The sum rule allows us to compute the marginal distribution of a variable X given the joint distribution over variables X and Y, by marginalizing out Y.

Using the law of total expectation, we have $\mu_X = \mathbb{E}_X[\varphi_X(X)] = \mathbb{E}_Y \mathbb{E}_{X|Y}[\varphi_X(X)|Y]$. If $\mathscr{U}_{X|Y}$ is a bounded conditional mean operator, $\phi_Y(Y)$ is Bochner integrable such that $\mathbb{E}_Y[\|\phi_Y(Y)\|_{\mathscr{H}_\mathscr{Y}}] < \infty$, and the reproducing kernel Hilbert spaces $\mathscr{H}_\mathscr{X}$ and $\mathscr{H}_\mathscr{Y}$ are separable, then it follows from Proposition 1.1 that

$$\mu_X = \mathbb{E}_Y[\mathscr{U}_{X|Y} \phi_Y(Y)] = \mathscr{U}_{X|Y} \mathbb{E}_Y[\phi_Y(Y)] = \mathscr{U}_{X|Y} \mu_Y.$$

Kernel product rule. To construct the kernel product rule, we consider the tensor product feature map $\varphi_X(X) \otimes \phi_Y(Y)$. We can factorize the embedding $\mu_{XY} = \mathbb{E}_{XY}[\varphi_X(X) \otimes \phi_Y(Y)]$ in two ways using the law of total expectation, assuming that the expectation and conditional mean operators commute

$$\mathbb{E}_Y \mathbb{E}_{X|Y}[\varphi_X(X) \otimes \phi_Y(Y)|Y] = \mathscr{U}_{X|Y} \mathbb{E}_Y[\phi_Y(Y) \otimes \phi_Y(Y)],$$
$$\mathbb{E}_X \mathbb{E}_{Y|X}[\phi_Y(Y) \otimes \varphi_X(X)|X] = \mathscr{U}_{Y|X} \mathbb{E}_X[\varphi_X(X) \otimes \varphi_X(X)].$$

Let $\mu_Y^{\otimes} := \mathbb{E}_Y[\phi_Y(Y) \otimes \phi_Y(Y)]$, and $\mu_X^{\otimes} := \mathbb{E}_X[\varphi_X(X) \otimes \varphi_X(X)]$. Then we can rewrite the above as

$$\mu_{XY} = \mathscr{U}_{X|Y} \mu_Y^{\otimes} = \mathscr{U}_{Y|X} \mu_X^{\otimes}.$$

Kernel Bayes' rule. The kernel Bayes' rule (KBR) is a nonparametric method which allows for Bayesian inference in the absence of a parametric model or likelihood. In KBR we embed the prior and likelihood in an RKHS via the kernel mean embedding and cross-covariance operator respectively, and use the sum and product rules to manipulate the embeddings in the RKHS.

The presentation of KBR shown here is that given in Muandet et al. [2017], which provides a concise summary of the original work [Fukumizu et al., 2013]. Our aim is to compute the embedding of the posterior of $Y|X$ given a prior distribution $\Pi(Y)$. We obtain the embedding of the posterior distribution as $\mu_{Y|X=x} = \mathscr{U}_{Y|X}^{\Pi} \varphi_X(x)$, where we use a superscript Π to denote dependence on the prior $\Pi(Y)$. More precisely, we have

$$\mu_{Y|X=x} = \mathscr{U}_{Y|X}^{\Pi} \varphi_X(x) = \left(\left(\mathscr{C}_{XX}^{\Pi}\right)^{\dagger} \left(\mathscr{C}_{YX}^{\Pi}\right)^{*} \right)^{*} \varphi_X(x),$$

where the cross-covariance operators depend on the prior, and are given by

$$\mathscr{C}_{YX}^{\Pi} = (\mathscr{U}_{X|Y} \mathscr{C}_{YY}^{\Pi})^{T}, \quad \mathscr{C}_{XX}^{\Pi} = \mathscr{U}_{(XX)|Y} \mu_Y^{\Pi}.$$

These results follow from the product and sum rule respectively, where we have replaced the input feature map $\varphi_X(X)$ with the tensor product feature map $\varphi_X(X) \otimes \varphi_X(X)$. The embeddings μ_Y^{Π} and $\mathscr{C}_{YY}^{\Pi}$ are simply the embedded prior distribution corresponding to the feature maps $\phi_Y(Y)$ and $\phi_Y(Y) \otimes \phi_Y(Y)$.

1.2.2.4 Properties of the Bochner integral

The kernel mean embedding of a probability distribution is defined in terms of the Bochner integral, which is an extension of the Lebesgue integral that allows for elements in a complete normed vector space. Several useful properties hold for the Bochner integral, and we describe one here that will be used throughout this thesis. The following is a non-trivial property that we use to commute expectations and operators.

Proposition 1.1 (Da Prato and Zabczyk [2014, Proposition 1.6]). *Let H_1, H_2 be two separable Hilbert spaces, $T : H_1 \to H_2$ be a bounded linear operator and Z be a random variable taking values in H_1 such that $\mathbb{E}[\|Z\|_{H_1}] < \infty$. Then $\mathbb{E}[T(Z)] = T(\mathbb{E}[Z])$.*

The following lemma formalizes a result used in Chapter 3, which follows from Proposition 1.1.

Lemma 1.3. *Let Y_1, Y_2 be random variables on $\mathscr{Y}$ and let X_2 be a random variable on $\mathscr{X}$ such that Y_1 and Y_2 are conditionally independent given X_2. Let k denote a bounded kernel function on $\mathscr{Y}$ with an associated separable RKHS $\mathscr{H}_\mathscr{Y}$ and canonical feature map $\phi_Y : \mathscr{Y} \to \mathscr{H}_\mathscr{Y}$. Then*

$$\mathbb{E}_{Y_1Y_2|X_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)|X_2] = \mathbb{E}_{Y_1|X_2}[\phi_Y(Y_1)|X_2] \otimes \mathbb{E}_{Y_2|X_2}[\phi_Y(Y_2)|X_2].$$

Proof. It follows from the law of total expectation that

$$\mathbb{E}_{Y_1Y_2|X_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)|X_2] = \mathbb{E}_{Y_1|X_2}\left[\mathbb{E}_{Y_2|X_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)|X_2]\right].$$

Let $y_1 \in \mathscr{Y}$ and define the rank-one linear operator $T_{y_1} : \mathscr{H}_\mathscr{Y} \to \mathscr{H}_\mathscr{Y}^{\otimes 2}$ as $T_{y_1}(z) = \phi_Y(y_1) \otimes z$ for $z \in \mathscr{H}_\mathscr{Y}$.

The kernel function on $\mathscr{Y}$ is bounded, and therefore for all $y_1 \in \mathscr{Y}$ the operator T_{y_1} is bounded as for all $z \in \mathscr{H}_\mathscr{Y}$,

$$\|T_{y_1}(z)\|_{\mathscr{H}_\mathscr{Y}^{\otimes 2}} = \|\phi_Y(y_1) \otimes z\|_{\mathscr{H}_\mathscr{Y}^{\otimes 2}} = \|\phi_Y(y_1)\|_{\mathscr{H}_\mathscr{Y}} \|z\|_{\mathscr{H}_\mathscr{Y}} = \sqrt{k(y_1, y_1)}\|z\|_{\mathscr{H}_\mathscr{Y}}.$$

Define the $\mathscr{H}_\mathscr{Y}$-valued random variable $Z := \phi_Y(Y_2)$, then it follows from the separability of $\mathscr{H}_\mathscr{Y}$ and Proposition 1.1 that

$$\mathbb{E}_{Y_2|X_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)|X_2] = \mathbb{E}_{Z|X_2}[T_{Y_1}(Z)|X_2] = T_{Y_1}(\mathbb{E}_{Z|X_2}[Z|X_2]) = \phi_Y(Y_1) \otimes \mathbb{E}_{Y_2|X_2}[\phi_Y(Y_2)|X_2].$$

Repeating the steps outlined above, one can see that

$$\mathbb{E}_{Y_1|X_2}\left[\phi_Y(Y_1) \otimes \mathbb{E}_{Y_2|X_2}[\phi_Y(Y_2)|X_2]\right] = \mathbb{E}_{Y_1|X_2}\left[\phi_Y(Y_1)|X_2\right] \otimes \mathbb{E}_{Y_2|X_2}[\phi_Y(Y_2)|X_2].$$

Combining the above, we have the following

$$\begin{aligned}\mathbb{E}_{Y_1Y_2|X_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)|X_2] &= \mathbb{E}_{Y_1|X_2}\left[\mathbb{E}_{Y_2|X_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)|X_2]\right] \\ &= \mathbb{E}_{Y_1|X_2}\left[\phi_Y(Y_1) \otimes \mathbb{E}_{Y_2|X_2}[\phi_Y(Y_2)|X_2]\right] \\ &= \mathbb{E}_{Y_1|X_2}\left[\phi_Y(Y_1)|X_2\right] \otimes \mathbb{E}_{Y_2|X_2}[\phi_Y(Y_2)|X_2],\end{aligned}$$

which concludes the proof. ■

1.2.2.5 Tensor product spaces

Throughout this thesis we often work with tensor products of kernel functions, and tensor products of Hilbert spaces. We discuss two ways in which these spaces can be constructed.

We can define a tensor product Hilbert space by defining its inner product in terms of the inner products defined on the component Hilbert spaces, extending the inner product by linearity, and defining the tensor product Hilbert space to be the completion with respect to the inner product. Alternatively, we can use the additional structure of reproducing kernel Hilbert spaces to simplify the construction. Given two reproducing kernel Hilbert spaces we can define a new kernel function to be the tensor product of the kernels of the component spaces, and by definition this new kernel function has an associated RKHS: the tensor product RKHS. As we primarily work with RKHSs, this latter approach suffices.

Tensor products of Hilbert spaces. Let H_1 and H_2 denote two Hilbert spaces over $\mathscr{X}$ and $\mathscr{Y}$, with inner products $\langle\cdot,\cdot\rangle_1$ and $\langle\cdot,\cdot\rangle_2$ respectively. We define the tensor product of H_1 and H_2, denoted by $H_1 \otimes H_2$, to be the completion of the following inner product which is defined on the elementary tensors of $H_1 \otimes H_2$ and extended to the entire space by linearity:

$$\langle x_1 \otimes y_1, x_2 \otimes y_2 \rangle_{H_1 \otimes H_2} = \langle x_1, x_2 \rangle_{H_1} \langle y_1, y_2 \rangle_{H_2},$$

for $x_1, x_2 \in H_1$, $y_1, y_2 \in H_2$.

Tensor products of kernels. Let $\mathscr{H}_1$ and $\mathscr{H}_2$ denote two reproducing kernel Hilbert spaces over $\mathscr{X}$ and $\mathscr{Y}$, with kernels and inner products k_1, $\langle\cdot,\cdot\rangle_1$ and k_2, $\langle\cdot,\cdot\rangle_2$ respectively. We can define a new kernel function $k := k_1 \otimes k_2$ such that

$$k(x_1, y_1, x_2, y_2) = k_1(x_1, x_2) k_2(y_1, y_2),$$

and this kernel function is associated with the tensor product RKHS $\mathscr{H}_1 \otimes \mathscr{H}_2$.

Multi-linear operators. Let $\mathscr{H}_1, \dots, \mathscr{H}_M$ be a collection of M reproducing kernel Hilbert spaces, then we define the tensor product RKHS generated by the collection to be $\mathscr{H}_1 \otimes \cdots \otimes \mathscr{H}_M$, and use the shorthand $\otimes_{m=1}^M \mathscr{H}_m$ to denote the space. For $f_m \in \mathscr{H}_m$, $m \in \{1, \dots, M\}$, $\otimes_{m=1}^M f_m$ is an element in the tensor product RKHS $\otimes_{m=1}^M \mathscr{H}_m$, and $\otimes_{m=1}^M f_m$ is also the multi-linear operator defined as

$$\left(\otimes_{i=1}^M f_m\right)(g_1, \dots, g_M) = \prod_{m=1}^M \langle f_m, g_m \rangle_{\mathscr{H}_m},$$

for $g_m \in \mathscr{H}_m$, $m \in \{1, \dots, M\}$. Furthermore, the tensor product defines a family of rank-one linear operators $\times_n : \mathscr{H}_m \to \otimes_{m=1, m \neq n}^M \mathscr{H}_m$, for $n \in \{1, \dots, M\}$, as

$$\left(\otimes_{m=1}^M f_m\right) \times_n g_n = \langle f_n, g_n \rangle_{\mathscr{H}_n} \otimes_{m=1, m \neq n}^M f_m.$$

The notation $\times_n$ emphasizes which RKHS the operator acts upon, and provides clarity when working with more than two RKHSs. When the operator $\times_n$ is not used, the tensor product is defined such that the right-most component is being acted upon. For example, $(f_1 \otimes f_2)(g_2) = f_1 \langle f_2, g_2 \rangle_{\mathscr{H}_2}$, and $(f_1 \otimes f_2 \otimes f_3)(g_3) = (f_1 \otimes f_2)\langle f_3, g_3 \rangle_{\mathscr{H}_3}$.

CHAPTER 2

Density recovery from kernel mean embeddings

2.1 Introduction

Kernel mean embeddings have enabled a framework for performing nonparametric statistical inference using probabilistic intuition, where learning algorithms embed probability distributions into reproducing kernel Hilbert spaces and produce consistent estimators of complex distributions as embeddings. These estimators can often be computed with access to a finite set of observations. For characteristic kernel functions, the kernel mean embedding is injective, meaning all of the information of the embedded distribution is captured by its embedding. An open question has long been what information can be recovered from an estimated kernel mean embedding. It is clear that we can estimate expectations of functions in the RKHS with an estimated kernel mean embedding [Smola et al., 2007], however it is not obvious what exact properties of the embedded distribution we can recover. In this chapter we generalize and prove a remark made in Song et al. [2014]: that densities can be recovered from embeddings at no cost (with zero computation). Our result applies to all consistent estimated embeddings, such as those obtained via the kernel sum, product, and Bayes' rule, and conditional mean embeddings. In the case that the empirical kernel mean embedding is used, we recover the standard kernel density estimator. We show that one can use this result to consistently estimate conditional densities using conditional mean embeddings.

2.1.1 Literature review

To perform inference using estimated kernel mean embeddings, we must extract information regarding the embedded distribution from the embedding. Let $\hat{\mu}_{\mathbb{P}}$ denote an estimated embedding of the distribution $\mathbb{P}$ in an RKHS $\mathscr{H}$, then we can estimate expectations of functions in the RKHS: $\hat{\mathbb{E}}[f(X)] = \langle f, \hat{\mu}_{\mathbb{P}} \rangle_{\mathscr{H}}$ where $X \sim \mathbb{P}$. Kernel functions are well-known to produce a rich class of functions, for example the RKHS associated with the Gaussian kernel on $\mathbb{R}^d$ is dense in the space of continuous functions on $\mathbb{R}^d$ over compacts [Sriperumbudur et al., 2011]. However, the Gaussian RKHS does not contain the identity function and thus we cannot estimate the mean of $\mathbb{P}$ from its embedding. Moreover, Minh [2010] show that arbitrary polynomials do not belong to the Gaussian RKHS, a result that was extended to a larger class of RKHSs in Dette and Zhigljavsky [2021], and thus we cannot recover the moments of $\mathbb{P}$. Nor can we recover quantiles or tail probabilities from the embedding, as the functions belonging to the RKHS associated with a continuous kernel function are also continuous [Paulsen and Raghupathi, 2016]. Hence, whilst the RKHS specifies a rich class of functions, recovering information from embeddings remains a complicated task.

Several alternative approaches have been proposed to recover information from estimated kernel mean embeddings. These include distributional pre-image learning [Kwok and Tsang, 2004, Smola et al., 2007, Song, 2008], kernel herding [Welling, 2009, Chen et al., 2010], and the estimation of Radon-Nikodym derivatives [Schuster et al., 2020]. We discuss these approaches below.

To recover information from the estimated embedding, several authors consider the distributional pre-image problem where they aim to find a probability distribution in the input space that is mapped to a kernel mean embedding similar to the estimated embedding. [Song et al., 2008] specifies a parametric family of distributions and their RKHS embeddings, and minimizes the RKHS norm between the estimated embedding and the embedding of the parametric distributions. This approach may perform poorly when the parametric assumption is not accurate, and the optimization problem is not convex for each parameter. Additionally, when a distributional pre-image is learned, it's not clear to what degree the learned pre-image estimates the underlying distribution.

Rather than estimate the underlying embedded distribution, one can draw samples from the embedded distribution. Kernel herding [Chen et al., 2010] generalizes the herding algorithm [Welling, 2009] to continuous spaces, and produces pseudo-samples deterministically. Expectations of functions in the RKHS can be computed using the samples, and it is shown that the error in expectation estimation decreases at rate $O(T^{-1})$ for $T \geq 1$ samples. The asymptotic moments of the samples are shown to match the empirical moments of the data.

Schuster et al. [2020] propose a method to estimate densities from kernel mean embeddings by specifying a reference measure and estimating a Radon-Nikodym derivative via an inverse problem. This approach assumes that the embedded distribution is absolutely continuous with respect to the reference measure, that the Radon-Nikodym derivative belongs to the RKHS, and that the probability distribution is defined on a compact. The author's demonstrate that their density estimator produces state-of-the-art results, and has favourable performance compared to modern machine learning methods such as deep neural networks when estimating conditional densities.

Kanagawa and Fukumizu [2014] proposed a method to recover the density using a nonparametric estimator, by showing that expectations can be computed for functions in a Besov space, a space larger than the RKHS. Unfortunately there was a mistake in their proof of Theorem 1, and the results do not hold. The authors later showed that expectations can be computed for functions in a power of the RKHS, which is an interpolation space between the RKHS and L^2 [Kanagawa et al., 2016], although this does not allow for a nonparametric density estimator.

The density estimator that is proposed in this chapter was briefly mentioned in a remark in Song et al. [2014]. Let $\mathscr{H}_{\mathscr{Y}}$ denote an RKHS on $\mathscr{Y}$ associated with a normalized Gaussian kernel function, and let $\mu_{Y|X=x}$ be the kernel mean embedding of the distribution of $Y|X=x$ in $\mathscr{H}_{\mathscr{Y}}$. The authors note that the conditional density can be estimated via $\hat{p}(y|x) = \hat{\mathbb{E}}_{Y|X=x}[k(Y,y)] = \langle k(y,\cdot), \mu_{Y|X=x} \rangle_{\mathscr{H}_{\mathscr{Y}}}$. They note that the density can be estimated from the estimated embedding. However, there is little justification for the remark, and it seems to have not been recognized by the broader literature. For example, the review paper by Muandet et al. [2017] provides an in-depth discussion of recovering information from kernel mean embeddings; yet, it does not mention that densities can be recovered. We formalize the remark, showing that the density estimator is applicable to all estimated embeddings which are consistent in probability, and that the only requirement on the kernel function used is that it can be written as a smoothing kernel.

2.2 Density estimation

Let $\mathbb{P} \in M_+^1(\mathbb{R}^d)$, where we recall that $M_+^1(\mathbb{R}^d)$ denotes the space of probability measures on $\mathbb{R}^d$, and let $k^\gamma : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ denote a reproducing kernel function on $\mathbb{R}^d$ with hyperparameter $\gamma > 0$. Let $\mathscr{H}_\gamma$ denote the RKHS associated with reproducing kernel k^γ, then the kernel mean embedding of $\mathbb{P}$ in $\mathscr{H}_\gamma$ is given by

$$\mu_{\mathbb{P}}^\gamma = \int_{\mathbb{R}^d} k^\gamma(x,\cdot)\mathbb{P}(dx).$$

Under Assumption 1.1 on k^γ, the embedding is well defined for all $\gamma > 0$.

Throughout the following, all random variables are defined on $(\Omega, \mathscr{F}, \mathbb{P})$, and we say that a sequence of random variables $X_n = o_{\mathbb{P}}(1)$ if for all $\epsilon > 0$, $\limsup_{n\to\infty} \mathbb{P}(|X_n| \geq \epsilon) = 0$. Denote by $\hat{\mu}_{\mathbb{P},n}^\gamma$ an estimate of $\mu_{\mathbb{P}}^\gamma$ defined by

$$\hat{\mu}_{\mathbb{P},n}^\gamma = \sum_{i=1}^n W_{\gamma,n,i} k^\gamma(X_{\gamma,n,i},\cdot), \tag{2.1}$$

for $\{W_{\gamma,n,i}\}_{i=1}^n$ a collection of $\mathbb{R}$-valued random variables, and $\{X_{\gamma,n,i}\}_{i=1}^n$ a collection of $\mathbb{R}^d$-valued random variables. We assume that $\hat{\mu}_{\mathbb{P},n}^\gamma$ is a consistent estimator of $\mu_{\mathbb{P}}^\gamma$ in probability, that is $\|\hat{\mu}_{\mathbb{P},n}^\gamma - \mu_{\mathbb{P}}^\gamma\|_{\mathscr{H}_\gamma} = o_{\mathbb{P}}(1)$.

Estimated kernel mean embedding are weighted sums of kernel functions, corresponding to the general formulation presented in Equation (2.1). Detailed examples of kernel mean embeddings and their empirical counterparts are given in Section 1.2. For example, let $x_1,\dots,x_n$ be i.i.d. samples from $\mathbb{P} \in M_+^1\left(\mathbb{R}^d\right)$. Then the empirical kernel mean embedding of $\mathbb{P}$ is $\hat{\mu}^\gamma_{\mathbb{P},n} := n^{-1}\sum_{i=1}^n k^\gamma(x_i,\cdot)$, where $W_{\gamma,n,i} = n^{-1}$ for $i = 1,\dots,n$.

We also assume that $\mathbb{P}(dx) = p(x)dx$ for some $p:\mathbb{R}^d \to [0,\infty)$, and we consider the following estimator of p:

$$q_n = \sum_{i=1}^n W_{\gamma_n,n,i}\bar{k}^{\gamma_n}(X_{\gamma_n,n,i},\cdot), \tag{2.2}$$

where $(\gamma_n)_{n\geq 1}$ is a sequence in $(0,\infty)$, and a bar is used to denote that the kernel function is normalized as follows

$$\bar{k}^\gamma(x,x') = \frac{k^\gamma(x,x')}{\int_{\mathbb{R}^d} k^\gamma(x,z)dz}, \quad \forall \gamma > 0.$$

If k^γ is a Gaussian kernel function, then $\bar{k}^\gamma(x,x') = (2\pi\gamma^2)^{-d/2}k^\gamma(x,x')$ for all $\gamma > 0$. The reproducing kernel Hilbert spaces associated with k^γ and $\bar{k}^\gamma$ are isometrically isomorphic. The spaces consist of the same functions, and the inner products on the spaces are related by a scaling factor of $(2\pi\gamma^2)^{d/2}$ due to the normalization term.

Our results hold for several widely-used kernels. We only require that the kernel function can be expressed in terms of a smoothing kernel, and that the degree of smoothing is controlled via the kernel hyperparameter.

Assumption 2.1. *There exists a continuous and bounded function $K:\mathbb{R}^d \to [0,\infty)$ such that*

$$k^\gamma(x,x') = K\left(\frac{x-x'}{\gamma}\right),$$

for all $x,x' \in \mathscr{X}$. The function K is called a smoothing kernel.

This assumption holds for the Laplace, Gaussian, and Matérn kernel functions, and throughout the following we assume that the kernel associated with the kernel mean embedding and its estimate satisfies Assumption 2.1.

The following theorem provides a guarantee on the density estimator q_n under standard assumptions on the kernel mean embedding. We use $C_0(\mathbb{R}^d)$ to denote the space of continuous functions on $\mathbb{R}^d$ that vanish at infinity.

Theorem 2.1. *Assume that $p \in C_0\left(\mathbb{R}^d\right)$, that $\int_{\mathbb{R}^d} K(x)dx < \infty$, and that $(\gamma_n)_{n\geq 1}$ is such that*

$$\lim_{n\to\infty}\gamma_n = 0, \quad \frac{1}{\gamma_n^d}\left\|\hat{\mu}^{\gamma_n}_{\mathbb{P},n} - \mu^{\gamma_n}_{\mathbb{P}}\right\|^2_{\mathscr{H}_{\gamma_n}} = o_{\mathbb{P}}(1).$$

Then $\|q_n - p\|_\infty = o_{\mathbb{P}}(1)$.

Remark 2.1. *If $\|\hat{\mu}^\gamma_{\mathbb{P},n} - \mu^\gamma_{\mathbb{P}}\|^2_{\mathscr{H}_\gamma} = o_{\mathbb{P}}(1)$ for all $\gamma > 0$ small enough and if $\gamma_n \to 0$ sufficiently slowly, then the condition $\gamma_n^{-d}\|\hat{\mu}^{\gamma_n}_{\mathbb{P},n} - \mu^{\gamma_n}_{\mathbb{P}}\|^2_{\mathscr{H}_{\gamma_n}} = o_{\mathbb{P}}(1)$ is satisfied.*

Theorem 2.1 states that given a consistent kernel mean embedding estimator, one may consistently estimate the embedding's underlying density. Our result is applicable to all estimated kernel mean embeddings that are consistently estimated in probability. We provide two examples below, and develop an application to conditional mean embeddings in the following section.

Example 2.1 (Independent and identically distributed data)**.** *Let $x_1,\dots,x_n$ denote n i.i.d. samples from $\mathbb{P} \in M_+^1(\mathcal{X})$, then $\hat{\mu}_{\mathbb{P},n}^{\gamma} := n^{-1}\sum_{i=1}^n k^{\gamma}(x_i,\cdot)$. It follows from the boundedness of the kernel function and the law of large numbers in Banach spaces [Hoffmann-Jørgensen and Pisier, 1976] that for all $\gamma>0$, $\|\hat{\mu}_{\mathbb{P},n}^{\gamma}-\mu_{\mathbb{P}}^{\gamma}\|_{\mathcal{H}_\gamma}^2 = o_{\mathbb{P}}(1)$. Hence there exists a sequence $(\gamma_n)_{n\ge1}$ with $\lim_{n\to\infty}\gamma_n=0$ such that $\gamma_n^{-d}\|\hat{\mu}_{\mathbb{P},n}^{\gamma_n}-\mu_{\mathbb{P}}^{\gamma_n}\|_{\mathcal{H}_{\gamma_n}}^2 = o_{\mathbb{P}}(1)$. Therefore $\|q_n-p\|_\infty = o_{\mathbb{P}}(1)$ by Theorem 2.1.*

Example 2.2 (Time series data)**.** *Let $y_1,\dots,y_n$ denote n observations from a hidden Markov model satisfying Assumptions 3.2 to 3.4 (sufficient conditions for identifiability), and let $\mathbb{P}$ denote the distribution of an observation. Then $\hat{\mu}_{\mathbb{P},n}^{\gamma} := n^{-1}\sum_{i=1}^n k^{\gamma}(y_i,\cdot)$, and it follows from the first concentration inequality given in Lemma 3.11 that for all $\gamma>0$, $\|\hat{\mu}_{\mathbb{P},n}^{\gamma}-\mu_{\mathbb{P}}^{\gamma}\|_{\mathcal{H}_\gamma}^2 = o_{\mathbb{P}}(1)$. Hence there exists a sequence $(\gamma_n)_{n\ge1}$ with $\lim_{n\to\infty}\gamma_n=0$ such that $\gamma_n^{-d}\|\hat{\mu}_{\mathbb{P},n}^{\gamma_n}-\mu_{\mathbb{P}}^{\gamma_n}\|_{\mathcal{H}_{\gamma_n}}^2 = o_{\mathbb{P}}(1)$. It therefore follows from Theorem 2.1 that $\|q_n-p\|_\infty = o_{\mathbb{P}}(1)$.*

The proof of Theorem 2.1 requires the following lemma that is proved in Section 2.7. The lemma states that when a function is smoothed by a convolution, the smoothed function uniformly converges to the original function as the degree of smoothing tends towards zero.

Lemma 2.1. *Let $\nu \in M_+^1(\mathbb{R}^d)$, $(\sigma_n)_{n\ge1}$ be a sequence in $(0,\infty)$ such that $\lim_{n\to\infty}\sigma_n=0$, $f\in C_0(\mathbb{R}^d)$, and*

$$f_n(x) = \int_{\mathbb{R}^d} f(x+\sigma_n y)\nu(dy), \quad \forall x\in\mathbb{R}^d,\ \forall n\ge1.$$

Then $\lim_{n\to\infty}\|f_n-f\|_\infty = 0$.

Proof of Theorem 2.1. Let $\bar{K}:\mathbb{R}^d\to[0,\infty)$ be defined by

$$\bar{K}(x) = \frac{K(x)}{\int_{\mathbb{R}^d}K(x')dx'}, \quad x\in\mathbb{R}^d,$$

and note that for all $\gamma>0$ we have

$$\bar{k}^{\gamma}(x,x') = \gamma^{-d}\bar{K}\left(\frac{x-x'}{\gamma}\right), \quad x,x'\in\mathbb{R}^d. \tag{2.3}$$

Then, for all $n\ge1$,

$$\|q_n-p\|_\infty \le \sup_{x'\in\mathbb{R}^d}\left|\sum_{i=1}^n W_{\gamma_n,n,i}\bar{k}^{\gamma_n}(X_{\gamma_n,n,i},x') - \mathbb{E}_{\mathbb{P}}\left[\bar{k}^{\gamma_n}(X,x')\right]\right| \\ + \sup_{x'\in\mathbb{R}^d}\left|\mathbb{E}_{\mathbb{P}}\left[\bar{k}^{\gamma_n}(X,x')\right]-p(x')\right|. \tag{2.4}$$

Note that for all $x' \in \mathbb{R}^d$,

$$\begin{aligned}
\left|\mathbb{E}_{\mathbb{P}}[\bar{k}^{\gamma_n}(X,x')] - p(x')\right| &= \left|\int_{\mathbb{R}^d} \bar{k}^{\gamma_n}(x,x')p(x)dx - p(x')\right| \\
&= \left|\gamma_n^{-d}\int_{\mathbb{R}^d} \bar{K}\left(\frac{x-x'}{\gamma_n}\right)p(x)dx - p(x')\right| \qquad (2.5)\\
&= \left|\int_{\mathbb{R}^d} p(x'+\gamma_n z)\bar{K}(z)dz - p(x')\right|.
\end{aligned}$$

Under the assumption that $\lim_{n\to\infty}\gamma_n = 0$, $p \in C_0(\mathbb{R}^d)$, and $\bar{K}(z)dz \in M_+^1(\mathbb{R}^d)$, it follows from Lemma 2.1 that

$$\lim_{n\to\infty}\sup_{x'\in\mathbb{R}^d}\left|\mathbb{E}_{\mathbb{P}}\left[\bar{K}^{\gamma_n}\left(X,x'\right)\right] - p(x')\right| = 0.$$

On the other hand, it follows from Equation (2.3) that

$$\bar{k}^{\gamma_n}(x,x') = \gamma_n^{-d}C_K k^{\gamma_n}(x,x') \quad x,x' \in \mathbb{R}^d,$$

where we have used the shorthand $C_K = 1/\int_{\mathbb{R}^d} K(x)dx$, and thus for all $n \geq 1$ we have

$$\begin{aligned}
&\sup_{x'\in\mathbb{R}^d}\left|\sum_{i=1}^n W_{\gamma_n,n,i}\bar{k}^{\gamma_n}\left(X_{\gamma_n,n,i},x'\right) - \mathbb{E}_{\mathbb{P}}\left[\bar{k}^{\gamma_n}\left(X,x'\right)\right]\right| \\
&= \gamma_n^{-d}C_K\sup_{x'\in\mathbb{R}^d}\left|\sum_{i=1}^n W_{\gamma_n,n,i}k^{\gamma_n}\left(X_{\gamma_n,n,i},x'\right) - \mathbb{E}_{\mathbb{P}}\left[k^{\gamma_n}\left(X,x'\right)\right]\right| \qquad (2.6)\\
&= \gamma_n^{-d}C_K\sup_{x'\in\mathbb{R}^d}\left|\left\langle\hat{\mu}_{\mathbb{P},n}^{\gamma_n} - \mu_{\mathbb{P}}^{\gamma_n}, k^{\gamma_n}(x',\cdot)\right\rangle_{\mathcal{H}_{\gamma_n}}\right| \\
&\leq \gamma_n^{-d}C_K\left\|\hat{\mu}_{\mathbb{P},n}^{\gamma_n} - \mu_{\mathbb{P}}^{\gamma_n}\right\|_{\mathcal{H}_{\gamma_n}}\left\|k^{\gamma_n}\right\|_\infty,
\end{aligned}$$

where $\|k^{\gamma_n}\|_\infty < \infty$ as per Assumption 2.1.

Using Equations (2.4) to (2.6) and under the assumptions of the lemma it follows that

$$\limsup_{n\to\infty}\mathbb{P}(\|q_n - p\|_\infty \geq \epsilon) \leq \limsup_{n\to\infty}\mathbb{P}\left(\gamma_n^{-d}\left\|\hat{\mu}_{\mathbb{P},n}^{\gamma_n} - \mu_{\mathbb{P}}^{\gamma_n}\right\|_{\mathcal{H}_{\gamma_n}} \geq \frac{\epsilon}{2C_K\|k^{\gamma_n}\|_\infty}\right) = 0, \quad \forall\epsilon > 0,$$

and the proof is complete. ■

Remark 2.2 (Connection to kernel density estimation)**.** *Suppose that we observe n i.i.d. samples $x_1,\ldots,x_n$ from a distribution $\mathbb{P}$ on $\mathbb{R}^d$, and estimate the embedding of $\mathbb{P}$ via the sample mean. Then $\hat{\mu}_{\mathbb{P},n}^{\gamma_n} = \frac{1}{n}\sum_{i=1}^n k^{\gamma_n}(x_i,\cdot)$, and the density estimator q_n is*

$$q_n(x) = \frac{1}{n}\sum_{i=1}^n \bar{k}^{\gamma_n}(x_i,x) = \frac{1}{n\gamma_n^d}\sum_{i=1}^n \bar{K}\left(\frac{x_i - x}{\gamma_n}\right),$$

where $\bar{K}(x) := K(x)/\int_{\mathbb{R}^d} K(x')dx'$, $x \in \mathbb{R}^d$. This is exactly the kernel density estimator of the distribution $\mathbb{P}$ given n i.i.d. samples and smoothing kernel $\bar{K}$.

2.2.1 Practical considerations

Our main theorem shows that the density estimator q_n converges uniformly to the density p in probability. However, without additional assumptions on the random variables $\{W_{\gamma,n,i}\}_{i=1}^n$ (such as non-negativity) q_n is not necessarily a density. This is a standard problem in nonparametric density estimation that is easily fixed by taking the positive part of the estimator [Tsybakov, 2008]. For practical applications, ensuring that the estimator is a density is often desirable. We can define a probability density function that has the same consistency properties, under additional assumptions.

Proposition 2.1. *Let $q_n^+(x) := \max(q_n(x), 0)$, for $x \in \mathbb{R}^d$. If $\|q_n - p\|_\infty = o_{\mathbb{P}}(1)$, then we have $\|q_n^+ - p\|_\infty = o_{\mathbb{P}}(1)$. Assume that the distribution $\mathbb{P}$ has compact support $D \subset \mathbb{R}^d$, and define the following probability density function*

$$p_n(x) := \frac{q_n^+(x)\mathbb{1}_D(x)}{\int_D q_n^+(x')dx'}, \quad x \in \mathbb{R}^d, \tag{2.7}$$

where $\mathbb{1}_D(x)$ takes value 1 when $x \in D$ and 0 otherwise. If $\|q_n - p\|_\infty = o_{\mathbb{P}}(1)$, then $\|p_n - p\|_\infty = o_{\mathbb{P}}(1)$.

Proof. We first prove the claim that $\|q_n - p\|_\infty = o_{\mathbb{P}}(1)$ implies $\|q_n^+ - p\|_\infty = o_{\mathbb{P}}(1)$. Suppose that $q_n(x) \geq 0$ for some $x \in \mathbb{R}^d$, then $q_n^+(x) = q_n(x)$ and therefore $|q_n^+(x) - p(x)| = |q_n(x) - p(x)|$. On the other hand, if $q_n(x) < 0$ then it follows that $|q_n^+(x) - p(x)| < |q_n(x) - p(x)|$. Combining the two inequalities, we have $|q_n^+(x) - p(x)| \leq |q_n(x) - p(x)|$, $\forall x \in \mathbb{R}^d$. It follows that $\|q_n^+ - p\|_\infty \leq \|q_n - p\|_\infty$ and therefore $\|q_n - p\|_\infty = o_{\mathbb{P}}(1)$ implies $\|q_n^+ - p\|_\infty = o_{\mathbb{P}}(1)$.

We now prove the second claim. It was shown above that $\|q_n^+ - p\|_\infty = o_{\mathbb{P}}(1)$, and therefore for $\epsilon > 0$ there exists an $N_\epsilon \in \mathbb{N}$ such that for all $n \geq N_\epsilon$, $\sup_{x \in \mathbb{R}^d} |q_n^+(x) - p(x)| < \epsilon$ with high probability. Let λ denote the Lebesgue measure on $\mathbb{R}^d$, then $\forall n \geq N_\epsilon$

$$\begin{aligned} \left| \int_D q_n^+(x)dx - 1 \right| &= \left| \int_D q_n^+(x)dx - \int_D p(x)dx \right| \\ &\leq \int_D \left| q_n^+(x) - p(x) \right| dx \\ &\leq \int_D \sup_{x \in \mathbb{R}^d} \left| q_n^+(x) - p(x) \right| dx \\ &\leq \lambda(D)\epsilon. \end{aligned} \tag{2.8}$$

Hence $\int_D q_n^+(x)dx$ converges in probability to 1, and it follows from $\|q_n^+ - p\|_\infty = o_{\mathbb{P}}(1)$ that $\|p_n - p\|_\infty = o_{\mathbb{P}}(1)$. ■

2.2.2 Hyperparameter selection

Our theorem guarantees that there exists a sequence of hyperparameters such that the density estimator uniformly converges to the truth in probability.

In practice, we only have access to a fixed set of samples, and we require a way to tune the hyperparameter. As highlighted in Remark 2.2, our density estimator is closely related to kernel density estimation, and hyperparameter selection is a well-studied task in this setting. To tune the hyperparameter in a data-driven fashion, we follow Silverman [1986] and minimize the integrated squared error using leave-one-out cross-validation. Several review papers consider this a favourable approach [Park and Marron, 1990, Cao et al., 1994, Jones et al., 1996].

The integrated squared error is defined as $\mathrm{ISE}(\tilde{\gamma}_n) = \int [q_n(x) - p(x)]^2 \, dx$. We estimate the ISE using the data, and ignore the term independent of $q_n(x)$ as it is independent of the hyperparameter, using the following score function $M(\gamma_n) = \int q_n(x)^2 dx - 2\int q_n^{-i}(x)p(x)dx$, where q_n^{-i} denotes the density estimator obtained when using all observations except the i-th. We can easily compute the score function via

$$M(\gamma_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{n,i}w_{n,j}\bar{k}^{\sqrt{2}\gamma_n}(x_i,x_j) - 2\sum_{i=1}^{n}\sum_{j\neq i} w_{n,i}w_{n,j}^{-i}\bar{k}^{\gamma_n}(x_i,x_j), \tag{2.9}$$

where w_n^{-i} are the $n-1$ weights obtained from an embedding learned without the i-th observation.

To tune the hyperparameter we minimize the score function with respect to γ_n. To evaluate the score we must recompute the embedding n times, and so there may be considerable computational cost. If the cost of obtaining the embedding is $O(n^\alpha)$, and we evaluate M over a grid of m hyperparameters then the additional cost of this method is $O(m(n(n-1)^\alpha + n^2))$. For $\alpha > 2$ and n large enough, the additional cost is approximately $O(mn^{1+\alpha})$.

2.3 Application to conditional mean embeddings

Conditional mean embedding is used to embed conditional distributions into reproducing kernel Hilbert spaces given samples from a joint distribution. Conditional mean embedding is well studied and is known to produce estimated embeddings which are consistent in probability, under appropriate assumptions. Song et al. [2009] showed that under Assumption 1.2, the estimated conditional mean embedding is consistent in probability. Grünewälder et al. [2012] later demonstrated that the conditional mean embedding can be estimated by minimizing an objective function over a vector-valued RKHS, however the analysis only holds when the RKHS the embedding belongs to has finite dimension. Recently, Li et al. [2022] derived consistency results when the embedding space is infinite dimensional, and they consider a misspecified scenario in which the conditional mean operator does not belong to the same space as the estimator.

We consider random variables X and Y taking values in $\mathscr{X}$ and $\mathscr{Y} = \mathbb{R}^d$, $d \geq 1$ respectively. Let $\mathscr{H}_\mathscr{X}$ be an RKHS on $\mathscr{X}$, and let $\mathscr{H}_\gamma$ be an RKHS on $\mathscr{Y}$ with reproducing kernel function $k^\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ which has hyperparameter $\gamma > 0$ and satisfies Assumption 2.1. Conditional mean embedding is discussed in greater detail in Section 1.2. We denote the conditional mean embedding of the distribution of $Y|X = x$, $x \in \mathscr{X}$, in $\mathscr{H}_\gamma$ by $\mu^\gamma_{Y|X=x}$. Suppose we observe n i.i.d.

samples $(x_i, y_i)_{i=1}^n$ from the joint distribution of (X, Y), then the empirical estimator of the conditional mean embedding in $\mathscr{H}_\gamma$ is

$$\hat{\mu}^\gamma_{Y|X=x} = \sum_{i=1}^n w_{n,i}(x) k^\gamma(y_i, \cdot) \tag{2.10}$$

where $w_n(x) = (K_X + n\lambda I_n)^{-1} l_x$, for $\lambda > 0$ a regularization parameter, K_X an $n \times n$ matrix with (i,j)-th value $[K_X]_{i,j} = l(x_i, x_j)$, and $l_x = [l(x_1, x), \ldots, l(x_n, x)]$. Throughout the following we use $\mathbb{P}_X$ to denote the marginal distribution of the random variable X.

The consistency results of Li et al. [2022] are given for the conditional mean operator, and the following corollary shows that this implies consistency of the conditional mean embedding. Consistency of the conditional mean operator is given under the sufficient conditions Assumption 1.1, and assumptions (EVD), (EMB), and (SRC). (EVD) is an assumption on the eigenvalue decay of the cross-covariance operator $\mathscr{C}_{XX}$ (that the decay is at least polynomial with degree $p \in (0,1]$), (EMB) is an assumption on the RKHS on $\mathscr{X}$, and (SRC) assumes that the conditional mean operator belongs to a space which captures the misspecified and well specified scenario. We refer the reader to Li et al. [2022] for the details of these assumptions.

Of the sufficient conditions for consistency of the empirical conditional mean embeddings, Assumption 1.1, (EVD), (EMB), and (SRC), only Assumption 1.1 depends upon the kernel function on $\mathscr{Y}$. For popular kernel functions such as the Gaussian and Laplace kernels, Assumption 1.1 is satisfied for all $\gamma > 0$ trivially.

Corollary 2.1 (Consistency of estimated conditional mean embeddings)**.** *Assume Assumption 1.1 holds for all $\gamma > 0$, (EVD), (EMB), and (SRC). Then there exists a set $A \subseteq \mathscr{X}$ such that $\mathbb{P}_X(A) = 1$ and*

$$\sup_{x \in A} \left\| \hat{\mu}^\gamma_{Y|X=x} - \mu^\gamma_{Y|X=x} \right\|^2_{\mathscr{H}_\mathscr{Y}} = o_\mathbb{P}(1), \quad \forall \gamma > 0.$$

The proof of Corollary 2.1 is given in Section 2.7.2.

The empirical conditional mean embedding in Equation (2.10) is of the same form as Equation (2.1), and thus we have the following density estimator of the conditional density function $p(y|x)$ associated with the conditional distribution of $Y|X$:

$$q_n(y|x) = \sum_{i=1}^n w_{n,i}(x) \bar{k}^{\gamma_n}(y_i, y), \quad x \in \mathscr{X},\ y \in \mathscr{Y}.$$

The following lemma provides a consistency result for the conditional density estimator, which follows from Theorem 2.1 and the consistency of the empirical conditional mean embedding.

Lemma 2.2. *Assume that Assumptions 1.1 and 2.1 hold for all $\gamma > 0$, and additionally assume (EVD), (EMB), and (SRC). Assume that $p(\cdot|x) \in C_0(\mathbb{R}^d)$ for $\mathbb{P}_X$-almost all $x \in \mathscr{X}$ and that $\int_{\mathbb{R}^d} K(y)dy < \infty$. Then there exists a sequence $(\gamma_n)_{n \geq 1}$ with $\lim_{n \to \infty} \gamma_n = 0$ and a set $A \subseteq \mathscr{X}$ with $\mathbb{P}_X(A) = 1$ such that*

$$\sup_{x \in A} \left\| q_n^{\gamma_n}(\cdot|x) - p(\cdot|x) \right\|_\infty = o_\mathbb{P}(1).$$

Proof. Let $\mu^{\gamma}_{Y|X=x}$ be the $\mathscr{H}_\gamma$-valued conditional mean embedding of $Y|X=x$ for $x \in \mathscr{X}$, and let $\hat{\mu}^{\gamma}_{Y|X=x}$ denote its empirical estimate, for $\gamma > 0$. Assumption 1.1, (EVD), (EMB), and (SRC) hold for all $\gamma > 0$, and therefore it follows from Corollary 2.1 that for all $\gamma > 0$, there exists a set $A_\gamma \subseteq \mathscr{X}$ with $\mathbb{P}_X(A_\gamma) = 1$ such that $\sup_{x \in A_\gamma} \|\hat{\mu}^{\gamma}_{Y|X=x} - \mu^{\gamma}_{Y|X=x}\|_{\mathscr{H}_\gamma} = o_{\mathbb{P}}(1)$. Let $A^\star := \cap_{k \in \mathbb{N}} A_{1/k}$ then $\mathbb{P}_X(A^\star) = 1$ and there exists a sequence $(\gamma_n)_{n\geq 1}$ decreasing towards zero sufficiently slowly, such that $\sup_{x \in A^\star} \gamma_n^{-d} \|\hat{\mu}^{\gamma_n}_{Y|X=x} - \mu^{\gamma_n}_{Y|X=x}\|_{\mathscr{H}_{\gamma_n}} = o_{\mathbb{P}}(1)$. It then follows from Theorem 2.1 that $\sup_{x \in A^\star} \|q_n^{\gamma_n}(\cdot|x) - p(\cdot|x)\|_\infty = o_{\mathbb{P}}(1)$. ■

2.3.1 Hyperparameter selection

The integrated squared error is defined as $\mathrm{ISE}(\gamma_n) = \int [q_n(y \mid x) - p(y \mid x)]^2 \, dy p(x) dx$. As discussed in Section 2.2.2, we focus on minimizing the terms which depend on γ_n: we minimize the score function $M(\gamma_n) = \int q_n(y \mid x)^2 dy p(x) dx - 2 \int q_n(y \mid x) p(y,x) dy dx$. This function can be estimated consistently by $\hat{M}(\gamma_n) = n^{-1} \sum_{i=1}^n \int (q_n^{-i}(y \mid x_i))^2 dy - 2n^{-1} \sum_{i=1}^n q_n^{-i}(y_i \mid x_i)$ [Fan and Yim, 2004]. The estimated score can be computed as

$$\hat{M}(\gamma_n) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ \sum_{j \neq i} \sum_{k \neq i} w_{n,j}^{-i}(x_i) w_{n,k}^{-i}(x_i) \bar{k}^{\sqrt{2}\gamma_n}(y_j, y_k) - 2 \sum_{j \neq i} w_{n,j}^{-i}(x_i) \bar{k}^{\gamma_n}(y_i, y_j) \Big\}, \tag{2.11}$$

where w_n^{-i} are the $n-1$ weights obtained from an embedding learned without the i-th observation.

To tune the hyperparameter we minimize the estimated score function with respect to γ_n. To evaluate the function we must recompute the embedding n times using $n-1$ observations, and the cost of computing the empirical conditional mean embedding is $O((n-1)^3)$. If we evaluate M over a grid of m bandwidths then the additional cost of this method is $O(m(n(n-1)^3 + n^2))$, and thus the additional cost is approximately $O(mn^4)$.

This is also the approach used to tune the hyperparameters for the Nadaraya-Watson conditional density estimator [Fan and Yim, 2004], however alternative approaches have been proposed with considerably lower computational cost, such as the cross-validated likelihood method of Holmes et al. [2007].

2.3.2 Comparison to existing methods

2.3.2.1 Conditional density operators

Schuster et al. [2020] propose a method of recovering a density from a conditional mean embedding using an alternative density estimator, and they refer to this as a conditional density operator (CDO). Their density estimator requires the construction of a reference measure ρ, typically taken to be a uniform distribution over the range of the observed data, defined such that the true distribution $\mathbb{P}$ is absolutely continuous with respect to ρ. They assume that the Radon-Nikodym derivative (i.e. the density) $\frac{d\mathbb{P}}{d\rho}$ belongs to an RKHS, and that $\mathbb{P}$ has compact support. Suppose we observe i.i.d. samples $(x_i, y_i)_{i=1}^n$ from a joint distribution (X, Y), and aim to estimate the conditional density $p(y|x)$ associated with the conditional distribution of $Y|X$.

Let l be a kernel function on $\mathscr{X}$ and k a kernel function on $\mathscr{Y}$, then the CDO density estimator can be computed by drawing m i.i.d. samples, $z_1,\dots,z_m$, from the reference measure ρ on $\mathscr{Y}$ and computing

$$\hat{p}_n^{CDO}(y|x) = \sum_{j=1}^{m} \tilde{w}_j(x) k(z_i,\cdot), \qquad \tilde{w}(x) = m^{-2}(K_\rho + \alpha' I_m)^{-2} K_{\rho Y} w_n(x),$$

where $w_n(x) = (K_X + n\lambda I_n)^{-1} l_x$, $K_{\rho Y}$, K_ρ, K_X are kernel matrices with $\left[K_{\rho Y}\right]_{i,j} = k(z_i, y_j)$, $\left[K_\rho\right]_{i,j} = k(z_i, z_j)$, and $[K_X]_{i,j} = l(x_i, x_j)$, and $l_x = [l(x_1,x),\dots,l(x_n,x)]$. In comparison, our estimator is $q_n(y|x) = \sum_{j=1}^{n} w_{n,i}(x)\bar{k}^\gamma(y_i,y)$. Our estimators are very closely related: they both estimate the density associated with an estimated conditional mean embedding. Both estimators are weighted sums of kernel functions on $\mathscr{Y}$, and the weights used by the CDO density estimator are a function of the weights used by our density estimator. Computing the weights of the CDO estimator requires an additional $m \times m$ matrix inversion, which adds both computational cost and numerical instability. The additional computations required are primarily functions of the samples from the reference measure, and thus it is unlikely that they will improve the quality of the estimator. In their experiments Schuster et al. [2020] take $m = \lfloor n \rfloor$.

2.3.2.2 Kernel conditional density estimation

Kernel density estimation [Rosenblatt, 1969, Parzen, 1962] is a nonparametric method to estimate the probability density function of a continuous random variable, given access to i.i.d. samples. A kernel density estimator is a normalized sum of smoothing kernels: let $(x_i)_{i=1}^n$ be i.i.d. samples taking values in $\mathbb{R}^d$ from a distribution with density function p, then

$$\hat{p}(x) = \frac{1}{n\gamma^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\gamma}\right), \quad x \in \mathbb{R}^d.$$

They produce a consistent estimator under the assumption that $\gamma \to 0$ and $n\gamma^d \to \infty$ as $n \to \infty$ [Devroye and Lugosi, 2001]. The consistency is often subject to an assumption on the smoothness of the density p, for example Nadaraya [1965] and Giné and Guillou [2002] assume that p is uniformly continuous.

Conditional densities can also be estimated nonparametrically using kernel conditional density estimation (KCDE) which has been studied and extended in several works [Rosenblatt, 1969, Fan et al., 1996, Hyndman et al., 1996, Fan and Yim, 2004]. The density estimator uses smoothing kernels which interpolate over both domains. For example, let $(x_i, y_i)_{i=1}^n$ be i.i.d. samples from a joint distribution on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, then the conditional density associated with the conditional distribution $Y|X$ is estimated as

$$\hat{p}(y|x) = \frac{\sum_{i=1}^{n} K^{\gamma_x, d_x}(x, x_i) K^{\gamma_y, d_y}(y, y_i)}{\sum_{i=1}^{n} K^{\gamma_x, d_x}(x, x_i)}, \quad x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y} \tag{2.12}$$

where $K^{\alpha,d}(z,z') := \alpha^{-d} K((z - z')/\alpha)$ for $\alpha > 0$ and $z, z' \in \mathbb{R}^d$. In the case that $d_x = d_y = 1$ the KCDE estimator is consistent when $\gamma_x, \gamma_y \to 0$ and $n\gamma_x\gamma_y \to \infty$ as $n \to \infty$ under the assumption that the conditional density of $Y|X = x$ and the marginal density of X have continuous second derivatives and are square integrable and $\mathbb{E}(Y|X = \cdot)$ has continuous second derivative [Hyndman et al., 1996]. The hyperparameters can be tuned by minimizing a cross-validated estimator of the integrated squared error, although this can be computationally expensive [Fan and Yim, 2004]. Alternative methods have been proposed such as the cross-validated likelihood proposed in Holmes et al. [2007].

The KCDE estimator given in Equation (2.12) is motivated by Bayes' rule: $p(y|x) = p(x, y)/p(x)$. The denominator is a kernel density estimator of the marginal density $p(x)$, and the numerator is a kernel density estimator of the joint density $p(x, y)$. The KCDE requires that densities on both domains are estimated, whereas our density estimator using kernel mean embeddings avoids estimating the marginal $p(x)$. This is particularly beneficial when $p(x)$ is hard to model with a kernel density estimator. For example, if X is distributed according to a Beta distribution with shape parameters $\alpha, \beta < 1$ then the density is unbounded at 0 and 1. This violates the minimal assumption of boundedness required for consistency of the kernel density estimator [Jiang, 2017]. It is also possible that the consistency holds, but estimating the density remains challenging. For example, in the case that $p(x)$ is a mixture of densities with different variances, choosing an optimal kernel hyperparameter may be particularly difficult and KCDE may not perform well.

2.4 Experiments

We showcase the effectiveness of our estimator in various settings, demonstrating its ability to recover multi-modal densities in multiple dimensions, and its ability to learn complex patterns that would be difficult to model parametrically. Whilst our estimator is applicable to all embeddings, in this section we demonstrate its use in recovering densities from conditional mean embeddings. Throughout this section we refer to our method as KMDE (kernel mean density estimation).

We evaluate the performance of estimators using the *maximum* absolute difference (MAD) and the mean squared error (MSE). In our experiments we evaluate the estimated density and true density functions over a sequence of points uniformly spaced over the range of the observed data, and compute the MAD and MSE. These metrics are of particular interest as the mean squared error approximates the integrated squared error minimized by our hyperparameter selection procedure (Section 2.2.2), and the maximum absolute difference approximates the supremum norm of the difference between the estimated density and true density functions which is used in our consistency result (Theorem 2.1).

2.4.1 Simulated data

We simulate data from several models, which we describe below.

Model 1 (Gaussian triangle)**.** *We first specify three points in $\mathbb{R}^3$ which are the vertices of a triangle. We uniformly sample a point along an edge of the triangle, and then draw a sample from a 3D*

Gaussian distribution centred at the point with covariance matrix $\sigma^2 I_3$. This sampling procedure corresponds to sampling from a distribution where a Gaussian sphere is placed at every point along the edges of the triangle.

A set of 1500 samples can be seen in Figure 2.1, and this example is particularly interesting as the conditional density of $Y,Z|X=x$ is structurally different for different values of x, ranging from near-uniform to bimodal to unimodal.

Model 2 (Beta)**.** *Let* $\mathrm{Beta}(\alpha,\beta)$ *denote a Beta distribution with shape parameters* $\alpha,\beta>0$. *For* $i\in\{1,\ldots,n\}$,

$$X_i \overset{i.i.d.}{\sim} \mathrm{Beta}(\alpha,\beta), \quad Y_i|X_i=x\sim N(x,0.1^2).$$

Model 3 (Bimodal)**.** *Samples are generated as follows. For* $i\in\{1,\ldots,n\}$, $X_i \overset{i.i.d.}{\sim} U[0,1]$ *where* $U[a,b]$ *denotes a uniform distribution between a and b. Then*

$$Y_i|X_i=x\sim\begin{cases} N(-2.5,1^2), & \textit{with probability } x \\ N(2.5,1^2), & \textit{with probability } 1-x \end{cases},$$

where $N(\mu,\sigma^2)$ *denotes a Gaussian distribution with mean* μ *and variance* σ^2.

For Model 3, the conditional distribution $Y|X=x$ is unimodal when x is either 0 or 1 with means 2.5 and -2.5 respectively, and the conditional distribution is bimodal for $x\in(0,1)$.

Model 4 (Dirichlet mixture)**.** *For* $d\geq1$ *and for* $i\in\{1,\ldots,n\}$, $X_i \overset{i.i.d.}{\sim} \mathrm{Dir}([0.1,\ldots,0.1])$, *where* $\mathrm{Dir}(\alpha)$ *denotes a Dirichlet distribution over a d-dimensional simplex parametrized by the d-dimensional vector* α *of positive reals. Let* $x\in\mathbb{R}^d$, *then*

$$Y_i|X_i=x\sim\begin{cases} N(1,0.1^2) & \textit{with probability} & x_1 \\ N(2,0.1^2) & \textit{with probability} & x_2 \\ & \vdots & \\ N(d,0.1^2) & \textit{with probability} & x_d \end{cases}$$

where x_j *denotes the* j*-th element of* $x\in\mathbb{R}^d$.

Model 3 with $\alpha,\beta<1$ and Model 4 are particularly interesting as the densities associated with X are unbounded and consistency results for kernel density estimators do not typically hold in this setting.

Model 5 (Linear)**.** *Let* d_x *be a positive integer. For* $i\in\{1,\ldots,n\}$, $X_i \overset{i.i.d.}{\sim} N(\mu,\Sigma)$ *where* μ *is a* d_x*-dimensional zero vector and* $\Sigma=\sigma^2 I_{d_x}$ *where* I_{d_x} *is a* d_x*-dimensional identity matrix. Then*

$$Y_i|X_i=x\sim N\left(\frac{1}{d_x}1_{d_x}^{\mathrm{T}}x,0.1^2\right),$$

where 1_{d_x} *denotes the* d_x*-dimensional one vector.*

Model 6 (Linear, Toeplitz)**.** *Let d be a positive integer. For $i \in \{1,\dots,n\}$, $X_i \overset{i.i.d.}{\sim} N(\mu,\Sigma)$ where μ is a d-dimensional zero vector and Σ is a symmetric Toeplitz matrix with the first column the d-dimensional vector of equidistant points between 0.1 and 1 inclusive. When $d = 1$, $\Sigma = 1$. Then*

$$Y_i|X_i = x \sim N\left(\frac{1}{d}1_d^{\mathrm{T}}x, 0.5^2\right),$$

where 1_d denotes the d-dimensional one vector.

Model 7 (Mixture)**.** *For $i \in \{1,\dots,n\}$, the random variables X_i are sampled as follows*

$$X_i \overset{i.i.d.}{\sim} \begin{cases} N(-4,2^2), & \textit{with probability } 0.4 \\ N(0,0.2^2), & \textit{with probability } 0.2 \\ N(4,1^2), & \textit{with probability } 0.4 \end{cases},$$

and $Y_i|X_i = x \sim N(x,1^2)$.

Model 8 (MVN)**.** *For $i \in \{1,\dots,n\}$, $Z_i \overset{i.i.d.}{\sim} N(\mu,\Sigma)$ where μ and Σ are defined as*

$$\mu = [0,0,0], \quad \Sigma = \begin{bmatrix} 1 & 0.9 & 0.3 \\ 0.9 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix}.$$

Then $Z_i = [Z_{i,1}, Z_{i,2}, Z_{i,3}]$, and we define $X_i = Z_{i,3}$ and $Y_i = [Z_{i,1}, Z_{i,2}]$.

The conditional distribution of $Y|X$ for Model 8 is a 2D multivariate normal distribution.

Model 9 (Non-linear type 1)**.** *For $i \in \{1,\dots,n\}$, the random variables X_i are sampled from a multivariate normal distribution with mean vector μ and covariance matrix Σ where μ is a 10-dimensional vector of zeros and Σ is a symmetric Toeplitz matrix with first column ranging from 1 to 0.1. Then $Y_i|X_i = x \sim N(\mu_x, 0.5^2)$ with*

$$\mu_x = \cos(\pi x_1 x_2) + \sin(2\pi x_3 x_4) + \sum_{j=5}^{8} x_j + x_9^2 + x_{10}^2.$$

Model 10 (Non-linear type 2)**.** *For $i \in \{1,\dots,n\}$, the random variables X_i are sampled from a multivariate normal distribution with mean vector μ and covariance matrix Σ where μ is a 10-dimensional vector of zeros and Σ is a symmetric Toeplitz matrix with first column ranging from 1 to 0.1. Then $Y_i|X_i = x \sim N(\mu_x, 0.5^2)$ with*

$$\mu_x = \sum_{j=1}^{5} x_j + x_6^2 + x_7^2.$$

Model 11 (Cauchy)**.** *Let $\mathrm{Cauchy}(x_0,\gamma)$ denote a Cauchy distribution with location parameter $x_0 \in \mathbb{R}$ and scale parameter $\gamma > 0$. For $i \in \{1,\dots,n\}$, $X_i \overset{i.i.d.}{\sim} \mathrm{Cauchy}(0,\gamma)$, and $Y_i|X_i = x \sim N(x,1^2)$.*

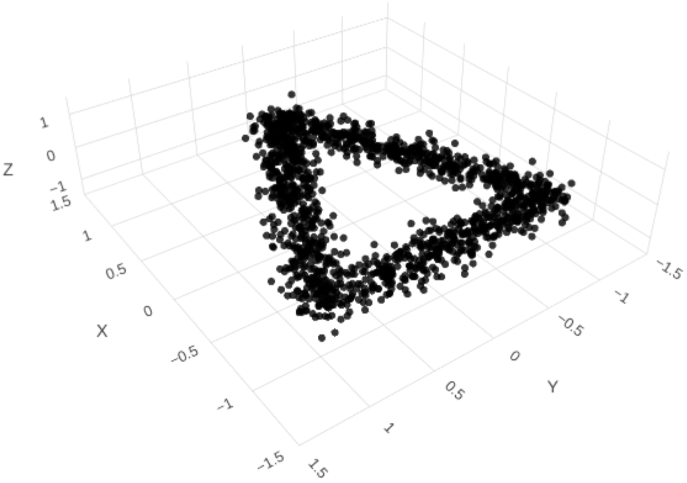


Figure 2.1: A scatterplot of 1500 samples from a Gaussian triangle.

2.4.2 Illustrative example

We start with an example demonstrating conditional density estimation using data generated from the Gaussian triangle described in Model 1. We generate 1500 observations $(x_i, y_i, z_i)_{i=1}^{1500}$ and estimate the 2D conditional density $Y, Z|X = x$ for $x \in \{-1, -0.3, 0.3, 1\}$. The conditional mean embedding can be estimated as described in Sections 1.2.2.1 and 2.3, resulting in the Tikhonov-regularized estimate

$$\hat{\mu}_{YZ|X=x} = \sum_{i=1}^{1500} w_i(x) k([y_i, z_i], \cdot),$$

where k is a kernel defined on $\mathbb{R}^2$, $w_i(x) = (K_X + 1500\lambda I_n)^{-1} l_x$, and $l_x = [l(x_1, x), \ldots, l(x_{1500}, x)]$. The density associated with the estimated embedding $\hat{\mu}_{YZ|X=x}$ can then be estimated via

$$q_n(y, z|x) = \sum_{i=1}^{1500} w_i(x) \bar{k}([y_i, z_i], [y, z]), \quad y, z \in \mathbb{R}.$$

We set the regularization parameter λ to be 0.1, and use Gaussian kernel functions for l and k. For the former we set the kernel hyperparameter via the median heuristic, and for the latter we tune the hyperparameter by numerically minimizing the estimated score function defined in Equation (2.11) for each x.

Figure 2.2 shows the 2D conditional densities $p(y, z|x)$ for $x \in \{-1, -0.3, 0.3, 1\}$ alongside their estimates produced using conditional mean embeddings and the density estimator. The estimator is very similar to the true density across all values of x displayed. Notably, the estimator accurately reflects the inherent characteristics of the true densities, capturing whether they are near-uniform, unimodal, or bimodal.

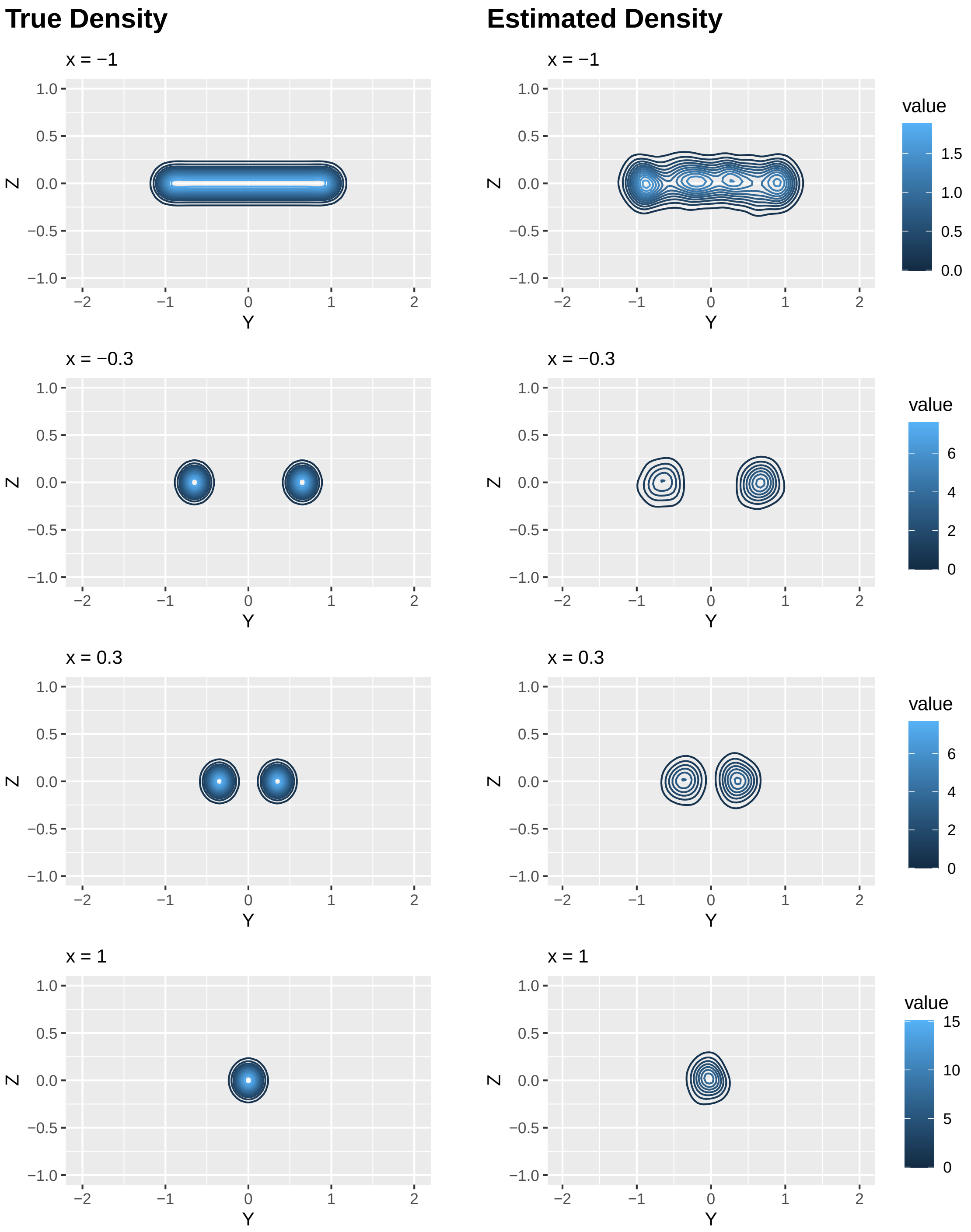


Figure 2.2: True and estimated conditional densities for the Gaussian triangle model. Each row presents the true conditional density $Y, Z|X = x$ for a given x (left) and its corresponding estimate using the conditional mean embedding (right), with the contour values indicated by their colour as explained by the legend to the right of each plot.

2.4.3 Comparison to CDO

We compare our method to the conditional density operator approach [Schuster et al., 2020] in estimating 1D conditional densities in the Gaussian triangle model. We draw 1500 samples $(x_i, y_i, z_i)_{i=1}^{1500}$ from Model 1, and consider estimating the conditional density of $Y|X = x, Z = 0$, for $x \in \{-1, -0.3, 0.3, 1\}$. For a fair comparison, we use Gaussian kernel functions for both methods with hyperparameter chosen via the median heuristic, and for the CDO specific hyperparameters we use the default values suggested in Schuster et al. [2020]. For each $x \in \{-1, -0.3, 0.3, 1\}$ we estimate the conditional density using both methods and compute the MAD and MSE between the truth and the estimate. We repeat this process 100 times for different samples from Model 1.

We provide a visual comparison between our method KMDE and the existing approach CDO in Figure 2.3. We compare the distributions of the MAD and MSE in estimating the conditional density $p(y|x,z)$ for $z = 0$ and across $x \in \{-1, -0.3, 0.3, 1\}$. The first row of the figure analyzes the MAD whilst the second analyzes the MSE. For each value of x a separate boxplot displays the distribution of the associated metric over the 100 repetitions, where a smaller value is better. It is evident from Figure 2.3 that our method outperforms the existing method across all values of x for both metrics.

Figure 2.4 shows several estimated 1-dimensional conditional densities corresponding to $Y|X = x, Z = 0$ across $x \in \{-1, -0.3, 0.3, 1\}$, for our method KMDE and the alternative CDO. As above, all hyperparameters are set to their default values.

2.4.4 Comparison to KCDE

We compare our method to kernel conditional density estimation (KCDE), also known as the Nadaraya-Watson conditional density estimator, across a wide range of simulated datasets. We generate 1000 observations $(x_i, y_i)_{i=1}^{1000}$ from each model that we consider, and estimate the conditional densities associated with $Y|X = x$. Both KMDE and KCDE place a kernel on the domain of X and Y, and for both methods we use Gaussian kernels and we tune the kernel hyperparameters by minimizing the estimated score function given in Equation (2.11). For our method, we set the estimated conditional mean embedding regularization parameter to $\lambda = 0.001$. Both methods produce estimates of the conditional density $p(y|x)$, and we illustrate the estimated conditional densities in Figures 2.5 to 2.7 for several values of x for Model 3 and Model 4 (with $d = 4$ and $d = 10$) respectively.

We note that while the Gaussian smoothing kernel is a popular and classical choice of smoothing kernel, it may not be optimal for estimating density functions that are sufficiently smooth (see Section 1.2 of Tsybakov [2008]).

We first demonstrate that whilst our hyperparameter selection minimizes the integrated squared error and our consistency result holds for the supremum norm of the difference between the estimated density and the true density, our estimator performs well in terms of both the MAD and the MSE. Figure 2.8 shows the distributions of the MAD and the MSE for our method

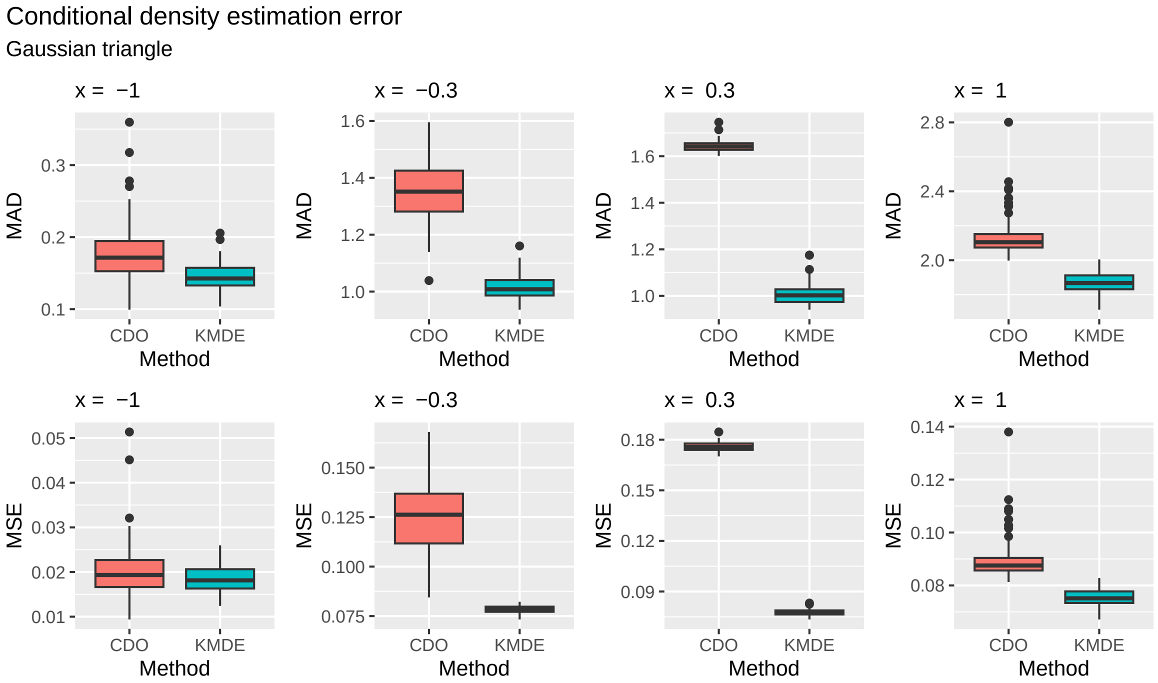


Figure 2.3: Comparison of the distributions of the MAD and MSE for our method KMDE and the existing method CDO when estimating the conditional density $p(y|x,z)$ for $z = 0$ and across $x \in \{-1,-0.3,0.3,1\}$ for Model 1. The four columns correspond to different values of x, and the first and second rows compare the distributions of the MAD and MSE respectively over 100 trials.

KMDE and the existing method KCDE when estimating the conditional density $p(y|x)$ over 50 values of x uniformly sampled from the observations $(x_i)_{i=1}^{1000}$, for data generated from Model 2 with $\alpha = \beta = 0.1$. The figure shows that our method outperforms the existing method for both metrics.

To compare the two approaches systematically across a range of datasets, for each simulated dataset we compute the MAD between their estimated densities and the true densities across a range of values of x. For the values of x we sample a set of 50 values uniformly from the observations $(x_i)_{i=1}^{1000}$. We then perform a two-sample t-test wherein we test the null hypothesis that there is no difference in the MAD between the two methods. Results that are significant at the 0.05 level are reported in Table 2.1. Our results compare the methods over a wide-range of different datasets, with the dimension of X ranging from 1 up to 12, and the dimension of Y up to 2. For approximately 89% (17/19) of the significant results, our method KMDE outperforms KCDE.

The kernel conditional density estimator estimates the density associated with the random variable X (and our method does not), and thus our method outperforms KCDE when the density for X is hard to model with a kernel density estimator. For example, in the case that $X \sim \text{Beta}(0.1,0.1)$, the density is bimodal with modes at 0 and 1, and the density spikes towards

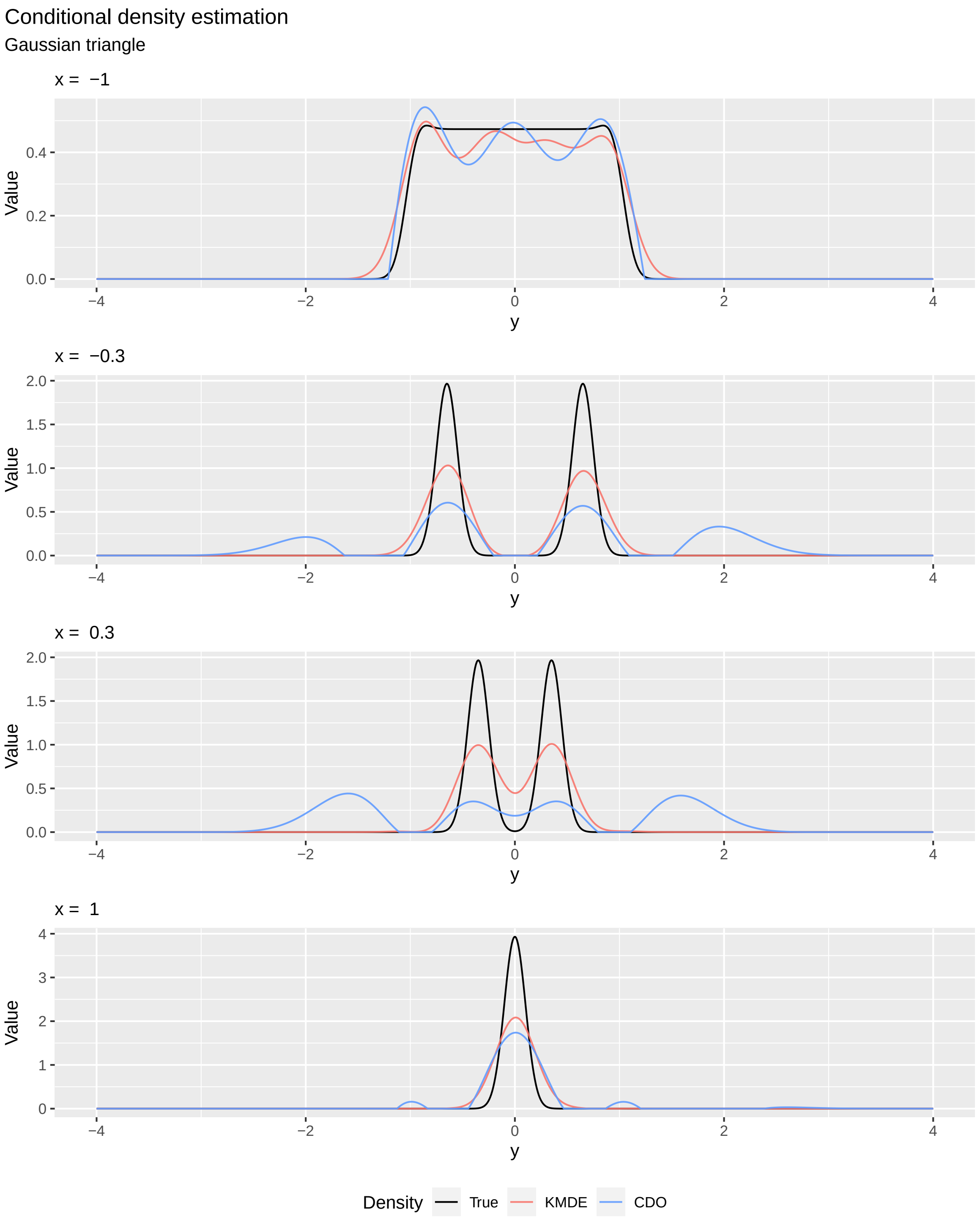


Figure 2.4: Several estimated 1D conditional densities of $Y|X = x, Z = 0$ across $x \in \{-1, -0.3, 0.3, 1\}$ for Model 1 using our method `KMDE` and an existing method `CDO`. Each plot shows the true density in black over a sequence of y for the value of x which is specified in the plot title.

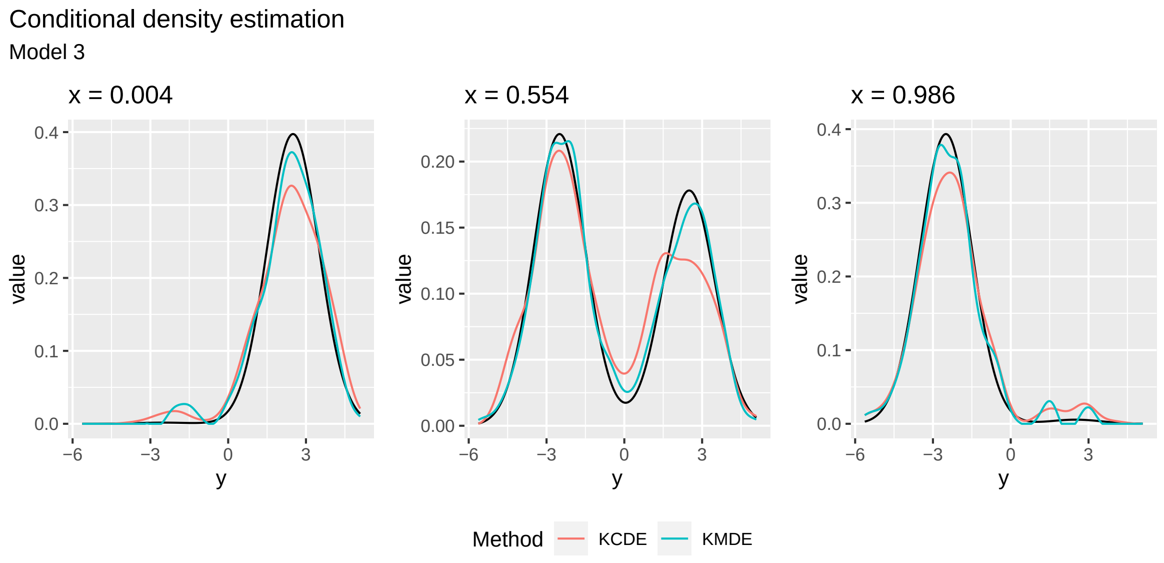


Figure 2.5: Comparison of estimated conditional densities for Model 3 using our method `KMDE` and an existing method `KCDE`. Each plot shows the true density $p(y|x)$ in black over a sequence of y for the value of x which is specified in the plot title.

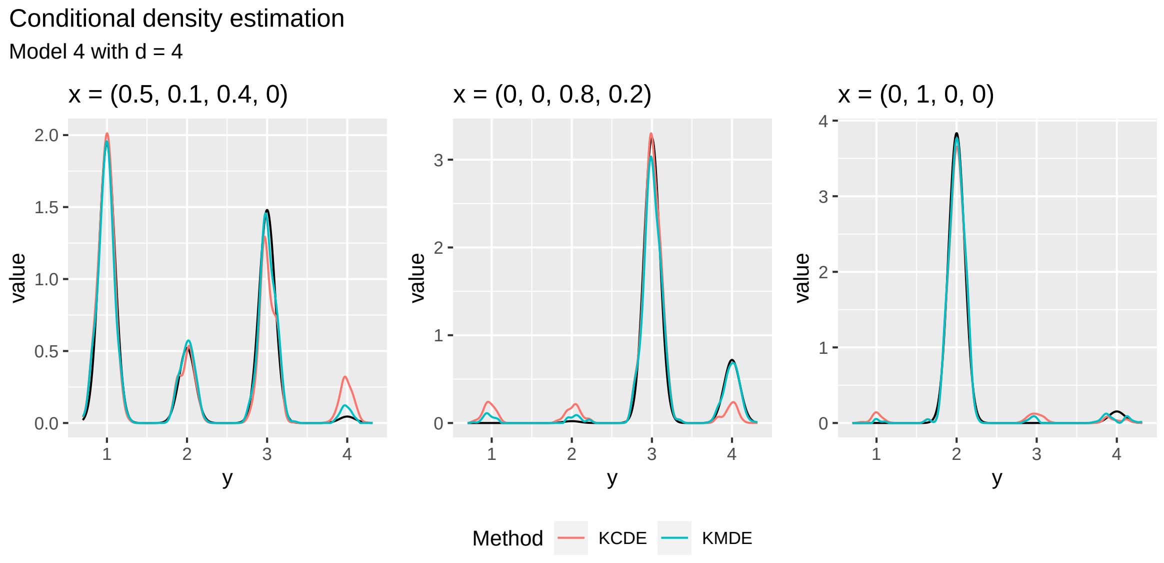


Figure 2.6: Comparison of estimated conditional densities for Model 4 with $d = 4$ using our method `KMDE` and an existing method `KCDE`. Each plot shows the true density $p(y|x)$ in black over a sequence of y for the value of x which is specified in the plot title.

the boundaries of the domain and decreases towards the center of the domain at 0.5. The density in this case is unbounded and the consistency results for kernel density estimation do not hold. Kernel density estimators will face several difficulties in estimating such a density. Firstly, kernel

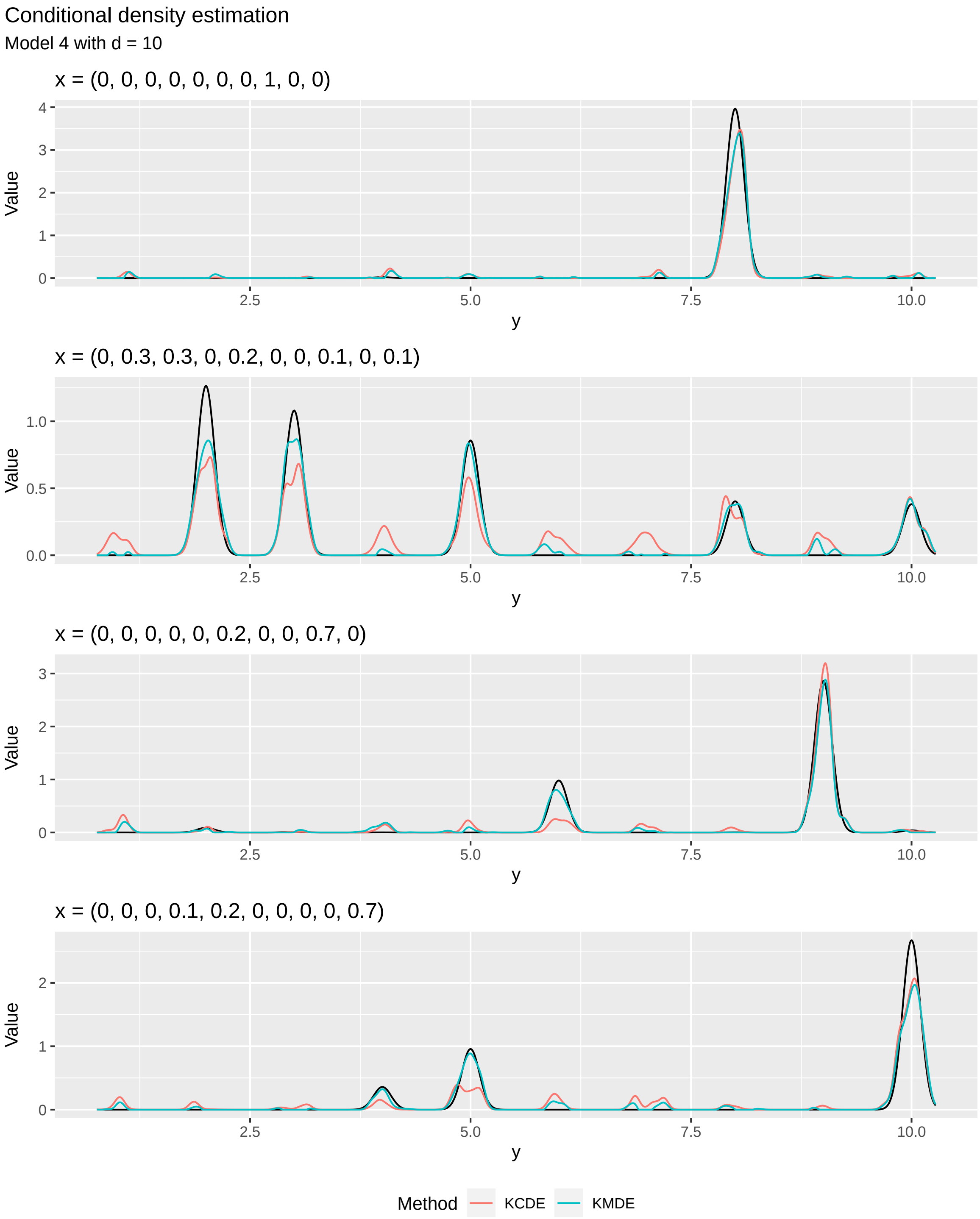


Figure 2.7: Comparison of estimated conditional densities for Model 4 with $d = 10$ using our method KMDE and an existing method KCDE. Each plot shows the true density $p(y|x)$ in black over a sequence of y for the value of x which is specified in the plot title.

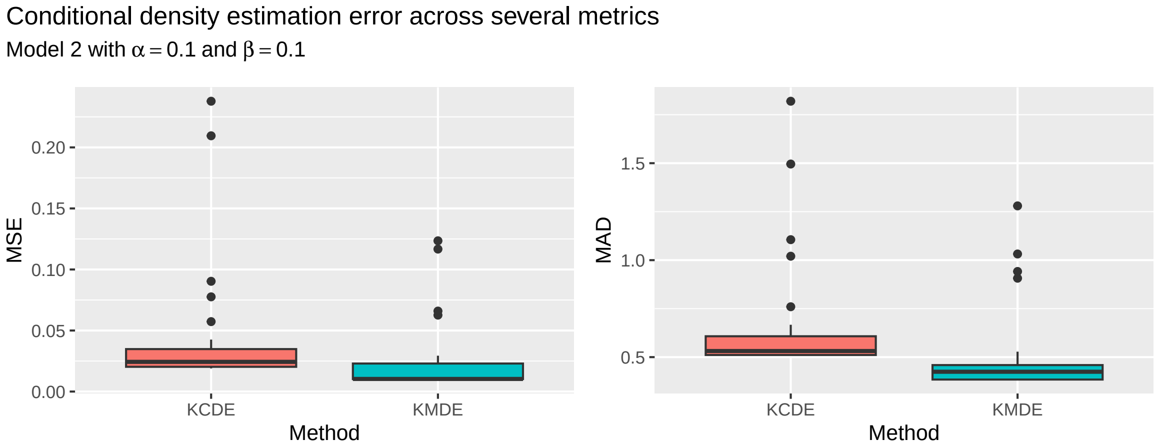


Figure 2.8: Comparison of the distributions of the MAD and MSE for our method KMDE and the existing method KCDE when estimating the conditional density $p(y|x)$ over 50 values of x for Model 2 with $\alpha = \beta = 0.1$.

density estimators are biased towards the boundary of the domain, so the density estimator is likely to underestimate the density towards its modes. Secondly, tuning the bandwidth is particularly difficult as one would require a small bandwidth to capture the modes of the density and a large bandwidth to capture the smooth behaviour between the modes. Finally, a majority of samples will take values close to 0 and 1, and very few samples will be close to 0.5, making the behaviour between the modes particularly hard to model. For a similar reasoning, our method outperforms the kernel conditional density estimator when X is sampled from a mixture of distributions with different variances, as the choice of bandwidth is not straightforward. Our claim is further supported by the fact that KCDE outperformed our method for both non-linear models where the distribution of X is high-dimensional but simple to model with a kernel density estimator.

Comparison over increasing dimension. Density estimation becomes significantly harder as the dimension of the data increases — a phenomenon referred to as the curse of dimensionality. For a fixed number of observations, as the dimension of the space increases the available data becomes more sparse, and the convergence rate of nonparametric density estimators slows. This problem cannot be avoided, as the optimal convergence rate of nonparametric density estimators decreases as a function of the dimension [Stone, 1980]. Hence, we expect our estimator KMDE to outperform KCDE as the dimension of the data increases, as KCDE estimates the density of the random variable X whereas KMDE does not.

In the following we compare our method KMDE to KCDE when the dimension of the random variable X increases. We simulate 1000 observations from Model 5 for d_x ranging from 1 to 200, and evaluate performance in estimating the 1D condition density $Y|X = x$. We follow the

d_x	d_y	**Model**	**p-value**	**Preferred**
1	1	Beta, $\alpha = \beta = 0.1$	<0.001***	KMDE
1	1	Bimodal	0.001**	KMDE
1	1	Cauchy, $\gamma = 0.5$	0.017*	KMDE
1	1	Cauchy, $\gamma = 2.5$	0.007**	KMDE
2	1	Dirichlet	<0.001***	KMDE
3	1	Dirichlet	<0.001***	KMDE
4	1	Dirichlet	<0.001***	KMDE
5	1	Dirichlet	<0.001***	KMDE
6	1	Dirichlet	<0.001***	KMDE
7	1	Dirichlet	<0.001***	KMDE
8	1	Dirichlet	0.010*	KMDE
9	1	Dirichlet	0.003**	KMDE
10	1	Dirichlet	<0.001***	KMDE
3	1	Linear Toeplitz	0.026*	KMDE
5	1	Linear Toeplitz	0.008**	KMDE
1	1	Mixture	0.014*	KMDE
1	2	MVN	0.002**	KMDE
10	1	Nonlinear type 1	0.008**	KCDE
10	1	Nonlinear type 2	<0.001***	KCDE

Table 2.1: The outcome of several hypothesis tests across different models and data dimensions in which we test the hypothesis that the MAD in estimating conditional densities using our method `KMDE` and the existing method `KCDE` is the same. The columns d_x and d_y denote the dimension of the X variable and Y variable respectively. In the p-value column, a single asterisk (*) indicates significance at the 0.05 level, double asterisks (**) at the 0.01 level, and triple asterisks (***) at the 0.001 level. Only significant results are shown in this table, and the Preferred column states which method achieved the lowest mean MAD.

same procedure described above. For each dataset, we tune the two parameters of KMDE and KCDE by minimizing the estimated score function defined in Equation (2.11). For the optimal hyperparameters which minimize the estimated score, we estimate the conditional density of $Y|X = x$ for 50 values of x uniformly sampled from the observed $(x_i)_{i=1}^{1000}$, and we compute the MAD between the estimated conditional density and the truth.

Figure 2.9 shows the mean MAD (averaged over the 50 values of x) for both methods as the dimension of the random variable X increases. As the dimension increases, the error of both methods increases, however the error for KMDE increases at a slower rate and is consistently less than the error of KCDE for $d_x > 3$. This significant improvement in performance stems from the fact that KMDE does not estimate the density of the random variable X. It is also evident that as the dimension increases the variance of the mean MAD increases for KCDE. We perform two-sample t-tests for each value of d_x, testing the null hypothesis that the MAD of the two methods is equal, and we report the results in Table 2.2. For d_x equal to 2, we find that KCDE outperforms our method KMDE; this is not surprising as the underlying model has simple dynamics, and estimating the conditional mean embedding induces numerical error as the estimator requires the inversion of a 1000 × 1000 matrix. The results for d_x equal to 1 and 3 to 4 were not significant, and for all $d_x \geq 5$ our method outperforms KCDE with the results becoming more significant as dimension increases. For all $d_x \geq 6$ the tests are significant at the 0.001 level.

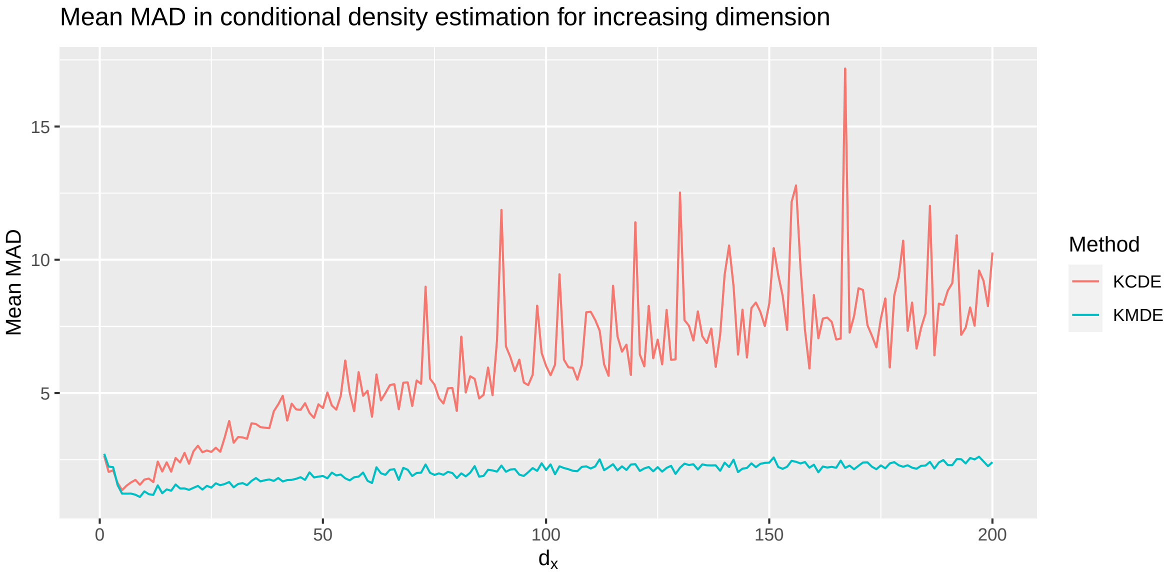


Figure 2.9: Comparison of the mean MAD between our method KMDE and the existing method KCDE when estimating the conditional density of $Y|X$ for increasing dimension d_x in Model 5. The mean MAD is the maximum absolute difference in estimating the density of $Y|X = x$, averaged over a set of 50 values of x.

d_x	d_y	**Model**	**p-value**	**Preferred**
2	1	Linear	0.023*	KCDE
5	1	Linear	0.029*	KMDE
6–100	1	Linear	<0.001***	KMDE

Table 2.2: The outcome of several hypothesis tests across different data dimensions in which we test the hypothesis that the MAD in estimating conditional densities using our method KMDE and the existing method KCDE is the same. The columns d_x and d_y denote the dimension of the X variable and Y variable respectively. In the p-value column, a single asterisk (*) indicates significance at the 0.05 level, double asterisks (**) at the 0.01 level, and triple asterisks (***) at the 0.001 level. Only significant results are shown in this table, and the Preferred column states which method achieved the lowest mean MAD.

2.5 Future work

The density estimator proposed in this work and its consistency is a self-contained contribution. Nevertheless, it presents several opportunities for further study. One promising direction involves studying the rate of convergence of the density estimator and potentially developing a theoretical result for the rate. Future research could also investigate the use of alternative kernel functions for the density estimator, such as higher-order kernel functions, which may improve the performance of the density estimator. Additionally, there is also potential to develop alternative hyperparameter selection procedures. It would be preferable to select the kernel hyperparameter in a manner motivated by the consistency result, by minimizing the supremum norm of the difference between the density estimator and the underlying density.

Furthermore, the proposed density estimator can be used in future works using kernel mean embeddings. The output of statistical methods using kernel mean embeddings is typically an estimated kernel mean embedding of a distribution of interest; it would be preferable to directly output an estimate of the distribution of interest. This is made possible by the density estimator we propose, allowing for end-to-end statistical modelling using kernel mean embeddings.

2.6 Conclusion

Embedding a probability distribution into a reproducing kernel Hilbert space is a relatively simple and well-studied task, and for many kernel functions this mapping is injective. The mapping from an RKHS embedding to a probability distribution is non-trivial, and understanding precisely what information can be recovered from such embeddings has remained an open question in the kernel mean embedding literature. Existing methods for estimating the embedded distribution involve non-convex optimization problems and parametric assumptions, with no guarantee on the recovered distribution.

In this work we have shown that given an estimated embedding, the density of the embedded distribution can be recovered at no cost, provided that the estimated embedding is consistent in probability. Furthermore, the density estimator is uniformly consistent in probability. Kernel mean embeddings offer a versatile framework for nonparametric statistical modelling. They facilitate the estimation of (embedded) distributions, using probabilistic intuition via the kernel sum, product, and Bayes' rule. Our research extends this framework, introducing the potential for comprehensive end-to-end modelling using kernel mean embeddings.

We showed that our estimator can be used to estimate conditional densities without parametric assumptions via the use of conditional mean embeddings. In this setting, we compared our estimator to conditional density operators [Schuster et al., 2020] and kernel conditional density estimators. Our experiments show that our method outperforms the existing kernel-based methods in terms of the maximum absolute difference in density estimation across a wide range of simulated datasets, and that the improvement in performance obtained by our method increases as the dimension of the conditioning variable increases.

2.7 Supplementary

2.7.1 Proof of Lemma 2.1

Proof of Lemma 2.1. Let $\epsilon > 0$ and note that, since $f \in C_0(\mathbb{R}^d)$, there exists an $a_\epsilon \in \mathbb{R}$ such that $f(x) < \epsilon/4$ for all $x \in \mathbb{R}^d$ such that $\|x\| \geq a_\epsilon$.

Next, let $x \in \mathbb{R}^d$ be such that $\|x\| \geq 2a_\epsilon$, and let

$$A_{n,\epsilon}(x) = \left\{ y \in \mathbb{R}^d \mid \|x + \sigma_n y\| \geq a_\epsilon \right\}, \quad \forall n \geq 1,$$

and remark that if $Y \sim \nu$ then for all $n \geq 1$ we have

$$\begin{aligned}\Pr\left(Y \in A_{n,\epsilon}(x)\right) &= \Pr(\|x + \sigma_n Y\| \geq a_\epsilon) \\ &\geq \Pr(\|x\| - \sigma_n \|Y\| \geq a_\epsilon) \\ &\geq \Pr(\sigma_n \|Y\| \leq a_\epsilon) \\ &=: p_n(\epsilon),\end{aligned}$$

where the first inequality holds by the reverse triangle inequality. Since $\lim_{n\to\infty} \sigma_n = 0$ it follows that

$$\lim_{n\to 0} p_n(\epsilon) = 1, \tag{2.13}$$

and thus

$$\begin{aligned}&\limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d:\|x\|\geq 2a_\epsilon} \left| \int_{\mathbb{R}^d} (f(x+\sigma_n y) - f(x))\,\nu(dy) \right| \\ &\leq \limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d:\|x\|\geq 2a_\epsilon} \int_{A_{n,\epsilon}(x)} |f(x+\sigma_n y) - f(x)|\,\nu(dy) \\ &\qquad + \limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d:\|x\|\geq 2a_\epsilon} \int_{A^c_{n,\epsilon}(x)} |f(x+\sigma_n y) - f(x)|\,\nu(dy) \\ &\leq \limsup_{n\to\infty} \left[\frac{\epsilon}{4} p_n(\epsilon) + \left(\|f\|_\infty + \frac{\epsilon}{4} \right)(1 - p_n(\epsilon)) \right] \\ &= \epsilon/4.\end{aligned} \tag{2.14}$$

To proceed further let $x \in \mathbb{R}^d$ be such that $\|x\| < 2a_\epsilon$,

$$\tilde{A}_{n,\epsilon}(x) = \{ y \in \mathbb{R}^d \mid \|x + \sigma_n y\| \leq 3a_\epsilon \}, \quad \forall n \geq 1,$$

and remark that if $Y \sim \nu$ then

$$\Pr\left(Y \in \tilde{A}_{n,\epsilon}(x)\right) = \Pr(\|x + \sigma_n Y\| \leq 3a_\epsilon) \geq \Pr(\|x\| + \sigma_n \|Y\| \leq 3a_\epsilon) \geq p_n(\epsilon).$$

In addition, as the function f is continuous on $\mathbb{R}^d$, it is uniformly continuous on the compact set $K_\epsilon := \{z \in \mathbb{R}^d \mid \|z\| \le 3a_\epsilon\}$. Therefore, there exists a continuous function $w_\epsilon : [0,\infty) \to [0,\infty)$ such that $w_\epsilon(0) = 0$ and such that

$$|f(x) - f(x')| \le w_\epsilon(\|x - x'\|), \quad \forall x, x' \in K_\epsilon,$$

and thus

$$\begin{aligned}
&\sup_{x\in\mathbb{R}^d:\|x\|<2a_\epsilon} \left| \int_{\mathbb{R}^d} (f(x+\sigma_n y) - f(x))\, \nu(dy) \right| \\
&\le \sup_{x\in\mathbb{R}^d:\|x\|<2a_\epsilon} \int_{\tilde{A}_{n,\epsilon}} |f(x+\sigma_n y) - f(x)|\, \nu(dy) \\
&\quad + \sup_{x\in\mathbb{R}^d:\|x\|<2a_\epsilon} \int_{\tilde{A}^c_{n,\epsilon}} |f(x+\sigma_n y) - f(x)|\, \nu(dy) \\
&\le \int_{\mathbb{R}^d} w_\epsilon(\sigma_n \|y\|) \nu(dy) + 2\|f\|_\infty (1 - p_n(\epsilon)).
\end{aligned} \tag{2.15}$$

Without loss of generality we can assume that w_ϵ is bounded (for example, by taking w_ϵ such that $w_\epsilon(\delta) = w_\epsilon(\delta_\epsilon)$ for all $\delta \ge \delta_\epsilon := \sup_{x,x'\in K_\epsilon} \|x - x'\|$). Therefore, using Equations (2.13) and (2.15) and the dominated convergence theorem, it follows that

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d:\|x\|<2a_\epsilon} \left| \int_{\mathbb{R}^d} (f(x+\sigma_n y) - f(x)) \nu(dy) \right| \le 0. \tag{2.16}$$

Therefore, using Equations (2.14) and (2.16), we have

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{x\mathbb{R}^d} \left| \int_{\mathbb{R}^d} (f(x+\sigma_n y) - f(x))\, \nu(dy) \right| \\
&\le \limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d:\|x\|\ge 2a_\epsilon} \left| \int_{\mathbb{R}^d} (f(x+\sigma_n y) - f(x))\, \nu(dy) \right| \\
&\quad + \limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d:\|x\|<2a_\epsilon} \left| \int_{\mathbb{R}^d} (f(x+\sigma_n y) - f(x))\, \nu(dy) \right| \\
&\le \frac{\epsilon}{4}.
\end{aligned}$$

As $\epsilon > 0$ is arbitrary, this concludes the proof. ■

2.7.2 Assumptions and proof of Corollary 2.1

The following uses results from Li et al. [2022], and so we briefly adopt their notation. We keep the following to a high-level and refer the reader to the original paper for the details. Let $\mathscr{H}_\mathscr{X}$ and $\mathscr{H}_\mathscr{Y}$ denote RKHSs on $\mathscr{X}$ and $\mathscr{Y}$. Li et al. [2022] consider a misspecified setting where the conditional mean embedding operator belongs to the space of Hilbert-Schmidt operators mapping from an interpolation space between $\mathscr{H}_\mathscr{X}$ and L_2, to $\mathscr{H}_\mathscr{Y}$. Let $[\mathscr{G}]^\beta$, $\beta \ge 0$ denote an interpolation space between $\mathscr{H}_\mathscr{X}$ and L_2; the spaces are such that $[\mathscr{G}]^\beta \hookrightarrow [\mathscr{G}]^\alpha$ for $0 < \alpha < \beta$. Let $F^\star(x)$ denote the kernel mean embedding of the conditional distribution $Y|X = x$ in $\mathscr{H}_\mathscr{Y}$, for some $x \in \mathscr{X}$, and let $\hat{F}^{\lambda}(x)$ denote the empirical estimated conditional mean embedding with regularization parameter $\lambda > 0$. In the well-specified setting, one has $F^\star \in \mathcal{G}$, the space of Hilbert-Schmidt operators mapping from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}}$.

The consistency result of Li et al. [2022] depends on their assumptions (1)-(3), (EVD), (EMB), and (SRC). The additional assumptions required are detailed below:

(EVD) Eigenvalue decay: Let $(\mu_i)_{i \in I}$ denote the non-increasing sequence eigenvalues of $\mathcal{C}_{XX}$ where I is an at most countable index set. For some constants $c_2 > 0$ and $p \in (0,1]$ and for all $i \in I$,

$$\mu_i \le c_2 i^{-1/p}.$$

(EMB) Embedding property: For $\alpha \in (p,1]$, the inclusion map $I_\pi^{\alpha,\infty} : [\mathcal{H}]_X^\alpha \hookrightarrow L_\infty(\pi)$ is continuous, and there is a constant $A > 0$ such that

$$\left\| I_\pi^{\alpha,\infty} \right(\|_{[\mathcal{H}]_X^\alpha \to L_\infty(\pi)} \le A.$$

(SRC) Source condition: There exists $0 < \beta \le 2$ such that

$$F^\star \in [\mathcal{G}]^\beta.$$

Assumptions (1)-(3) are equivalent to Assumption 1.1. (EVD) is an assumption on the eigenvalue decay of the cross-covariance operator $\mathcal{C}_{XX}$ (that the decay is at least polynomial with degree $p \in (0,1]$), (EMB) is an assumption on the interpolation spaces, and (SRC) assumes that $F^\star \in [\mathcal{G}]^\beta$ for some $0 < \beta \le 2$. Theorem 2 of Li et al. [2022] states that under Assumption 1.1, and assumptions (EVD), (EMB), and (SRC), $\|[\hat{F}^\lambda] - F^\star\|_\alpha^2 = o_{\mathbb{P}}(1)$, for $0 \le \alpha \le 1$ with $\alpha < \beta$, where $\|\cdot\|_\alpha$ denotes the α-norm.

The following lemma allows us to relate the RKHS norm and the norm associated with the interpolation space. The proof mirrors Lemma 4 of Li et al. [2022] in its approach, though the conclusion is different.

Lemma 2.3. *For any $F \in [\mathcal{G}]^\alpha$ and $x \in \mathcal{X}$, we have $\|F(x)\|_{\mathcal{H}_{\mathcal{Y}}}^2 \le \|F\|_\alpha^2 \|k_{\mathbb{P}_X}^\alpha\|_\infty$, where $\|k_{\mathbb{P}_X}^\alpha\|_\infty < \infty$ under assumption (EMB), for $\mathbb{P}_X$-almost all $x \in \mathcal{X}$.*

Proof. As $F \in [\mathcal{G}]^\alpha$, there exists $a_{ij} \in \ell_2(I \times J)$ such that for $\mathbb{P}_X$-almost all $x \in \mathcal{X}$,

$$F(x) = \sum_{i \in I} \sum_{j \in J} a_{ij} d_j \mu_i^{\alpha/2} [e_i](x),$$

where $(d_j)_{j\in J}$ is any orthonormal basis of $\mathscr{H}_{\mathscr{Y}}$. Then

$$\begin{aligned}
\|F(x)\|^2_{\mathscr{H}_{\mathscr{Y}}} &= \| \sum_{j\in J} \left(\sum_{i\in I} a_{ij} \mu_i^{\alpha/2} [e_i](x) \right) d_j \|^2_{\mathscr{H}_{\mathscr{Y}}} \\
&= \sum_{j\in J} \left(\sum_{i\in I} a_{ij} \mu_i^{\alpha/2} [e_i](x) \right)^2 \\
&\le \sum_{j\in J} \sum_{i\in I} a_{ij}^2 \sum_{i\in I} \mu_i^{\alpha} [e_i]^2(x) \\
&\le \|k^{\alpha}_{\mathbb{P}_X}\|_\infty \|F\|^2_\alpha,
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality and $\|F\|^2_\alpha = \sum_{i\in I}\sum_{j\in J} a_{ij}^2$. ■

Corollary 2.1 uses the above lemma to state that if the estimated conditional mean embedding operators are consistent, then the estimated conditional mean embeddings are consistent. In the final part of the following proof we translate the result from the notation of Li et al. [2022] to the notation used throughout this thesis. We note that when $\mathscr{H}_{\mathscr{Y}}$ is an RKHS with kernel function $k^\gamma : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$, the embedding of the distribution of $Y|X = x$ in $\mathscr{H}_{\mathscr{Y}}$ is $F^\star(x) = \mu^\gamma_{Y|X=x}$, and the estimated embedding in $\mathscr{H}_{\mathscr{Y}}$ with regularization $\lambda > 0$ is $\hat{F}^\lambda(x) = \hat{\mu}^\gamma_{Y|X=x}$.

Proof of Corollary 2.1. Under assumption (SRC), $F^\star \in [\mathscr{G}]^\beta$ for some $0 < \beta \le 2$, and thus $F^\star \in [\mathscr{G}]^\alpha$ for $\alpha < \beta$. It follows from Lemma 2.3 that $\|\hat{F}^\lambda(x) - F^\star(x)\|^2_{\mathscr{H}_{\mathscr{Y}}} \le \|\hat{F}^\lambda - F^\star\|^2_\alpha \|k^\alpha_{\mathbb{P}_X}\|_\infty$ for $\mathbb{P}_X$-almost all $x \in \mathscr{X}$. Assumption (EMB) is equivalent to the assumption that $\|k^\alpha_{\mathbb{P}_X}\|_\infty \le C$ for some constant $C > 0$ and $\alpha \in (p, 1]$ (see Theorem 9 of Fischer and Steinwart [2020]). Hence $\|\hat{F}^\lambda(x) - F^\star(x)\|^2_{\mathscr{H}_{\mathscr{Y}}} = o_{\mathbb{P}}(1)$, and equivalently $\|\hat{\mu}^\gamma_{Y|X=x} - \mu^\gamma_{Y|X=x}\|^2_{\mathscr{H}_{\mathscr{Y}}} = o_{\mathbb{P}}(1)$ for $\mathbb{P}_X$-almost all $x \in \mathscr{X}$. Thus there exists a set $A \subseteq \mathscr{X}$ with $\mathbb{P}_X(A) = 1$ such that

$$\sup_{x\in A} \|\hat{\mu}^\gamma_{Y|X=x} - \mu^\gamma_{Y|X=x}\|^2_{\mathscr{H}_{\mathscr{Y}}} = o_{\mathbb{P}}(1).$$

■

Chapter 3

Hidden Markov Models

3.1 Introduction

In this chapter we develop a nonparametric method for learning the parameters of an HMM via the use of kernel mean embeddings. We embed the joint distributions of (Y_1, Y_2) and (Y_1, Y_2, Y_3) into reproducing kernel Hilbert spaces, and estimate the RKHS embeddings of the observation distributions. The embeddings can then be manipulated via simple linear algebra to obtain estimates of the stationary distribution and transition matrix. We propose two methods to perform inference on the hidden states in the filtering problem. The first method we propose recovers the observation densities from their embeddings following Chapter 2, and then uses the forward algorithm. For the second method, we propose a novel alternative kernel Bayes' rule to estimate the embedded filtering distributions using the embedded observation distributions, avoiding the intermediary density estimation step of the first method. We derive an estimator of the HMM order for our setting, which consistently estimates the HMM order almost surely. We conclude the chapter with a comparison to existing nonparametric methods for HMMs.

The method we propose builds upon the works of Anandkumar et al. [2012] and De Castro et al. [2017]. Anandkumar et al. [2012] developed a method for learning parametric hidden Markov models, and more generally multi-view models, using the spectral method of Chang [1996] and a sample of three consecutive observations. They prove that their method is consistent and provide non-asymptotic convergence guarantees. De Castro et al. [2017] built upon Anandkumar et al. [2012] and developed a method for estimating a nonparametric HMM by projecting observations onto an approximation space. They prove that their method is consistent, and provide controls on the filtering and smoothing errors in terms of the risk of the estimated HMM parameters. Lehéricy [2019] proved that these estimators can be used to consistently estimate the order of the model almost surely.

Novelty. Our work is novel on several fronts. We develop a new nonparametric method for estimating the parameters of a hidden Markov model, and we derive concentration inequalities that highlight the method's improvement upon the existing spectral method of De Castro et al. [2017]. We formulate a novel alternative kernel Bayes' rule, which allows for applications of Bayes' rule to kernel mean embeddings without access to samples. This allows for inference in the filtering problem using the estimated parameters learned by our nonparametric method. We provide a kernel-based estimator of the HMM order and prove its almost-sure consistency. Finally, we demonstrate that our method outperforms existing nonparametric methods across a range of simulated datasets.

3.1.1 Setting and assumptions

Throughout the chapter we use the notation defined in Sections 1.1 and 1.2.1.1. We also require the following assumptions.

Assumption 3.1. *The kernel function defined on $\mathcal{Y}$ is characteristic, and the associated RKHS is separable.*

Assumption 3.2. *The transition matrix Q has full rank, and the hidden process Markov chain is irreducible and aperiodic.*

Assumption 3.3. *The initial distribution π is the stationary distribution.*

Assumption 3.4. *The set of observation densities F is linearly independent.*

Assumption 3.1 is used to ensure that the kernel mean embeddings are unique, and many popular kernels such as the Gaussian, Laplace, and Matérn kernel defined on a locally compact Hausdorff space such as $\mathbb{R}^d$, $d \geq 1$, are characteristic [Sriperumbudur et al., 2011]. Assumptions 3.2 to 3.4 are sufficient to obtain identifiability for nonparametric models upon observing a sequence of three consecutive observations, see Gassiat et al. [2016]. Assumptions 3.2 and 3.3 also ensure that $\pi_i > 0$ for $i = 1, \dots, K$.

3.2 Problem formulation

A hidden Markov model is characterized by the set of observation densities, F, the transition matrix of the hidden process, Q, and the stationary distribution of the hidden process, π. Statistical estimation of hidden Markov models focuses on estimating the set of parameters (F, Q, π) given a sequence of observations. In this chapter we develop a nonparametric procedure to estimate the parameters (O_2, Q, π), where O_2 denotes the set of observation distributions embedded in an

RKHS. We also address the estimation of the HMM order, a crucial yet often challenging aspect of model estimation.

As discussed in Chapter 1, several inference problems arise when working with hidden Markov models. We focus on the filtering task wherein we aim to estimate the posterior distributions of the current hidden state X_t given a sequence of observations $y_{1:t}$ for $t \geq 1$.

3.3 Nonparametric estimation via the kernel spectral method

The method we propose to estimate the parameters of the hidden Markov model relies upon several preliminary lemmas. To facilitate a better understanding and to guide the reader through this section, we provide a brief outline of the section below.

An outline. Firstly, Section 3.3.1 defines the kernel mean embeddings of the distributions of two and three consecutive observations, and Section 3.3.2 shows that these embeddings can be decomposed in terms of the embedded observation densities, O_2. Section 3.3.3 leverages the decomposition to define a quantity termed the observable operator. This quantity is key to defining a linear system, defined in Lemma 3.4, which allows for the estimation of O_2 using a set of observations from the HMM. Lastly, Section 3.3.4 provides an expression for the remaining parameters Q and π in terms of O_2. These expressions allow for the estimation of all HMM parameters.

3.3.1 Embeddings

To derive a method for estimating the parameters of a hidden Markov model satisfying Assumptions 3.1 to 3.4 we start by embedding the distributions of two and three consecutive observations into reproducing kernel Hilbert spaces. The use of three consecutive observations is sufficient for identifiability of the nonparametric model following Gassiat et al. [2016].

Let $\mathcal{H}_{\mathcal{Y}}$ be a reproducing kernel Hilbert space with kernel function $k(y,y') = \langle \phi_Y(y), \phi_Y(y') \rangle_{\mathcal{H}_{\mathcal{Y}}}$, where $\phi_Y : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ is the canonical feature mapping associated with the kernel function. We embed the distribution of Y_1 in $\mathcal{H}_{\mathcal{Y}}$ via the kernel mean embedding

$$\mu_{Y_1} = \mathbb{E}_{Y_1}[\phi_Y(Y_1)] \in \mathcal{H}_{\mathcal{Y}},$$

and we embed the joint distributions of (Y_1, Y_2) and (Y_1, Y_2, Y_3) into the tensor product RKHSs $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}}$, and $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}}$ via the cross-covariance operators

$$\mathscr{C}_{Y_1,Y_2} = \mathbb{E}_{Y_1,Y_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)], \qquad \mathscr{C}_{Y_1,Y_2,Y_3} = \mathbb{E}_{Y_1,Y_2,Y_3}[\phi_Y(Y_1) \otimes \phi_Y(Y_2) \otimes \phi_Y(Y_3)].$$

These embeddings exist under Assumption 1.1, see Lemma 1.2. To simplify notation in the following, we define $\mu_1 = \mu_{Y_1}$, $\mathscr{C}_{1,2} = \mathscr{C}_{Y_1,Y_2}$, and $\mathscr{C}_{1,2,3} = \mathscr{C}_{Y_1,Y_2,Y_3}$, and we define $\mathcal{H}_{\mathcal{Y}}^{\otimes v} = \otimes_{i=1}^{v} \mathcal{H}_{\mathcal{Y}}$ for $v = 2,3$.

For $v \in \{1,2,3\}$, we define O_v to be a row vector of K elements in $\mathcal{H}_{\mathcal{Y}}$, denoted $O_v \in \mathcal{H}_{\mathcal{Y}}^K$, with j-th element $\mathbb{E}[\phi_Y(Y_v)|X_2 = j]$, $j = 1,\dots,K$. O_v denotes the kernel mean embedding of the observation Y_v given the hidden state X_2, and we note that O_2 is of particular interest as it contains the embeddings of the observation distributions. The j-th element of the row vector O_2 corresponds to the embedding of $(Y_2|X_2 = j)$ in $\mathcal{H}_{\mathcal{Y}}$ for $j = 1,\dots,K$.

The quantities O_1, O_2, and O_3 are closely related as shown in the following lemma. This observation follows naturally from the fact that a hidden Markov model can be considered a multi-view model as shown in Figure 1.2, an observation which motivated previous parametric spectral HMMs [Hsu et al., 2012, Anandkumar et al., 2012].

Lemma 3.1. *We can express O_1 and O_3 in terms of O_2, and the HMM parameters Q and π as follows:*

$$O_1 = O_2 \operatorname{diag}(\pi) Q \operatorname{diag}(Q\pi)^{-1}, \quad O_3 = O_2 Q^{\mathrm{T}}.$$

Proof. Consider the i-th element of O_1. It follows from the law of total expectation and the conditional independence implied by the hidden Markov model's structure that

$$\begin{aligned}
[O_1]_i &= \mathbb{E}[\phi_Y(Y_1)|X_2 = i] \\
&= \mathbb{E}_{X_1|X_2=i}\left[\mathbb{E}[\phi_Y(Y_1)|X_2 = i, X_1]|X_2 = i\right] \\
&= \mathbb{E}_{X_1|X_2=i}\left[\mathbb{E}[\phi_Y(Y_1)|X_1]|X_2 = i\right] \\
&= \sum_{j=1}^{K} \mathbb{E}[\phi_Y(Y_1)|X_1 = j]\mathbb{P}(X_1 = j|X_2 = i) \\
&= [O_2 \operatorname{diag}(\pi) Q \operatorname{diag}(\pi Q)^{-1}]_i,
\end{aligned}$$

where we have used Bayes' rule to express

$$\begin{aligned}
\mathbb{P}(X_1 = j|X_2 = i) &= \frac{\mathbb{P}(X_2 = i|X_1 = j)\mathbb{P}(X_1 = j)}{\mathbb{P}(X_2 = i)} \\
&= \frac{Q_{j,i}\pi_j}{[\pi Q]_i} \\
&= [\operatorname{diag}(\pi) Q \operatorname{diag}(\pi Q)^{-1}]_{j,i}.
\end{aligned}$$

Similarly, we can represent the i-th element of O_3 as follows

$$\begin{aligned}
[O_3]_i &= \mathbb{E}[\phi_Y(Y_3)|X_2 = i] \\
&= \mathbb{E}_{X_3|X_2=i}\left[\mathbb{E}[\phi_Y(Y_3)|X_2 = i, X_3]|X_2 = i\right] \\
&= \mathbb{E}_{X_3|X_2=i}\left[\mathbb{E}[\phi_Y(Y_3)|X_3]|X_2 = i\right] \\
&= \sum_{j=1}^{K} \mathbb{E}[\phi_Y(Y_3)|X_3 = j]\mathbb{P}(X_3 = j|X_2 = i) \\
&= [O_2 Q^{\mathrm{T}}]_i.
\end{aligned}$$

■

3.3.2 Decomposition

The key to our method is contained in the following lemma, which shows that the embeddings of (Y_1, Y_2) and (Y_1, Y_2, Y_3) can be decomposed into various quantities of interest.

In the following we treat $\mathscr{C}_{1,2,3} \in \mathscr{H}_{\mathscr{Y}}^{\otimes 3}$ as a rank-one linear operator, and use the notation $\times_2$ introduced in Section 1.2.2.5 to emphasize that the inner product is taken over the second dimension, such that

$$\mathscr{C}_{1,2,3} \times_2 : \mathscr{H}_{\mathscr{Y}} \to \mathscr{H}_{\mathscr{Y}}^{\otimes 2}, \quad \mathscr{C}_{1,2,3} \times_2 : f \mapsto \mathbb{E}_{Y_1,Y_2,Y_3}[\left(\phi_Y(Y_1) \otimes \phi_Y(Y_3)\right) \langle \phi_Y(Y_2), f \rangle_{\mathscr{H}_{\mathscr{Y}}}].$$

That is, the cross-covariance operator can be considered an operator which maps a function $f \in \mathscr{H}_{\mathscr{Y}}$ to an element of the tensor product RKHS $\mathscr{H}_{\mathscr{Y}}^{\otimes 2}$ by taking an inner product over the second dimension.

Lemma 3.2. *The embeddings of* (Y_1, Y_2) *and* (Y_1, Y_2, Y_3) *in* $\mathscr{H}_{\mathscr{Y}}^{\otimes 2}$ *and* $\mathscr{H}_{\mathscr{Y}}^{\otimes 3}$ *can be decomposed as*

$$\mathscr{C}_{1,3} = O_1 \operatorname{diag}(\pi Q) O_3^{\mathrm{T}}, \qquad \textit{and} \qquad \mathscr{C}_{1,2,3} \times_2 f = O_1 \operatorname{diag}\left(O_2^{\mathrm{T}} f\right) \operatorname{diag}(\pi Q) O_3^{\mathrm{T}},$$

for any $f \in \mathscr{H}_{\mathscr{Y}}$.

Proof. The first decomposition follows from the law of total expectation and the conditional independence structure specified by the hidden Markov model,

$$\begin{aligned} \mathscr{C}_{1,3} &= \mathbb{E}_{Y_1,Y_3}[\phi_Y(Y_1) \otimes \phi_Y(Y_3)] \\ &= \mathbb{E}_{X_2}\left[\mathbb{E}_{Y_1,Y_3|X_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_3)|X_2]\right] \\ &= \mathbb{E}_{X_2}\left[\mathbb{E}_{Y_1|X_2}[\phi_Y(Y_1)|X_2] \otimes \mathbb{E}_{Y_3|X_2}[\phi_Y(Y_3)|X_2]\right] \\ &= \sum_{i=1}^{K} \left[\mathbb{E}_{Y_1|X_2}[\phi_Y(Y_1)|X_2 = i] \otimes \mathbb{E}_{Y_3|X_2}[\phi_Y(Y_3)|X_2 = i]\right][\pi Q]_i \\ &= O_1 \operatorname{diag}(\pi Q) O_3^{\mathrm{T}}, \end{aligned}$$

where we have used Lemma 1.3 and the final line follows from the definition of O_v as a row vector in the RKHS. The second equation follows similarly,

$$\begin{aligned} \mathscr{C}_{1,2,3} \times_2 f &= \mathbb{E}_{Y_1,Y_2,Y_3}[\left(\phi_Y(Y_1) \otimes \phi_Y(Y_3)\right) \langle \phi_Y(Y_2), f \rangle_{\mathscr{H}_{\mathscr{Y}}}] \\ &= \mathbb{E}_{X_2}\left[\mathbb{E}_{Y_1,Y_2,Y_3|X_2}[\left(\phi_Y(Y_1) \otimes \phi_Y(Y_3)\right) \langle \phi_Y(Y_2), f \rangle_{\mathscr{H}_{\mathscr{Y}}} |X_2]\right] \\ &= \mathbb{E}_{X_2}\left[\left(\mathbb{E}_{Y_1|X_2}[\phi_Y(Y_1)|X_2] \otimes \mathbb{E}_{Y_3|X_2}[\phi_Y(Y_3)|X_2]\right) \mathbb{E}_{Y_2|X_2}[\langle \phi_Y(Y_2), f \rangle_{\mathscr{H}_{\mathscr{Y}}} |X_2]\right] \\ &= \sum_{i=1}^{K} \left(\mathbb{E}_{Y_1|X_2}[\phi_Y(Y_1)|X_2 = i] \otimes \mathbb{E}_{Y_3|X_2}[\phi_Y(Y_3)|X_2 = i]\right) \mathbb{E}_{Y_2|X_2}[\langle \phi_Y(Y_2), f \rangle_{\mathscr{H}_{\mathscr{Y}}} |X_2 = i][\pi Q]_i \\ &= O_1 \operatorname{diag}\left(O_2^{\mathrm{T}} f\right) \operatorname{diag}(\pi Q) O_3^{\mathrm{T}}. \end{aligned}$$

■

Lemma 3.2 states that the cross-covariance operators $\mathscr{C}_{1,2}$ and $\mathscr{C}_{1,2,3}$, which are easily estimated via the sample mean given a set of samples of (Y_1, Y_2, Y_3), can be written in terms of the embedded observation densities O_2.

3.3.3 The observable operator

The following lemma defines a quantity termed the observable operator. The operator is observable in the sense that it can be estimated from the HMM observations, and as an operator it can be used to define a representation of the HMM referred to as an observable operator model [Jaeger, 2000, Hsu et al., 2012]. We do not make use of this representation here; instead we focus on the spectral properties of the observable operator described in the following lemma.

Lemma 3.3. *For $v \in \{1,2,3\}$, let $U_v \in \mathscr{H}_{\mathscr{Y}}^K$ be a row vector such that $U_v^\mathrm{T} O_v \in \mathbb{R}^{K\times K}$ is invertible. Then $U_1^\mathrm{T}\mathscr{C}_{1,3}U_3$ is invertible under Assumptions 3.2 and 3.3. The observable operator $\mathscr{B}_{1,2,3} : \mathscr{H}_{\mathscr{Y}} \to \mathbb{R}^{K\times K}$ is defined by*

$$\mathscr{B}_{1,2,3}(f) := U_1^\mathrm{T}(\mathscr{C}_{1,2,3} \times_2 f)U_3(U_1^\mathrm{T}\mathscr{C}_{1,3}U_3)^{-1},$$

and satisfies

$$\mathscr{B}_{1,2,3}(f) = (U_1^\mathrm{T}O_1)\operatorname{diag}\left(O_2^\mathrm{T} f\right)(U_1^\mathrm{T}O_1)^{-1}, \quad \forall f \in \mathscr{H}_{\mathscr{Y}}.$$

Proof. Under Assumptions 3.2 and 3.3, $\pi_i > 0$ for $i = 1,\dots,K$, and so $\operatorname{diag}(\pi)$ is invertible. As the transition matrix Q has full rank, it follows from Lemma 3.2 that $U_1^\mathrm{T}\mathscr{C}_{1,3}U_3$ is invertible. Using the decompositions stated in Lemma 3.2, we can show that

$$\begin{aligned}
\mathscr{B}_{1,2,3}(f) &= U_1^\mathrm{T}(\mathscr{C}_{1,2,3} \times_2 f)U_3(U_1^\mathrm{T}\mathscr{C}_{1,3}U_3)^{-1} \\
&= U_1^\mathrm{T}O_1 \operatorname{diag}\left(O_2^\mathrm{T} f\right)\operatorname{diag}(\pi Q)O_3^\mathrm{T}U_3(U_1^\mathrm{T}O_1\operatorname{diag}(\pi Q)O_3^\mathrm{T}U_3)^{-1} \\
&= U_1^\mathrm{T}O_1 \operatorname{diag}\left(O_2^\mathrm{T} f\right)(U_1^\mathrm{T}O_1)^{-1}(U_1^\mathrm{T}O_1)\operatorname{diag}(\pi Q)O_3^\mathrm{T}U_3(U_1^\mathrm{T}O_1\operatorname{diag}(\pi Q)O_3^\mathrm{T}U_3)^{-1} \\
&= U_1^\mathrm{T}O_1 \operatorname{diag}\left(O_2^\mathrm{T} f\right)\left(U_1^\mathrm{T}O_1\right)^{-1}.
\end{aligned}$$

In the above, the operator $\mathscr{C}_{1,3}$ is applied element-wise such that $\mathscr{C}_{1,3}U_3 \in \mathscr{H}_{\mathscr{Y}}^K$. U_1 is a row vector in $\mathscr{H}_{\mathscr{Y}}^K$, and we interpret U_1^T to be the equivalent column vector in $\mathscr{H}_{\mathscr{Y}}^K$. Thus, $U_1^\mathrm{T}\mathscr{C}_{1,3}U_3 \in \mathbb{R}^{K\times K}$. Following the same reasoning we see that $U_1^\mathrm{T}(\mathscr{C}_{1,2,3} \times_2 f)U_3 \in \mathbb{R}^{K\times K}$. Hence, $\mathscr{B}_{1,2,3}$ is an operator mapping from the RKHS $\mathscr{H}_{\mathscr{Y}}$ to $\mathbb{R}^{K\times K}$. ■

Lemma 3.3 shows that for any $f \in \mathscr{H}_{\mathscr{Y}}$, the matrices $\mathscr{B}_{1,2,3}(f)$ and $\operatorname{diag}\left(O_2^\mathrm{T} f\right)$ are closely related. In particular, the columns of $U_1^\mathrm{T}O_1$ are eigenvectors of $\mathscr{B}_{1,2,3}(f)$ with associated eigenvalues $O_2^\mathrm{T} f \in \mathbb{R}^K$. The eigenvectors are only defined up to a scaling, and thus we cannot directly use this fact to estimate O_2. We apply the observable operator to several inputs, and obtain O_2 as the solution to a linear system. The linear system is defined in the following lemma.

Lemma 3.4. *Let $\Theta \in \mathbb{R}^{K\times K}$ be an invertible matrix, and let $\theta_i^\mathrm{T} \in \mathbb{R}^K$ be its i-th row. Then for $i = 1,\dots,K$, $U_2\theta_i \in \mathscr{H}_{\mathscr{Y}}$, and we let $\lambda_{i,1},\dots,\lambda_{i,K}$ denote the eigenvalues of $\mathscr{B}_{1,2,3}(U_2\theta_i)$, in the order specified by the matrix of right eigenvectors $U_1^\mathrm{T}O_1$. Let $L \in \mathbb{R}^{K\times K}$ denote the matrix with (i,j)-th entry $\lambda_{i,j}$, then*

$$\Theta U_2^\mathrm{T} O_2 = L.$$

Hence, the RKHS vector $O_2 \in \mathscr{H}_{\mathscr{Y}}^K$ is the solution to a linear system.

Proof. It follows from Lemma 3.3 that for $f = U_2\theta_i \in \mathscr{H}_{\mathscr{Y}}$, for all $i \in \{1,\dots,K\}$

$$\begin{aligned}(U_1^{\mathrm{T}}O_1)^{-1}\mathscr{B}_{1,2,3}(U_2\theta_i)(U_1^{\mathrm{T}}O_1) &= \operatorname{diag}\left(O_2^{\mathrm{T}}U_2\theta_i\right)\\ &= \operatorname{diag}\left(\left\langle O_2^{\mathrm{T}}U_2e_1,\theta_i\right\rangle,\dots,\left\langle O_2^{\mathrm{T}}U_2e_K,\theta_i\right\rangle\right)\\ &= \operatorname{diag}\left(\lambda_{i,1},\dots,\lambda_{i,K}\right).\end{aligned}$$

Let L denote the matrix with (i,j)-th entry $\lambda_{i,j}$, then it follows that

$$L = \begin{bmatrix}\langle O_2^{\mathrm{T}}U_2e_1,\theta_1\rangle & \cdots & \langle O_2^{\mathrm{T}}U_2e_K,\theta_1\rangle\\ \vdots & \ddots & \vdots\\ \langle O_2^{\mathrm{T}}U_2e_1,\theta_K\rangle & \cdots & \langle O_2^{\mathrm{T}}U_2e_K,\theta_K\rangle\end{bmatrix} = \Theta U_2^{\mathrm{T}}O_2.$$

■

It follows from Lemma 3.3 that for $\Theta \in \mathbb{R}^{K\times K}$ an invertible matrix, and $\theta_i^{\mathrm{T}} \in \mathbb{R}^K$ its i-th row, for any $i \in \{1,\dots,K\}$ the matrix $\mathscr{B}_{1,2,3}(U_2\theta_i)$ has eigenvectors $\left(U_1^{\mathrm{T}}O_1\right)^{-1}$. These eigenvectors simultaneously diagonalize the matrices $\mathscr{B}_{1,2,3}(U_2\theta_i)$ for all $i = 1,\dots,K$, and the corresponding eigenvalues form the rows of the matrix L in Lemma 3.4, which defines a linear system in terms of O_2.

3.3.4 The transition matrix and stationary distribution

The following lemma provides an expression for the stationary distribution π and the transition matrix Q in terms of the embedded observation densities O_2. The lemma provides a way to estimate the HMM parameters given an estimate of the embedded observation densities, and allows for estimation of all of the hidden Markov model's parameters (F,Q,π).

Theorem 3.1. *Let $U \in \mathscr{H}_{\mathscr{Y}}^K$ be such that $U^{\mathrm{T}}O_2 \in \mathbb{R}^{K\times K}$ is invertible. Then the stationary distribution π and transition matrix Q can be expressed as follows*

$$\pi = (U^{\mathrm{T}}O_2)^{-1}U^{\mathrm{T}}\mu_1 \qquad \textit{and} \qquad Q = (U^{\mathrm{T}}O_2\operatorname{diag}(\pi))^{-1}U^{\mathrm{T}}\mathscr{C}_{1,2}U(O_2^{\mathrm{T}}U)^{-1},$$

where μ_1 denotes the kernel mean embedding of the marginal distribution of Y_1 in $\mathscr{H}_{\mathscr{Y}}$.

Proof. It follows from the law of total expectation that the kernel mean embedding of the distribution of Y_1 in $\mathscr{H}_{\mathscr{Y}}$ can written as follows

$$\mu_1 := \mathbb{E}[\phi_Y(Y_1)] = \sum_{i=1}^{K}\mathbb{E}[\phi_Y(Y_1)\,|\,X_1 = i]\mathbb{P}(X_1 = i) = O_2\pi,$$

where we use the fact that $\mathbb{E}[\phi_Y(Y_1)\,|\,X_1] = \mathbb{E}[\phi_Y(Y_2)\,|\,X_2] = O_2 \in \mathscr{H}^K$. As $U^{\mathrm{T}}O_2$ is an invertible $K\times K$ matrix, we have $\pi = (U^{\mathrm{T}}O_2)^{-1}U^{\mathrm{T}}\mu_1$.

Analogously, we apply the law of total expectation and conditional independence to the cross-covariance operator of (Y_1, Y_2), and Lemma 1.3 to obtain

$$\begin{aligned}
\mathscr{C}_{1,2} &:= \mathbb{E}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)] \\
&= \mathbb{E}_{X_1 X_2}\left\{\mathbb{E}[\phi_Y(Y_1) \otimes \phi_Y(Y_2) \mid X_1, X_2]\right\} \\
&= \mathbb{E}_{X_1 X_2}\{\mathbb{E}[\phi_Y(Y_1) \mid X_1] \otimes \mathbb{E}[\phi_Y(Y_2) \mid X_2]\} \\
&= \sum_{i=1}^{K} \sum_{j=1}^{K} \{\mathbb{E}[\phi_Y(Y_1) \mid X_1 = i] \otimes \mathbb{E}[\phi_Y(Y_2) \mid X_2 = j]\} \mathbb{P}(X_2 = j, X_1 = i) \\
&= \sum_{i=1}^{K} \sum_{j=1}^{K} \{\mathbb{E}[\phi_Y(Y_1) \mid X_1 = i] \otimes \mathbb{E}[\phi_Y(Y_2) \mid X_2 = j]\} Q_{i,j} \pi_i \\
&= \sum_{i=1}^{K} \sum_{j=1}^{K} [O_2]_i Q_{i,j} \pi_i [O_2^{\mathrm{T}}]_j \\
&= O_2 \operatorname{diag}(\pi) Q O_2^{\mathrm{T}}.
\end{aligned}$$

Once again as $U^{\mathrm{T}} O_2$ is invertible, we may write $Q = (U^{\mathrm{T}} O_2 \operatorname{diag}(\pi))^{-1} U^{\mathrm{T}} \mathscr{C}_{1,2} U (O_2^{\mathrm{T}} U)^{-1}$. ■

3.3.5 The choice of U_1, U_2, and U_3

In this section we discuss how to choose U_1, U_2, and U_3, such that the assumptions required in the preceding lemmas are satisfied. The following lemma shows that we can define the row vector $U \in \mathscr{H}_{\mathscr{Y}}^K$ to be the left singular vectors of the operator $\mathscr{C}_{1,3}$ and upon setting $U_1 = U_2 = U_3 = U$, $U_v^{\mathrm{T}} O_v$ is invertible for $v = 1, 2, 3$.

Lemma 3.5. *Let $U \in \mathscr{H}_{\mathscr{Y}}^K$ be a row vector containing the K leading left singular vectors of the cross-covariance operator $\mathscr{C}_{1,3} \in \mathscr{H}_{\mathscr{Y}}^{\otimes 2}$ as ordered according to their corresponding singular values. If Assumptions 3.1 to 3.4 hold, then $U^{\mathrm{T}} O_v \in \mathbb{R}^{K \times K}$ is invertible for $v = 1, 2, 3$.*

Proof. We first note that under Assumptions 3.1 and 3.4, the elements of O_2 are linearly independent, and under Assumptions 3.2 and 3.3 the matrix Q is full rank, the elements of π are positive, and $\pi Q = \pi$. Lemma 3.1 shows that $O_1 = O_2 \operatorname{diag}(\pi) Q^{\mathrm{T}} \operatorname{diag}(Q\pi)^{-1}$ and $O_3 = O_2 Q$. It follows that the matrix $\operatorname{diag}(\pi) Q^{\mathrm{T}} \operatorname{diag}(Q\pi)^{-1}$ has full rank. As O_1 and O_3 are transformations of O_2 by full rank matrices, the elements of O_1 and O_3 are also linearly independent, and O_1, O_2, and O_3 span the same K-dimensional subspace of $\mathscr{H}_{\mathscr{Y}}$.

Recall from Lemma 3.2 that $\mathscr{C}_{1,3} = O_1 \operatorname{diag}(\pi) O_3^{\mathrm{T}}$, which implies that $\mathscr{C}_{1,3}$ has rank K as all elements of π are positive under Assumptions 3.2 and 3.3, and the elements of O_1 and O_3 are linearly independent under Assumptions 3.1 and 3.4. As $\mathscr{C}_{1,3}$ has rank K, it has the singular value decomposition [Mollenhauer et al., 2020]

$$\mathscr{C}_{1,3} = \sum_{i=1}^{K} \sigma_i (u_i \otimes v_i),$$

where $\{u_i\}_{i=1}^K$ and $\{v_i\}_{i=1}^K$ are orthonormal systems of left and right singular vectors, and $\{\sigma_i\}_{i=1}^K$ is the set of positive singular values. Let $U \in \mathscr{H}_{\mathscr{Y}}^K$ and $V \in \mathscr{H}_{\mathscr{Y}}^K$ be row-vectors formed by the left and right singular vectors of $C_{1,3}$, and let σ denote the vector of singular values. Then

$$\mathscr{C}_{1,3} = U \operatorname{diag}(\sigma) V^{\mathrm{T}} = O_1 \operatorname{diag}(\pi) O_3^{\mathrm{T}},$$

and hence U and O_1 span the same K-dimensional subspace of $\mathscr{H}_{\mathscr{Y}}$. The elements of U and O_1 are linearly independent, and thus $U^{\mathrm{T}} O_1 \in \mathbb{R}^{K\times K}$ is invertible.

As the elements of O_1, O_2, and O_3 are linearly independent and span the same K-dimensional subspace of $\mathscr{H}_{\mathscr{Y}}$, it follows that $U^{\mathrm{T}} O_v$ is invertible for $v = 1,2,3$. ■

Remark 3.1. *Lemma 3.5 also holds when $U \in \mathscr{H}_{\mathscr{Y}}^K$ is a row vector containing the K leading left or right singular vectors (sorted according to their corresponding singular values) of the embedding of (Y_i, Y_j) in $\mathscr{H}_{\mathscr{Y}}^{\otimes 2}$ for any $i, j \in 1,2,3$. The proof is identical to the above.*

3.3.6 Estimation of the HMM parameters

Lemma 3.4 and Theorem 3.1 describe how to estimate the HMM parameters $(\hat{O}_2, \hat{Q}, \hat{\pi})$ given a sequence of observations. Suppose we observe a sequence of realizations of the observable process, $(Y_t)_{t=1}^{n+2}$, for $n \geq 1$, then we define $Y^{(1)} = (Y_1, \ldots, Y_n)$, $Y^{(2)} = (Y_2, \ldots, Y_{n+1})$, and $Y^{(3)} = (Y_3, \ldots, Y_{n+2})$. We can compute quantities such as the empirical cross-covariance operator $\hat{\mathscr{C}}_{1,3}$ using $(Y_i^{(1)}, Y_i^{(3)})_{i=1}^n$ as a sample from the joint distribution of (Y_1, Y_3). We note that the pairs and triples of samples used throughout the following are not independent, and hence the accuracy of estimated quantities such as $\hat{\mathscr{C}}_{1,3}$ is influenced by mixing properties of the Markov chain. This is accounted for in our concentration inequalities detailed in Lemma 3.11.

To estimate the HMM parameters, we first develop empirical versions of the quantities used in the preceding lemmas. The reproducing kernel Hilbert space on $\mathscr{Y}$ may be infinite dimensional, and in the following we repeatedly use the reproducing property to replace inner products between feature mappings with kernel evaluations. We illustrate with a simple example how the empirical cross-covariance operator acting on a function can be computed with finite cost.

Example 3.1. *Suppose that we observe a sequence of random variables $(Y_1, \ldots, Y_{n+2})$. We form the feature vectors $\Phi_1 = [\phi_Y(Y_1), \ldots, \phi_Y(Y_n)]$ and $\Phi_3 = [\phi_Y(Y_3), \ldots, \phi_Y(Y_{n+2})]$ which are n-dimensional row vectors in $\mathscr{H}_{\mathscr{Y}}$. The cross-covariance operator $\mathscr{C}_{1,3}$ is estimated via the sample mean, and hence the empirical cross-covariance operator $\hat{\mathscr{C}}_{1,3} : \mathscr{H}_{\mathscr{Y}} \to \mathscr{H}_{\mathscr{Y}}$ is a sum of rank-one linear operators $\hat{\mathscr{C}}_{1,3} = \sum_{i=1}^n \frac{1}{n}(\phi_Y(Y_i) \otimes \phi_Y(Y_{i+2}))$. For any $f \in \mathscr{H}_{\mathscr{Y}}$ we have*

$$\begin{aligned}\hat{\mathscr{C}}_{1,3}(f) &= \sum_{i=1}^n \frac{1}{n} \phi_Y(Y_i) \langle \phi_Y(Y_{i+2}), f \rangle_{\mathscr{H}_{\mathscr{Y}}} \\ &= \sum_{i=1}^n \alpha_i \phi_Y(Y_i) \in \mathscr{H}_{\mathscr{Y}},\end{aligned}$$

where we have used the reproducing property in the second line so $\langle \phi_Y(Y_{i+2}), f \rangle_{\mathcal{H}_\mathcal{Y}} = f(Y_{i+2})$*, and defined* $\alpha_i = \frac{1}{n} f(Y_{i+2})$*. Hence, we see that whilst the RKHS may be infinite dimensional, operations on these spaces reduce to a function evaluation. In the case that* $f = \phi_Y(y)$ *for* $y \in \mathcal{Y}$*, we have* $\alpha_i = n^{-1} k(y, Y_{i+2})$*. Using the feature vector notation specified above, this can be rewritten as* $\hat{\mathscr{C}}_{1,3}(f) = \Phi_1 \Phi_3^{\mathrm{T}} f$*, and* $\Phi_3^{\mathrm{T}} f = \alpha \in \mathbb{R}^n$*.*

When working with empirical RKHS operators as in Example 3.1, we can simplify expressions using the following rules which follow from the reproducing property. For $A \in \mathcal{H}_\mathcal{Y}^n$ and $B \in \mathcal{H}_\mathcal{Y}^m$ row vectors in $\mathcal{H}_\mathcal{Y}$, $A^{\mathrm{T}} B \in \mathbb{R}^{n \times m}$. If $C \in \mathbb{R}^{n \times m}$ then $ACB^{\mathrm{T}} \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Y}$. For example, consider the feature vectors $\Phi_1 = [\phi_Y(Y_1), \ldots, \phi_Y(Y_n)] \in \mathcal{H}_\mathcal{Y}^n$ and $\Phi_3 = [\phi_Y(Y_3), \ldots, \phi_Y(Y_{n+2})] \in \mathcal{H}_\mathcal{Y}^n$, then $n^{-1} \Phi_1 \Phi_3^{\mathrm{T}} = \hat{\mathscr{C}}_{1,3} \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Y}$, whereas $\Phi_1^{\mathrm{T}} \Phi_3 = K_{1,3} \in \mathbb{R}^{n \times n}$ for $K_{1,3}$ the $n \times n$ real-valued matrix with (i,j)-th element $k(y_i, y_{j+2})$. We can think of $\Phi_1 \Phi_3^{\mathrm{T}}$ as the outer product between elements, and $\Phi_1^{\mathrm{T}} \Phi_3$ as the inner product between elements.

Empirical quantities. We now derive the empirical quantities required to estimate the HMM parameters. To simplify our notation we denote by $K_{v,w}$, for $v, w \in \{1,2,3\}$, the matrix of kernel evaluations over $Y^{(v)}$ and $Y^{(w)}$. When $v = m$ we simplify the subscript such that $K_{v,v} = K_v$. For example, the (i,j)-th element of $K_{1,3}$ is $[K_{1,3}]_{i,j} = k(Y_i, Y_{j+2})$, and the (i,j)-th element of K_1 is $[K_1]_{i,j} = k(Y_i, Y_j)$, where $i, j \in \{1, \ldots, n\}$.

We define U to be the row vector containing the top K right singular vectors of the cross-covariance operator $\mathscr{C}_{1,3}$, and we derive its empirical estimate computing the singular value decomposition of $\hat{\mathscr{C}}_{1,3}$. To do this, we solve an auxiliary problem as follows.

Proposition 3.1 (Mollenhauer et al. [2020])**.** *Let* $\hat{\mathscr{C}}_{1,3} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ *denote the empirical RKHS operator* $\hat{\mathscr{C}}_{1,3} = \Phi_1 B \Phi_3^{\mathrm{T}}$ *with* $\operatorname{rank}(\hat{\mathscr{C}}_{1,3}) = r := \min(K, n)$*, where* $\Phi_1 = [\phi_Y(Y_1), \ldots, \phi_Y(Y_n)]$ *and* $\Phi_3 = [\phi_Y(Y_3), \ldots, \phi_Y(Y_{n+2})]$*, and* $B \in \mathbb{R}^{n \times n}$ *is the diagonal matrix with entries* n^{-1}*. Assume that the multiplicity of each singular value of* $\hat{\mathscr{C}}_{1,3}$ *is 1. Then the SVD of* $\hat{\mathscr{C}}_{1,3}$ *is given by*

$$\hat{\mathscr{C}}_{1,3} = \sum_{i=1}^{r} \lambda_i^{1/2} (u_i \otimes v_i),$$

where

$$\begin{aligned} v_i &:= (w_i^{\mathrm{T}} K_{\Phi_3} w_i)^{-1/2} \Phi_3 w_i, \\ u_i &:= \lambda_i^{-1/2} \hat{\mathscr{C}}_{1,3} v_i, \end{aligned}$$

with the non-zero eigenvalues $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$ *of the matrix* $N^{-2} K_{\Phi_1} K_{\Phi_3} \in \mathbb{R}^{N \times N}$ *counted with their multiplicities and corresponding eigenvectors* $w_1, \ldots, w_r \in \mathbb{R}^N$*. In the above,* K_{Φ_1} *is the* $N \times N$ *Gram matrix with* (i,j)*-th element* $k(y_i, y_j)$*, and similarly* $[K_{\Phi_1}]_{i,j} = k(y_{i+2}, y_{j+2})$*, for* $i, j \in \{1, \ldots, n\}$*.*

Hence, let $\hat{W}$ denote the matrix which has column vectors $\hat{w}_1, \ldots \hat{w}_K$, which are the leading K right singular vectors of $n^{-2} K_1 K_3$ ordered according to their singular values. Define the diagonal matrix $\hat{D} = \operatorname{diag}\big((\hat{w}_1^{\mathrm{T}} K_3 \hat{w}_1)^{-1/2}, \ldots, (\hat{w}_K^{\mathrm{T}} K_3 \hat{w}_K)^{-1/2}\big)$, then the empirical estimator of U is $\hat{U} = \Phi_3 \hat{W} \hat{D}$.

The empirical observable operator acting upon $\hat{U}\theta_i$ is defined as

$$\hat{\mathscr{B}} : \hat{U}\theta_i \mapsto \hat{U}^{\mathrm{T}}(\mathscr{C}_{1,2,3} \times_2 \hat{U}\theta_i)\hat{U}(\hat{U}^{\mathrm{T}}\mathscr{C}_{1,3}\hat{U})^{-1}, \quad \hat{\mathscr{B}} : \mathscr{H}_{\mathscr{Y}} \to \mathbb{R}^{K \times K}.$$

We can rewrite the empirical observable operator $\hat{\mathscr{B}}$ in terms of kernel matrices as follows

$$\begin{aligned}
\hat{\mathscr{B}}(\hat{U}\theta_i) :&= \hat{U}^{\mathrm{T}}(\hat{\mathscr{C}}_{1,2,3} \times_2 \hat{U}\theta_i)\hat{U}(\hat{U}^{\mathrm{T}}\hat{\mathscr{C}}_{1,3}\hat{U})^{-1}, \\
&= \left(\Phi_3 \hat{W}\hat{D}\right)^{\mathrm{T}} (\hat{\mathscr{C}}_{1,2,3} \times_2 \hat{U}\theta_i)\left(\Phi_3 \hat{W}\hat{D}\right) \left[\left(\Phi_3 \hat{W}\hat{D}\right)^{\mathrm{T}} N^{-1} \Phi_1 \Phi_3^{\mathrm{T}} \left(\Phi_3 \hat{W}\hat{D}\right)\right]^{-1}, \\
&= \hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} (\hat{\mathscr{C}}_{1,2,3} \times_2 \hat{U}\theta_i)\Phi_3 \hat{W}\hat{D} \left[n^{-1} \hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} \Phi_1 \Phi_3^{\mathrm{T}} \Phi_3 \hat{W}\hat{D}\right]^{-1}, \\
&= \hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} \Phi_1 \operatorname{diag}\left(\Phi_2^{\mathrm{T}} \Phi_3 \hat{W}\hat{D}\theta_i\right) \Phi_3^{\mathrm{T}} \Phi_3 \hat{W}\hat{D} \left[\hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} \Phi_1 \Phi_3^{\mathrm{T}} \Phi_3 \hat{W}\hat{D}\right]^{-1}, \\
&= \hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_{3,1} \operatorname{diag}\left(K_{2,3} \hat{W}\hat{D}\theta_i\right) K_3 \hat{W}\hat{D} \left[\hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_{3,1} K_3 \hat{W}\hat{D}\right]^{-1}.
\end{aligned}$$

We can now form estimates of the HMM parameters $\hat{O}_2$, $\hat{Q}$, and $\hat{\pi}$. Suppose we have an empirical estimate of the matrix of eigenvalues $L \in \mathbb{R}^{K \times K}$ defined in Lemma 3.4, then the estimator of O_2 is $\hat{O}_2 = \Phi_3 \hat{W}\hat{D}\Theta\hat{L} \in \mathscr{H}_{\mathscr{Y}}^K$. We define intermediary estimators of the stationary distribution and transition matrix as follows

$$\begin{aligned}
\tilde{\pi} &= (\hat{U}^{\mathrm{T}} \hat{O}_2)^{-1} \hat{U}^{\mathrm{T}} \hat{\mu}_1 \\
&= (D^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} \Phi_3 \hat{W}\hat{D}\Theta\hat{L})^{-1} D^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} n^{-1} \Phi_1 1_n \\
&= n^{-1} (D^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_3 \hat{W}\hat{D}\Theta\hat{L})^{-1} D^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_{3,1} 1_n \\
\tilde{Q} &= (\hat{U}^{\mathrm{T}} \hat{O}_2 \operatorname{diag}(\hat{\pi}))^{-1} \hat{U}^{\mathrm{T}} \hat{\mathscr{C}}_{1,2} \hat{U} (\hat{O}_2^{\mathrm{T}} \hat{U})^{-1} \\
&= (D^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} \Phi_3 \hat{W}\hat{D}\Theta\hat{L} \operatorname{diag}(\hat{\pi}))^{-1} D^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} n^{-1} \Phi_1 \Phi_2^{\mathrm{T}} \Phi_3 \hat{W}\hat{D} (\hat{L}^{\mathrm{T}} \Theta^{\mathrm{T}} \hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} \Phi_3^{\mathrm{T}} \Phi_3 \hat{W}\hat{D})^{-1} \\
&= n^{-1} (D^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_3 \hat{W}\hat{D}\Theta\hat{L} \operatorname{diag}(\hat{\pi}))^{-1} D^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_{3,1} K_{2,3} \hat{W}\hat{D} (\hat{L}^{\mathrm{T}} \Theta^{\mathrm{T}} \hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_3 \hat{W}\hat{D})^{-1}.
\end{aligned}$$

The intermediary estimators are motivated by Theorem 3.1, however when population quantities are replaced by estimates the intermediary transition matrix $\tilde{Q}$ may not be well defined as a probability matrix. Thus, to estimate the transition matrix we project $\tilde{Q}$ onto the convex set of transition matrices. We define $\hat{Q} = \Pi_{TM}(\tilde{Q})$, where Π_{TM} denotes the projection onto the convex set of transition matrices. We then define the estimated stationary distribution, $\hat{\pi}$, to be the stationary distribution corresponding to the estimated transition matrix $\hat{Q}$.

Choosing Θ. The embedded observation densities are obtained as the solution to the linear system specified in Lemma 3.4. Thus, it is important to choose Θ that ensures that the system is well-conditioned. Anandkumar et al. [2012] suggest that in the absence of prior knowledge, one can sample Θ from the $\Delta^{(K-1)\times(K-1)}$ simplex. Lehéricy [2018] note that improved performance can be obtained by sampling a set of Θ's and keeping that which maximizes

$$\min_i \min_{j_1 \neq j_2} |L_{i,j_1} - L_{i,j_2}|, \quad i, j_1, j_2 = 1, \ldots, K \tag{3.1}$$

This works well when our assumptions are satisfied and the problem is well-specified. However, the separation of the entries of L is directly related to the linear independence of the observation densities (see Lemmas 10, 16, and 34 of De Castro et al. [2017]). Therefore when the observation densities are close to not being linearly independent, finding a Θ which maximizes the above objective becomes increasingly difficult. To avoid this problem, we propose to choose Θ by maximizing the objective function Equation (3.1) via an optimization procedure. We use the Cayley transform to parametrize the set of $K \times K$ orthonormal matrices by a real-valued vector $\beta \in \mathbb{R}^{K(K-1)}$, and maximize the objective via an unconstrained optimization method such as BFGS [Nocedal and Wright, 1999]. There is no guarantee that this will attain a global maxima, so we propose several initializations of β, optimize, and keep that which maximizes the objective.

Computational cost and implementation. The method requires the truncated SVD of an $n \times n$ matrix, several inversions of a $K \times K$ matrix, and matrix multiplication. The truncated SVD can be computed with cost $O(Kn^2)$ using Krylov methods such as the Lanczos algorithm. A version for implementation is given in Algorithm 1.

Algorithm 1 Kernel spectral method for HMMs

```
Input: An observed sequence $(Y_1, \dots, Y_{n+2})$, number of hidden states $K$, kernel function $k : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$.
Output: Estimated HMM parameters $\hat{O}_2, \hat{\pi}, \hat{Q}$.
Compute kernel matrices $K_1, K_3, K_{2,3}, K_{3,1}$.
Compute the $K$ leading eigenvectors $\hat{w}_1, \dots \hat{w}_K$ of $\frac{1}{n^2} K_1 K_3$, and compute $\hat{W} = [\hat{w}_1, \dots \hat{w}_K]$ and set $\hat{D} = \mathrm{diag}\left((\hat{w}_1^{\mathrm{T}} K_3 \hat{w}_1)^{-1/2}, \dots, (\hat{w}_K K_3 \hat{w}_K)^{-1/2}\right)$. Then compute $\hat{U} = \Phi_3 \hat{W} \hat{D}$.
Choose $\Theta$. Compute $\hat{\mathscr{B}}(\hat{U}\theta_i) = \hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_{3,1} \,\mathrm{diag}\left(K_{2,3} \hat{W} \hat{D} \theta_i\right) K_3 \hat{W} \hat{D} \left[\hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_{3,1} K_3 \hat{W} \hat{D}\right]^{-1}$, for $i = 1, \dots, K$, where $\hat{\mathscr{B}}(\hat{U}\theta_i)$ are $K \times K$ matrices.
Compute the matrix $\hat{R}$ that diagonalizes $\hat{\mathscr{B}}(U\theta_1)$. For $i = 1, \dots, K$ compute the diagonal matrix $\hat{R}^{-1} \mathscr{B}(\hat{U}^{\mathrm{T}} \theta_i) \hat{R} = \mathrm{diag}\left(\hat{\lambda}_{i,1}, \dots, \hat{\lambda}_{i,K}\right)$ and set $[\hat{L}]_{i,\cdot} = [\hat{\lambda}_{i,1}, \dots, \hat{\lambda}_{i,K}]$, where $[\hat{L}]_{i,\cdot}$ denotes the $i$-th row of $\hat{L}$.
Compute the estimated embedded observation densities $\hat{O}_2 = \Phi_3 \hat{W} \hat{D} \Theta \hat{L}$.
Compute the estimate of the stationary distribution and transition matrix

    $\tilde{\pi} = n^{-1} (D^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_3 \hat{W} \hat{D} \Theta \hat{L})^{-1} D^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_{3,1} 1_n$
    $\hat{Q} = \Pi_{TM} \left( n^{-1} (D^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_3 \hat{W} \hat{D} \Theta \hat{L} \,\mathrm{diag}(\tilde{\pi}))^{-1} D^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_{3,1} K_{2,3} \hat{W} \hat{D} (\hat{L}^{\mathrm{T}} \Theta^{\mathrm{T}} \hat{D}^{\mathrm{T}} \hat{W}^{\mathrm{T}} K_3 \hat{W} \hat{D})^{-1} \right),$

where $\Pi_{TM}$ denotes the projection onto the convex set of transition matrices. Then set $\hat{\pi}$ to be the stationary distribution associated with $\hat{Q}$.
```

3.4 Inference: the filtering problem

In this section we focus on the filtering problem, and discuss how to approach the problem given the output of the kernel spectral algorithm described in Section 3.3.

Algorithm 1 outputs estimates of the transition matrix and stationary distribution, and estimated embeddings of the observation densities in an RKHS. We therefore propose two methods: one in which we estimate the observation densities from their estimated embeddings, and another in which we work directly with the embeddings. We compare the two approaches when applied to simulated data in Section 3.6. Both methods induce errors in different ways: the first requires an additional density estimation procedure while the second requires that we approximate the conditional mean embedding as it may not be well defined, by introducing a regularized matrix inversion.

The filtering problem. Throughout the following we assume that we observe a sequence of p realizations of the observable process, denoted $y_{1:p}$ for $p \geq 1$ an integer. We aim to compute the posterior distributions of $X_t | y_{1:t}$ for $1 \leq t \leq p$.

3.4.1 Density estimation and the forward algorithm

In this approach to the filtering problem, we estimate the observation densities from their RKHS embeddings and use the forward algorithm to obtain the filtering distributions. De Castro et al. [2017] provide uniform consistency bounds on the filtering distributions in terms of the risk of the estimators $(\hat{F}, \hat{Q}, \hat{\pi})$, and thus whilst estimating the observation densities induces additional error, the errors do not propagate and accumulate in the filtering problem due to the forgetting properties of the model.

Density estimation. Given the estimated embeddings of the observation densities, $\hat{O}_2$, we recover the underlying densities using the estimator defined in Chapter 2. Let $A := \hat{W}\hat{D}\Theta\hat{L} \in \mathbb{R}^{n\times K}$, and let $a_{i,j}$ denote the (i,j)-th element of A for $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,K\}$. Then we have $\hat{\mu}_{Y_2|X_2=v} = \sum_{i=1}^{n} a_{i,v} k(Y_{i+2},\cdot)$, which is of the form Equation (2.1), and we therefore estimate the density as

$$\hat{f}(y|x=v) = \sum_{i=1}^{n} a_{i,v} \bar{k}^{\tilde{\gamma}_n}(Y_{i+1},\cdot), \quad v = 1,\dots,K.$$

The forward algorithm. As the set of values that the hidden state can take is finite, the filtering distributions can be computed analytically using the forward procedure of the forward-backward algorithm [Baum et al., 1970]. The forward algorithm provides a simple method for performing Bayesian updates whilst taking advantage of the conditional independence specified by the HMM. Let $x \in \mathscr{X}$, and (F, Q, π) denote the HMM parameters, then the filtering distributions are updated recursively using the equations

$$p(X_1 = x|y_1) = \frac{f_x(y_1)\pi(x)}{\sum_{i=1}^{K} f_i(y_1)\pi_i},$$

$$p(X_t = x|y_{1:t}) = \frac{\sum_{i=1}^{K} f_x(y_t) Q_{i,x} p(X_{t-1} = i|y_{1:(t-1)})}{\sum_{j,k=1}^{K} f_j(y_t) Q_{k,j} p(X_{t-1} = k|y_{1:(t-1)})},$$

for $1 \le t \le p$. We estimate the filtering distributions using the forward algorithm wherein the population HMM parameters (F, Q, π) are replaced by their empirical counterparts $(\hat{F}, \hat{Q}, \hat{\pi})$. We predict the hidden state at time t via $\hat{X}_t = \operatorname{argmax}_{x \in \mathscr{X}} \hat{p}(X_t = x | y_{1:t})$.

3.4.2 An alternative kernel Bayes' rule

In this section we approach the filtering problem by working directly with the estimated embedded observation densities. We derive a set of recursive equations similar to the forward algorithm, using the estimated HMM $(\hat{O}_2, \hat{Q}, \hat{\pi})$ to estimate the posterior $X_t | y_{1:t}$.

Filtering in hidden Markov models using kernel mean embeddings has been studied in several papers. Song et al. [2009] implements a version of Bayes' rule for the filtering problem by a heuristic approximation, and Fukumizu et al. [2013] generalize this by providing an estimator for the posterior obtained via kernel Bayes' rule and prove its consistency. Both of these methods infer the filtering distributions using a set of paired observations $(X_t, Y_t)_{t \ge 1}$ to model the transition and observation distributions for state-space models. Song et al. [2010] generalize a spectral algorithm for hidden Markov models using an observable operator representation [Hsu et al., 2012], however they show that their error in estimating the embedding of the predictive distribution of $Y_{t+1} | y_{1:t}$ grows linearly with t, for $t \ge 1$. Nishiyama et al. [2020] introduce a model-based kernel sum rule which requires a model rather than a set of samples, and they show that the resulting estimated embedding is consistent. The model-based kernel sum rule is used in addition to the sample-based kernel Bayes' rule of Fukumizu et al. [2013] to infer the filtering distributions. Access to a set of paired samples of the observable and *unobservable* process is an unrealistic assumption, and in the following we derive an alternative kernel Bayes' rule which does not require access to paired samples.

In a probabilistic framework, the filtering task is divided into two steps: the prediction and the update. Suppose at time $t \ge 1$ we have an estimate of the posterior distribution $X_t | y_{1:t}$. The prediction step estimates the next hidden state by marginalizing over the current hidden state. That is,

$$p(X_{t+1} \mid y_{1:t}) = \int p(X_{t+1} \mid X_t) p(X_t \mid y_{1:t}) dX_t.$$

In the update step we update our beliefs given a new observation y_{t+1} via Bayes' rule:

$$p(X_{t+1} \mid y_{1:(t+1)}) = \frac{p(y_{t+1} | X_{t+1}) p(X_{t+1} | y_{1:t})}{\int p(y_{t+1} | X_{t+1}) p(X_{t+1} | y_{1:t}) dX_{t+1}}.$$

We develop an analogue to this procedure using kernel mean embeddings, recursively updating a belief state $\mu_{X_t | y_{1:t}}$, which is the embedding of the distribution $X_t | y_{1:t}$ in an RKHS $\mathscr{H}_{\mathscr{X}}$ on $\mathscr{X} = \{1, \dots, K\}$ with kernel function $l : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ and canonical feature map $\varphi_X(x) = l(x, \cdot) \in \mathscr{H}_{\mathscr{X}}$ for $x \in \mathscr{X}$.

The prediction step can be accomplished using the kernel sum rule [Song et al., 2009], which gives

$$\mu_{X_{t+1} | y_{1:t}} = \mathscr{U}_{X_{t+1} | X_t} \mu_{X_t | y_{1:t}}. \tag{3.2}$$

To update our beliefs we then use kernel Bayes' rule Fukumizu et al. [2013] to compute $\mu_{X_{t+1}|y_{1:(t+1)}}$, where the prior $p(X_{t+1}|y_{1:t})$ has the embedding $\mu_{X_{t+1}|y_{1:t}}$ given by the prediction step. We compute the embedding of the posterior as

$$\mu_{X_{t+1}|y_{1:(t+1)}} = \mathscr{U}^{\pi}_{X_{t+1}|Y_{t+1}} \phi_Y(y_{t+1}) = ((\mathscr{C}^{\pi}_{Y_{t+1}Y_{t+1}})^{\dagger}(\mathscr{C}^{\pi}_{X_{t+1}Y_{t+1}})^{*})^{*} \phi_Y(y_{t+1}), \tag{3.3}$$

where a superscript π is used to emphasize dependence on the prior. The cross-covariance operators can be computed in terms of the prior as follows

$$\begin{aligned}
\mathscr{C}^{\pi}_{X_{t+1}Y_{t+1}} &= \mu^{\pi}_{X_{t+1}Y_{t+1}} = \mathscr{U}_{X_{t+1}Y_{t+1}|X_{t+1}} \mu_{X_{t+1}|y_{1:t}} = ((\mathscr{C}_{X_{t+1}X_{t+1}})^{\dagger} \mathscr{C}^{*}_{X_{t+1}Y_{t+1}X_{t+1}})^{*} \mu_{X_{t+1}|y_{1:t}} \\
\mathscr{C}^{\pi}_{Y_{t+1}Y_{t+1}} &= \mu^{\pi}_{Y_{t+1}Y_{t+1}} = \mathscr{U}_{Y_{t+1}Y_{t+1}|X_{t+1}} \mu_{X_{t+1}|y_{1:t}} = ((\mathscr{C}_{X_{t+1}X_{t+1}})^{\dagger} \mathscr{C}^{*}_{Y_{t+1}Y_{t+1}X_{t+1}})^{*} \mu_{X_{t+1}|y_{1:t}},
\end{aligned}$$

which follows from the kernel sum rule, and $A^{\dagger}$ denotes the Moore-Penrose inverse and A^{*} the adjoint of the operator A.

3.4.2.1 HMM cross-covariance operators

To implement the filtering procedure described by Equations (3.2) and (3.3) we must estimate several cross-covariance operators. Cross-covariance operators are often estimated using samples, however as the hidden process is unobservable, we cannot use the standard empirical estimator. The following lemma shows that cross-covariance operators of the hidden and observable process can be decomposed in terms of the HMM parameters.

In the following let $\Psi \in \mathscr{H}^{K}_{\mathscr{X}}$ denote a row vector with i-th element $\varphi_X(i)$ for $i = 1, \dots, K$. We denote by K_X the $K \times K$ kernel matrix with (i,j)-th element $[K_X]_{i,j} = l(i,j)$.

Lemma 3.6. *The cross-covariance operators of the joint distributions* (X_t, Y_t), (X_{t+1}, X_t), *and* (X_t, X_t) *can be decomposed in terms of the HMM parameters* (O_2, Q, π) *as follows*

$$\mathscr{C}_{X_tY_t} = \Psi \operatorname{diag}\left(\pi Q^{t-1}\right) O_2^{\mathrm{T}}, \quad \mathscr{C}_{X_{t+1}X_t} = \Psi Q^{\mathrm{T}} \operatorname{diag}\left(\pi Q^{t-1}\right) \Psi^{\mathrm{T}}, \quad \mathscr{C}_{X_tX_t} = \Psi \operatorname{diag}\left(\pi Q^{t-1}\right) \Psi^{\mathrm{T}}.$$

Proof. We first decompose $\mathscr{C}_{X_tY_t}$. It follows from the law of total expectation that

$$\mathscr{C}_{X_tY_t} = \mathbb{E}_{X_tY_t}[\varphi_X(X_t) \otimes \phi_Y(Y_t)] = \mathbb{E}_{X_t}[\varphi_X(X_t) \otimes \mathbb{E}_{Y_t|X_t}[\phi_Y(Y_t)|X_t]],$$

and noting that X_t takes values in $\mathscr{X} = \{1, \dots, K\}$, we expand the expectation over X_t as follows

$$\begin{aligned}
\mathscr{C}_{X_tY_t} &= \sum_{i=1}^{K} [\varphi_X(i) \otimes \mathbb{E}_{Y_t|X_t}[\phi_Y(Y_t)|X_t = i]] \mathbb{P}(X_t = i) \\
&= \Psi \operatorname{diag}([\mathbb{P}(X_t = 1), \dots, \mathbb{P}(X_t = K)]) O_2^{\mathrm{T}} \\
&= \Psi \operatorname{diag}\left(\pi Q^{t-1}\right) O_2^{\mathrm{T}},
\end{aligned}$$

where $\Psi = [\varphi_X(1), \dots, \varphi_X(K)]$ is a row vector in $\mathcal{H}_{\mathcal{X}}^K$. The second equation follows similarly,

$$\begin{aligned}
\mathscr{C}_{X_{t+1}X_t} &= \mathbb{E}_{X_{t+1}X_t}[\varphi_X(X_{t+1}) \otimes \varphi_X(X_t)] \\
&= \sum_{i=1}^{K} \sum_{j=1}^{K} \left[\varphi_X(i) \otimes \varphi_X(j)\right] \mathbb{P}(X_{t+1} = i | X_t = j) \mathbb{P}(X_t = j) \\
&= \sum_{i=1}^{K} \sum_{j=1}^{K} \left[\varphi_X(i) \otimes \varphi_X(j)\right] Q_{j,i} [\pi Q^{t-1}]_j \\
&= \Psi Q^{\mathrm{T}} \operatorname{diag}\left(\pi Q^{t-1}\right) \Psi^{\mathrm{T}}.
\end{aligned}$$

The final equation can be derived by expanding the expectation over X_t as follows

$$\begin{aligned}
\mathscr{C}_{X_t X_t} &= \mathbb{E}_{X_t}[\varphi_X(X_t) \otimes \varphi_X(X_t)] \\
&= \sum_{i=1}^{K} \left[\varphi_X(i) \otimes \varphi_X(i)\right] \mathbb{P}(X_t = i) \\
&= \sum_{i=1}^{K} \left[\varphi_X(i) \otimes \varphi_X(i)\right] [\pi Q^{t-1}]_i \\
&= \Psi \operatorname{diag}\left(\pi Q^{t-1}\right) \Psi^{\mathrm{T}}.
\end{aligned}$$

■

3.4.2.2 Filtering

The representations of the cross-covariance operators given in Lemma 3.6 allow us to estimate the embeddings of distributions involving the hidden process, without directly observing the hidden states. We use this to develop a filtering procedure.

Several assumptions are required for the conditional mean embedding to be well defined, and in some settings these assumptions are rather strong. For this reason, when considering the conditional mean embedding of the hidden process given a sequence of observations, $X_t | y_{1:t}$, we replace the pseudo-inverse with a regularized inverse. That is, rather than consider $\mu_{X_{t+1}|y_{1:(t+1)}}$ specified in Equation (3.3), we provide an update rule for

$$\begin{aligned}
\mu^{\mathrm{reg}}_{X_{t+1}|y_{1:(t+1)}} &:= ((\mathscr{C}^{\pi,\mathrm{reg}}_{Y_{t+1}Y_{t+1}} + \lambda \mathscr{I}_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}})^{\dagger} (\mathscr{C}^{\pi,\mathrm{reg}}_{X_{t+1}Y_{t+1}})^{*})^{*} \phi_Y(y_{t+1}) \\
&= \mathscr{C}^{\pi,\mathrm{reg}}_{X_{t+1}Y_{t+1}} (\mathscr{C}^{\pi,\mathrm{reg}}_{Y_{t+1}Y_{t+1}} + \lambda \mathscr{I}_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}})^{-1} \phi_Y(y_{t+1}),
\end{aligned}$$

where $\lambda > 0$ is a regularization parameter, $\mathscr{I}_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}}$ denotes the identity operator on the tensor product RKHS $\mathcal{H}_{\mathcal{Y}}^{\otimes 2}$, and a superscript reg is used to emphasize that the cross-covariance operators are computed using the regularized prior. The regularized embedding $\mu^{\mathrm{reg}}_{X_{t+1}|y_{1:(t+1)}}$ is an approximation of the embedding $\mu_{X_{t+1}|y_{1:(t+1)}}$ and is well-studied and known to be consistent under appropriate assumptions [Grünewälder et al., 2012, Fukumizu, 2015, Li et al., 2022].

However, the setting we consider is sufficiently nice that several strong properties hold, allowing for direct computation of the conditional mean embeddings conditioned on the hidden process, such as $X_{t+1}|X_t$ and $Y_t|X_t$. These properties stem from the fact that the latent space is finite dimensional, and thus only a finite RKHS on $\mathscr{X}$ is required; these properties are described in the following lemma.

Lemma 3.7. *For $\mathscr{X} = \{1,\dots,K\}$, let $l : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be the Kronecker-delta kernel function defined such that for $x,x' \in \mathscr{X}$, $l(x,x') = 1$ when $x = x'$ and $l(x,x') = 0$ otherwise. Let $\mathscr{H}_{\mathscr{X}}$ be the reproducing kernel Hilbert space associated with l, and let $\mathscr{Z}$ be a topological space such that $\mathscr{X} \subseteq \mathscr{Z}$. Then under Assumptions 3.2 and 3.3*

- *For any bounded function $g : \mathscr{Z} \to \mathbb{R}$ we have $\mathbb{E}[g(X_{t+1})|X_t = \cdot] \in \mathscr{H}_{\mathscr{X}}$,*
- *$\varphi_X(x)$ belongs to the range of $\mathscr{C}_{X_tX_t}$ for all $x \in \mathscr{X}$,*
- *$\mathscr{C}_{X_tX_t}$ is injective.*

Proof. We first show that for any function $g : \mathscr{Z} \to \mathbb{R}$ we have $\mathbb{E}[g(X_{t+1})|X_t = \cdot] \in \mathscr{H}_{\mathscr{X}}$ where $\mathbb{E}[g(X_{t+1})|X_t = \cdot]$ is a function from $\mathscr{X}$ to $\mathbb{R}$. Let $f : \mathscr{X} \to \mathbb{R}$ be defined by $f = \mathbb{E}[g(X_{t+1})|X_t = \cdot]$, and let $f_i := f(i)$. We construct a function $h \in \mathscr{H}_{\mathscr{X}}$ such that $h \equiv f$: let $h(\cdot) := \sum_{i=1}^K f_i l(i,\cdot)$, then $h(x) = f(x)$ for all $x \in \mathscr{X}$ and $h \in \mathscr{H}_{\mathscr{X}}$ by construction. Hence $\mathbb{E}[g(X_{t+1})|X_t = \cdot] \in \mathscr{H}_{\mathscr{X}}$.

We now show that $\varphi_X(x)$ belongs to the range of $\mathscr{C}_{X_tX_t}$ for all $x \in \mathscr{X}$. Recall from Lemma 3.6 that $\mathscr{C}_{X_tX_t} = \Psi \operatorname{diag}\left(\pi \boldsymbol{Q}^{t-1}\right) \Psi^{\mathrm{T}}$ where $\Psi = [\varphi_X(1),\dots,\varphi_X(K)]$. Under Assumptions 3.2 and 3.3, $\pi \boldsymbol{Q}^{t-1} = \pi$ and $\pi_i > 0$ for all $i \in \mathscr{X}$. For any $x \in \mathscr{X}$ we can define the function $f_x = \Psi \alpha_x$ for $\alpha_x \in \mathbb{R}^K$ defined such that $\alpha_{x,i} = \pi_i^{-1}$, if $i = x$ and $\alpha_{x,i} = 0$ otherwise. Clearly, $f_x \in \mathscr{H}_{\mathscr{X}}$, and we have

$$\begin{aligned}\mathscr{C}_{X_tX_t} f_x &= \Psi \operatorname{diag}(\pi) \Psi^{\mathrm{T}} f_x \\ &= \Psi \operatorname{diag}(\pi) K_X \alpha_x \\ &= \sum_{i=1}^K \varphi_X(i) \pi_i \alpha_{x,i} \\ &= \varphi_X(x),\end{aligned}$$

where we have used the fact that $\Psi^{\mathrm{T}}\Psi = K_X = I_K$ for the Kronecker-delta kernel function. Hence for any $x \in \mathscr{X}$ there exists a function $f_x \in \mathscr{H}_{\mathscr{X}}$ such that $\mathscr{C}_{X_tX_t} f_x = \varphi_X(x)$ and thus $\varphi_X(x)$ belongs to the range of $\mathscr{C}_{X_tX_t}$ for all $x \in \mathscr{X}$.

Finally, we show that $\mathscr{C}_{X_tX_t} : \mathscr{H}_{\mathscr{X}} \to \mathscr{H}_{\mathscr{X}}$ is injective. Note that $\langle \mathscr{C}_{X_tX_t} f, f \rangle_{\mathscr{H}_{\mathscr{X}}} = \mathbb{E}_{X_t}[f(X_t)^2]$, and hence the operator is injective if for $f \in \mathscr{H}_{\mathscr{X}}$, $\mathbb{E}_{X_t}[f(X_t)^2] = 0$ implies that f is the zero function (injectivity of the operator is equivalent to the kernel of the operator being the zero function). Let $f \in \mathscr{H}_{\mathscr{X}}$, then f is not necessarily the zero function, and under Assumptions 3.2 and 3.3, $p(X_t = i) > 0$ for all $i \in \mathscr{X}$. Then $\mathbb{E}_{X_t}[f(X_t)^2] = \sum_{i=1}^K f(i)^2 p(X_t = i) = 0$ is only possible if $f \equiv 0$. Therefore, the cross-covariance operator $\mathscr{C}_{X_tX_t}$ is injective. ■

The following lemma provides an expression for the prediction step seen in Equation (3.2) in terms of the model parameters, using the proof techniques of Fukumizu et al. [2013]. We assume that the kernel function on $\mathscr{X}$ is the Kronecker-delta kernel function and derive an update rule for the regularized embedding $\mu^{\text{reg}}_{X_t|y_{1:t}}$.

It follows from the kernel sum rule that $\mu_{X_{t+1}|y_{1:t}} = \mathscr{U}_{X_{t+1}|X_t}\mu_{X_t|y_{1:t}}$ [Song et al., 2009, 2013], and by Lemma 3.7 for all $g : \mathscr{X} \to \mathbb{R}$ we have $\mathbb{E}[g(X_{t+1})|X_t = \cdot] \in \mathscr{H}_{\mathscr{X}}$, and thus the conditional mean embedding as defined in Klebanov et al. [2020] (see Section 1.2.2.2) is well defined and such that $\mu_{X_{t+1}|y_{1:t}} = (\mathscr{C}^{\dagger}_{X_tX_t}\mathscr{C}^{*}_{X_{t+1}X_t})^{*}\mu_{X_t|y_{1:t}}$. Hence, to derive an update rule for the regularized embedding we define

$$\mu^{\text{reg}}_{X_{t+1}|y_{1:t}} := (\mathscr{C}^{\dagger}_{X_tX_t}\mathscr{C}^{*}_{X_{t+1}X_t})^{*}\mu^{\text{reg}}_{X_t|y_{1:t}}.$$

Additionally, the conditional mean embeddings $\mathscr{U}_{(X_{t+1}Y_{t+1})|X_{t+1}} = (\mathscr{C}_{X_{t+1}X_{t+1}})^{\dagger}(\mathscr{C}^{*}_{X_{t+1}Y_{t+1}X_{t+1}})^{*}$ and $\mathscr{U}_{(Y_{t+1}Y_{t+1})|X_{t+1}} = ((\mathscr{C}_{X_{t+1}X_{t+1}})^{\dagger}\mathscr{C}^{*}_{Y_{t+1}Y_{t+1}X_{t+1}})^{*}$ are well defined, and the regularized cross-covariance operators are naturally defined as

$$\mathscr{C}^{\pi,\text{reg}}_{X_{t+1}Y_{t+1}} := (\mathscr{C}_{X_{t+1}X_{t+1}})^{\dagger}(\mathscr{C}^{*}_{X_{t+1}Y_{t+1}X_{t+1}})^{*}\mu^{\text{reg}}_{X_{t+1}|y_{1:t}}, \tag{3.4}$$

$$\mathscr{C}^{\pi,\text{reg}}_{Y_{t+1}Y_{t+1}} := ((\mathscr{C}_{X_{t+1}X_{t+1}})^{\dagger}\mathscr{C}^{*}_{Y_{t+1}Y_{t+1}X_{t+1}})^{*}\mu^{\text{reg}}_{X_{t+1}|y_{1:t}}. \tag{3.5}$$

Another advantage of placing the Kronecker-delta kernel function on $\mathscr{X}$ is that it allows us to make predictions over the hidden states. Suppose at time point $t \geq 1$ we have $\mu_{X_t|y_{1:t}}$, then $\mu_{X_t|y_{1:t}}(x) = \mathbb{E}[\delta(X_t, x)|y_{1:t}] = \mathbb{P}(X_t = x|y_{1:t})$, and thus given observations $y_{1:t}$ we can predict the hidden state X_t by estimating $\hat{\mu}_{X_t|y_{1:t}}$, and setting $\hat{X}_t = \operatorname{argmax}_{x\in\mathscr{X}} \hat{\mu}_{X_t|y_{1:t}}(x)$.

Lemma 3.8 (Prediction step). *Suppose that $\mu^{reg}_{X_t|y_{1:t}} = \Psi\alpha_t$ for some $\alpha_t \in \mathbb{R}^K$, and additionally define $\mu^{reg}_{X_{t+1}|y_{1:t}} := (\mathscr{C}^{\dagger}_{X_tX_t}\mathscr{C}^{*}_{X_{t+1}X_t})^{*}\mu^{reg}_{X_t|y_{1:t}}$. Then under Assumptions 3.2 and 3.3*

$$\mu^{reg}_{X_{t+1}|y_{1:t}} = \Psi Q^{\mathrm{T}}\alpha_t. \tag{3.6}$$

Proof. We define the updated regularized embedding $\mu^{\text{reg}}_{X_{t+1}|y_{1:t}} \in \mathscr{H}_{\mathscr{X}}$ to be

$$\mu^{\text{reg}}_{X_{t+1}|y_{1:t}} := (\mathscr{C}^{\dagger}_{X_tX_t}\mathscr{C}^{*}_{X_{t+1}X_t})^{*}\mu^{\text{reg}}_{X_t|y_{1:t}}.$$

By Lemma 3.7, $\varphi_X(x)$ belongs to the range of $\mathscr{C}_{X_tX_t}$ for $x \in \mathscr{X}$, and $\mathscr{C}_{X_tX_t}$ is injective. Together, these two conditions ensure that

$$\mu^{\text{reg}}_{X_{t+1}|y_{1:t}} = (\mathscr{C}^{\dagger}_{X_tX_t}\mathscr{C}^{*}_{X_{t+1}X_t})^{*}\mu^{\text{reg}}_{X_t|y_{1:t}} = \mathscr{C}_{X_{t+1}X_t}\mathscr{C}^{-1}_{X_tX_t}\mu^{\text{reg}}_{X_t|y_{1:t}}.$$

Let $h := (\mathscr{C}_{X_tX_t})^{-1}\mu^{\text{reg}}_{X_t|y_{1:t}}$, then there exists a $\beta \in \mathbb{R}^K$ such that $h = \Psi\beta + h_{\perp}$ where $h_{\perp}$ is orthogonal to $\operatorname{span}\Psi$. It is straightforward to see that $\mathscr{C}_{X_tX_t}h = \mu^{\text{reg}}_{X_t|y_{1:t}}$. Expanding the left-hand side gives

$$\mathscr{C}_{X_tX_t}h = \mathscr{C}_{X_tX_t}(\Psi\beta + h_{\perp}) = \Psi\operatorname{diag}\left(\pi Q^{t-1}\right)\Psi^{\mathrm{T}}\Psi\beta,$$

and thus $\Psi\operatorname{diag}\left(\pi Q^{t-1}\right)\Psi^{\mathrm{T}}\Psi\beta = \Psi\alpha_t$. Left-multiplying by Ψ^{T} and noting that $\Psi^{\mathrm{T}}\Psi = I_K$, the $K \times K$ identity matrix, we obtain

$$\beta = (\operatorname{diag}\left(\pi Q^{t-1}\right))^{-1}\alpha_t.$$

Recall that $\mu_{X_{t+1}|y_{1:t}} = \mathscr{C}_{X_{t+1}X_t} h$, and thus

$$\begin{aligned}
\mu^{\text{reg}}_{X_{t+1}|y_{1:t}} &= \mathscr{C}_{X_{t+1}X_t} h \\
&= \Psi Q^{\mathrm{T}} \operatorname{diag}\left(\pi Q^{t-1}\right) \Psi^{\mathrm{T}} \Psi \beta \\
&= \Psi Q^{\mathrm{T}} \operatorname{diag}\left(\pi Q^{t-1}\right) (\operatorname{diag}\left(\pi Q^{t-1}\right))^{-1} \alpha_t \\
&= \Psi Q^{\mathrm{T}} \alpha_t,
\end{aligned}$$

where the final line uses the fact that $\pi Q^{t-1} = \pi$ under Assumption 3.3, and $\operatorname{diag}(\pi)$ is invertible under Assumption 3.2. ■

We now show how the update step, Equation (3.3), can be computed using the model parameters. We start by developing expressions for the regularized cross-covariance operators $\mathscr{C}^{\pi,\text{reg}}_{X_{t+1}Y_{t+1}}$ and $\mathscr{C}^{\pi,\text{reg}}_{Y_{t+1}Y_{t+1}}$ in terms of the model parameters and the embedded prior $\mu^{\text{reg}}_{X_t|y_{1:t}}$.

Lemma 3.9. *Suppose that $\mu^{reg}_{X_t|y_{1:t}} = \Psi \alpha_t$ for some $\alpha_t \in \mathbb{R}^K$, and let $\mathscr{C}^{\pi,reg}_{X_{t+1}Y_{t+1}}$ and $\mathscr{C}^{\pi,reg}_{X_{t+1}Y_{t+1}}$ be defined as in Equations (3.4) and (3.5). Then*

$$\begin{aligned}
\mathscr{C}^{\pi,reg}_{X_{t+1}Y_{t+1}} &= \sum_{i=1}^{K} \mu_{i,t+1} \varphi_X(i) \otimes \mathbb{E}[\phi_Y(Y_{t+1}) | X_{t+1} = i] = \Psi \Lambda^{t+1} O_2^{\mathrm{T}} \\
\mathscr{C}^{\pi,reg}_{Y_{t+1}Y_{t+1}} &= \sum_{i=1}^{K} \mu_{i,t+1} \mathbb{E}[\phi_Y(Y_{t+1}) | X_{t+1} = i] \otimes \mathbb{E}[\phi_Y(Y_{t+1}) | X_{t+1} = i] = O_2 \Lambda^{t+1} O_2^{\mathrm{T}},
\end{aligned}$$

where $\Lambda^{t+1} := \operatorname{diag}\left(\mu_{t+1}\right)$, and under Assumptions 3.2 and 3.3 the shared parameters $\mu_{t+1} \in \mathbb{R}^K$ can be expressed

$$\mu_{t+1} = Q^{\mathrm{T}} \alpha_t.$$

Proof. The regularized cross-covariance operator is defined as

$$\mathscr{C}^{\pi,\text{reg}}_{X_{t+1}Y_{t+1}} := (\mathscr{C}_{X_{t+1}X_{t+1}})^{\dagger} (\mathscr{C}^{*}_{X_{t+1}Y_{t+1}X_{t+1}})^{*} \mu^{\text{reg}}_{X_{t+1}|y_{1:t}}.$$

It follows from Lemma 3.7 that the cross-covariance operator $\mathscr{C}_{X_{t+1}X_{t+1}}$ is injective and by definition $\varphi_X(x) \in \mathscr{H}_{\mathscr{X}}$ for all $x \in \mathscr{X}$, and thus

$$\mathscr{C}^{\pi,\text{reg}}_{X_{t+1}Y_{t+1}} = (\mathscr{C}_{X_{t+1}X_{t+1}})^{\dagger} (\mathscr{C}^{*}_{X_{t+1}Y_{t+1}X_{t+1}})^{*} \mu^{\text{reg}}_{X_{t+1}|y_{1:t}} = \mathscr{C}_{X_{t+1}Y_{t+1}X_{t+1}} (\mathscr{C}_{X_{t+1}X_{t+1}})^{-1} \mu^{\text{reg}}_{X_{t+1}|y_{1:t}}.$$

Let $h := (\mathscr{C}_{X_{t+1}X_{t+1}})^{-1} \mu_{X_{t+1}|y_{1:t}} = \Psi \beta + h_{\perp}$, for some $\beta \in \mathbb{R}^K$ and $h_{\perp}$ orthogonal to $\operatorname{span} \Psi$. Then

$$\begin{aligned}
\mathscr{C}_{X_{t+1}X_{t+1}} h &= \mathscr{C}_{X_{t+1}X_{t+1}} (\Psi \beta + h_{\perp}) \\
&= \Psi \operatorname{diag}\left(\pi Q^t\right) \Psi^{\mathrm{T}} \Psi \beta,
\end{aligned}$$

which is equal to $\mu_{X_{t+1}|y_{1:t}} = \Psi Q^{\mathrm{T}} \alpha_t$ by Lemma 3.8. Left-multiplying by Ψ^{T} and noting that $\Psi^{\mathrm{T}} \Psi = I_K$, the $K \times K$ identity matrix, we find

$$\beta = (\operatorname{diag}\left(\pi Q^t\right))^{-1} Q^{\mathrm{T}} \alpha_t.$$

It can be shown that $\mu_{t+1} = \operatorname{diag}\left(\pi Q^t\right)\beta$, and hence

$$\mu_{t+1} = \operatorname{diag}\left(\pi Q^t\right)\left(\operatorname{diag}\left(\pi Q^t\right)\right)^{-1} Q^{\mathrm{T}} \alpha_t = Q^{\mathrm{T}} \alpha_t,$$

where the final equality follows from the fact that $\pi Q^{t-1} = \pi$ under Assumption 3.3, and $\operatorname{diag}(\pi)$ is invertible under Assumption 3.2. To conclude we note that $\mathscr{C}^{\pi,\mathrm{reg}}_{X_{t+1}Y_{t+1}}$ and $\mathscr{C}^{\pi,\mathrm{reg}}_{Y_{t+1}Y_{t+1}}$ share the same coefficients. ■

The following lemma completes the use of kernel Bayes' rule by expressing the embedded updated filtering distribution in terms of the model parameters, and providing an update rule for the weights of $\mu^{\mathrm{reg}}_{X_{t+1}|y_{1:(t+1)}}$.

Lemma 3.10. *Let μ_{t+1} denote the coefficients of $\mathscr{C}^{\pi,reg}_{Y_{t+1}Y_{t+1}}$ and $\mathscr{C}^{\pi,reg}_{X_{t+1}Y_{t+1}}$, then under Assumptions 3.1 and 3.4, $\mu^{reg}_{X_{t+1}|y_{1:(t+1)}} = \Psi\alpha_{t+1}$ where the coefficients $\alpha_{t+1} \in \mathbb{R}^K$ are given by*

$$\alpha_{t+1} = \Lambda^{t+1} O_2^{\mathrm{T}} O_2 (\Lambda^{t+1} O_2^{\mathrm{T}} O_2 + \lambda I_K)^{-1} (O_2^{\mathrm{T}} O_2)^{-1} O_2^{\mathrm{T}} \phi_Y(y_{t+1}), \tag{3.7}$$

where $\Lambda^{t+1} := \operatorname{diag}\left(\mu_{t+1}\right)$.

Proof. Recall that $\mu^{\mathrm{reg}}_{X_{t+1}|y_{1:(t+1)}} = \mathscr{C}^{\pi,\mathrm{reg}}_{X_{t+1}Y_{t+1}} (\mathscr{C}^{\pi,\mathrm{reg}}_{Y_{t+1}Y_{t+1}} + \lambda \mathscr{I}_{\mathscr{H}^{\otimes 2}_{\mathscr{Y}}})^{-1} \phi_Y(y_{t+1})$, and define the quantity $h := (\mathscr{C}^{\pi,\mathrm{reg}}_{Y_{t+1}Y_{t+1}} + \lambda \mathscr{I}_{\mathscr{H}^{\otimes 2}_{\mathscr{Y}}})^{-1} \phi_Y(y_{t+1}) = O_2\beta + h_\perp$, for $\beta \in \mathbb{R}^K$ and $h_\perp$ orthogonal to $\operatorname{span}(O_2)$. Then using the representation of $\mathscr{C}^{\pi,\mathrm{reg}}_{Y_{t+1}Y_{t+1}}$ from Lemma 3.9, we have

$$\begin{aligned}
(\mathscr{C}^{\pi,\mathrm{reg}}_{Y_{t+1}Y_{t+1}} + \lambda \mathscr{I}_{\mathscr{H}^{\otimes 2}_{\mathscr{Y}}}) h &= (\mathscr{C}^{\pi,\mathrm{reg}}_{Y_{t+1}Y_{t+1}} + \lambda \mathscr{I}_{\mathscr{H}^{\otimes 2}_{\mathscr{Y}}})(O_2\beta + h_\perp) \\
&= (O_2 \Lambda^{t+1} O_2^{\mathrm{T}} + \lambda \mathscr{I}_{\mathscr{H}^{\otimes 2}_{\mathscr{Y}}})(O_2\beta + h_\perp) \\
&= O_2 (\Lambda^{t+1} O_2^{\mathrm{T}} O_2 + \lambda I_K)\beta + \lambda h_\perp,
\end{aligned}$$

which is equal to $\phi_Y(y_{t+1})$. Hence, left-multiplying by O_2^{T} and noting that $O_2^{\mathrm{T}} O_2$ is invertible by Assumptions 3.1 and 3.4, we have $\beta = (\Lambda^{t+1} O_2^{\mathrm{T}} O_2 + \lambda I_K)^{-1} (O_2^{\mathrm{T}} O_2)^{-1} O_2^{\mathrm{T}} \phi_Y(y_{t+1})$, and therefore

$$\begin{aligned}
\mu^{\mathrm{reg}}_{X_{t+1}|y_{1:(t+1)}} &= \mathscr{C}^{\pi,\mathrm{reg}}_{X_{t+1}Y_{t+1}} h \\
&= \Psi \Lambda^{t+1} O_2^{\mathrm{T}} O_2 \beta \\
&= \Psi \Lambda^{t+1} O_2^{\mathrm{T}} O_2 (\Lambda^{t+1} O_2^{\mathrm{T}} O_2 + \lambda I_K)^{-1} (O_2^{\mathrm{T}} O_2)^{-1} O_2^{\mathrm{T}} \phi_Y(y_{t+1}).
\end{aligned}$$

Finally we conclude that $\mu^{\mathrm{reg}}_{X_{t+1}|y_{1:(t+1)}} = \Psi\alpha_{t+1}$ where

$$\alpha_{t+1} = \Lambda^{t+1} O_2^{\mathrm{T}} O_2 (\Lambda^{t+1} O_2^{\mathrm{T}} O_2 + \lambda I_K)^{-1} (O_2^{\mathrm{T}} O_2)^{-1} O_2^{\mathrm{T}} \phi_Y(y_{t+1}).$$

■

3.4.2.3 Practical implementation

We now combine the above lemmas to provide a set of recursive equations allowing for inference in the filtering problem. As one would do in a probabilistic setting, we compute the filtering distributions via prediction and update steps described by Equations (3.2) and (3.3), and implement these steps in the RKHS via the recursive equations given in Equations (3.6) and (3.7).

We sequentially estimate the embedding of $X_t|y_{1:t}$ for $t \geq 1$. We initialize the procedure by estimating $\mu^{\text{reg}}_{X_1|y_1} = \mathscr{C}_{X_1Y_1}(\mathscr{C}_{Y_1Y_1} + \lambda \mathscr{I}_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}})^{-1}\phi_Y(y_1)$, which corresponds to computing the weight vector $\alpha_1 = \text{diag}(\pi)O_2^{\text{T}}O_2(\text{diag}(\pi)O_2^{\text{T}}O_2 + \lambda I_K)^{-1}O_2^{\text{T}}\phi_Y(y_1)$. The procedure is summarized as

$$
\begin{aligned}
\alpha_1 &= \text{diag}(\pi)O_2^{\text{T}}O_2(\text{diag}(\pi)O_2^{\text{T}}O_2 + \lambda I_K)^{-1}O_2^{\text{T}}\phi_Y(y_1) \\
\mu_{t+1} &= Q^{\text{T}}\alpha_t \\
\alpha_{t+1} &= \Lambda^{t+1}O_2^{\text{T}}O_2(\Lambda^{t+1}O_2^{\text{T}}O_2 + \lambda I_K)^{-1}(O_2^{\text{T}}O_2)^{-1}O_2^{\text{T}}\phi_Y(y_{t+1}).
\end{aligned}
$$

At time point t we have $\mu_{X_t|y_{1:t}} = \Psi\alpha_t$, and recall that when the kernel on $\mathscr{X}$ is the Kronecker-delta kernel we have $\mu_{X_t|y_{1:t}}(x) = \mathbb{E}[\delta(X_t, x)|y_{1:t}] = \mathbb{P}(X_t = x|y_{1:t})$. Hence, given observations $y_{1:t}$ we can predict the hidden state X_t as $\hat{X}_t = \text{argmax}_{x\in\mathscr{X}}\,\hat{\mu}^{\text{reg}}_{X_t|y_{1:t}}(x)$.

The alternative kernel Bayes' rule described above avoids the typical $O(p^3)$ computational complexity typically required in kernel methods — each iteration of the above filtering procedure costs $O(K^3)$ and it is typically assumed that $K \ll p$. Compared to using the forward-backward algorithm in Section 3.4.1, this approach has the added benefit that any characteristic kernel on $\mathscr{Y}$ can be used.

A practical implementation of the algorithm is described in Algorithm 2. To simplify the presentation we define $\hat{B} = \hat{W}\hat{D}\Theta\hat{L} \in \mathbb{R}^{n\times K}$ so that $\hat{O}_2 = \Phi_3\hat{B}$.

Algorithm 2 Filtering via the alternative KBR

```
Input: Estimated HMM parameters ($\hat{O}_2 = \Phi_3\hat{B}, \hat{Q}, \hat{\pi}$), regularization parameter $\lambda > 0$, observations $y_{1:p}$.
Output: Filtering predictions $(\hat{X}_t)_{t=1}^{p}$.
Compute $\hat{N}_2 := \hat{B}^{\text{T}}K_3\hat{B}$
Compute $\hat{N}(y) := \hat{B}^{\text{T}}\Phi_3(y)$
Initialize $\hat{\alpha}_1 = \text{diag}(\hat{\pi})\hat{N}_2(\text{diag}(\hat{\pi})\hat{N}_2 + \lambda I_K)^{-1}\hat{N}(y_1)$
for $t = 1$ to $p - 1$ do
    Compute $\hat{\mu}_{t+1} = \hat{Q}^{\text{T}}\hat{\alpha}_t$, set $\hat{\Lambda}^{t+1} = \text{diag}\left(\hat{\mu}_{t+1}\right)$
    Update $\hat{\alpha}_{t+1} = \hat{\Lambda}^{t+1}\hat{N}_2(\hat{\Lambda}^{t+1}\hat{N}_2 + \lambda I_K)^{-1}\hat{N}_2^{-1}\hat{N}(y_{t+1})$
    Predict $\hat{X}_{t+1} = \text{argmax}_{x\in\mathscr{X}}\,\hat{\mu}^{\text{reg}}_{X_{t+1}|y_{1:t+1}}(x) = \Psi(x)\hat{\alpha}_{t+1}$
end for
```

3.5 Order estimation

In this section we provide a consistent estimator of the HMM order using kernel mean embeddings. Consistent order estimation for the nonparametric method proposed by De Castro et al. [2017]

was studied in Lehéricy [2019], and in the following we repeat their analysis adapted to our setting. Lehéricy [2019]'s consistency result demonstrates a trade-off between the approximation complexity and the sample size, whereas our order estimator does not have such a trade-off. The consistency of our estimator depends only on the transition matrix of the hidden process.

The estimator is motivated by the following simple observation. It is easily shown that (as in the proof of Theorem 3.1)

$$\mathscr{C}_{1,2} = O_2 \operatorname{diag}(\pi) Q O_2^{\mathrm{T}}, \tag{3.8}$$

and hence under Assumptions 3.1 to 3.4, the cross-covariance operator $\mathscr{C}_{1,2}$ has rank K.

Hence, by studying the singular values of the empirical cross-covariance operator we may be able to infer K. The empirical cross-covariance operator $\hat{\mathscr{C}}_{1,2}$ approximates the population version $\mathscr{C}_{1,2}$, and given n samples the empirical operator will have n singular values which decay towards zero. Equation (3.8) hints that the singular values of $\mathscr{C}_{1,2}$ are closely related to the rank of the transition matrix Q and the linear independence of the observation densities. When the transition matrix is close to not being full rank, several of the singular values of $\mathscr{C}_{1,2}$ shrink towards zero, making them difficult to differentiate from the noisy singular values of the empirical operator $\hat{\mathscr{C}}_{1,2}$.

We present the main consistency result below. In the following lemma we denote by $|\cdot|$ the number of elements in a set and by $\sigma_1(A) \geq \sigma_2(A) \geq \cdots$ the singular values of the operator A.

Theorem 3.2 (Order estimator consistency). *Let $\hat{K}(C) = |\{i \geq 1 \mid \sigma_i(\hat{\mathscr{C}}_{1,2}) > C\sqrt{\log(n)/n}\}|$. Under Assumptions 3.1 to 3.4 there exists $C_0 = C_0(Q,k)$ and $n_0 = n_0(Q,k,O_2)$ such that for all $C \geq C_0$ and $n \geq n_0 C^2(1+\log(C))$, $\mathbb{P}(\hat{K}(C) \neq K) \leq n^{-2}$, such that $\hat{K}(C) \to K$ almost surely.*

A data-driven estimator. The consistency result states that our order estimator is consistent almost surely for any C larger than some C_0 which depends on the properties of the underlying Markov chain. Choosing an appropriate C can be difficult in practice, and hence we propose a data-driven estimator of the HMM order. Our estimator is motivated by the fact that if our assumptions hold, the leading K singular values of the empirical cross-covariance operator should be significantly larger than the remaining $n-K$ noisy singular values which decrease towards zero. Thus, we compute the leading $\hat{K}_{\max} \geq 1$ singular values of the empirical cross-covariance operator $\hat{\mathscr{C}}_{1,3}$ and search for an 'elbow' in the sequence of singular values. We use the Kneedle algorithm [Satopaa et al., 2011] to determine the point of maximum curvature in the sequence. This type of heuristic is widely-used in statistical techniques which analyze eigenvalues, such as principle component analysis.

3.5.1 Proof of Theorem 3.2

To prove Theorem 3.2, we first derive a set of concentration inequalities for the cross-covariance operators. Consider consecutive observations of the hidden Markov chain $Z_t := (Y_t, Y_{t+1}, Y_{t+2})$, for

$1 \leq t \leq n$ which satifies Assumptions 3.2 to 3.4 . We adapt the proof of Lemma 27 of De Castro et al. [2017] and use results from Paulin [2015].

Lemma 3.11 (Concentration inequalities). *There exists a constant $C^\star$ which depends on the transition matrix Q such that for any n and $u > 0$, for A_k the bound of the kernel on $\mathcal{Y}$ such that* $\sup_{y\in\mathcal{Y}} k(y,y) \leq A_k$*,*

$$\mathbb{P}\left(\|\hat{\mu}_1 - \mu_1\|_{\mathcal{H}_\mathcal{Y}} \geq C^\star \frac{A_k^{1/2}}{\sqrt{n}}(1+u)\right) \leq \exp\left(-u^2\right)$$

$$\mathbb{P}\left(\|\hat{\mathscr{C}}_{1,3} - \mathscr{C}_{1,3}\|_{\mathcal{H}_\mathcal{Y}^{\otimes 2}} \geq C^\star \frac{A_k}{\sqrt{n}}(1+u)\right) \leq \exp\left(-u^2\right)$$

$$\mathbb{P}\left(\|\hat{\mathscr{C}}_{1,2,3} - \mathscr{C}_{1,2,3}\|_{\mathcal{H}_\mathcal{Y}^{\otimes 3}} \geq C^\star \frac{A_k^{3/2}}{\sqrt{n}}(1+u)\right) \leq \exp\left(-u^2\right).$$

Proof. We first define the pseudo-spectral gap of the hidden process which applies to non-reversible Markov chains, and the mixing time [Paulin, 2015]. Let G_{ps} denote the pseudo-spectral gap of the hidden process $(X_n)_{n\geq 1}$ defined as

$$G_{\text{ps}} = \max_{k\geq 1}\left\{G\left(\operatorname{diag}(\pi)^{-1}(Q^{\mathrm{T}})^k \operatorname{diag}(\pi) Q^k\right)/k\right\},$$

where $G(A)$ denotes the spectral gap of the transition matrix A defined to be

$$G(A) = \begin{cases} 1 - \max\{\lambda \mid \lambda \neq 1\}, & \text{if eigenvalue 1 has multiplicity 1} \\ 0, & \text{otherwise} \end{cases}$$

and T_{mix} denotes the mixing time of the hidden process which we define to be

$$T_{\text{mix}} = \frac{1 + 3\log(2) - \log(\pi_{\min})}{G_{\text{ps}}}.$$

We derive the concentration inequality for $\|\hat{\mathscr{C}}_{1,3} - \mathscr{C}_{1,3}\|_{\mathcal{H}_\mathcal{Y}^{\otimes 2}}$.

Set $\zeta_{1,3}(Z_1,\dots,Z_n) = \|\hat{\mathscr{C}}_{1,3}(Z_1,\dots,Z_n) - \mathscr{C}_{1,3}\|_{\mathcal{H}_\mathcal{Y}^{\otimes 2}}$, where $\hat{\mathscr{C}}_{1,3}(Z_1,\dots,Z_n)$ is used to highlight the dependence of the estimator $\hat{\mathscr{C}}_{1,3}$ on the sample $(Z_1,\dots,Z_n)$. To proceed via McDiarmid's inequality, we compute the difference upon changing the i-th coordinate,

$$c_i := \sup_{z_i, z_i' \in \mathcal{Y}^3} \left|\zeta_{1,3}(z_1,\dots,z_i,\dots,z_n) - \zeta_{1,3}(z_1,\dots,z_i',\dots,z_n)\right|$$

which by the inverse triangle inequality is bounded as follows

$$c_i \leq \sup_{z_i, z_i' \in \mathcal{Y}^3} \|\hat{\mathscr{C}}_{1,3}(z_1,\dots,z_i,\dots,z_n) - \hat{\mathscr{C}}_{1,3}(z_1,\dots,z_i',\dots,z_n)\|_{\mathcal{H}_\mathcal{Y}^{\otimes 2}}.$$

We then obtain

$$c_i \leq \frac{1}{n} \sup_{z_i, z_i' \in \mathcal{Y}^3} \|\phi_Y(y_i) \otimes \phi_Y(y_{i+2}) - \phi_Y(y_i') \otimes \phi_Y(y_{i+2}')\|_{\mathcal{H}_\mathcal{Y}^{\otimes 2}}.$$

The kernel on $\mathscr{H}_{\mathscr{Y}}$ is bounded by A_k such that $\sup_{y\in\mathscr{Y}} k(y,y) \leq A_k$, and thus we have for any $y, y' \in \mathscr{Y}$, $\|\phi_Y(y) \otimes \phi_Y(y')\|_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} = \|\phi_Y(y)\|_{\mathscr{H}_{\mathscr{Y}}} \|\phi_Y(y')\|_{\mathscr{H}_{\mathscr{Y}}} = \sqrt{k(y,y)k(y',y')} \leq A_k$. Then $c_i \leq 2n^{-1}A_k$, and hence $\|c\|_2^2 \leq 4A_k^2 n^{-1}$. McDiarmid's inequality then yields for any $u > 0$ (using Equations 2.8 and 2.9 of Paulin [2015]),

$$\mathbb{P}\left(\|\hat{\mathscr{C}}_{1,3} - \mathscr{C}_{1,3}\|_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} \geq \mathbb{E}\left[\|\hat{\mathscr{C}}_{1,3} - \mathscr{C}_{1,3}\|_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}\right] + u\right) \leq \exp\left(-\frac{nu^2}{18T_{mix}A_k^2}\right).$$

Lemma 28 of De Castro et al. [2017] can be rewritten to state that

$$\mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{n}\left[(\phi_Y(Y_i) \otimes \phi_Y(Y_{i+2})) - \mathscr{C}_{1,3}\right]\right]^2 \leq \frac{4}{nG_{ps}} \mathbb{E}[(\phi_Y(Y_1) \otimes \phi_Y(Y_3)) - \mathscr{C}_{1,3}]^2,$$

and it follows that $\mathbb{E}\left[\|\hat{\mathscr{C}}_{1,3} - \mathscr{C}_{1,3}\|_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}\right] \leq \left(\frac{4A_k^2}{nG_{ps}}\right)^{1/2}$. Combining the above, we have for any $u > 0$,

$$\mathbb{P}\left(\|\hat{\mathscr{C}}_{1,3} - \mathscr{C}_{1,3}\|_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} \geq \left(\frac{4A_k^2}{nG_{ps}}\right)^{1/2} + u\right) \leq \exp\left(-\frac{nu^2}{18T_{mix}A_k^2}\right).$$

Let $u' := \left(\frac{nu^2}{18T_{mix}A_k^2}\right)^{1/2}$, then $u^2 = \frac{18T_{mix}A_k^2}{nu^2}$ and $T_{mix} = \kappa^2/\pi_{min}$ for $\kappa = \sqrt{1 + \log(8/\pi^\star_{min}})$. Then we have the following concentration inequality for any $u' > 0$

$$\mathbb{P}\left(\|\hat{\mathscr{C}}_{1,3} - \mathscr{C}_{1,3}\|_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} \geq \left(\frac{4A_k^2}{nG_{ps}}\right)^{1/2} (1 + \sqrt{4.5}u'\kappa)\right) \leq \exp\left(-(u')^2\right). \tag{3.9}$$

As $u' > 0$ and $\kappa > 1$, the above implies that

$$\mathbb{P}\left(\|\hat{\mathscr{C}}_{1,3} - \mathscr{C}_{1,3}\|_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} \geq C^\star \frac{A_k}{\sqrt{n}}(1 + u')\right) \leq \exp\left(-(u')^2\right), \tag{3.10}$$

where $C^\star = \sqrt{4/G_{ps}}$. The other inequalities follow almost identically. ■

Comparing the concentration inequalities given in Lemma 3.11 to those in Appendix E of De Castro et al. [2017], we see that when embedding consecutive observations the existing method possesses a trade-off between the approximation complexity and the sample size. Their term which grows with approximation complexity is replaced by A_k in our setting, which is typically equal to 1 for popular kernels such as the Gaussian and Laplace kernels.

The following lemma describes an inequality similar to Weyl's inequality for compact operators on separable Hilbert spaces.

Lemma 3.12 (Gohberg et al. [1990, Corollary 1.6])**.** *Let H be a separable Hilbert space, and let A and B be compact operators on H. Then for all $i \geq 1$ we have*

$$|\sigma_i(A) - \sigma_i(B)| \leq \sigma_1(A - B).$$

We now prove the consistency of the order estimator by adapting the proof of Lehéricy [2019, Theorem 13].

Proof of Theorem 3.2. For any Hilbert-Schmidt operator A, we have $\|A\|_{HS}^2 = \sum_{i\geq 1}\sigma_i(A)^2$. Hence $\sigma_1(A) \leq \|A\|_{HS}$. Under Assumption 1.1, the cross-covariance operator $\mathscr{C}_{1,2}$ and its empirical estimate $\hat{\mathscr{C}}_{1,2}$ are well defined as Hilbert-Schmidt operators. Setting $u = \sqrt{2\log(n)}$, Lemma 3.11 implies that with probability $1-n^2$,

$$\sigma_1(\mathscr{C}_{1,2} - \hat{\mathscr{C}}_{1,2}) \leq C\sqrt{\frac{log(n)}{n}},$$

for $C \geq C_0 := 2\sqrt{2}A_k C^\star$. It follows from Lemma 3.12 that with probability at least $1-n^{-2}$, for all $1 \leq i \leq K$, $\sigma_i(\hat{\mathscr{C}}_{1,2}) > \sigma_K(\mathscr{C}_{1,2}) - C\sqrt{\log(n)/n}$, and for all $i > K$, $\sigma_i(\hat{\mathscr{C}}_{1,2}) < C\sqrt{\log(n)/n}$. Note that if n and C are such that $2C\sqrt{\log(n)/n} < \sigma_K(\mathscr{C}_{1,2})$, then the number of singular values greater than $C\sqrt{\log(n)/n}$ is equal to $K^\star$ with probability at least $1-n^{-2}$. It can be shown that for $n \geq n_0 C^2(1+\log(C))$, where $n_0 = 12/\sigma_K(\mathscr{C}_{1,2})^2$, this condition is satisfied. ■

3.6 Experiments

In this section we apply our method to several simulated and synthetic datasets. We evaluate the method's performance in recovering the model parameters, hidden state estimation in the filtering problem, and order estimation. We analyze the performance as the order of the underlying HMM increases, and also on a synthetic HMM created using the MNIST data set [LeCun et al., 2010].

Metrics. As the labels are unobserved, the HMM parameters are only identifiable up to permutations of the hidden state labels. Therefore, in the following we compute errors using permutation invariant measures. Let $\mathscr{S}(\mathscr{X})$ denote the set of permutations of the set $\mathscr{X}$, and let $\tau \in \mathscr{S}(\mathscr{X})$ be a permutation of the hidden state labels. Then we define the permuted estimators $\hat{Q}_\tau$ and $\hat{\pi}_\tau$ to be $[\hat{Q}_\tau]_{i,j} = [\hat{Q}]_{\tau(i),\tau(j)}$ and $[\hat{\pi}_\tau]_i = [\hat{\pi}]_{\tau(i)}$ for $i,j \in \{1,\ldots,K\}$. We then compute the total mean squared error (TMSE) in estimating the HMM parameters as

$$\text{TMSE}((Q,\pi),(\hat{Q},\hat{\pi})) := \inf_{\tau\in\mathscr{S}(\mathscr{X})}\left(\|\pi - \hat{\pi}_\tau\|_2^2 + \|Q - \hat{Q}_\tau\|_F^2\right), \tag{3.11}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Similarly, let $\hat{X}_t$ denote an estimated hidden state at time $t \geq 1$ and let $\hat{X}_{\tau,t}$ denote the estimated hidden state when the hidden labels are permuted by τ, such that if $\hat{X}_t = i$, then $\hat{X}_{\tau,t} = \tau(i)$, for $i \in \mathscr{X}$. Then to compute the accuracy in predicting hidden states in the filtering problem, we use the permutation-invariant accuracy measure

$$\text{Accuracy}(\hat{X}_{1:p}, X_{1:p}) := \sup_{\tau\in\mathscr{S}(\mathscr{X})}\left(\frac{1}{p}\sum_{t=1}^{p} 1_{\hat{X}_{\tau,t}=X_t}\right), \tag{3.12}$$

where 1_x denotes the indicator function which is equal to 1 when x is true and 0 otherwise.

3.6.1 Simulated and synthetic datasets

Let $f_{\alpha,\beta} := f(x;\alpha,\beta)$ denote the probability density function of a beta distribution on $[0,1]$ with shape parameters $\alpha,\beta > 0$, and let $g_{\mu,\sigma^2} := f(x;\mu,\sigma^2)$ denote the probability density function of a Gaussian distribution with mean μ and variance σ^2 for $\mu \in \mathbb{R}$ and $\sigma > 0$. The simulated datasets we consider are described by the HMM (F,Q,π), where the model parameters are as follows

Model 12.

$$F = \begin{pmatrix} f_{2,5}, & f_{4,3} \end{pmatrix}, \quad Q = \begin{pmatrix} 0.4 & 0.6 \\ 0.8 & 0.2 \end{pmatrix}, \quad \pi = \begin{pmatrix} \frac{4}{7} & \frac{3}{7} \end{pmatrix} \tag{3.13}$$

Model 13.

$$F = \begin{pmatrix} f_{2,5}, & f_{4,2}, & f_{4,4} \end{pmatrix}, \quad Q = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.07 & 0.13 & 0.8 \end{pmatrix}, \quad \pi = \begin{pmatrix} \frac{47}{120} & \frac{11}{40} & \frac{1}{3} \end{pmatrix} \tag{3.14}$$

Model 14.

$$F = \begin{pmatrix} f_{0.5,0.5}, & f_{2,2} \end{pmatrix}, \quad Q = \begin{pmatrix} 0.1 & 0.9 \\ 0.7 & 0.3 \end{pmatrix}, \quad \pi = \begin{pmatrix} \frac{7}{16} & \frac{9}{16} \end{pmatrix} \tag{3.15}$$

Model 12 provides a simple example where the Markov chain has good mixing properties, and the observation densities are well separated. This model was considered in De Castro et al. [2017], where it is demonstrated that their method performs well. Model 13 provides an example where the observation densities are not well separated, which may lead to poor performance as the observation densities may not be linearly independent, violating Assumption 3.4. This setting was considered in Lehéricy [2019] as an example in which the spectral method of De Castro et al. [2017] fails to recover the order of the HMM. Model 14 provides an example where one of the observation densities is hard to estimate as it is unbounded near the closure of the domain. In this setting we expect our method which avoids density estimation to outperform both the existing spectral method and the kernel method where filtering is performed via density estimation and the forward-backward procedure.

We also generate a synthetic dataset using a hidden Markov model and the MNIST dataset. The MNIST dataset [LeCun et al., 2010] is a collection of handwritten digits, size-normalized and centred in a 28 × 28 grayscale image. The observations are the pixel intensities of the image, making the observation space $[0,255]^{28\times 28}$, accompanied with their label which takes values between 0 and 9. We flatten the observations into vectors and scale each observation to have mean zero and standard variance so that the observation space is $\mathscr{Y} = \mathbb{R}^{784}$. In the following we define a hidden Markov model using the MNIST data set by defining the latent space to be a subset of possible digits, $\mathscr{X} \subset \{0,\dots,9\}$, and sampling $Y_t|X_t = x$ uniformly from the set of observations with label $x \in \mathscr{X}$. To define the HMM we must also define the transition matrix and stationary distribution. We consider the following models, where the transition matrices are chosen so that the Markov chain has good mixing properties,

Model 15. *In this case $K = 2$ and $\mathscr{X} = \{2,7\}$. We specify*

$$Q = \begin{pmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{pmatrix}, \quad \pi \approx \begin{pmatrix} 0.53 & 0.47 \end{pmatrix} \tag{3.16}$$

Model 16. *In this case $K = 3$ and $\mathscr{X} = \{2,4,7\}$. We specify*

$$Q = \begin{pmatrix} 0.1 & 0.3 & 0.6 \\ 0.45 & 0.25 & 0.3 \\ 0.4 & 0.6 & 0 \end{pmatrix}, \quad \pi \approx \begin{pmatrix} 0.32 & 0.37 & 0.31 \end{pmatrix} \tag{3.17}$$

Model 17. *In this case $K = 4$ and $\mathscr{X} = \{1,2,4,7\}$. We specify*

$$Q = \begin{pmatrix} 0.1 & 0.3 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.3 & 0.1 \\ 0.45 & 0.3 & 0 & 0.25 \\ 0.15 & 0.3 & 0.35 & 0.2 \end{pmatrix}, \quad \pi \approx \begin{pmatrix} 0.27 & 0.27 & 0.22 & 0.24 \end{pmatrix} \tag{3.18}$$

This setting provides an interesting example where a parametric hidden Markov model is hard to specify, and we show that impressive performance can be obtained in a complicated unsupervised learning problem.

3.6.2 An example

We start with an illustrative example using simulated data generated from Model 12 to estimate the observation densities using the kernel method proposed in Section 3.3 and the density estimator defined in Chapter 2. We sample a sequence of 5000 observations and estimate the HMM parameters using Algorithm 1. We use a Gaussian kernel function and use the median heuristic to choose the kernel hyperparameter. Given the estimated embedded observation distributions, we estimate the underlying densities with the kernel hyperparameter 0.5 times the median heuristic. The true densities and their estimates are shown in Figure 3.1. No parameter tuning is performed.

3.6.3 Filtering

In this section we assess the performance of our method in estimating the parameters of a hidden Markov model and hidden state estimation for the filtering problem. We assess performance by simulating data from Models 12 to 14, and compare our method's performance to that of the nonparametric spectral HMM of De Castro et al. [2017].

In the following experiments we sample a sequence of n observations from Models 12 to 14, for n varying between 500 and 20000, and use Algorithm 1 to estimate the hidden Markov model parameters. We use the Gaussian kernel function on $\mathscr{Y}$ and set the kernel hyperparameter using the median heuristic. We then sample an additional sequence of $p = 1000$ observations, and

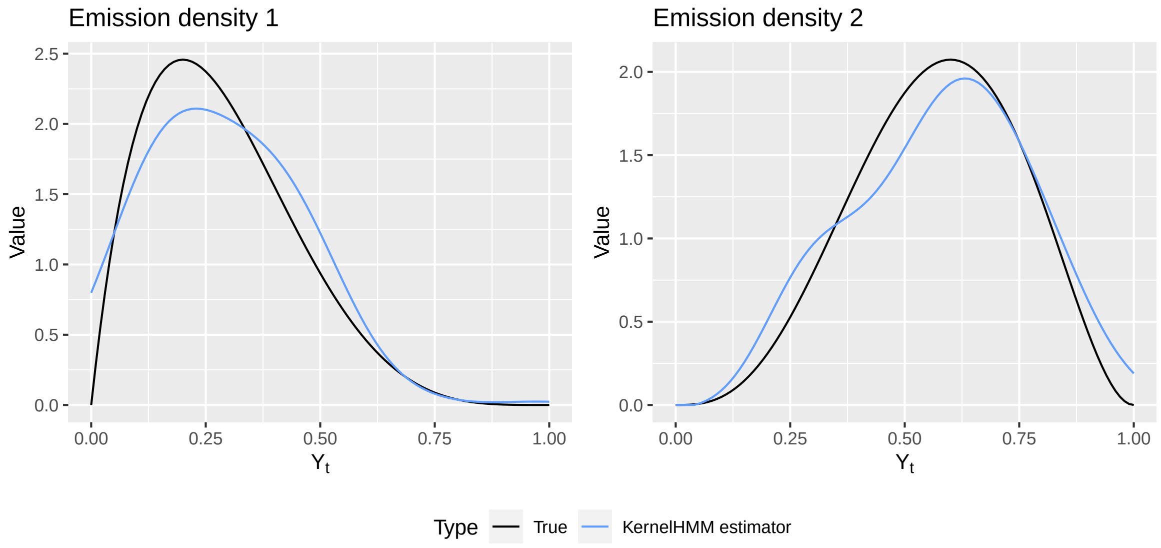


Figure 3.1: HMM observation densities and their estimates. Observation density 1 and 2 are in black and correspond to Beta distributions with parameters $(2,5)$ and $(4,3)$ respectively. The kernel estimators overlaid in blue are estimated from the estimated embeddings output from Algorithm 1, using 5000 observations and the Gaussian kernel function with hyperparameter chosen via the median heuristic.

use the two filtering procedures discussed in Section 3.4 to predict the hidden states from the estimated posterior distributions. We compute the total mean squared error in estimating the HMM parameters Q and π using Equation (3.11), and evaluate the accuracy in predicting hidden states in the filtering problem using Equation (3.12). We repeat this procedure 10 times and average the results which can be seen in Figures 3.2 and 3.3, where `KernelHMM` is used to denote our method, and `ProjHMM` is used to denote the nonparametric method proposed by De Castro et al. [2017].

For the implementation of the existing nonparametric method, we use a trigonometric basis, and fit models for basis dimensions $M = 5,\dots,20$. For each n and M we average the TMSE and accuracy over 10 runs, and to avoid tuning the parameter M we keep the model which produces the best performance on the test set. Hence, our experiments correspond to the best-case scenario for the existing nonparametric method. Whilst the kernel method only uses a maximum of 20000 observations, we fit the existing method for up to 200000 observations to demonstrate that the existing method can achieve similar results to our method, however more observations are required. The computational cost of the kernel method scales quadratically with n, and hence we limit our analysis of the method to 20000 observations.

In Figure 3.2, we see that for Model 12 our parameter estimators outperform the existing method, and for the existing method to obtain similar performance approximately 10 times more data is required. Both methods perform poorly for data generated from Model 13, as expected,

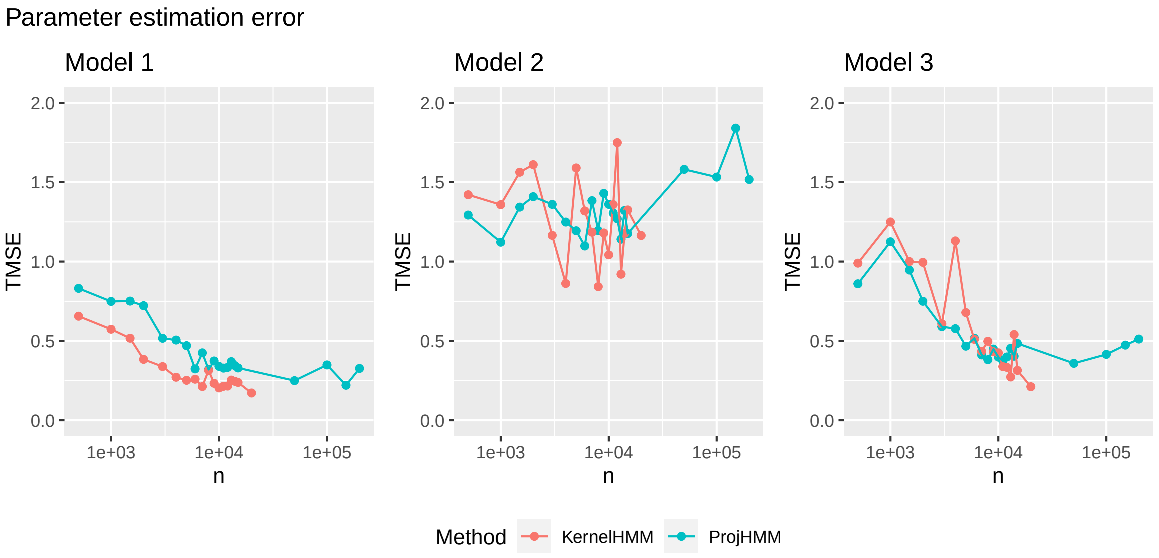


Figure 3.2: The total mean squared error in estimating HMM parameters Q and π for both the proposed method, KernelHMM, and the existing nonparametric method, ProjHMM, proposed by De Castro et al. [2017]. The error is averaged over 10 runs for a varying number of observations (n), and the TMSE is plot on a $\log_{10}$ scale. The figure demonstrates the performance of our method (KernelHMM) and the comparative method (ProjHMM) when applied to different simulated data sets.

as the model violated the model assumptions. Both methods perform similarly for Model 14, however our method obtains the best performance as n increases beyond 10000.

In Figure 3.3 we use `KernelHMM + FB` to denote the kernel algorithm followed by the forwards procedure and `KernelHMM + KBR` to denote the kernel algorithm followed by the alternative kernel Bayes' rule procedure. We refer to the existing nonparametric method and the forwards procedure as `ProjHMM + FB`. We see that the kernel method followed by the kernel Bayes' rule consistently produces the best accuracy, and we believe this is because it avoids density estimation. Our parameter estimates input to the forward algorithm generally outperform the existing method.

3.6.4 Order estimation

In this section we evaluate the performance of the order estimator proposed in Section 3.5 for Models 12 to 14. We use a Gaussian kernel function, with a fixed hyperparameter across all experiments. We generate n observations from the model, for n varying between 500 and 10000, and estimate the order of the HMM by the following procedure, motivated by Theorem 3.2. Assuming that $K \ll n$, we define a positive integer $\hat{K}_{\max}$ and compute the leading $\hat{K}_{\max}$ singular values of the empirical cross-covariance operator $\hat{\mathscr{C}}_{1,2}$ via a truncated singular value decomposition. We then estimate K to be $\hat{K} = |\{i \geq 1 \mid \sigma_i(\hat{\mathscr{C}}_{1,2}) > C\sqrt{\log(n)/n}\}|$ for some $C > 0$, where σ_i denotes

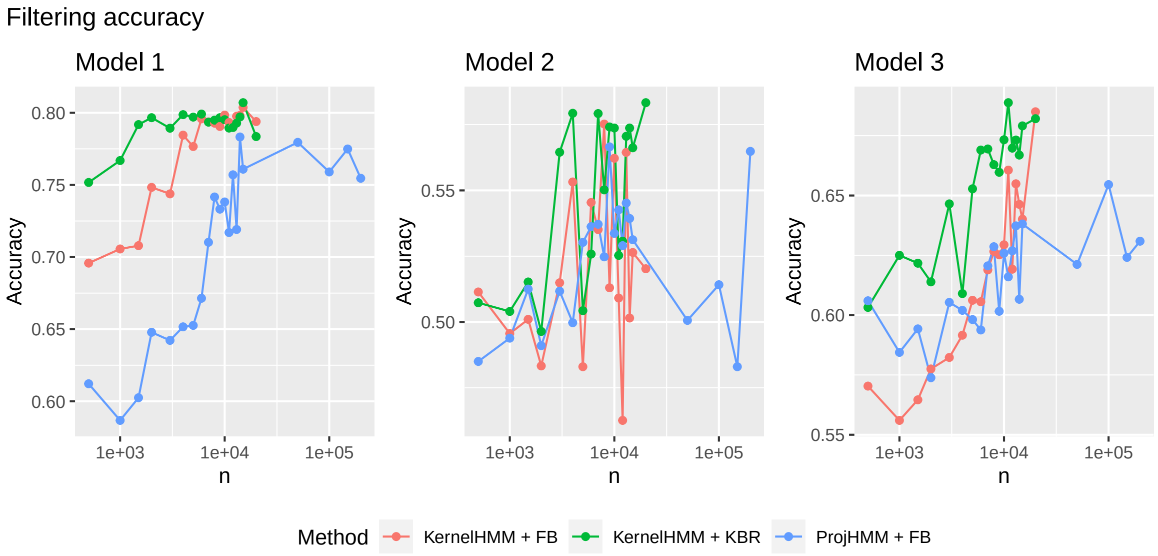


Figure 3.3: The filtering accuracy of the proposed method, KernelHMM, and the existing nonparametric method, ProjHMM, proposed by De Castro et al. [2017]. The accuracy is computed over a test set of $p = 1000$ observations and is averaged over 10 runs for a varying number of training observations (n). The accuracy is plot on a $\log_{10}$ scale. The figure demonstrates the filtering performance of our method (KernelHMM) and the comparative method (ProjHMM) when applied to different simulated data sets.

the i-th singular value. We repeat the experiment 10 times for each n and report the accuracy over n for several values of C in Figure 3.4. We see that for a large enough value of C, the order estimator correctly estimates the order almost surely as n grows.

As an appropriate value of C depends on the underlying Markov chain, we also evaluate the performance of the data-driven estimator proposed in Section 3.5. We compute the leading 100 singular values of the empirical operator and use the Kneedle algorithm to determine the point of maximum curvature in the sequence of singular values. Our order estimate is the point of maximum curvature. As in the previous experiment, we predict the HMM order for varying n between 50 and 5000, and repeat the experiment 10 times. The order estimation accuracy for this data-driven estimator can be seen in Figure 3.5. We see that for each model the accuracy increases over n, although not monotonically. Note that for Model 13, Lehéricy [2019] report that a similar order estimator using the existing nonparametric method of De Castro et al. [2017] has an accuracy of 0 for $n = 7500, 19998, 30000$, and 0.1 for $n = 49998$. Model 13 has three states and thus Figure 3.5 shows that our estimator does significantly better than guessing.

Using the truncated SVD, both order estimation procedures have cost $O(\hat{K}_{\max} n^2)$.

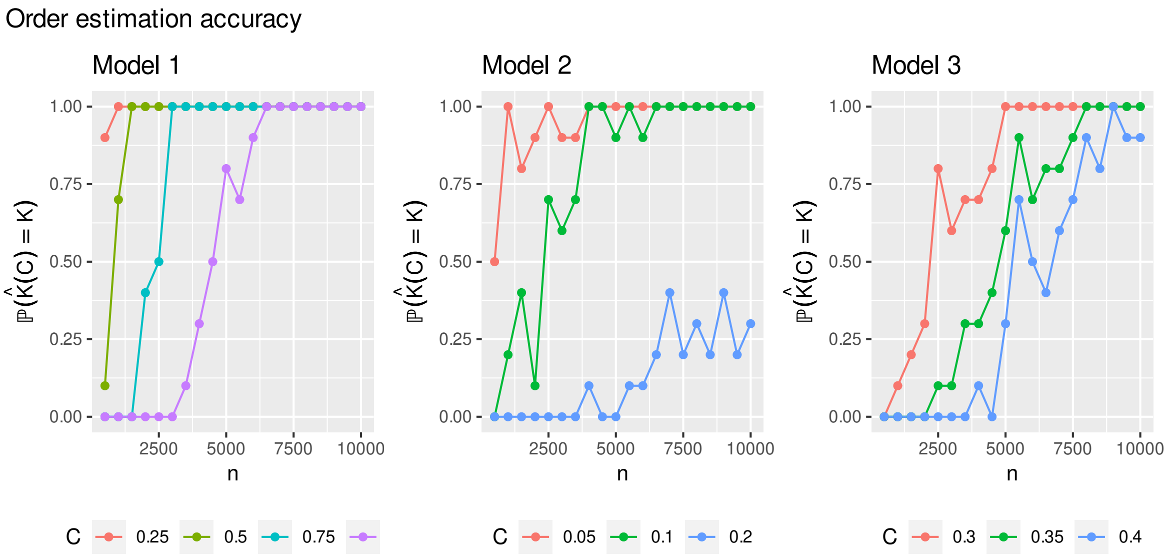


Figure 3.4: The accuracy in estimating the HMM order using the order estimator suggested in Theorem 3.2, for varying values of C and an increasing number of samples n.

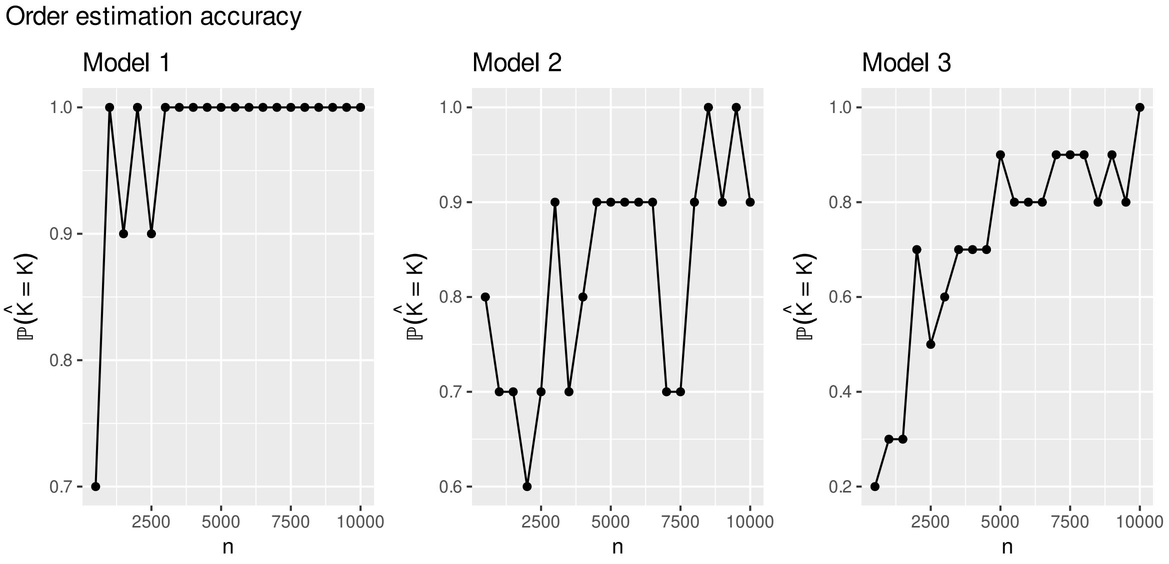


Figure 3.5: The accuracy in estimating the HMM order using the data-driven order estimator proposed in Section 3.5, for an increasing number of samples n.

3.6.5 MNIST HMM

In this section we apply our method to the synthetic MNIST HMM datasets defined in Models 15 to 17 for a varying number of observations. The observations are 28×28 pixel images, and hence the observation space is high-dimensional: $\mathscr{Y} = \mathbb{R}^{784}$. We place a Gaussian kernel function on the observation space, and define the bandwidth to be $\gamma := \alpha \gamma_{\text{med}}$ where γ_{med} denotes the median heuristic, and α is a parameter to be tuned. For each dataset (a combination of $K \in \{2,3,4\}$ and $n \in \{5000, 10000, 20000\}$) we estimate the parameters of a nonparametric HMM following Algorithm 1, for $\alpha = 1, 2, \dots, 10$. As the simultaneous diagonalization can be sensitive to perturbations when K is large, we fit the model 5 times. To choose which model to keep, we use a validation dataset of 1000 observations, and consider two possible objectives to minimize over the validation set, one which retains the unsupervised nature of the problem, and one which assumes access to labelled data for the validation set.

In the unsupervised approach, we assume that we only have access to observations. We use the estimated model to iteratively predict the observation at time $t+1$ given observations 1 to t for $t = 1, \dots, 999$, and measure performance over the validation set by the mean squared error obtained in the prediction task. We then keep the model which minimizes the MSE and evaluate its performance in the filtering task over a test set of 1000 observations. While performance in the prediction task and filtering tasks are correlated, an increase in performance in prediction may not necessarily correspond to an increase in performance in filtering. In the semi-supervised approach, we assume that the validation set has access to the hidden labels. In this case, we use the validation set to directly evaluate the model's performance in the filtering problem. As in the unsupervised scenario, we evaluate performance in the filtering problem over the test set, and present the results in Table 3.1.

Our method performs well in both the unsupervised and semi-supervised settings. In both cases, for $K = 2$ and $K = 3$ our method has an accuracy of at least 0.9 for each value of n tested. When $K = 4$ our method's performance is not as impressive, particularly for $n = 5000$, although we emphasize that in this setting a naïve estimator will achieve 0.25 accuracy whereas ours obtains 0.434 and 0.608 in the unsupervised and semi-supervised settings respectively. Note that when $K = 4$, performance decreases as n grows from $n = 10000$ to $n = 20000$ in the unsupervised setting; as described above, this is because the model is not tuned using labelled observations but rather via an auxiliary problem (the observation prediction problem). In the semi-supervised setting wherein the model is tuned using labelled samples, the performance over the test set consistently improves over n when $K = 4$.

We plot the expected value of $Y_t | X_t = x$ for $x \in \{2,7\}$ and $x \in \{2,4,7\}$ in Figures 3.6 and 3.7, in addition to several observations. This represents the 'average' observed image for a given hidden state captured by the estimated kernel mean embeddings.

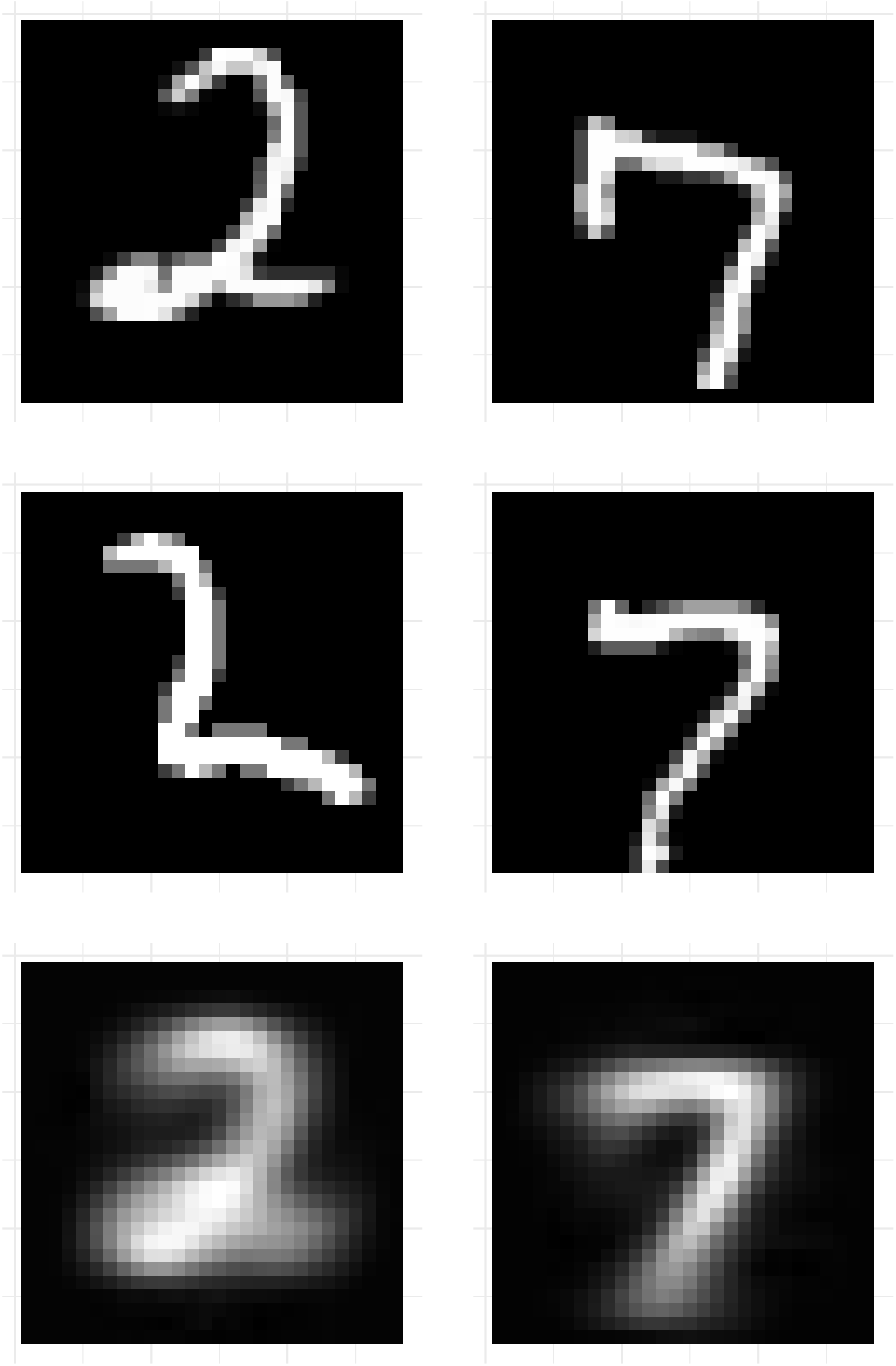


Figure 3.6: The first two rows contain samples of Y_t given $X_t = x$ for $x = 2,7$. The final row shows the expected value of $Y_2|X_2 = x$, for $x \in \{2,7\}$, obtained from the estimated embedding of the observation distribution estimated by applying Algorithm 1 to 20000 observations from the MNIST HMM data generated from Model 15.

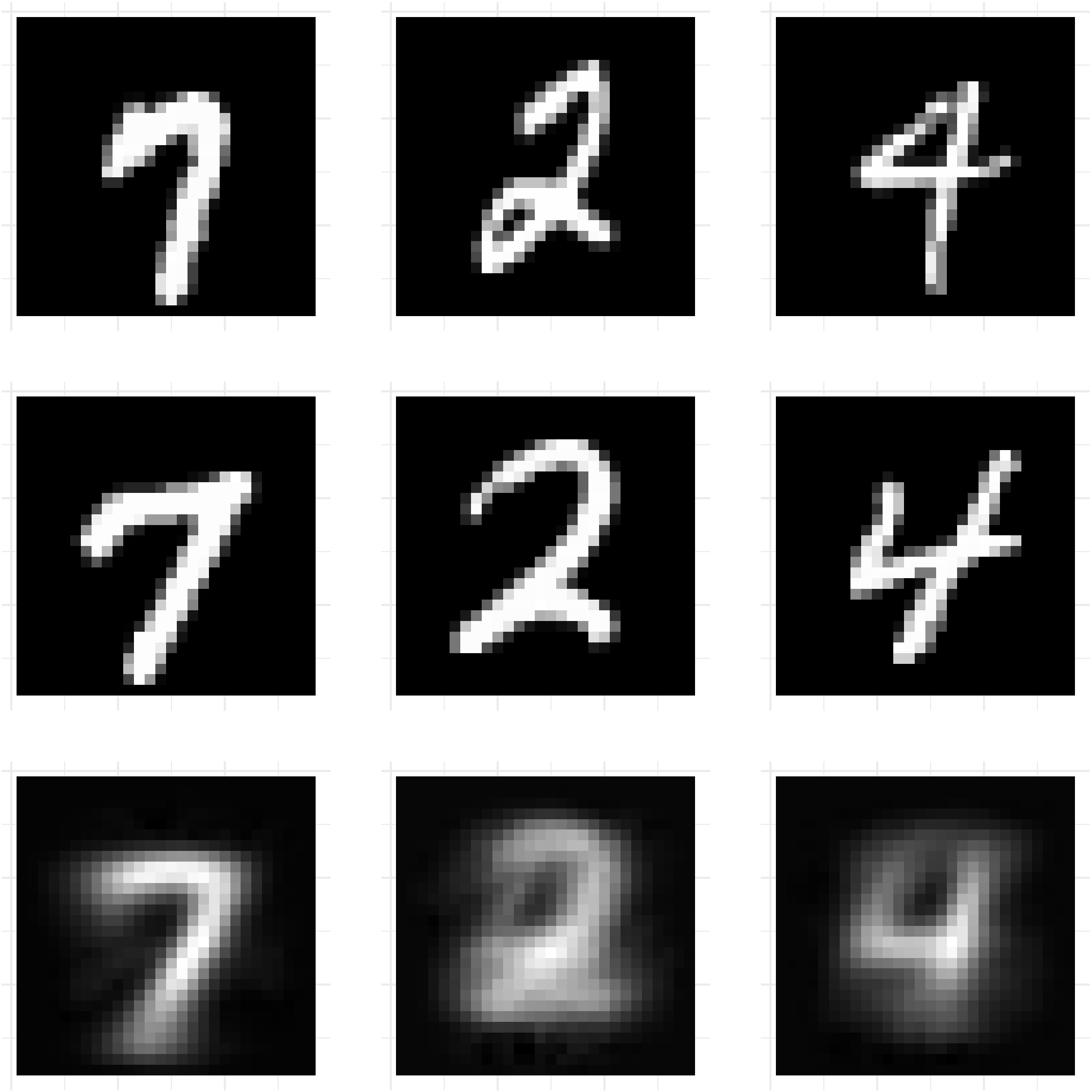


Figure 3.7: The first two rows contain samples of Y_t given $X_t = x$ for $x = 2,4,7$. The final row shows the expected value of $Y_2|X_2 = x$, for $x \in \{2,4,7\}$, obtained from the estimated embedding of the observation distribution estimated by applying Algorithm 1 to 20000 observations from the MNIST HMM data generated from Model 16.

	Unsupervised			**Semi-supervised**		
K / n	5000	10000	20000	5000	10000	20000
2	0.942	0.967	0.960	0.945	0.969	0.961
3	0.910	0.914	0.921	0.907	0.913	0.921
4	0.434	0.656	0.477	0.608	0.752	0.786

Table 3.1: Accuracy in predicting the hidden states in the filtering problem for the MNIST data sets. The unsupervised approach performs model selection by minimizing MSE in predicting observations on a validation data set, whereas the semi-supervised approach performs model selection by maximizing accuracy in predicting hidden states on a validation data set which has labelled data.

Analysis of the estimated embeddings. Our method demonstrates impressive performance when applied to the MNIST HMM dataset, as shown above. The output of Algorithm 1 is a set of K kernel mean embeddings, characterized by a set of K weight vectors which assign a weight to each observation in the training data. We analyze the estimated embeddings for the models obtained when $K = 2$ and $K = 3$ with $n = 20000$, by studying the weight vectors and using the visual nature of the MNIST HMM dataset. Our aim is to gain intuition on the information captured by the kernel mean embeddings.

Let $\hat{B}$ be the $n \times K$ matrix of estimated embedding weights, with the i-th column representing the weight vector for the estimated embedding of $Y|X = i$. We can express $\hat{B}$ as $[\hat{b}_1,\dots,\hat{b}_K]$, where $\hat{b}_i$ is the n-element column vector corresponding to the embedding of $Y|X = i$, for $i \in \{1,\dots,K\}$.

We first consider the estimated embeddings obtained for $K = 2$ and $n = 20000$. Figure 3.8 presents a colour-coded histogram of $\hat{b}_1$ and $\hat{b}_2$, where the colour corresponds to a specific label for the observed data. The figure shows that $\hat{b}_1$ assigns a positive weight to samples with label 7, and a negative weight to samples with label 2, whereas $\hat{b}_2$ does the opposite by assigning positive weight to samples with label 2 and negative weight to samples with label 7.

This example provides some intuition behind the density estimator proposed in Chapter 2. Recall that the density estimator is a linear combination of kernel functions with weights given by the estimated kernel mean embedding, and consider the density estimator with weights $\hat{b}_1$. If the density estimator is evaluated at a sample with label 2, then the kernel function will take a large value for components associated with observations with label 2, and these weights will be positive. Hence the density estimator with weights $\hat{b}_1$ will take a large value when evaluated at a sample with label 2.

Figure 3.8 highlights that the weights have a strong discriminative ability. The weights have a bimodal distribution where the modes differentiate the samples according to their unobserved labels, however several weights take values between the modes. To visualize which observations the embeddings strongly and weakly discriminate, we divide the weights $\hat{b}_1$ and $\hat{b}_2$ into 9 bins. We visualize 25 unique observations associated with weights taking values in bins 1, 5, and 9 in

Figure 3.9. Observations associated with weights taking values in bins 1 and 9 are observations the embedding is most certain of, and observations associated with weights belonging to bin 5 represent the observations that the embedding is uncertain of. Several observations belonging to the fifth bins are hard to distinguish.

We now consider the estimated embeddings obtained for $K = 3$ and $n = 20000$. A preliminary analysis of the histograms of the weight vectors $\hat{b}_1$, $\hat{b}_2$, and $\hat{b}_3$ shown in Figure 3.10 shows that this setting is more nuanced than the case where $K = 2$. As there are more than 2 classes, the embeddings cannot simply assign positive weights to one class and negative weights to another class. However, it is clear that the weights have discriminative power. Figure 3.11 shows a matrix of scatterplots for a sample of 3000 weights. The figure emphasizes that when the rows of $\hat{B}$ are treated as elements in $\mathbb{R}^3$, the weights are capable of differentiating between observations with labels 2, 4, and 7. To substantiate this claim, we divide the 20000 observations into a training set of 14000 observations and a test set of 6000 observations, and fit a k-nearest neighbour classifier to the weights. The classifier obtains an accuracy of 0.925 over the test data, and thus the embedding weights effectively differentiate between the three types of observations without access to the labels. One might consider the observations associated with weights closest to the centroids of the 3 clusters to be observations which the classifier is most certain of, and in Figure 3.12 we show the 9 unique observations which are closest to the centroids. On the other hand, there are several observations where the 3 nearest neighbours of the associated weights correspond to observations with 3 different labels, and the k-nearest neighbour classifier can only guess the label of such observations; these observations are those that the embedding weights cannot discriminate between. We visualize 25 such unique observations in Figure 3.13.

3.7 Comparison to existing methods

The method proposed by De Castro et al. [2017] is similar to ours in that it also projects observations onto another space and extends the method proposed by Anandkumar et al. [2012]. Hence, several of our results such as the concentration inequalities in Lemma 3.11 are directly comparable to those of De Castro et al. [2017]. The aforementioned method projects observations onto an M-dimensional space that is dense in L_2, and several of their results demonstrate a trade-off between sample size n and approximation space dimension M. For example, our concentration inequalities depend on the RKHS via the bound of the kernel on $\mathscr{Y}$ (which is often 1), whereas the concentration inequalities provided in De Castro et al. [2017] depend on an increasing function of M. In our experiments, Section 3.6, we found that our method outperforms the existing method with significantly less data in terms of estimating the HMM parameters and filtering accuracy.

Compared to the existing method, ours has an additional benefit in that direct estimation of the observation densities can be avoided. The alternative kernel Bayes' rule proposed in Section 3.4 allows us to perform inference with the embedded observation distributions, and

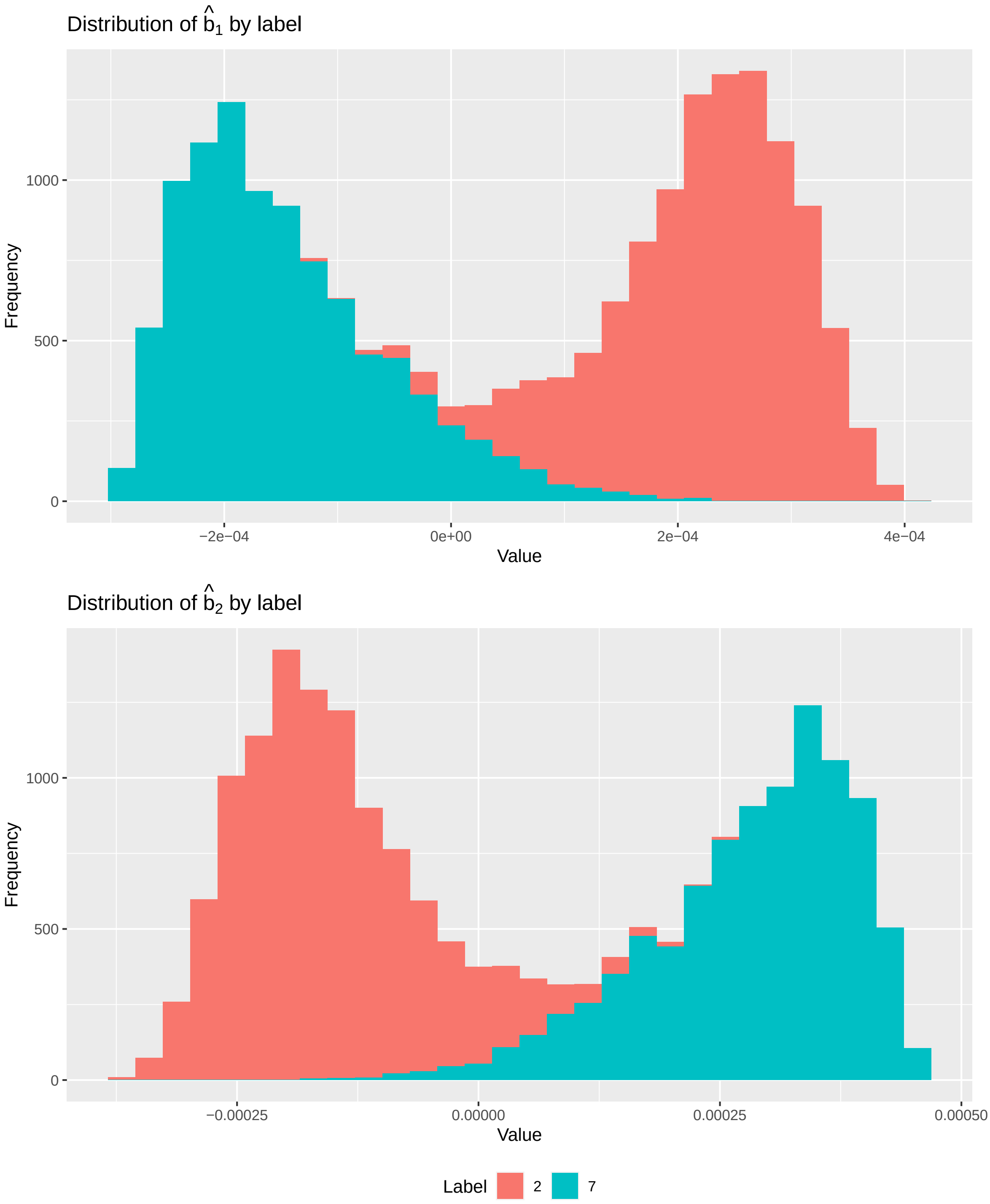
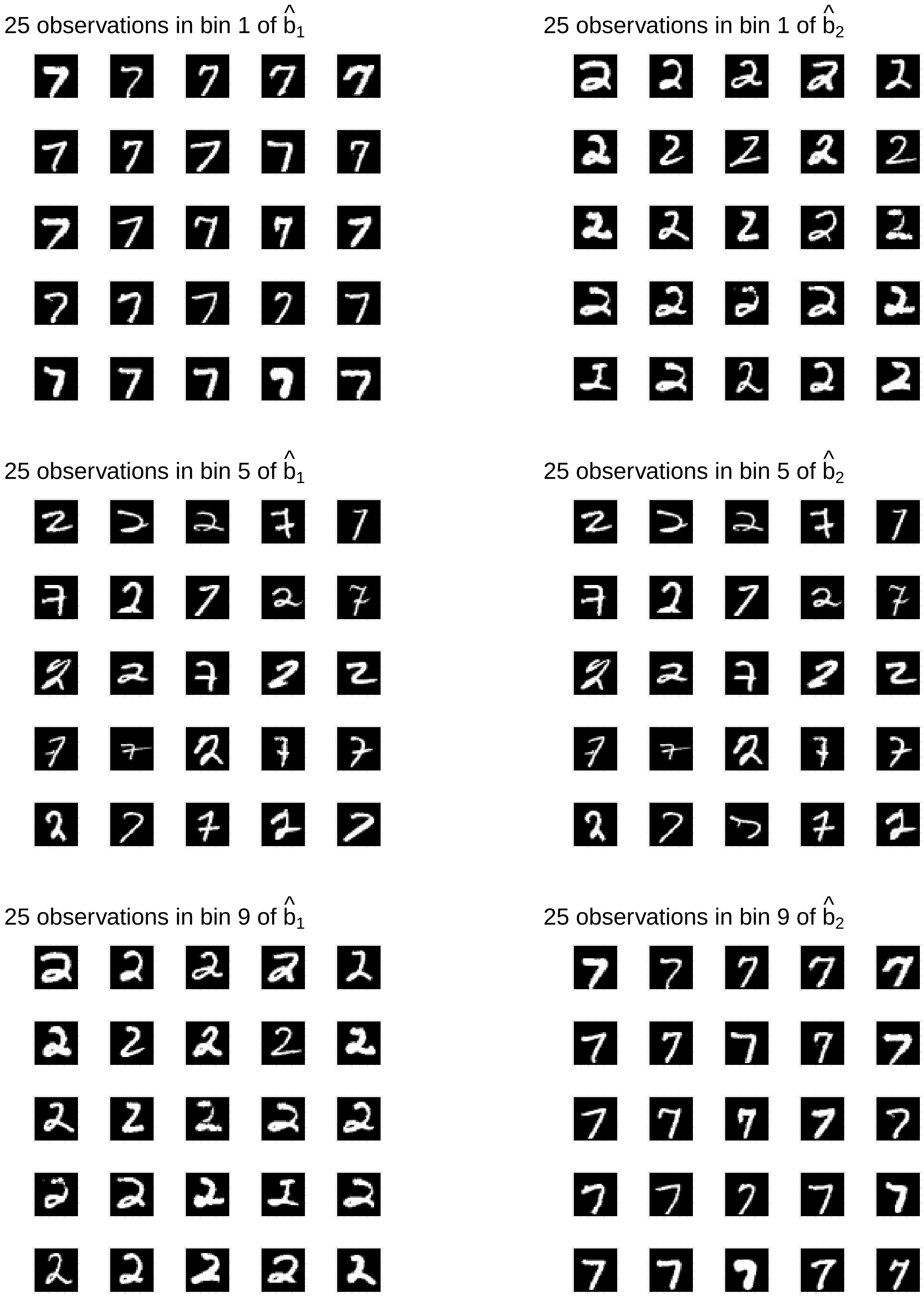


Figure 3.8: Colour-coded histograms of the embedding weights $\hat{b}_1$ and $\hat{b}_2$ obtained by applying Algorithm 1 to 20000 observations from the MNIST HMM data generated from Model 16. Each weight corresponds to an observation and the colours indicate the unobserved label associated with the observation.



25 observations in bin 1 of $\hat{b}_1$

25 observations in bin 1 of $\hat{b}_2$

25 observations in bin 5 of $\hat{b}_1$

25 observations in bin 5 of $\hat{b}_2$

25 observations in bin 9 of $\hat{b}_1$

25 observations in bin 9 of $\hat{b}_2$

Figure 3.9: 25 unique observations associated with weights in bins 1, 5, and 9 from the estimated embedding weights $\hat{b}_1$ and $\hat{b}_2$ obtained by applying Algorithm 1 to 20000 observations from the MNIST HMM data generated from Model 16.

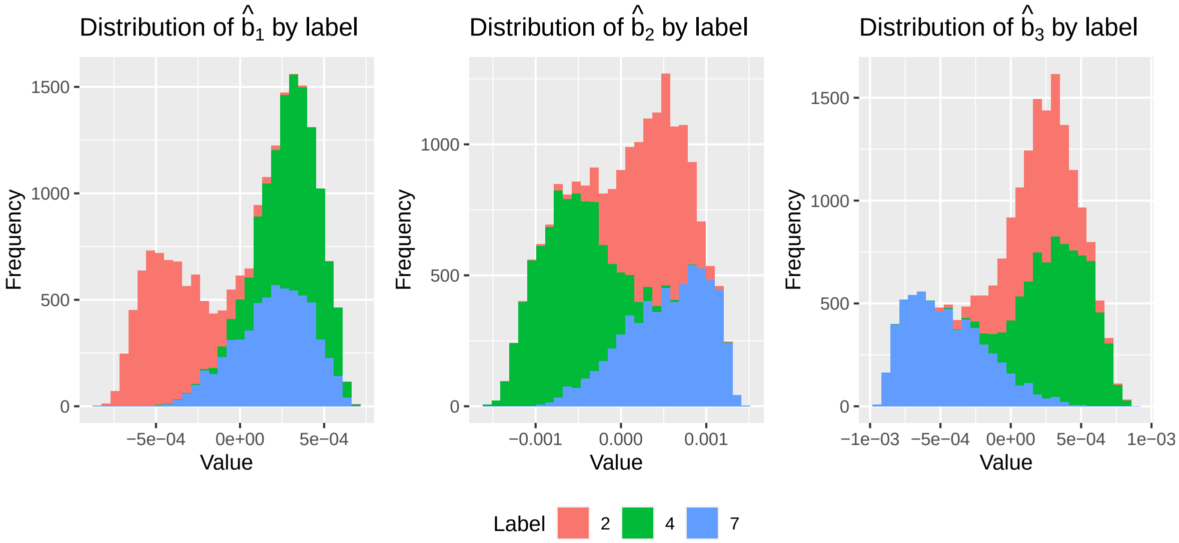


Figure 3.10: Colour-coded histograms of the embedding weights $\hat{b}_1$ and $\hat{b}_2$ obtained by applying Algorithm 1 to 20000 observations from the MNIST HMM data generated from Model 16. Each weight corresponds to an observation and the colours indicate the unobserved label associated with the observation.

we found in our experiments that this produces the best results. This is not surprising: density estimation is well-known to be a difficult task, and estimation of kernel mean embeddings is easier than estimation of the distributions or densities themselves.

Lehéricy [2019] developed an estimator of the HMM order using the HMM parameter estimators proposed by De Castro et al. [2017]. Lehéricy [2019] showed that their order estimator is dependent upon the linear independence of the observation densities when projected onto the approximation space. We showed in Section 3.5 that our order estimator depends on the linear independence of the observation distribution when embedded in the RKHS. For an appropriate choice of kernel function (such as a characteristic kernel function), the linear independence of observation distributions is preserved upon embedding the distributions into the RKHS. In Section 3.6 we saw that our order estimator successfully recovers the HMM order in a setting where the existing method fails.

An observable operator representation of HMMs was used for estimating discrete HMMs by Hsu et al. [2012] and continuous HMMs by [Song et al., 2010]. The latter work also uses kernel mean embeddings, and use m triples of three consecutive observations to learn an observable operator representation, which given a sequence of observations $y_{1:t}$ allows for the estimation of the embedding of $Y_{t+1}|y_{1:t}$. However, their theory shows that their estimate of the embedding of $Y_{t+1}|y_{1:t}$ worsens as the number of observations $y_{1:t}$ increases. The uniform consistency results of De Castro et al. [2017] show that this is not the case when parametric and nonparametric hidden Markov models are estimated. There also exists an additional kernel method which uses

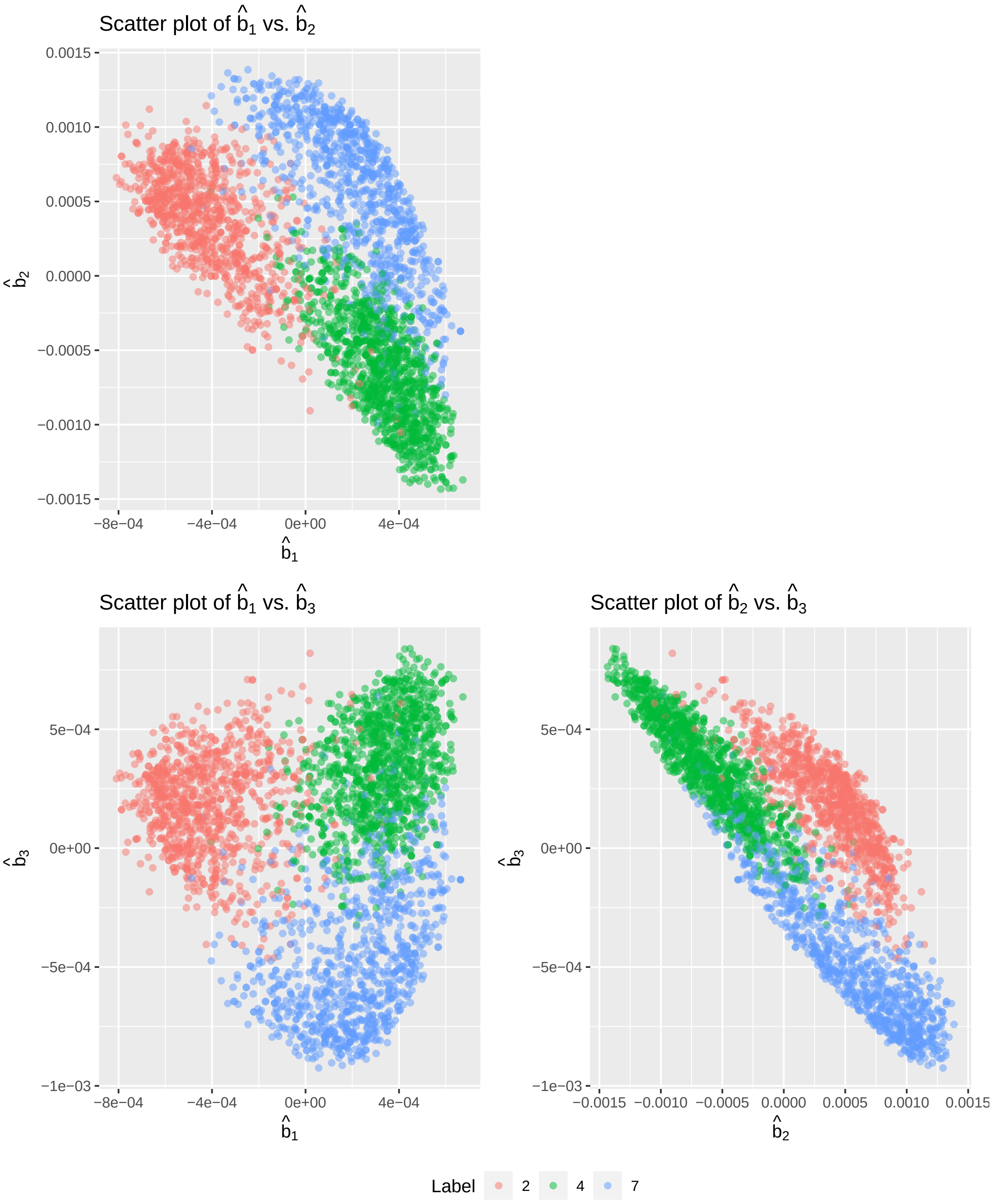


Figure 3.11: A matrix of scatterplots for the embedding weights $\hat{b}_1$, $\hat{b}_2$, and $\hat{b}_3$ obtained by applying Algorithm 1 to 20000 observations from the MNIST HMM data generated from Model 17.

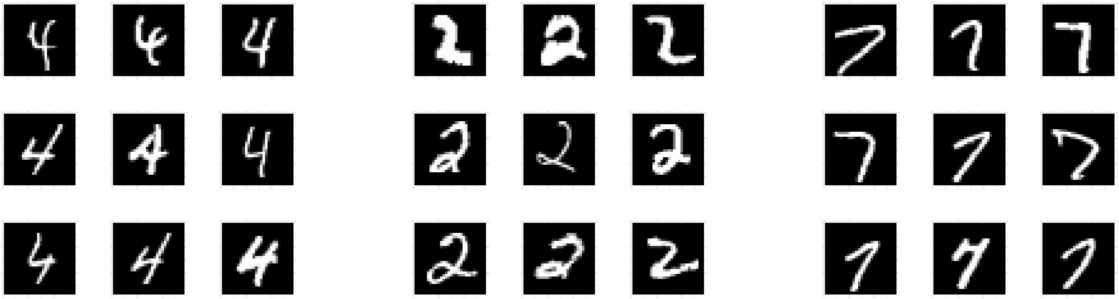


Figure 3.12: 27 unique observations that the embedding weights clearly differentiate. The observations' weights are those closest to the centroids determined by a k-nearest neighbours classifier applied to the matrix of embedding weights, obtained by applying Algorithm 1 to 20000 observations from the MNIST HMM data generated from Model 17.

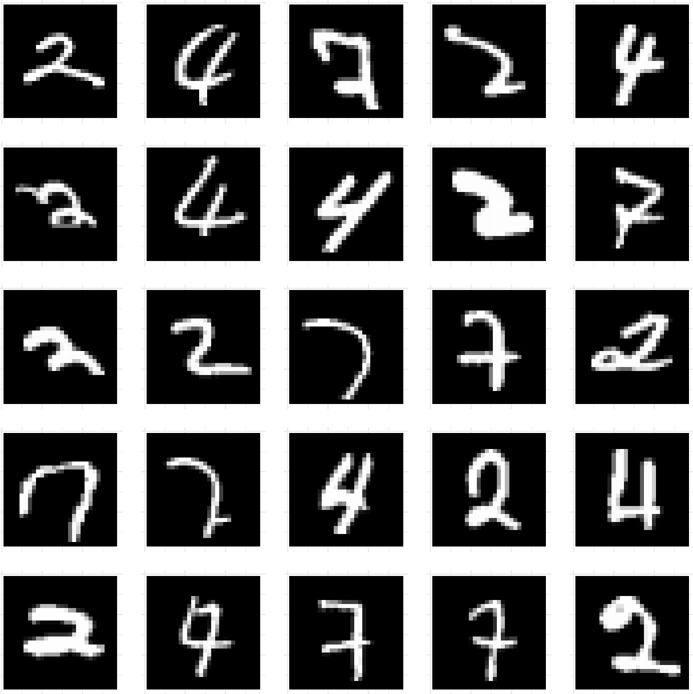


Figure 3.13: 25 unique observations that the embeddings have difficulty differentiating. For each image, the associated weights of their three nearest neighbours correspond to three distinct labels.

spectral decompositions to learn the parameters of a latent variable model, proposed by Song et al. [2014]. Our method differs from theirs in terms of the technique used to obtain the spectral decomposition: ours uses simultaneous diagonalization of several matrices, whilst the other uses the tensor power method to compute an orthogonal tensor decomposition. The differences between the two decompositions are discussed in Anandkumar et al. [2014] and Janzamin et al. [2019].

3.8 Future work

Future research could establish the consistency of the proposed method. The concentration inequalities derived in Lemma 3.11 provide the first step towards this aim, and a consistency result may be produced by adapting the consistency results of De Castro et al. [2017]. It is expected that consistency results can be produced for all estimated HMM parameters $\hat{O}_2$, $\hat{Q}$, and $\hat{\pi}$.

The alternative kernel Bayes' rule developed in this chapter is particularly interesting as it allows for applications of kernel Bayes' rule where samples are replaced by a model. The procedure is easily generalized to the setting in which the hidden states take continuous values, and hence future research could develop theoretical results for this *model-based* kernel Bayes' rule. This could enable nonparametric Bayesian inference in models with latent processes. For example, one could develop nonparametric inference procedures for state-space models where the latent process is modelled with a parametric model and the observation process is modelled nonparametrically as is done in the following chapter.

3.9 Conclusion

Hidden Markov models are popular statistical models for modelling time series, which often employ parametric assumptions that can lead to poor performance when applied to complex problems. For this reason there has been a surge of interest in nonparametric hidden Markov models, although it was not known until recently that such models are identifiable. The identifiability results of Gassiat et al. [2016] led towards a consistent method for estimating a nonparametric HMM [De Castro et al., 2017]. We have proposed a new method for estimating a nonparametric HMM using kernel mean embeddings, which is motivated by the identifiability results of Gassiat et al. [2016], and is shown to outperform the existing method of De Castro et al. [2017].

As the method we propose is closely related to the existing method, several of our results are directly comparable. Comparison of the concentration inequalities for the estimated embedded observations reveals that our method overcomes a certain trade-off present in the existing method. The concentration inequalities derived in Section 3.5 provide the first step towards proving that our method produces a consistent estimator and further work could prove this.

The estimated HMM can be used to perform inference, such as in the filtering task, and we have proposed a novel alternative kernel Bayes' rule to perform inference without density estimation. Our experiments show that this approach outperforms the alternative wherein we estimate the observation densities and use the forward algorithm. We are not aware of any other method that can perform inference in an HMM without estimating the underlying densities. Furthermore, our experiments also show that our method outperforms the existing nonparametric method with an order of magnitude less data. In particular, we found that our order estimator successfully recovers the HMM order in settings where the existing method fails.

Our experiments also studied the scenario in which the observation space is very high-dimensional; a setting where specifying a parametric model and estimating the observation density is very difficult. Our method estimates the embedded observation distributions, which are a weighted sum of kernels functions over the observations. In our experiments we analyzed the information captured by the embedding by studying the embedding weights, where we found that the weight vectors have strong discriminative power in differentiating observations based on their labels, which were unobserved during training.

To conclude, we have proposed a new nonparametric method for estimating a hidden Markov model. Our method is motivated by recent identifiability results, and we have shown that it outperforms an existing related method with far less data. Our proposal of an alternative kernel Bayes' rule allows for inference in the HMM without estimating the observation densities, which further improves performance.

Chapter 4

State-space models

4.1 Introduction

We develop a nonparametric method for learning the distributions which characterize a state-space model, whilst making minimal assumptions on the underlying process. The method we propose relies on two facts. Firstly, without making additional assumptions, the model is non-identifiable from a sequence of observations: there exist multiple state-space models which produce the same distribution over the observations. Secondly, the embedding of two consecutive observations in an RKHS can be decomposed into RKHS embeddings of the observation process and Markov transition process. The former allows us to fix an invariant distribution of our choice and sample latent states, and the latter allows us to form an optimization problem that does not require paired samples, which we use to learn embeddings of the observation distribution and Markov transition distribution. Hence, whilst the model is non-identifiable without further assumptions, we can exploit this to learn a state-space representation which adequately models our observed data.

The method we propose can be applied to problems where it may be difficult to specify a parametric model. This may be due to a lack of domain expertise, or the non-linear nature of a system; specifying a parametric model capable of modelling natural phenomena can be very difficult. There are various problems one may focus on when using state-space models, such as the filtering, smoothing, and prediction problems. As the model is non-identifiable we focus only on the prediction problem in which we predict Y_{t+1} given observations $Y_{1:t}$ for $t \geq 1$, and we do not perform inference on the learned latent process.

Many of the existing nonparametric latent variable models assume that the latent space takes finitely many values (a hidden Markov model). Generalizing to the continuous state-space poses significant difficulties, and there are few nonparametric methods in this setting. A popular class of models are Gaussian process state-space models [Frigola et al., 2014, Eleftheriadis et al., 2017], which assume that the underlying system satisfies

$$X_{t+1} = f(X_t) + \epsilon_{f,t}, \quad Y_{t+1} = g(X_{t+1}) + \epsilon_{g,t},$$

where f and g are potentially non-linear functions, and $\epsilon_{f,t}$ and $\epsilon_{g,t}$ are Gaussian noise. The functions f and g are modelled using Gaussian processes, however learning these functions is a very difficult task due to the models non-identifiability. One often has to assume that one function is an identity function, and variational inference is used to estimate the state-space model. Gassiat et al. [2020] propose two nonparametric methods for state-space models, where it is assumed that $(X_t)_{t\geq 1}$ is a stationary Markov chain and $Y_t = X_t + \epsilon_t$ where ϵ_t is i.i.d. noise with unknown distribution. They prove that in this setting nonparametric models are identifiable with respect to the distribution of the latent variables and the noise. Both of these methods assume that the relationship between the signal and noise is additive. In comparison, we assume that the underlying system is time-homogeneous and satisfies

$$X_{t+1} = f(X_t, U_t), \quad Y_{t+1} = g(X_{t+1}, V_t),$$

where U_t and V_t are white noise processes, and f and g are potentially non-linear functions. We denote by M and O the Markov transition kernel and the observation distribution, and we aim to estimate the kernel mean embeddings of M and O. Using the estimated kernel mean embeddings we can estimate their associated densities following Chapter 2, and we can then conduct inference using a particle filter.

Chapter outline. Section 4.3 develops the theory required to motivate our proposed method: Section 4.3.1 discuses the non-identifiability of the nonparametric state-space model without additional assumptions, and Section 4.3.2 shows that the kernel mean embedding of the distribution of two consecutive observations can be decomposed in terms of various quantities of interest. Section 4.4 uses our theory to propose a procedure to estimate the embedded state-space representation. Section 4.5 shows how the estimated model may be used to perform inference in the prediction problem using the density estimator proposed in Chapter 2 and a particle filter. Section 3.6 applies the proposed method to several simulated datasets. Section 4.7 discusses future avenues for research in this setting, and Section 4.8 concludes the chapter.

4.2 Problem formulation

A state-space model is characterized by the Markov transition of the hidden process M, the observation distribution O, and the invariant distribution of the hidden process π. We refer to the collection of these quantities, (O, M, π), as the state-space representation. Statistical estimation of a state-space model revolves around estimating a state-space representation using a sequence of observations. In this chapter we develop a nonparametric method to estimate an embedded state-space representation $(\mathscr{U}_{Y|X}, \mathscr{U}_{X_2|X_1}, \pi)$, where $\mathscr{U}_{Y|X}$ and $\mathscr{U}_{X_2|X_1}$ denote RKHS embeddings of O and M respectively.

One may want to perform inference using an estimated model. We propose a method of performing inference in the prediction task wherein we aim to estimate the next observation Y_{t+1} given a sequence of observations $y_{1:t}$ for $t \geq 1$.

4.3 Theory

4.3.1 Identifiability of state-space models

Throughout this chapter, we use the phrase 'state-space representation'; by this we mean a combination of the probability kernels and functions that define the state-space model: the initial distribution of the latent variables π, the Markov transition kernel M, and the observation distribution O. The state-space representation (O, M, π) fully defines the SSM.

The lemma presented below illustrates that, when provided with only a sequence of observations $(Y_t)_{t\geq 1}$, there exist several state-state representations that are indistinguishable from one another.

Lemma 4.1 (Non-uniqueness of the state-space representation). *Let $(Y_t)_{t\geq 1}$ be a sequence of random variables on $\mathscr{Y}$ with state-space representation (O, M, π) with π the invariant distribution of M. Assume that the Rosenblatt transformation of π, F_π^{-1}, exists. Then, for any $\tilde{\pi}$ such that $F_{\tilde{\pi}}^{-1}$ exists, $(Y_t)_{t\geq 1}$ also has state-space representation $(\tilde{O}, \tilde{M}, \tilde{\pi})$ for some $\tilde{M}$ with invariant distribution $\tilde{\pi}$.*

We further observe that when the initial distribution π is fixed, multiple state-space representations can still exist. Consequently, the state-space representation is non-identifiable for a given invariant distribution.

Lemma 4.2 (Non-uniqueness of the state-space representation for a given π). *Let $(Y_t)_{t\geq 1}$ be a sequence of random variables on $\mathscr{Y}$ with state-space representation (O, M, π) with π the invariant distribution of M. Assume that the Rosenblatt transformation of π, F_π^{-1}, exists. Then $(Y_t)_{t\geq 1}$ also has state-space representation $(\tilde{O}, \tilde{M}, \pi)$ for some $\tilde{M}$ with invariant distribution π, and $\tilde{O} \neq O$ and $\tilde{M} \neq M$.*

4.3.2 Decomposing RKHS embeddings

The structure of the state-space model allows us to specify a decomposition of the joint distribution of two consecutive observations in terms of the distributions $Y_t|X_t$ and $X_t|X_{t-1}$. We show that the embedding of (Y_t, Y_{t+1}) may be written in terms of RKHS operators $\mathscr{U}_{Y_t|X_t}$ and $\mathscr{U}_{X_{t+1}|X_t}$. This decomposition does not allow us to uniquely recover the RKHS operators, however non-unique operators which satisfy the decomposition must reconstruct the embedding of (Y_t, Y_{t+1}).

In a purely probabilistic setting, a similar decomposition can be derived by conditioning on the latent variable via the sum rule. In the following lemma we propose a sum rule for working with conditional mean operators.

Lemma 4.3 (Sum rule for conditional mean operators)**.** *Let X, Y, Z be integrable random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ such that Z depends on Y, Y depends on X, and Z is conditionally independent of X given Y. Let $\mathscr{H}_{\mathscr{X}}, \mathscr{H}_{\mathscr{Y}}, \mathscr{H}_{\mathcal{Z}}$ be separable RKHSs over topological spaces $\mathscr{X}$, $\mathscr{Y}$, and $\mathcal{Z}$ with bounded kernels, and assume that there exist bounded conditional mean operators $\mathscr{U}_{Z|Y} : \mathscr{H}_{\mathscr{Y}} \to \mathscr{H}_{\mathcal{Z}}$, $\mathscr{U}_{Y|X} : \mathscr{H}_{\mathscr{X}} \to \mathscr{H}_{\mathscr{Y}}$, and $\mathscr{U}_{Z|X} : \mathscr{H}_{\mathscr{X}} \to \mathscr{H}_{\mathcal{Z}}$. The conditional mean operators satisfy*

$$\mathscr{U}_{Z|X} = \mathscr{U}_{Z|Y}\mathscr{U}_{Y|X}.$$

Proof. Let φ_X, ϕ_Y, and ψ_Z denote the canonical feature maps associated with RKHSs $\mathscr{H}_{\mathscr{X}}$, $\mathscr{H}_{\mathscr{Y}}$, and $\mathscr{H}_{\mathcal{Z}}$ respectively. In the following we use the definition of the conditional mean embedding operator, that is, $\mathscr{U}_{Z|X} := \mathbb{E}_{Z|X}[\psi_Z(Z)|X]$, and the property $\mathscr{U}_{Z|Y}\phi_Y(y) = \mathbb{E}_{Z|Y}[\psi_Z(Z)|Y = y]$. Starting from this definition and applying the law of total expectation we see that

$$\begin{aligned}\mathscr{U}_{Z|X} &= \mathbb{E}_{Z|X}[\psi_Z(Z)|X]\\ &= \mathbb{E}_{Y|X}\{\mathbb{E}_{Z|Y}[\psi_Z(Z)|Y]|X\}\\ &= \mathbb{E}_{Y|X}[\mathscr{U}_{Z|Y}\phi_Y(Y)|X].\end{aligned}$$

To proceed we note that $\phi_Y(Y)$ is Bochner integrable under the assumption that

$$\mathbb{E}_Y(\|\phi_Y(Y)\|_{\mathscr{H}_{\mathscr{Y}}}) = \mathbb{E}_Y[\sqrt{k_Y(Y,Y)}] < \infty,$$

and hence it follows from Proposition 1.1 and the separability of the reproducing kernel Hilbert spaces that we can exchange the expectation and operator as follows

$$\begin{aligned}\mathscr{U}_{Z|X} &= \mathbb{E}_{Y|X}[\mathscr{U}_{Z|Y}\phi_Y(Y)|X]\\ &= \mathscr{U}_{Z|Y}\mathbb{E}_{Y|X}[\phi_Y(Y)|X]\\ &= \mathscr{U}_{Z|Y}\mathscr{U}_{Y|X},\end{aligned}$$

and the proof is complete. ∎

The following lemma describes a way to decompose the embedding of the joint distribution of (Y_1, Y_2) in terms of RKHS operators which are the embeddings of the Markov transition $X_{t+1}|X_t$ and the observation distribution $Y_t|X_t$.

Lemma 4.4. *Let $\mathscr{H}_{\mathscr{X}}$ and $\mathscr{H}_{\mathscr{Y}}$ denote RKHSs over $\mathscr{X}$ and $\mathscr{Y}$ respectively which satisfy Assumption 1.1. Let $\mu_{Y_1Y_2}$ be the kernel mean embedding of the joint distribution of (Y_1, Y_2) in the tensor* *product RKHS $\mathscr{H}_{\mathscr{Y}}^{\otimes 2}$, and $\mu_{X_1X_1}$ the embedding of the distribution of (X_1, X_1) in $\mathscr{H}_{\mathscr{X}}^{\otimes 2}$. Let $\mathscr{U}_{Y_1|X_1}$, $\mathscr{U}_{Y_2|X_2}$, and $\mathscr{U}_{X_2|X_1}$ be the bounded conditional mean embedding operators corresponding to the embeddings of the distributions of $(Y_1|X_1)$, $(Y_2|X_2)$, and $(X_2|X_1)$ respectively. We can decompose the kernel mean embedding of two consecutive observations in terms of these operators as follows*

$$\begin{aligned}\mu_{Y_1Y_2} &= (\mathscr{U}_{Y_1|X_1} \otimes \mathscr{U}_{Y_2|X_2}\mathscr{U}_{X_2|X_1})\mu_{X_1X_1}\\ &= \mathscr{U}_{Y_1|X_1}\mu_{X_1X_1}\mathscr{U}^*_{X_2|X_1}\mathscr{U}^*_{Y_2|X_2}.\end{aligned}$$

Proof. Let φ_X and ϕ_Y denote the canonical feature maps associated with RKHSs $\mathscr{H}_{\mathscr{X}}$ and $\mathscr{H}_{\mathscr{Y}}$ respectively. We decompose the kernel mean embedding of (Y_1, Y_2) by conditioning on X_1 as follows

$$\begin{aligned}\mu_{Y_1Y_2} &= \mathbb{E}_{Y_1Y_2}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)]\\ &= \mathbb{E}_{X_1}\{\mathbb{E}_{Y_1Y_2|X_1}[\phi_Y(Y_1) \otimes \phi_Y(Y_2)|X_1]\}\\ &= \mathbb{E}_{X_1}\{\mathbb{E}_{Y_1|X_1}[\phi_Y(Y_1)|X_1] \otimes \mathbb{E}_{Y_2|X_1}[\phi_Y(Y_2)|X_1]\}\\ &= \mathbb{E}_{X_1}\{\mathscr{U}_{Y_1|X_1}\varphi_X(X_1) \otimes \mathscr{U}_{Y_2|X_1}\varphi_X(X_1)\}\\ &= (\mathscr{U}_{Y_1|X_1} \otimes \mathscr{U}_{Y_2|X_1})\mathbb{E}_{X_1}[\varphi_X(X_1) \otimes \varphi_X(X_1)]\\ &= (\mathscr{U}_{Y_1|X_1} \otimes \mathscr{U}_{Y_2|X_1})\mu_{X_1X_1},\end{aligned}$$

where line three uses Lemma 1.3 and line five uses Proposition 1.1. Using the structure of a state-space model and Lemma 4.3, we can decompose the embedding of $Y_2|X_1$ as

$$\mathscr{U}_{Y_2|X_1} = \mathscr{U}_{Y_2|X_2}\mathscr{U}_{X_2|X_1},$$

which when combined with the previous decomposition gives

$$\mu_{Y_1Y_2} = (\mathscr{U}_{Y_1|X_1} \otimes \mathscr{U}_{Y_2|X_2}\mathscr{U}_{X_2|X_1})\mu_{X_1X_1}.$$

■

Remark: We can provide some intuition on the above decomposition by expanding the expectation of $\mu_{X_1X_1}$ as follows: $\mu_{Y_1Y_2} = \mathbb{E}_{X_1X_1}[\mathscr{U}_{Y_1|X_1}\varphi_X(X_1) \otimes \mathscr{U}_{Y_2|X_2}\mathscr{U}_{X_2|X_1}\varphi_X(X_1)]$. This is similar to the classical sum rule in probability, it is as if we are marginalizing out X_1.

4.4 Nonparametric estimation

Assume that we observe a sequence of observations $(Y_t)_{t=1}^T$ which we want to model with a nonparametric state-space model under minimal assumptions. We assume that the latent process is not observed and we exploit the fact that the model is non-identifiable. We specify a loss function which does not require paired observations, and motivated by the non-uniqueness of the invariant distribution (see Lemma 4.1) we sample a set of latent observations $(X_t)_{t=1}^T$ from a distribution of our choice $\pi^\star$. We then estimate the embedded Markov transition and observation distributions by minimizing a loss function motivated by Lemma 4.4.

We require that the distribution $\pi^\star$ is such that its Rosenblatt transformation exists. It is important to note that by specifying $\pi^\star$ we specify a class of state-space representations corresponding to the observed data (Lemma 4.2). The method we propose assumes that there exists a state-space representation such that the embeddings of the observation distribution and Markov transition belong to tensor product RKHSs characterized by the kernels we specify on $\mathscr{X}$ and $\mathscr{Y}$. Hence, there is an interdependence between the choice of $\pi^\star$ and the kernel functions used. The implicit assumption made in the following is that for a class of state-space representations induced by a given $\pi^\star$, there exists a state-space representation such that the RKHS embeddings of the Markov transition and observation distribution exist, and belong to a certain space determined by the kernel functions specified on $\mathscr{X}$ and $\mathscr{Y}$.

Throughout the following we use the notation specified in Section 1.2.1.2 and make Assumption 1.1 to ensure that the embeddings are well defined. We also assume that $\pi^\star$ has been specified, and that its Rosenblatt transformation exists.

4.4.1 Specifying a loss function

We first specify a loss function in terms of the RKHS operators $\mathscr{U}_{Y|X}$ and $\mathscr{U}_{X_2|X_1}$, which correspond to the RKHS embeddings of $Y_t|X_t$ and $X_{t+1}|X_t$ respectively. Our aim is to find operators which adequately model the observed data whilst remaining consistent with the structure of the state-space model. We use these requirements to motivate the following regularized loss function

$$\begin{aligned} L^{\pi^\star}(\mathscr{U}_{Y|X},\mathscr{U}_{X_2|X_1}) := {} & \|\mu_{Y_1Y_2} - \mathscr{U}_{Y|X}\mu_{XX}\mathscr{U}^*_{X_2|X_1}\mathscr{U}^*_{Y|X}\|^2_{\mathscr{H}_Y^{\otimes 2}} \\ & + \lambda_1\|\mu_Y - \mathscr{U}_{Y|X}\mu_X\|^2_{\mathscr{H}_Y} + \lambda_2\|\mu_X - \mathscr{U}_{X_2|X_1}\mu_X\|^2_{\mathscr{H}_X}, \end{aligned} \tag{4.1}$$

where λ_1 and λ_2 are positive scalars, and μ_X and μ_{XX} denote the embeddings of $\pi^\star$ in the reproducing kernel Hilbert spaces $\mathscr{H}_\mathscr{X}$ and $\mathscr{H}_\mathscr{X} \otimes \mathscr{H}_\mathscr{X}$ respectively.

The first term is motivated by Lemma 4.4 and measures the degree to which the embeddings model the observed data. In Chapter 3, motivated by recent identifiability results for nonparametric HMMs, the distribution of three consecutive observations was used. In this setting we do not have identifiability, and the invariant distribution is fixed meaning we only need to estimate the one-step observation distribution and Markov transition. For this reason we only use the distribution of two consecutive observations. One could use the distribution of more than two consecutive observations, however the additional terms would add significant complexity to the loss function; the loss landscape will become harder to optimize over and tensor algebra would be required, inducing additional computational cost. The second and third terms can be considered penalties on the consistency of the operators. The first consistency condition ensures that the operator corresponding to the observation distribution reproduces the distribution over the observations upon marginalizing over the latent states, and the second enforces the time-homogeneity of the Markov chain. They are equivalent to the following

$$p(y) = \int p(y|x)p(x)dx, \qquad p(x) = \int p(x_2|x_1)p(x_1)dx_1,$$

and follow directly from the kernel sum rule using the same intuition

$$\mu_Y = \mathbb{E}_Y[k(Y,\cdot)] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[k(Y,\cdot)|X]\right] = \mathscr{U}_{Y|X}\mathbb{E}_X\left[l(X,\cdot)\right] = \mathscr{U}_{Y|X}\mu_X.$$

Under the assumption that there exists a state-space representation such that the conditional mean operators are Hilbert-Schmidt (a sufficient condition for Assumption 1.2), the operators are minimizers of Equation (4.1). That is, let $\mathscr{U}^{\star}_{Y|X}$ and $\mathscr{U}^{\star}_{X_2|X_1}$ denote the embeddings of the observation distribution and the Markov transition, then

$$(4.2) \qquad \left(\mathscr{U}^{\star}_{Y|X}, \mathscr{U}^{\star}_{X_2|X_1}\right) \in \underset{\substack{\mathscr{U}_{Y|X}\in \mathrm{HS}(\mathscr{H}_{\mathscr{X}},\mathscr{H}_{\mathscr{Y}})\\ \mathscr{U}_{X_2|X_1}\in \mathrm{HS}(\mathscr{H}_{\mathscr{X}},\mathscr{H}_{\mathscr{X}})}}{\operatorname{argmin}} L^{\pi^\star}(\mathscr{U}_{Y|X}, \mathscr{U}_{X_2|X_1}),$$

where $\mathrm{HS}(H_1,H_2)$ denotes the space of Hilbert-Schmidt operators from H_1 to H_2, where H_1 and H_2 are Hilbert spaces.

To estimate the components of the loss function specified in Equation (4.1) we require samples of both the observable and latent process, and so to proceed we sample latent states from $\pi^\star$. Suppose we have observations $(Y_t)_{t=1}^T$, and a distribution $\pi^\star \in M_+^1(\mathscr{X})$, then we sample $(X_t)_{t=1}^T$ independently from $\pi^\star$. Given observations $(Y_t)_{t=1}^T$ and latent states $(X_t)_{t=1}^T$, not paired, the loss function given in Equation (4.1) can be computed empirically as

$$(4.3) \qquad \begin{aligned} \hat{L}^{\pi^\star}(\mathscr{U}_{Y|X}, \mathscr{U}_{X_2|X_1}) := {} & \|\hat{\mu}_{Y_1Y_2} - \mathscr{U}_{Y|X}\hat{\mu}_{XX}\mathscr{U}^*_{X_2|X_1}\mathscr{U}^*_{Y|X}\|^2_{\mathscr{H}_Y^{\otimes 2}} \\ & + \lambda_1\|\hat{\mu}_Y - \mathscr{U}_{Y|X}\hat{\mu}_X\|^2_{\mathscr{H}_Y} + \lambda_2\|\hat{\mu}_X - \mathscr{U}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathscr{H}_X}. \end{aligned}$$

4.4.2 A surrogate loss function

It is difficult to directly optimize the empirical loss function given in Equation (4.3) over the spaces of Hilbert-Schmidt operators, and thus we narrow our focus to a subspace of finite-rank operators. Given observations $(Y_t)_{t=1}^T$ and latent states $(X_t)_{t=1}^T$, not necessarily paired, we define finite-rank RKHS operators

$$(4.4) \qquad \hat{\mathscr{U}}^{W}_{Y|X} = \sum_{i,j=1}^{T} W_{ij}\left(\phi_Y(Y_i)\otimes\varphi_X(X_j)\right), \qquad \hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1} = \sum_{i,j=1}^{T} \tilde{W}_{ij}\left(\varphi_X(X_i)\otimes\varphi_X(X_j)\right),$$

where W and $\tilde{W}$ are $T \times T$ real matrices, and a superscript W and $\tilde{W}$ are used to emphasize the dependence of the estimators on W and $\tilde{W}$ respectively. The matrices W and $\tilde{W}$ can be thought of as relating the unpaired observations, and under the assumption that the weight matrices are full rank, the operators span their respective spaces. Our estimators are generalizations of those seen in the literature, for example if we observe paired data $(Y_t,X_t)_{t=1}^T$ and set $W := (K_X + T\delta I_T)^{-1}$ and $\tilde{W} := (K_X + T\delta I_T)^{-1}$, where $\delta > 0$ is a Tikhonov regularization parameter, then we recover the estimated conditional mean embedding seen in Song et al. [2009, 2013].

We define $H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{Y})$ and $H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{X})$ as being spanned by finite-rank operators of the form Equation (4.4):

$$\begin{aligned} H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{Y}) &:= \left\{ \sum_{i,j=1}^{T} w_{i,j}(\phi_Y(Y_i)\otimes\varphi_X(X_j)),\, w_{i,j}\in\mathbb{R} \right\}, \\ H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{X}) &:= \left\{ \sum_{i,j=1}^{T} \tilde{w}_{i,j}(\varphi_X(X_i)\otimes\varphi_X(X_j)),\, \tilde{w}_{i,j}\in\mathbb{R} \right\}. \end{aligned} \tag{4.5}$$

Under Assumption 1.1, $H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{Y})\subset \mathrm{HS}(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{Y})$ and $H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{X})\subset \mathrm{HS}(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{X})$, and hence we consider the related optimization problem wherein the operators are restricted to the spaces of finite-rank operators

$$\left(\left(\hat{\mathscr{U}}^W_{Y|X}\right)^\star,\left(\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}\right)^\star\right) \in \operatorname*{argmin}_{\substack{\mathscr{U}_{Y|X}\in H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{Y}) \\ \mathscr{U}_{X_2|X_1}\in H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{X})}} \hat{L}^{\pi^\star}(\mathscr{U}_{Y|X},\mathscr{U}_{X_2|X_1}). \tag{4.6}$$

For $\mathscr{U}_{Y|X}\in H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{Y})$ and $\mathscr{U}_{X_2|X_1}\in H_T(\mathscr{H}_\mathscr{X},\mathscr{H}_\mathscr{X})$, the empirical loss function in Equation (4.3) is equivalent to the following empirical surrogate loss function parametrized by $(W,\tilde{W})$

$$\begin{aligned} \hat{L}_S(W,\tilde{W}) := {} & \|\hat{\mu}_{Y_1Y_2} - \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_{XX}\left(\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}\right)^*\left(\hat{\mathscr{U}}^W_{Y|X}\right)^*\|^2_{\mathscr{H}_Y^{\otimes 2}} \\ & + \lambda_1\|\hat{\mu}_Y - \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_X\|^2_{\mathscr{H}_Y} + \lambda_2\|\hat{\mu}_X - \hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathscr{H}_X}. \end{aligned} \tag{4.7}$$

Therefore, the optimization problem specified in Equation (4.6) is equivalent to the following

$$(W^\star,\tilde{W}^\star) = \operatorname*{argmin}_{W,\tilde{W}\in\mathbb{R}^{T\times T}} \hat{L}_S(W,\tilde{W}), \tag{4.8}$$

The following lemma describes how the loss function can be computed using kernel evaluations and matrix multiplication. As seen in Section 1.2.2.1, empirical kernel mean embeddings such as $\hat{\mu}_{Y_1Y_2}$ can be computed as $\hat{\mu}_{Y_1Y_2} = (T-1)^{-1}\sum_{i=1}^{T-1}\phi_Y(Y_i)\otimes\phi_Y(Y_{i+1})$. In the following we denote by K a matrix of kernel evaluations and a subscript is used to denote the observations that produce the matrix. For example, $K_{Y_1Y_2}$ denotes the kernel matrix with (i,j)-th element $[K_{Y_1Y_2}]_{i,j} = k(Y_i,Y_{j+1})$ for $i,j\in\{1,\dots,T-1\}$. A subscript Y denotes that all T observations are used, and when one set of observations is used for both arguments only a single subscript is used, such as K_X.

Lemma 4.5 (Empirical surrogate loss). *Suppose we have data $(Y_t)_{t=1}^T$ and $(X_t)_{t=1}^T$, then let $\hat{\mu}_{Y_1Y_2}$, $\hat{\mu}_Y$ and $\hat{\mu}_X$ be empirical kernel mean embeddings, and let $\hat{\mathscr{U}}^W_{Y|X}$ and $\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}$ denote the finite-rank RKHS operators defined in Equation (4.4). The components of the empirical surrogate loss function defined in Equation (4.7) are computed as*

$$\begin{aligned} \|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathscr{H}_\mathscr{Y}^{\otimes 2}} &= \frac{1}{(T-1)^2}\mathrm{Tr}\left(K_{Y_1}K_{Y_2}\right) - 2\frac{1}{T-1}\mathrm{Tr}\left(K_{Y_1Y}\tilde{Z}K_{YY_2}\right) + \mathrm{Tr}(\tilde{Z}^\mathrm{T}K_Y\tilde{Z}K_Y), \\ \|\hat{\mu}_Y - \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_X\|^2_{\mathscr{H}_\mathscr{Y}} &= \frac{1}{T^2}1^\mathrm{T}K_Y1 - 2\frac{1}{T^2}1^\mathrm{T}K_YWK_X1 + \frac{1}{T^2}1^\mathrm{T}K_XW^\mathrm{T}K_YWK_X1, \\ \|\hat{\mu}_X - \hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathscr{H}_\mathscr{X}} &= \frac{1}{T^2}1^\mathrm{T}K_X1 - 2\frac{1}{T^2}1^\mathrm{T}K_X\tilde{W}K_X1 + \frac{1}{T^2}1^\mathrm{T}K_X\tilde{W}^\mathrm{T}K_X\tilde{W}K_X1, \end{aligned}$$

where $\tilde{Z} = \frac{1}{T}WK_XK_X\tilde{W}^\mathrm{T}K_XW^\mathrm{T}$, and 1 denotes the one-vector.

Let $W^\star$ and $\tilde{W}^\star$ be the minimizers of $\hat{L}_S(W,\tilde{W})$, then the estimated RKHS operators are

$$\hat{\mathscr{U}}^{W^\star}_{Y|X} = \Phi W^\star \Psi^{\mathrm{T}}, \qquad \hat{\mathscr{U}}^{\tilde{W}^\star}_{X_2|X_1} = \Psi \tilde{W}^\star \Psi^{\mathrm{T}},$$

where $\Phi := [\phi_Y(Y_1), \ldots, \phi_Y(Y_T)]$ and $\Psi := [\varphi_X(X_1), \ldots, \varphi_X(X_T)]$ are row-vectors in $\mathscr{H}_\mathscr{Y}$ and $\mathscr{H}_\mathscr{X}$ respectively.

4.4.3 Optimization

Minimization of the empirical surrogate loss specified in Equation (4.7) poses a complicated optimization problem. Indeed, in Section 4.3.1 we showed that upon fixing the invariant distribution π of the Markov chain, there are multiple possible state-state representations that produce the same distribution over the observed data. Thus, we proceed by optimizing the loss function numerically with respect to the matrices W and $\tilde{W}$.

We minimize the loss function using a first-order gradient-based algorithm. To minimize Equation (4.7) in such a way, we require the partial derivatives of the loss function with respect to the parameters, which are given in the following lemma.

Lemma 4.6. *Let $\hat{L}_S$ be the loss function defined in Equation* (4.7)*, parametrized by W and $\tilde{W}$. The partial derivatives of the loss are*

$$\frac{\hat{L}_S(W,\tilde{W})}{\partial W} = (A + A^{\mathrm{T}})W(B + B^{\mathrm{T}}) + \lambda_1 (C + \frac{2}{T^2} K_Y W K_X 1 1^{\mathrm{T}} K_X), \tag{4.9}$$

$$\frac{\hat{L}_S(W,\tilde{W})}{\partial \tilde{W}} = \frac{1}{T} K_X W^{\mathrm{T}} A^{\mathrm{T}} W K_X K_X + \lambda_2 [D + \frac{2}{T^2} K_X \tilde{W} K_X 1 1^{\mathrm{T}} K_X], \tag{4.10}$$

where we have defined the matrices $A := -\frac{2}{T-1} K_{YY_1} K_{Y_2 Y} + K_Y(\tilde{Z} + \tilde{Z}^{\mathrm{T}})K_Y$, $B := \frac{1}{T} K_X K_X \tilde{W}^{\mathrm{T}} K_X$, $C := -\frac{2}{T^2} K_Y 1 1^{\mathrm{T}} K_X$, and $D := -\frac{2}{T^2} K_X 1 1^{\mathrm{T}} K_X$.

To minimize the empirical surrogate loss function $\hat{L}_S$ with respect to the parameters W and $\tilde{W}$ we use coordinate descent wherein we take turns optimizing the function with respect to each parameter. When optimizing the loss with respect to a single parameter we perform several parameter updates using Adam [Kingma and Ba, 2014], a popular first-order gradient-based optimization algorithm which uses adaptive moment estimates to improve convergence. A significant advantage to using Adam is that each element of the parameter can be optimized adaptively. We refer to the steps alternating between optimizing W and $\tilde{W}$ as outer steps, and Adam steps optimizing the individual parameters as inner steps. For the Adam implementation, we follow the original paper and perform a specified number of steps with early stopping if a convergence criterion is met. For our convergence criterion we test whether the Frobenius norm (the L_2 norm) of the matrix of gradients is less than n times a tolerance hyperparameter.

Initialization. The landscape of the loss function used in the method is non-trivial, and there may exist multiple local minima depending on the difficulty of the problem. To initialize the weight matrices W and $\tilde{W}$ we assume that the sequences $(X_t)_{t=1}^T$ and $(Y_t)_{t=1}^T$ are paired, and use the empirical matrices for conditional mean operators given paired data: $W := (K_X + T\delta I_T)^{-1}$ and $\tilde{W} := (K_X + T\delta I_T)^{-1}$, where $\delta > 0$ is a Tikhonov regularization parameter. The quality of this initialization depends heavily on the initial pairing of the two sequences, and hence the initialization can be improved by first attempting to pair the observations. By initializing the matrices by first attempting to pair the data and then optimizing the loss function with respect to the matrices, we can think of the method as starting with an initial pairing and then learning a new pairing which better models the data, whilst ensuring the model is consistent with the structure implied by the state-space model. If the observed data takes values in $\mathbb{R}$, then a naïve initial pairing might involve sorting the observations and pairing observations and latent states according to their sorted order.

4.4.4 An algorithm

We now briefly summarize the above which defines an algorithm to learn an embedded state-space representation using only observations $(Y_t)_{t=1}^T$. We aim to estimate conditional mean operators corresponding to the RKHS embeddings of the observation distribution $Y_t|X_t$ and the Markov transition $X_{t+1}|X_t$, for $t \geq 1$. Motivated by the non-identifiability of the problem, we choose an invariant distribution which we sample $(X_t)_{t=1}^T$ from independently. We parametrize our estimators via $T \times T$ matrices W and $\tilde{W}$, and we minimize the empirical surrogate loss via coordinate descent with Adam updates. The minimizer of the empirical surrogate loss, $(W^\star, \tilde{W}^\star)$, corresponds to having learned RKHS embeddings of $Y_t|X_t$ and $X_{t+1}|X_t$, for $t \geq 1$.

Algorithm 3 Kernel state-space model

```
  Input: An observed sequence $(Y_t)_{t=1}^T$, regularization parameters $\lambda_1$ and $\lambda_2$, regularization parameter $\delta > 0$, invariant distribution $\pi^\star$, kernel functions $k : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$ and $l : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, optimization parameters $n_{\text{outer}}$ and $n_{\text{inner}}$.
  Output: Kernel mean embeddings of $Y_t|X_t$ and $X_t|X_{t-1}$, $\hat{\mathscr{U}}^{W}_{Y|X}$ and $\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}$ respectively.
Sample $(X_t)_{t=1}^T \overset{\text{i.i.d.}}{\sim} \pi^\star$.
Sort and pair the data $(Y_t)_{t=1}^T$ and $(X_t)_{t=1}^T$.
Initialize $W = (K_X + T\delta I_T)^{-1}$ and $\tilde{W} = (K_X + T\delta I_T)^{-1}$.
for $i = 1$ to $n_{\text{outer}}$ do
    Perform $n_{\text{inner}}$ Adam steps on $W$, using gradient function Equation (4.9).
    Update matrix $W$.
    Perform $n_{\text{inner}}$ Adam steps on $\tilde{W}$, using gradient function Equation (4.10).
    Update matrix $\tilde{W}$.
end for
Output $(W^\star, \tilde{W}^\star) = (W, \tilde{W})$.
```

4.4.5 Density recovery

The conditional mean operators $\hat{\mathscr{U}}^{W}_{Y|X}$ and $\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}$ can be thought of as defining a path of embeddings in their respective reproducing kernel Hilbert spaces: for any $x \in \mathscr{X}$ we can obtain the embedding of $Y_t|X_t = x$ and $X_{t+1}|X_t = x$ as follows

$$\hat{\mu}_{Y_t|X_t=x} = \hat{\mathscr{U}}^{W}_{Y|X}\varphi_X(x) = \Phi W \Psi^{\mathrm{T}}\varphi_X(x), \quad \hat{\mu}_{X_{t+1}|X_t=x} = \hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}\varphi_X(x) = \Psi\tilde{W}\Psi^{\mathrm{T}}\varphi_X(x).$$

Note that $\Psi^{\mathrm{T}}\varphi_X(x) = l_x = [l(X_1,x),\ldots,l(X_T,x)]$, and thus the embeddings $\hat{\mu}_{Y_t|X_t=x}$ and $\hat{\mu}_{X_{t+1}|X_t=x}$ are weighted sums of kernels on $\mathscr{Y}$, with respective weight vectors $w(x) = Wl_x$ and $\tilde{w}(x) = \tilde{W}l_x$. Following Chapter 2, the density functions can then be estimated as

$$\hat{g}(y|x) = \sum_{i=1}^{T} w_i(x)\bar{k}^{\gamma_Y}(Y_i,y), \quad \hat{f}(x'|x) = \sum_{i=1}^{T} \tilde{w}_i(x)\bar{l}^{\gamma_X}(X_i,x'),$$

for $y \in \mathscr{Y}$ and $x,x' \in \mathscr{X}$, where $\bar{l}$ and $\bar{k}$ denote the normalized kernel functions with hyperparameters $\gamma_Y,\gamma_X > 0$.

4.5 Inference: prediction

As the latent states are not observable, and the model is non-identifiable, we are primarily interested in the prediction problem. The aim of the prediction problem is to predict Y_{t+1} given observations $Y_{1:t}$, for $t \geq 1$. To make predictions we use the estimated state-space representation learned in Section 4.4 and a particle filter. The particle filter requires the density of the observation process and a way to sample from the Markov transition. The former can be obtained from the estimated embedding of the observation distribution as shown in Section 4.4.5, and the latter can be done via rejection sampling as shown below.

Rejection sampling from the Markov transition. As shown in Section 4.4.5, the density corresponding to the Markov transition M can be recovered as $\hat{f}(x'|x) = \sum_{i=1}^{T}\tilde{w}_i(x)\bar{l}^{\gamma_X}(X_i,x')$, for $x,x' \in \mathscr{X}$. This density estimate is a linear combination of probability densities, and thus if the weights are all positive, then we can easily sample from the corresponding distribution. However, the weights obtained from the proposed algorithm can be negative, and thus we propose a rejection sampling scheme in the following lemma.

Lemma 4.7. *Assume that $\rho(x) = \sum_{i=1}^{T}\tilde{w}_i\bar{l}^{\gamma_X}(X_i,x)$ is a probability density function for $x \in \mathscr{X}$ and $\tilde{w}_i \in \mathbb{R}$ for $i = 1,\ldots,T$, with distribution P. Then one can draw samples from P by drawing samples from the proposal $q(x) := M^{-1}\sum_{i=1}^{T}\tilde{w}_i^{+}\bar{l}^{\gamma_X}(X_i,x)$ for $x \in \mathscr{X}$, where $\tilde{w}_i^{+} = \max(0,\tilde{w}_i)$ and M is equal to the sum of the positive weights, and accepting samples with probability $q(x)/M\rho(x)$.*

Proof. Our proposal distribution is the normalized mixture of probability distributions corresponding to the positive weights. We require M such that $\rho(x) \leq Mq(x)$, where ρ is our target and

q is the proposal. Our target $\rho(x)$ can we written as

$$\rho(x) = \sum_{i=1}^{T} \tilde{w}_i \bar{l}^{\gamma_X}(X_i, x) = \sum_{i=1}^{T_1} \tilde{w}_i \bar{l}^{\gamma_X}(X_i, x) + \sum_{i=T_1+1}^{T} \tilde{w}_i \bar{l}^{\gamma_X}(X_i, x),$$

where $1 \leq T_1 < T$ and without loss of generality we assume that the first sum is over the non-negative weights and the second sum is over the negative weights. We write the proposal as

$$q(x) = \sum_{i=1}^{T_1} \frac{\tilde{w}_i}{\sum_{i=1}^{T_1} \tilde{w}_i} \bar{l}^{\gamma_X}(X_i, x) = W^+ \sum_{i=1}^{T_1} \tilde{w}_i \bar{l}^{\gamma_X}(X_i, x),$$

where $W^+ := (\sum_{i=1}^{T_1} \tilde{w}_i)^{-1}$. The above implies that $\rho(x) \leq \frac{1}{W^+} q(x)$, and so we choose $M = \frac{1}{W^+}$. ■

Combining the above, we use the following scheme to sample a particle from the Markov transition. We draw a sample from the proposal distribution with associated density $\rho(x)$. We then accept this sample with probability $q(x)/M\rho(x)$.

The lemma states that if the target distribution is a linear combination of probability distributions, where the mixture weights may be negative, then we can simply draw samples from the positive components and reject with a probability determined by a function evaluation. The lemma suggests a procedure to draw samples from the estimated Markov transition. In practice, it is possible that $\hat{f}$ is not a probability density function, however we use the rejection sampling procedure described by Lemma 4.7 as a heuristic. One could ensure that $\hat{f}$ is a valid density by following Section 2.2.1, although this could be computationally expensive.

4.6 Experiments

Our experiments evaluate the performance of our method over a set of simulated datasets. We also investigate the state-space representations learned during the optimization procedure, in order to provide intuition on the model-fitting procedure. Throughout the following we refer to our method as KSSM (a kernel state-space model).

Hyperparameter tuning. Our method has several hyperparameters which can be tuned, such as the regularization parameters λ_1 and λ_2, and the kernel function hyperparameters. To find the optimal set of hyperparameters we use the following procedure.

1. **Data Splitting.** We divide the observed data into three distinct sets: training, validation, and test sets. As our data is a time series we define the training set to be the first n_{train} observations, the next $n_{\text{validation}}$ to be the validation set, and the final n_{test} to be the test set.

2. **Hyperparameter gridsearch.** For every set of hyperparameters from a predefined grid:

- Use the training data and current hyperparameters to estimate the conditional mean operators via Algorithm 3.
- Estimate the observation densities and markov transition densities.
- Generate predictions over the training and validation datasets using a particle filter.
- Quantify the model's performance using the mean squared error (MSE) on the validation set.

3. **Optimal model selection.** Select the model that achieved the lowest MSE on the validation set as the optimal model.
4. **Performance evaluation.** Assess the model's predictive performance on the test dataset.

This high-level procedure offers a systematic approach to refining our model, ensuring that we select parameters that generalize well to new, unseen data.

4.6.1 Simulated datasets

We generate data from state-space models which follow the general framework outlined below. The latent process $(X_t)_{t\geq 1}$ and observation processes $(Y_t)_{t\geq 1}$ are defined as

$$\begin{aligned} X_{t+1} &= f(X_t, U_t), \\ Y_{t+1} &= g(X_{t+1}, V_t), \end{aligned}$$

where U_t and V_t are white noise processes and $f, g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. The models we consider are defined below.

Model 18 (Linear Gaussian)**.** *The latent state transition model and observation model are*

$$f(X_t, U_t) = \rho X_t + U_t, \quad g(X_{t+1}, V_t) = X_{t+1} + V_t,$$

where $U_t \sim N(0, \sigma_X^2)$ and $V_t \sim N(0, \sigma_Y^2)$ are white noise processes with $\sigma_X, \sigma_Y > 0$, and for stationarity we require $\rho \in (0,1)$. We set $\sigma_X = 0.6$, $\sigma_Y = 0.6$, and $\rho = 0.95$.

The linear Gaussian state-space model is a simple model with linear dynamics, with an additive structure between the signal and noise.

Model 19 (Bimodal)**.** *The latent state transition model and observation model are*

$$f(X_t, U_t) = \frac{1}{2} X_t + 25 \frac{X_t}{1 + X_t^2} + U_t, \quad g(X_{t+1}, V_t) = X_{t+1} + V_t,$$

where $U_t \sim N(0, \sigma_X^2)$ and $V_t \sim N(0, \sigma_Y^2)$ are white noise processes with $\sigma_X, \sigma_Y > 0$. We set $\sigma_X = \sqrt{10}$, and $\sigma_Y = 1$.

This model is inspired by Kitagawa [1987], modified to ensure stationarity. The latent state transition is highly non-linear, and the distribution over the latent states is bimodal.

Model 20 (Stochastic volatility)**.** *The latent state transition model and observation model are*

$$f(X_t, U_t) = \rho X_t + U_t, \quad g(X_{t+1}, V_t) = \left(\frac{1}{2} e^{X_{t+1}}\right)^{1/2} V_t,$$

where $U_t \sim N(0, \sigma_X^2)$ and $V_t \sim N(0, \sigma_Y^2)$ are white noise processes with $\sigma_X, \sigma_Y > 0$ and $\rho \in (0,1)$ for stationarity. We set $\sigma_X = 1, \sigma_Y = 1$, and $\rho = 0.95$.

The stochastic volatility model is of interest as the relationship between the signal and noise is multiplicative rather than additive. The model's name stems from its use in quantitative finance to model the stochastic volatility of an asset — the latent process captures the volatility's stochastic dynamics and the observable process models the asset's returns.

4.6.2 Application to simulated data

We evaluate the performance of our method when applied to data generated from Models 18 to 20, and compare the predictions made using our method followed by a particle filter to the predictions made by a particle filter using the true parametric model.

We sample 1200 observations from each model, and define the training set to be the first 800 observations, the validation set to be the next 200 observations, and the test set to be the final 200 observations. We sample latent states from a standard Gaussian distribution and as the data takes values in $\mathbb{R}$ we sort the observations and pair the observations and latent states based on their sorted order. We define a grid of hyperparameters over the surrogate loss regularization parameters λ_1 and λ_2, and the stepsize of the Adam optimizer. We set $n_{\text{outer}} = 10$ and $n_{\text{inner}} = 500$, $\delta = 0.5$, and we use the median heuristic for the kernel hyperparameters. We choose the optimal set of hyperparameters by minimizing the MSE in the prediction task over the validation data set, and then generate predictions over all observations for the optimal model. The predictions produced by the optimal model over the observed data can be seen in Figure 4.1. Our method, KSSM, followed by a particle filter is denoted `KSSM + PF` and the predictions made via the true model and a particle filter are denoted `True model + PF`. For the particle filter we use 500 particles and set the kernel hyperparameter of the density estimators to be 0.25 times the median heuristic. Note that for the stochastic volatility model we use the log-squared transform of the data, as is standard in many time-series applications modelling volatility. The predictions from our model capture the overall trend of the data for all examples. Our model is not able to predict data points towards the boundaries of the observed range, and we hypothesize that this problem may be solved by training with more data.

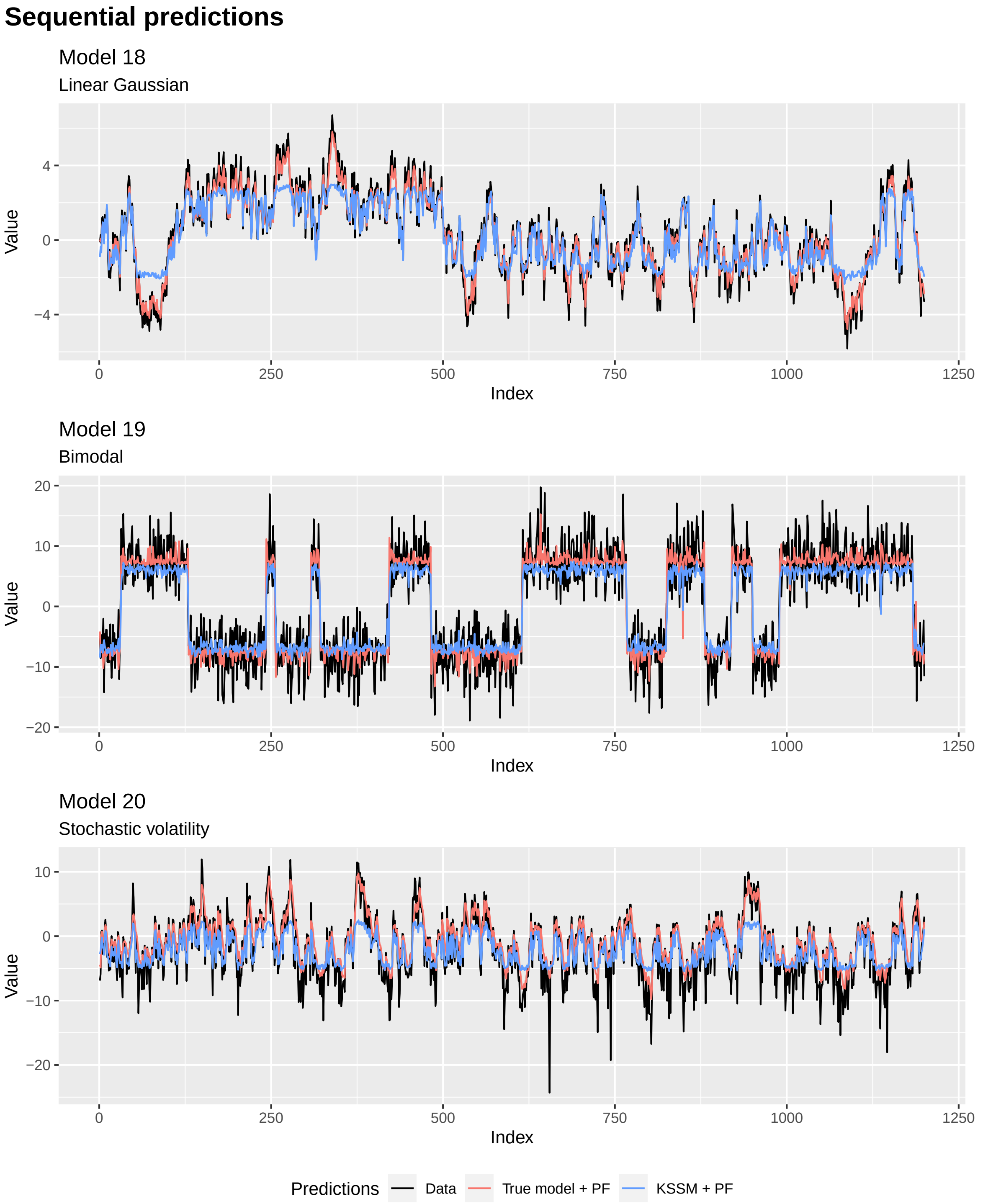


Figure 4.1: Sequential predictions for the three state-space models. The observed data (in black) is plotted alongside predictions from our proposed method `KSSM + PF`, and predictions using the true parametric model `True model + PF`. Our model was trained using the first 800 observations, and tuned using the following 200. The plots emphasize the comparative performance of our method against the benchmark of using the true model.

4.6.2.1 Analysis of the optimization procedure

As discussed in Section 4.4, one can think of the matrices $(W, \tilde{W})$ as relating pairs of observations. Our initial pairing of the latent and observed states induces a state-space representation that is represented by the initial estimated conditional mean operators, and through the optimization procedure we learn new pairings and new state-space representations that better model the observed data. In the following we investigate the optimization procedure by studying the evolution of the MSE in prediction and the value of the surrogate loss function throughout training. We also analyze the predictions made by the models learned throughout training.

For the optimal models found for the data simulated from Models 18 and 19, we re-initialize the weight matrices W and $\tilde{W}$ and optimize the empirical surrogate loss, pausing the optimization procedure every 100 inner steps. We recover the densities from the estimated intermediary embeddings, and produce predictions over the observed data using a particle filter. We record the MSE and the log objective, and continue training the model.

Figures 4.2 and 4.3 study the optimization procedure for the data generated from Models 18 and 19 respectively. The first row of plots show, from left to right: the mean squared error in the prediction task over the training, validation, and test sets, over the total number of steps taken in the optimization procedure; the log of the objective function, the empirical surrogate loss defined in Equation (4.7), over the total number of steps taken in the optimization procedure; and the log of each component of the empirical surrogate loss function over the total number of steps. The second part of the figure contains four plots showing the predictions made from the model after training for 0, 400, 5000, and 10000 total steps. In both Figures 4.2 and 4.3 we see that the predictions made prior to training are not trivial — the initial state-space representation captures some information contained within the data. After training for several hundred steps the mean squared errors increase rapidly, and the predictions become almost constant, whilst the log objective decreases rapidly. Further inspection shows that the second and third term of the log objective are the primary cause for this reduction, hinting that the initial state-space representation is not consistent with a state-space model. Beyond this point, the mean squared errors and log objective steadily decrease until they plateau. Our analysis shows that after initializing the matrices W and $\tilde{W}$, and estimating a corresponding state-space representation, the optimization procedure learns a new pairing and representation which better models the data and is consistent with a state-space model.

Figures 4.4 and 4.6 show the learned state-space representations for data generated from Models 18 and 19 respectively after taking 10000 total steps in the optimization procedure. For each plot the first row corresponds to the estimated observation densities and the second corresponds to the estimated Markov transition densities. For each row, the first five plots show estimated conditional densities for a specific value of the conditioning variable, and the right-most plot is a heatmap of the conditional density over a sequence of both variables. For example, the first row of Figure 4.4 shows estimated observation densities $\hat{g}(y|x)$ over y for

$x \in \{-1, -0.25, 0, 0.25, 1\}$, and the heatmap displays the value of $\hat{g}(y|x)$ over a sequences of y and x. In both cases the learned state-space representation captures the overall dynamics of the underlying data. Without making an assumption on the underlying data, both models have estimated the Markov transition distribution to be similar to a random walk, and the estimated observation densities are very different, capturing the nature of the underlying data and demonstrating the flexibility of our method. Figures 4.5 and 4.7 show the true state-space models for Models 18 and 19, and it's particularly interesting to compare these models to the learned representations seen in Figures 4.4 and 4.6. However, it is important to note that given the model's non-identifiability, the estimated representations are not estimates of the true model; they are estimates of a state-space representation which induces the correct distribution over the observed data.

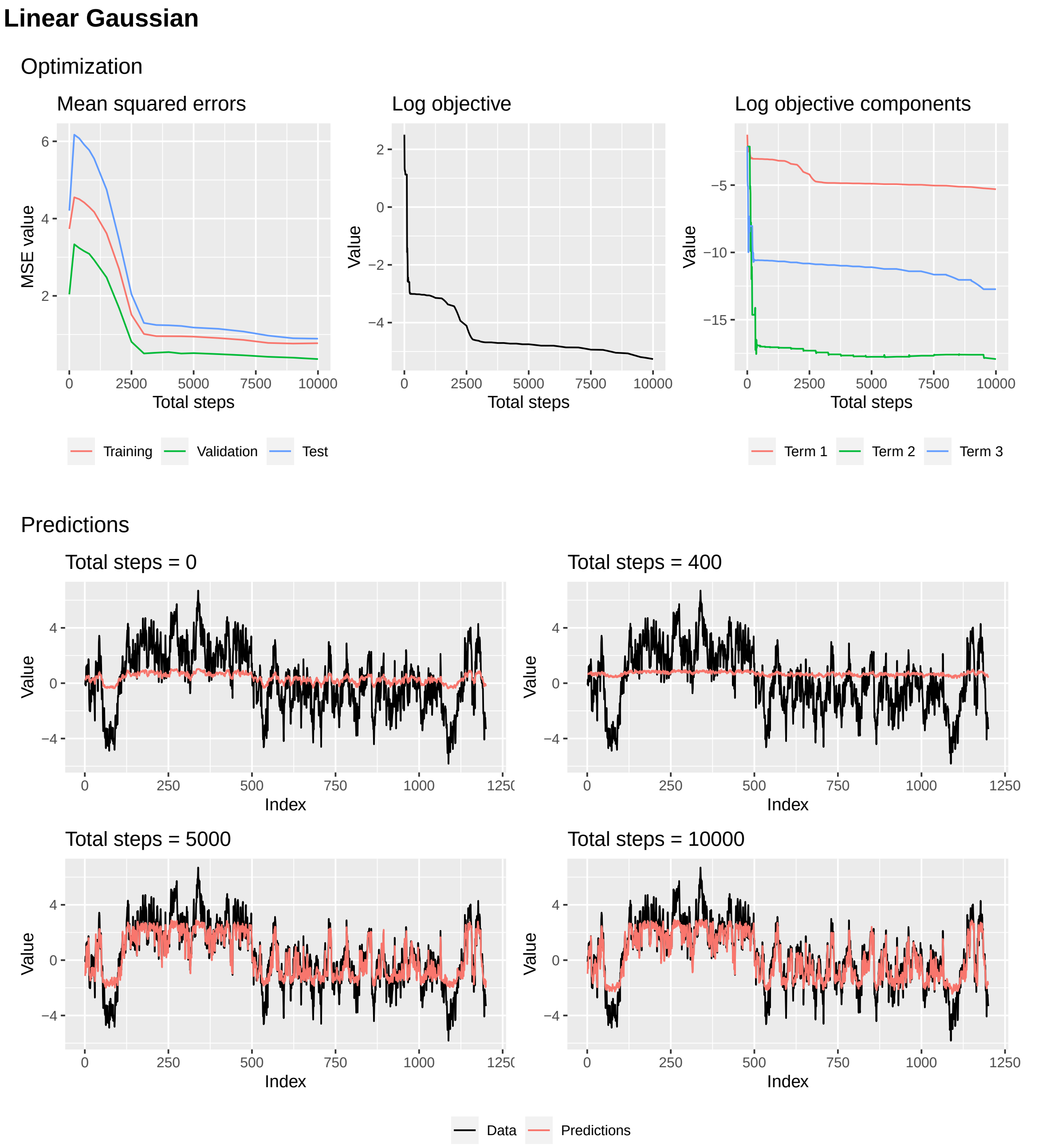


Figure 4.2: The first group of plots, the first row, shows the mean squared errors, the log objective, and the logged components of the objective function over an increasing number of total steps in the optimization procedure. The second group of plots show the observed data and the predictions made after taking a certain number of total steps in the optimization procedure.

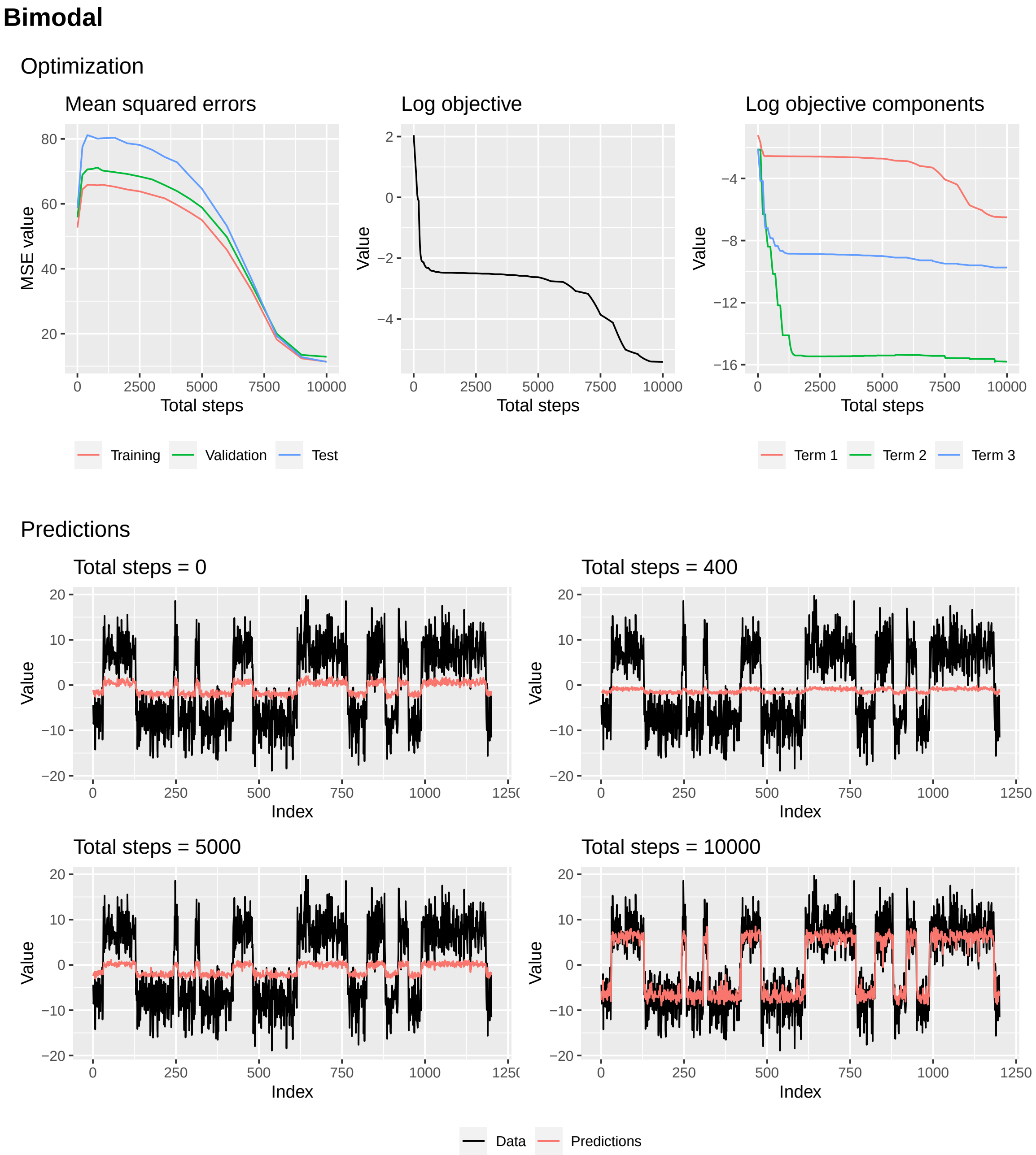


Figure 4.3: The first group of plots, the first row, shows the mean squared errors, the log objective, and the logged components of the objective function over an increasing number of total steps in the optimization procedure. The second group of plots show the observed data and the predictions made using a particle filter after taking a certain number of total steps in the optimization procedure.

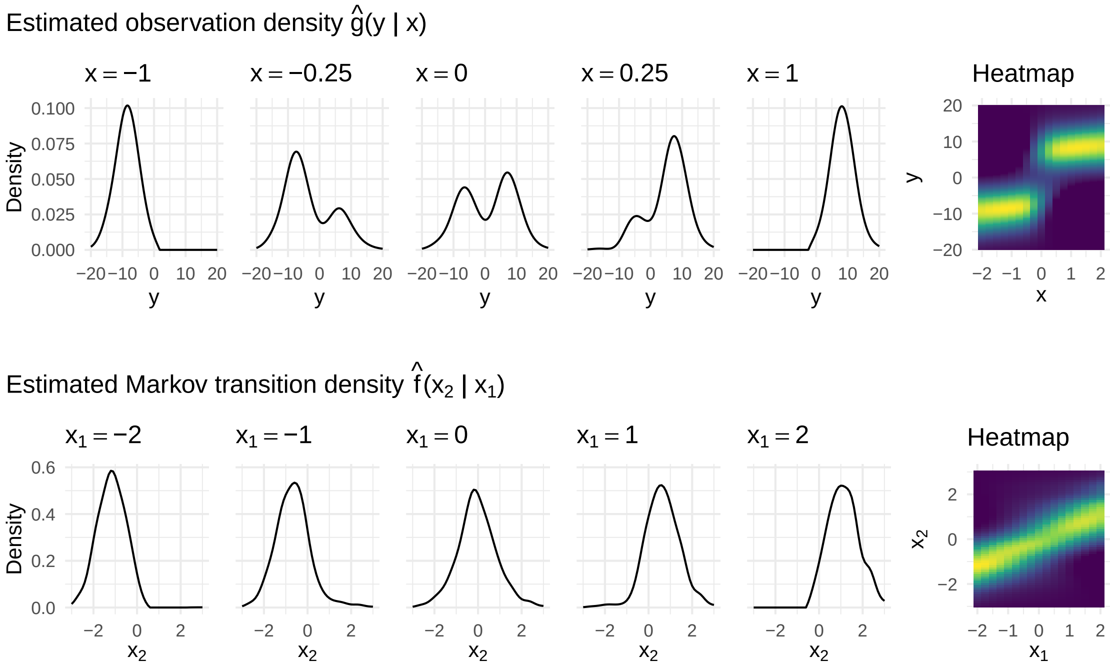


Figure 4.4: The state-space representation learned after training the kernel state-space model on the bimodal data sampled from Model 19. The first row contains the estimated observation densities $\hat{g}(y|x)$ over y for x equal to -1, -0.25, 0, 0.25, and 1, and a heatmap of $\hat{g}(y|x)$ over a sequence of y and x. The second row contains the estimated Markov transition densities $\hat{f}(x_2|x_1)$ over x_2 for x_1 equal to -2, -1, 0, 1, and 2, and a heatmap of $\hat{f}(x_2|x_1)$ over a sequence of x_2 and x_1.

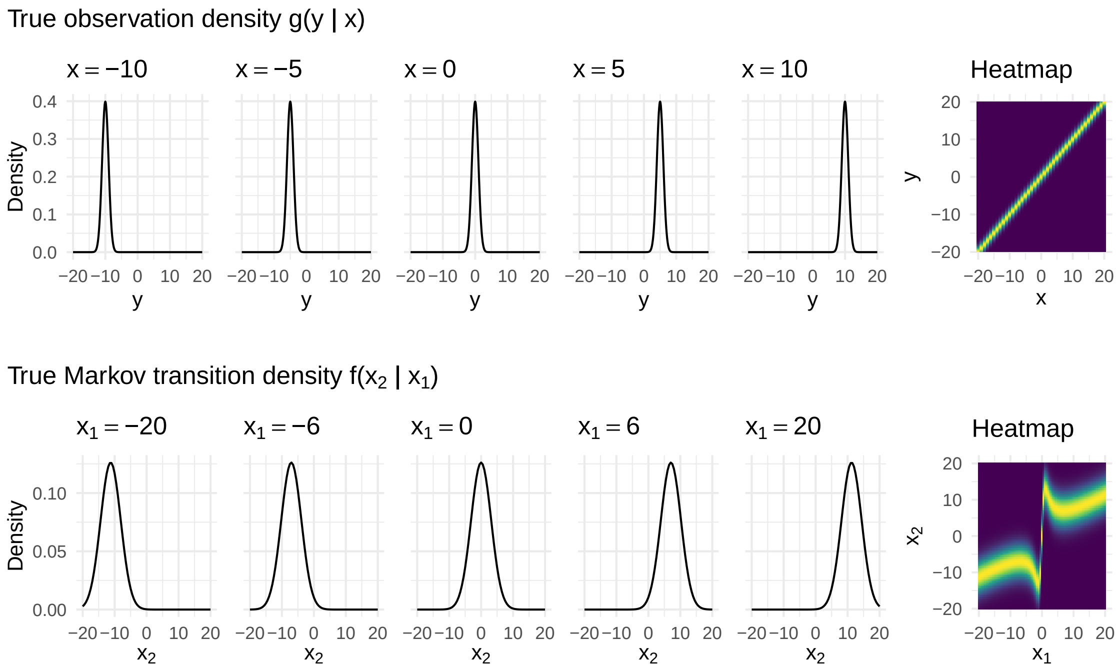


Figure 4.5: The true state-space model for Model 19. The first row contains the observation densities $\hat{g}(y|x)$ over y for x equal to -10, -5, 0, 5, and 10, and a heatmap of $g(y|x)$ over a sequence of y and x. The second row contains the Markov transition densities $f(x_2|x_1)$ over x_2 for x_1 equal to -20, -6, 0, 6, and 2, and a heatmap of $f(x_2|x_1)$ over a sequence of x_2 and x_1.

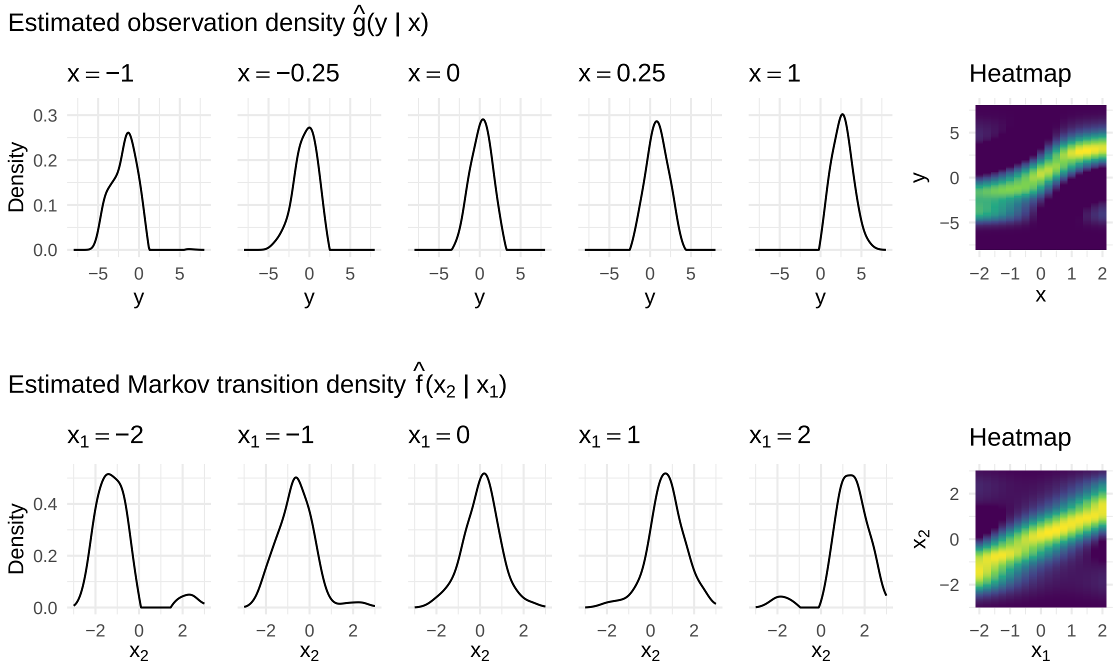


Figure 4.6: The state-space representation learned after training the kernel state-space model on the linear Gaussian data sampled from Model 18. The first row contains the estimated observation densities $\hat{g}(y|x)$ over y for x equal to -1, -0.25, 0, 0.25, and 1, and a heatmap of $\hat{g}(y|x)$ over a sequence of y and x. The second row contains the estimated Markov transition densities $\hat{f}(x_2|x_1)$ over x_2 for x_1 equal to -2, -1, 0, 1, and 2, and a heatmap of $\hat{f}(x_2|x_1)$ over a sequence of x_2 and x_1.

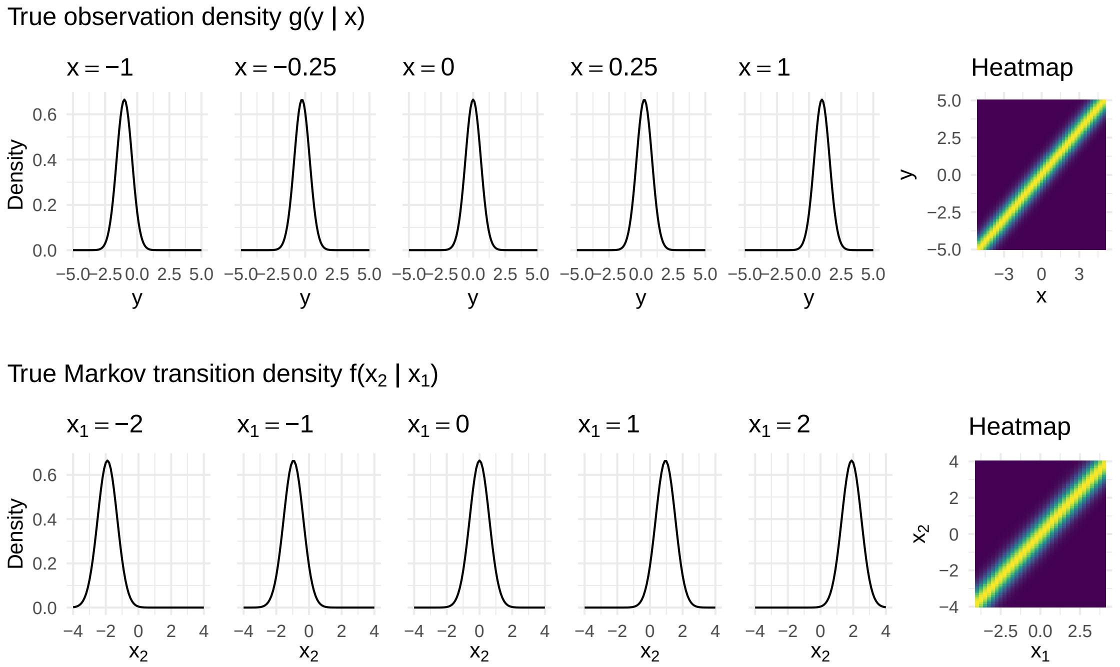


Figure 4.7: The true state-space model for Model 18. The first row contains the observation densities $g(y|x)$ over y for x equal to -1, -0.25, 0, 0.25, and 1, and a heatmap of $g(y|x)$ over a sequence of y and x. The second row contains the Markov transition densities $f(x_2|x_1)$ over x_2 for x_1 equal to -2, -1, 0, 1, and 2, and a heatmap of $f(x_2|x_1)$ over a sequence of x_2 and x_1.

4.7 Future work

In this chapter we make very general assumptions on the Markov transition and observation distribution of a state-space model, and develop a nonparametric method for their estimation. Future research could focus on incorporating additional model constraints, which may allow for identifiability of the underlying model and subsequently an estimation procedure with provable guarantees. Model constraints can be introduced by making a parametric assumption on one of the model's underlying processes; two possibilities are briefly discussed in the following.

A parametric assumption on the observation distribution was made in Gassiat et al. [2020]. The authors studied identifiability of nonparametric state-space models wherein $(X_t)_{t\geq 1}$ is a stationary Markov chain and $Y_t = X_t + \epsilon_t$ where ϵ_t is i.i.d. noise with unknown distribution. They show that under assumptions on the distribution of the latent variables and the noise, the state-space model is identifiable with respect to the distribution of the latent variables and the noise. The identifiability of such models allows for consistent estimation of the state-space representation, however incorporating such model constraints into our estimation procedure is non-trivial and could be the focus of future research.

One may also consider the scenario in which a parametric form is specified for the Markov transition, and no additional assumptions are made on the observation distribution. Our method is easily adapted to incorporate these constraints. Latent states can be sampled from the Markov process and the embedding of the Markov transition can be estimated via the conditional mean embedding. Algorithm 3 can then be simplified to solely estimate the embedded observation distribution. The additional parametric assumption simplifies the model estimation procedure, whilst the model remains more general than existing works. Notably, widely-used models with complex dynamics such as the stochastic volatility model are compatible with these assumptions. An important direction for future research would be to investigate the identifiability of such models, which may lead to a nonparametric estimation procedure with provable guarantees that is more general than existing approaches.

4.8 Conclusion

State-space models are a generalization of hidden Markov models where the latent space can take continuous values. This generalization presents significant difficulties in terms of model identification and computational inference. Even in the parametric setting, stringent assumptions are required in order to obtain analytic inference procedures, and generally one must rely upon computational techniques such as sequential Monte Carlo. For this reason there are very few existing nonparametric methods for state-space modelling, and in this work we have developed a new method for estimating state-space models using kernel mean embeddings.

The method we propose allows for the estimation of the observation distribution and Markov transition under minimal assumptions on the underlying model, using only a sample from the observable process. We showed that the model is non-identifiable and that there exist multiple state-space representations that produce the same distribution over the observed data, and that the kernel mean embedding of consecutive observations can be decomposed in terms of embeddings of the observation distribution and Markov transition. We used this to motivate a method to estimate the embedded model via the optimization of a regularized loss function. The state-space representation that we learn is not indicative of the true underlying model, however it models the data well enough that it cannot be distinguished from the true model.

Our experiments demonstrated that this proof of concept approach to modelling nonparametric state-space models works well in practice, and captures the overall dynamics of data simulated from several state-space models. An analysis of the optimization procedure demonstrated that whilst the procedure starts with a state-space representation, it learns a new representation entirely: the mean squared error rapidly increases after initialization before it steadily decreases as a new representation is learned. Future research may focus on improving the optimization procedure, and the incorporation of weak assumptions on the underlying model.

4.9 Proofs

4.9.1 Proofs for non-identifiability

We first prove a preliminary lemma which allows us to define a Markov kernel relating the two Markov processes in Lemmas 4.1 and 4.2.

Lemma 4.8. *Consider the setup of Lemma 4.1. Let $(\tilde{X}_t)_{t\geq 1}$ be a Markov chain on $\mathcal{X}$ with Markov kernel $\tilde{M}$, and define the Markov kernel $\tilde{M}'$ such that $\tilde{M}'(\tilde{x}, d\tilde{x}_{t+1}) = \tilde{M}(g^{-1}(\tilde{x}), d\tilde{x}_{t+1})$, for all $x \in \mathcal{X}$, for an invertible mapping g such that $g(X) \sim \pi(dx)$. Let $\tilde{p}$ denote distributions with respect to the state-space representation with Markov chain $(\tilde{X}_t)_{t\geq 1}$ and p distributions with respect to the original state-space representation. Then*

$$\tilde{p}(d\tilde{x}_{t+1}|y_{1:t}) = \int \tilde{M}'(x_t, d\tilde{x}_{t+1}) p(dx_t|y_{1:t}), \quad \forall t \geq 1.$$

Proof. We proceed with a proof by induction. Let the Markov kernel $\tilde{O}$ be defined as follows

$$\tilde{O}(\tilde{x}, dy) := O(g(\tilde{x}), dy), \quad \forall \tilde{x} \in \mathcal{X},$$

and note that the Markov kernel $\tilde{M}'$ relates the two Markov chains:

$$\tilde{M}(\tilde{x}, d\tilde{x}_{t+1}) = \tilde{M}'(g(\tilde{x}), d\tilde{x}_{t+1}), \quad \forall \tilde{x} \in \mathcal{X}.$$

Base case ($t = 1$):

$$\begin{aligned}\tilde{p}(d\tilde{x}_2|y_1) &= \int_{\mathscr{X}} \tilde{M}(\tilde{x}_1, d\tilde{x}_2)\tilde{p}(d\tilde{x}_1|y_1)\\ &= \int_{\mathscr{X}} \tilde{M}(\tilde{x}_1, d\tilde{x}_2)\tilde{O}(\tilde{x}_1, dy_1)\tilde{\pi}(d\tilde{x}_1)/p(dy_1)\\ &= \int_{\mathscr{X}} \tilde{M}'(g(\tilde{x}_1), d\tilde{x}_2)O(g(\tilde{x}_1), dy_1)\tilde{\pi}(d\tilde{x}_1)/p(dy_1)\\ &= \int_{\mathscr{X}} \tilde{M}'(x_1, d\tilde{x}_2)O(x_1, dy_1)\pi(dx_1)/p(dy_1)\\ &= \int_{\mathscr{X}} \tilde{M}'(x_1, d\tilde{x}_2)p(dx_1|y_1).\end{aligned}$$

Inductive hypothesis ($t = k$): We assume that the statement holds for $t = k$, that is

$$\tilde{p}(d\tilde{x}_k|y_{1:(k-1)}) = \int \tilde{M}'(x_{k-1}, d\tilde{x}_k)p(dx_{k-1}|y_{1:(k-1)}).$$

Inductive step ($t = k+1$):

$$\begin{aligned}\tilde{p}(d\tilde{x}_{k+1}|y_{1:k}) &= \int \tilde{M}(\tilde{x}_k, d\tilde{x}_{k+1})\tilde{p}(d\tilde{x}_k|y_{1:k})\\ &= \int \tilde{M}(\tilde{x}_k, d\tilde{x}_{k+1})\tilde{O}(\tilde{x}_k, dy_k)\tilde{p}(d\tilde{x}_k|y_{1:(k-1)})/\tilde{p}(dy_k|y_{1:(k-1)})\\ &= \int \tilde{M}(\tilde{x}_k, d\tilde{x}_{k+1})\tilde{O}(\tilde{x}_k, dy_k)\int \tilde{M}'(x_{k-1}, d\tilde{x}_k)p(dx_{k-1}|y_{1:(k-1)})/\tilde{p}(dy_k|y_{1:(k-1)})\\ &= \int \tilde{M}'(x_k, d\tilde{x}_{k+1})O(x_k, dy_k)p(dx_k|y_{1:(k-1)})/p(dy_k|y_{1:(k-1)})\\ &= \int \tilde{M}'(x_k, d\tilde{x}_{k+1})p(dx_k|y_{1:k}).\end{aligned}$$

Therefore the statement holds for $t = k+1$ under the assumption that it holds for $t = k$. By the Principle of Induction the statement holds for all $t \geq 1$. ■

Proof of Lemma 4.1. Let $(X_t)_{t\geq 1}$ be a Markov chain on $\mathscr{X}$ with Markov kernel M which has invariant distribution π, and let $(\tilde{X}_t)_{t\geq 1}$ be a Markov chain on $\mathscr{X}$ with a Markov kernel $\tilde{M}$. In the following we use a tilde to denote quantities associated with a state-space representation $(\tilde{O}, \tilde{M}, \tilde{\pi})$, and quantities without a tilde are associated with the original state-space representation (O, M, π). We use the Rosenblatt transformation to relate the quantities associated with the two Markov chains. Note that $F_{\tilde{\pi}}(\tilde{X}_1) \sim \mathscr{U}(0,1)^d$, and it therefore follows that $F_{\pi}^{-1}F_{\tilde{\pi}}(\tilde{X}_1) \sim \pi$.

We will prove that there exists an additional state-space representation such that under both state-space representations, the conditional distributions of the observed variables are the same, that is $\tilde{p}(dy_1) = p(dy_1)$, and $\tilde{p}(dy_t|y_{1:(t-1)}) = p(dy_t|y_{1:(t-1)})$ $\forall t \geq 2$. We proceed with a proof by induction.

Base case ($t = 1$): We define the Markov kernel $\tilde{O}$ as follows

$$\tilde{O}(\tilde{x}, dy) := O(F_{\pi}^{-1}F_{\tilde{\pi}}(\tilde{x}), dy), \quad \forall x \in \mathscr{X},$$

and therefore

$$\tilde{p}(dy_1) = \int_{\mathscr{X}} \tilde{O}(\tilde{x}, dy_1)\tilde{\pi}(d\tilde{x}) = \int_{\mathscr{X}} O(x, dy_1)\pi(dx) = p(dy_1).$$

Inductive hypothesis ($t = k$): We assume that $\tilde{p}(y_k|y_{1:(k-1)}) = p(y_k|y_{1:(k-1)})$.

Inductive step ($t = k+1$): We need to relate the two Markov chains $(X_t)_{t\geq 1}$ and $(\tilde{X}_t)_{t\geq 1}$. That is, we need to express $\tilde{p}(d\tilde{x}_{k+1}|y_{1:k})$ in terms of x_k. Following Lemma 4.8 with the invertible mapping $g := F_{\pi}^{-1}F_{\tilde{\pi}}$, we see that

$$\tilde{p}(d\tilde{x}_{t+1}|y_{1:t}) = \int \tilde{M}'(x_t, d\tilde{x}_{t+1})p(dx_t|y_{1:t}), \quad \forall t \geq 1,$$

for a Markov kernel $\tilde{M}'$ defined such that $\tilde{M}'(\tilde{x}, d\tilde{x}_{t+1}) = \tilde{M}(F_{\tilde{\pi}}^{-1}F_{\pi}(\tilde{x}), d\tilde{x}_{t+1})$, for all $\tilde{x} \in \mathscr{X}$. To relate M to $\tilde{M}'$, we let $\tilde{M}'$ be such that if $Z_x \sim \tilde{M}'(x, d\tilde{x}_{t+1})$ then $F_{\pi}^{-1}F_{\tilde{\pi}}(Z_x) \sim M(x, dx_{t+1})$. Using this, we can now prove that $\tilde{p}(dy_{k+1}|y_{1:k}) = p(dy_{k+1}|y_{1:k})$.

$$\begin{aligned}
\tilde{p}(dy_{k+1}|y_{1:k}) &= \int_{\mathscr{X}} \tilde{O}(\tilde{x}_{k+1}, dy_{k+1})\tilde{p}(d\tilde{x}_{k+1}|y_{1:k}) \\
&= \int_{\mathscr{X}} \tilde{O}(\tilde{x}_{k+1}, dy_{k+1}) \int_{\mathscr{X}} \tilde{M}'(x_k, d\tilde{x}_{k+1})p(dx_k|y_{1:k}) \\
&= \int_{\mathscr{X}} \left(\int_{\mathscr{X}} \tilde{O}(\tilde{x}_{k+1}, dy_{k+1})\tilde{M}'(x_k, d\tilde{x}_{k+1}) \right) p(dx_k|y_{1:k}) \\
&= \int_{\mathscr{X}} \left(\int_{\mathscr{X}} O(x_{k+1}, dy_{k+1})M(x_k, dx_{k+1}) \right) p(dx_k|y_{1:k}) \\
&= p(dy_{k+1}|y_{1:k}).
\end{aligned}$$

Therefore the statement holds for $t = k+1$ under the assumption that it holds for $t = k$. By the Principle of Induction the statement holds for all $t \geq 1$. The final statement to be shown is that $\tilde{\pi}$ is the invariant distribution of $\tilde{M}$. For any continuous and bounded function $\varphi : \mathscr{X} \to \mathbb{R}$ we have

$$\begin{aligned}
\int_{\mathscr{X}^2} \varphi(\tilde{x}_{t+1})\tilde{M}(\tilde{x}_t, d\tilde{x}_{t+1})\tilde{\pi}(d\tilde{x}_t) &= \int_{\mathscr{X}^2} \varphi(\tilde{x}_{t+1})\tilde{M}'(x_t, d\tilde{x}_{t+1})\pi(dx_t), \\
&= \int_{\mathscr{X}^2} \varphi(x_{t+1})M(x_t, dx_{t+1})\pi(dx_t), \\
&= \int_{\mathscr{X}} \varphi(x_{t+1})\pi(dx_{t+1}), \\
&= \int_{\mathscr{X}} \varphi(\tilde{x}_{t+1})\tilde{\pi}(d\tilde{x}_{t+1}),
\end{aligned}$$

and thus $\tilde{\pi}$ is the invariant distribution of $\tilde{M}$. ■

Proof of Lemma 4.2. Let $(X_t)_{t\geq 1}$ be a Markov chain on $\mathscr{X}$ with Markov kernel M which has invariant distribution π, and let $(\tilde{X}_t)_{t\geq 1}$ be a Markov chain on $\mathscr{X}$ with a Markov kernel $\tilde{M}$. In the following we use a tilde to denote quantities associated with a state-space representation $(\tilde{O}, \tilde{M}, \pi)$, and quantities without a tilde are associated with the original state-space representation (O, M, π).

We use the Rosenblatt transformations described above to relate the quantities associated with the two Markov chains. Note that $F_\pi(\tilde{X}_1) \sim \mathscr{U}(0,1)^d$, and define the invertible mapping $h : \mathscr{X} \to \mathscr{X}$, $h(\tilde{X}_t) := F_\pi^{-1}(1 - F_\pi(\tilde{X}_t))$.

We will prove that there exists an additional state-space representation with the same invariant distribution as the original state-space represenation, such that under both state-space representations the conditional distributions of the observed variables are the same. That is $\tilde{p}(dy_1) = p(dy_1)$, and $\tilde{p}(dy_t|y_{1:(t-1)}) = p(dy_t|y_{1:(t-1)})$ $\forall t \geq 2$. We proceed with a proof by induction.

Base case ($t = 1$): We can define the Markov kernel $\tilde{O}$ as follows

$$\tilde{O}(\tilde{x}, dy) := O(h(\tilde{x}), dy), \quad \forall \tilde{x} \in \mathscr{X},$$

and therefore

$$\tilde{p}(dy_1) = \int_{\mathscr{X}} \tilde{O}(\tilde{x}, dy_1)\pi(d\tilde{x}) = \int_{\mathscr{X}} O(h(\tilde{x}), dy)\pi(d\tilde{x}) = \int_{\mathscr{X}} O(x, dy_1)\pi(dx) = p(dy_1).$$

Inductive hypothesis ($t = k$): We assume that $\tilde{p}(y_k|y_{1:(k-1)}) = p(y_k|y_{1:(k-1)})$.

Inductive step ($t = k + 1$): We need to relate the two Markov chains $(X_t)_{t\geq 1}$ and $(\tilde{X}_t)_{t\geq 1}$. That is, we need to express $\tilde{p}(d\tilde{x}_{k+1}|y_{1:k})$ in terms of x_k. Following Lemma 4.8 with the invertible mapping $g := h$, we see that

$$\tilde{p}(d\tilde{x}_{t+1}|y_{1:t}) = \int \tilde{M}'(x_t, d\tilde{x}_{t+1})p(dx_t|y_{1:t}), \quad \forall t \geq 1,$$

for a Markov kernel $\tilde{M}'$ defined such that $\tilde{M}'(\tilde{x}, d\tilde{x}_{t+1}) = \tilde{M}(h^{-1}(\tilde{x}), d\tilde{x}_{t+1})$, for all $x \in \mathscr{X}$. To relate M to $\tilde{M}'$, we let $\tilde{M}'$ be such that if $Z_x \sim \tilde{M}'(x_t, d\tilde{x}_{t+1})$ then $h(Z_x) \sim M(x_t, dx_{t+1})$. Using this, we can now prove that $\tilde{p}(dy_{k+1}|y_{1:k}) = p(dy_{k+1}|y_{1:k})$.

$$\begin{aligned}
\tilde{p}(dy_{k+1}|y_{1:k}) &= \int_{\mathscr{X}} \tilde{O}(\tilde{x}_{k+1}, dy_{k+1})\tilde{p}(d\tilde{x}_{k+1}|y_{1:k}) \\
&= \int_{\mathscr{X}} \tilde{O}(\tilde{x}_{k+1}, dy_{k+1}) \int_{\mathscr{X}} \tilde{M}'(x_k, d\tilde{x}_{k+1})p(dx_k|y_{1:k}) \\
&= \int_{\mathscr{X}} \left(\int_{\mathscr{X}} \tilde{O}(\tilde{x}_{k+1}, dy_{k+1})\tilde{M}'(x_k, d\tilde{x}_{k+1}) \right) p(dx_k|y_{1:k}) \\
&= \int_{\mathscr{X}} \left(\int_{\mathscr{X}} O(h(\tilde{x}_{k+1}), dy_{k+1})\tilde{M}'(x_k, d\tilde{x}_{k+1}) \right) p(dx_k|y_{1:k}) \\
&= \int_{\mathscr{X}} \left(\int_{\mathscr{X}} O(x_{k+1}, dy_{k+1})M(x_k, dx_{k+1}) \right) p(dx_k|y_{1:k}) \\
&= p(dy_{k+1}|y_{1:k}).
\end{aligned}$$

Therefore the statement holds for $t = k + 1$ under the assumption that it holds for $t = k$. By the Principle of Induction the statement holds for all $t \geq 1$. The final statement to be shown is that π

is the invariant distribution of $\tilde{M}$. For any continuous and bounded function $\varphi : \mathscr{X} \to \mathbb{R}$ we have

$$\begin{aligned}\int_{\mathscr{X}^2} \varphi(\tilde{x}_{t+1})\tilde{M}(\tilde{x}_t, d\tilde{x}_{t+1})\pi(d\tilde{x}_t) &= \int_{\mathscr{X}^2} \varphi(\tilde{x}_{t+1})\tilde{M}'(x_t, d\tilde{x}_{t+1})\pi(dx_t),\\ &= \int_{\mathscr{X}^2} \varphi(x_{t+1})M(x_t, dx_{t+1})\pi(dx_t),\\ &= \int_{\mathscr{X}} \varphi(x_{t+1})\pi(dx_{t+1}),\\ &= \int_{\mathscr{X}} \varphi(\tilde{x}_{t+1})\pi(d\tilde{x}_{t+1}),\end{aligned}$$

and thus π is the invariant distribution of $\tilde{M}$. ■

4.9.2 Proofs for the proposed method

The following proof describes how the empirical surrogate loss function defined in Equation (4.7) can be computed.

Proof of Lemma 4.5. Recall that the empirical surrogate loss function is defined as

$$\begin{aligned}\hat{L}_S(W, \tilde{W}) := \|\hat{\mu}_{Y_1Y_2} - \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_{XX}\left(\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}\right)^*\left(\hat{\mathscr{U}}^W_{Y|X}\right)^*\|^2_{\mathscr{H}_Y^{\otimes 2}}\\ + \lambda_1\|\hat{\mu}_Y - \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_X\|^2_{\mathscr{H}_Y} + \lambda_2\|\hat{\mu}_X - \hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathscr{H}_X}.\end{aligned} \tag{4.11}$$

Given data $(Y_i)_{i=1}^T$ and $(X_i)_{i=1}^T$, let $Y^{(1)} := (Y_i)_{i=1}^{T-1}$ and $Y^{(2)} := (Y_i)_{i=2}^T$, then the empirical kernel mean embeddings of $\hat{\mu}_{Y_1Y_2}$, $\hat{\mu}_Y$ and $\hat{\mu}_X$ are

$$\hat{\mu}_{Y_1Y_2} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left(\phi_Y(Y_i^{(1)}) \otimes \phi_Y(Y_i^{(1)})\right), \quad \hat{\mu}_Y = \frac{1}{T}\sum_{i=1}^{T}\phi_Y(Y_i), \quad \hat{\mu}_X = \frac{1}{T}\sum_{i=1}^{T}\varphi_X(X_i).$$

The empirical surrogate loss function is formed by replacing all terms with their empirical estimates. Let $\tilde{\mu}_{Y_1Y_2} = \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_{XX}(\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1})^*(\hat{\mathscr{U}}^W_{Y|X})^*$, and define $\Phi = [\phi_Y(Y_1), \dots, \phi_Y(Y_T)]$ and $\Psi = [\varphi_X(X_1), \dots, \varphi_X(X_T)]$ to be row vectors in $\mathscr{H}_{\mathscr{Y}}$ and $\mathscr{H}_{\mathscr{X}}$ respectively. Using this notation, we have $\hat{\mathscr{U}}^W_{Y|X} = \Phi W \Psi^{\mathrm{T}}$, and $\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1} = \Psi\tilde{W}\Psi^{\mathrm{T}}$. We rewrite $\tilde{\mu}_{Y_1Y_2}$ as follows

$$\begin{aligned}\tilde{\mu}_{Y_1Y_2} &= \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_{XX}(\hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1})^*(\hat{\mathscr{U}}^W_{Y|X})^*\\ &= \Phi W \Psi^{\mathrm{T}}\Psi(\frac{1}{T}I_T)\Psi^{\mathrm{T}}\Psi\tilde{W}^{\mathrm{T}}\Psi^{\mathrm{T}}\Psi W^{\mathrm{T}}\Phi^{\mathrm{T}}\\ &= \Phi\left(\frac{1}{T}WK_XK_X\tilde{W}^{\mathrm{T}}K_XW^{\mathrm{T}}\right)\Phi^{\mathrm{T}}\\ &= \Phi\tilde{Z}\Phi^{\mathrm{T}},\end{aligned}$$

where K_X denotes the matrix of kernel evaluations over $(X_i)_{i=1}^T$, and $\tilde{Z} := \frac{1}{T}WK_XK_X\tilde{W}^{\mathrm{T}}K_XW^{\mathrm{T}}$.

The first term of the empirical surrogate loss function, $\|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathscr{H}_Y^{\otimes 2}}$, can be expressed in terms of inner products as follows

$$\begin{aligned}\|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} &= \left\langle\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}, \hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\right\rangle^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}\\ &= \left\langle\hat{\mu}_{Y_1Y_2}, \hat{\mu}_{Y_1Y_2}\right\rangle_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} - 2\left\langle\hat{\mu}_{Y_1Y_2}, \tilde{\mu}_{Y_1Y_2}\right\rangle_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} + \left\langle\tilde{\mu}_{Y_1Y_2}, \tilde{\mu}_{Y_1Y_2}\right\rangle_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}.\end{aligned}$$

We now show that the term $\left\langle \tilde{\mu}_{Y_1Y_2}, \tilde{\mu}_{Y_1Y_2} \right\rangle^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}$ can be expressed as a sum of kernel evaluations. We use the following property of the tensor product Hilbert spaces: Let $f \in H_1$ and $g \in H_2$, then

$$\begin{aligned}
\langle f \otimes g, f \otimes g \rangle_{H_1 \otimes H_2} &= \langle f, (f \otimes g) g \rangle_{H_1} \\
&= \left\langle f, f \langle g, g \rangle_{\mathscr{G}} \right\rangle_{H_1} \\
&= \langle f, f \rangle_{H_1} \langle g, g \rangle_{H_2} .
\end{aligned}$$

The term $\left\langle \tilde{\mu}_{Y_1Y_2}, \tilde{\mu}_{Y_1Y_2} \right\rangle^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}$ can be expressed

$$\begin{aligned}
\left\langle \tilde{\mu}_{Y_1Y_2}, \tilde{\mu}_{Y_1Y_2} \right\rangle^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} &= \left\langle \sum_{i,j=1}^{T} \tilde{Z}_{i,j} (\phi_Y(Y_i) \otimes \phi_Y(Y_j)), \sum_{k,l=1}^{T} \tilde{Z}_{k,l} (\phi_Y(Y_k) \otimes \phi_Y(Y_l)) \right\rangle_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} \\
&= \sum_{i,j,k,l=1}^{T} \tilde{Z}_{i,j} \tilde{Z}_{k,l} \left\langle (\phi_Y(Y_i) \otimes \phi_Y(Y_j)), (\phi_Y(Y_k) \otimes \phi_Y(Y_l)) \right\rangle_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} \\
&= \sum_{i,j,k,l=1}^{T} \tilde{Z}_{i,j} \tilde{Z}_{k,l} \left\langle \phi_Y(Y_i), (\phi_Y(Y_k) \otimes \phi_Y(Y_l)) \phi_Y(Y_j) \right\rangle_{\mathscr{H}_{\mathscr{Y}}} \\
&= \sum_{i,j,k,l=1}^{T} \tilde{Z}_{i,j} \tilde{Z}_{k,l} \left\langle \phi_Y(Y_i), \phi_Y(Y_k) \left\langle \phi_Y(Y_l), \phi_Y(Y_j) \right\rangle_{\mathscr{H}_{\mathscr{Y}}} \right\rangle_{\mathscr{H}_{\mathscr{Y}}} \\
&= \sum_{i,j,k,l=1}^{T} \tilde{Z}_{i,j} \tilde{Z}_{k,l} \left\langle \phi_Y(Y_i), \phi_Y(Y_k) \right\rangle_{\mathscr{H}_{\mathscr{Y}}} \left\langle \phi_Y(Y_l), \phi_Y(Y_j) \right\rangle_{\mathscr{H}_{\mathscr{Y}}} \\
&= \sum_{i,j,k,l=1}^{T} \tilde{Z}_{i,j} \tilde{Z}_{k,l} K(Y_i, Y_k) K(Y_l, Y_j) \\
&= \mathrm{Tr}(\tilde{Z}^{\mathrm{T}} K_Y \tilde{Z} K_Y).
\end{aligned}$$

Similarly, we have the following for $\left\langle \mu_{Y_1Y_2}, \tilde{\mu}_{Y_1Y_2} \right\rangle$ and $\left\langle \mu_{Y_1Y_2}, \mu_{Y_1Y_2} \right\rangle$

$$\begin{aligned}
\left\langle \hat{\mu}_{Y_1Y_2}, \tilde{\mu}_{Y_1Y_2} \right\rangle &= \left\langle \sum_{i=1}^{T-1} \frac{1}{T-1} (\phi_Y(Y_i^{(1)}) \otimes \phi_Y(Y_i^{(2)})), \sum_{j,k=1}^{T} \tilde{Z}_{j,k} (\phi_Y(Y_j) \otimes \phi_Y(Y_k)) \right\rangle_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} \\
&= \sum_{i=1}^{T-1} \sum_{j,k=1}^{T} \frac{1}{T-1} \tilde{Z}_{j,k} K(Y_i^{(1)}, Y_j) K(Y_k, Y_i^{(2)}) \\
&= \frac{1}{T-1} \mathrm{Tr}\left(K_{Y_1 Y} \tilde{Z} K_{Y Y_2} \right) \\
\left\langle \hat{\mu}_{Y_1Y_2}, \hat{\mu}_{Y_1Y_2} \right\rangle &= \left\langle \sum_{i=1}^{T-1} \frac{1}{T-1} (\phi_Y(Y_i^{(1)}) \otimes \phi_Y(Y_i^{(2)})), \sum_{j=1}^{T-1} \frac{1}{T-1} (\phi_Y(Y_j^{(1)}) \otimes \phi_Y(Y_j^{(2)})) \right\rangle_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} \\
&= \sum_{i,j=1}^{T-1} \frac{1}{(T-1)^2} K(Y_i^{(1)}, Y_j^{(1)}) K(Y_j^{(2)}, Y_i^{(2)}) \\
&= \frac{1}{(T-1)^2} \mathrm{Tr}\left(K_{Y_1} K_{Y_2} \right).
\end{aligned}$$

We combine the above to obtain

$$\begin{aligned}\|\hat{\mu}_{Y_1Y_2}-\tilde{\mu}_{Y_1Y_2}\|^2_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}} &= \langle \hat{\mu}_{Y_1Y_2},\hat{\mu}_{Y_1Y_2}\rangle_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}} - 2\langle \hat{\mu}_{Y_1Y_2},\tilde{\mu}_{Y_1Y_2}\rangle_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}} + \langle \tilde{\mu}_{Y_1Y_2},\tilde{\mu}_{Y_1Y_2}\rangle_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}}\\ &= \frac{1}{(T-1)^2}\operatorname{Tr}\left(K_{Y_1}K_{Y_2}\right) - 2\frac{1}{T-1}\operatorname{Tr}\left(K_{Y_1Y}\tilde{Z}K_{YY_2}\right) + \operatorname{Tr}(\tilde{Z}^{\mathrm{T}}K_Y\tilde{Z}K_Y)\end{aligned}$$

The second term of the empirical surrogate loss function, $\|\hat{\mu}_Y - \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X\|^2_{\mathcal{H}_Y}$, can be expressed as follows

$$\begin{aligned}\|\hat{\mu}_Y - \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X\|^2_{\mathcal{H}_{\mathcal{Y}}} &= \left\langle \hat{\mu}_Y - \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X, \hat{\mu}_Y - \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}\\ &= \langle \hat{\mu}_Y,\hat{\mu}_Y\rangle_{\mathcal{H}_{\mathcal{Y}}} - 2\left\langle \hat{\mu}_Y, \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}} + \left\langle \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X, \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}.\end{aligned}$$

As inner products between feature mappings simplify to evaluations of kernel functions, the individual terms can be computed as

$$\begin{aligned}\langle \hat{\mu}_Y,\hat{\mu}_Y\rangle_{\mathcal{H}_{\mathcal{Y}}} &= \frac{1}{T^2}1^{\mathrm{T}}K_Y1\\ \left\langle \hat{\mu}_Y, \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}} &= \left\langle \hat{\mu}_Y, \Phi W\Psi^{\mathrm{T}}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}\\ &= \frac{1}{T^2}1^{\mathrm{T}}K_YWK_X1\\ \left\langle \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X, \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}} &= \left\langle \Phi W\Psi^{\mathrm{T}}\hat{\mu}_X, \Phi W\Psi^{\mathrm{T}}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}\\ &= \frac{1}{T^2}1^{\mathrm{T}}K_XW^{\mathrm{T}}K_YWK_X1.\end{aligned}$$

Combining the above, the second component of the empirical surrogate loss is computed as

$$\|\hat{\mu}_Y - \hat{\mathcal{U}}^{W}_{Y|X}\hat{\mu}_X\|^2_{\mathcal{H}_{\mathcal{Y}}} = \frac{1}{T^2}1^{\mathrm{T}}K_Y1 - 2\frac{1}{T^2}1^{\mathrm{T}}K_YWK_X1 + \frac{1}{T^2}1^{\mathrm{T}}K_XW^{\mathrm{T}}K_YWK_X1.$$

The third term of the empirical surrogate loss function, $\|\hat{\mu}_X - \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathcal{H}_X}$, can be expressed as

$$\begin{aligned}\|\hat{\mu}_X - \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathcal{H}_{\mathcal{Y}}} &= \left\langle \hat{\mu}_X - \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X, \hat{\mu}_X - \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}\\ &= \langle \hat{\mu}_X,\hat{\mu}_X\rangle_{\mathcal{H}_{\mathcal{Y}}} - 2\left\langle \hat{\mu}_X, \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}} + \left\langle \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X, \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}.\end{aligned}$$

Each inner product can be computed as follows

$$\begin{aligned}\langle \hat{\mu}_X,\hat{\mu}_X\rangle_{\mathcal{H}_{\mathcal{Y}}} &= \frac{1}{T^2}1^{\mathrm{T}}K_X1\\ \left\langle \hat{\mu}_X, \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}} &= \left\langle \hat{\mu}_X, \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}\\ &= \left\langle \hat{\mu}_X, \Psi\tilde{W}\Psi^{\mathrm{T}}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}\\ &= \frac{1}{T^2}1^{\mathrm{T}}K_X\tilde{W}K_X1\\ \left\langle \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X, \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}} &= \left\langle \Psi\tilde{W}\Psi^{\mathrm{T}}\hat{\mu}_X, \Psi\tilde{W}\Psi^{\mathrm{T}}\hat{\mu}_X\right\rangle_{\mathcal{H}_{\mathcal{Y}}}\\ &= \frac{1}{T^2}1^{\mathrm{T}}K_X\tilde{W}^{\mathrm{T}}K_X\tilde{W}K_X1.\end{aligned}$$

Combining the above, the third component of the empirical surrogate loss function is computed as

$$\|\hat{\mu}_X - \hat{\mathscr{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathscr{H}_{\mathscr{Y}}} = \frac{1}{T^2}1^{\mathrm{T}}K_X 1 - 2\frac{1}{T^2}1^{\mathrm{T}}K_X\tilde{W}K_X 1 + \frac{1}{T^2}1^{\mathrm{T}}K_X\tilde{W}^{\mathrm{T}}K_X\tilde{W}K_X 1.$$

■

4.9.3 Partial derivatives of the empirical surrogate loss

In the following we derive the partial derivatives of the empirical surrogate loss function defined in Lemma 4.5.

Proof of Lemma 4.6. We first compute the partial derivative of the empirical surrogate loss function $\hat{L}_S$ with respect to $\tilde{W}$. It follows from Lemma 4.5 that the first component of the loss is given by

$$\|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} = \frac{1}{(T-1)^2}\operatorname{Tr}\left(K_{Y_1}K_{Y_2}\right) - 2\frac{1}{T-1}\operatorname{Tr}\left(K_{Y_1Y}\tilde{Z}K_{YY_2}\right) + \operatorname{Tr}(\tilde{Z}^{\mathrm{T}}K_Y\tilde{Z}K_Y),$$

where $\tilde{Z} := \frac{1}{T}WK_XK_X\widetilde{W}^{\mathrm{T}}K_XW^{\mathrm{T}}$. In the following, we write $\tilde{\mu}_{Y_1Y_2} = f(\tilde{Z}(W,\tilde{W}))$ to emphasize the dependence of $\tilde{Z}$ on W and $\tilde{W}$. We proceed with an application of the chain rule.

$$\begin{aligned}
\frac{\partial\|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial\tilde{W}_{i,j}} &= \sum_{k,l=1}^{T}\frac{\partial\|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial\tilde{Z}_{k,l}}\frac{\partial\tilde{Z}_{k,l}}{\partial\tilde{W}_{i,j}} \\
&= \sum_{k,l=1}^{T}\left[\frac{\partial\|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial\tilde{Z}}\right]_{k,l}\left[\frac{\partial\tilde{Z}}{\partial\tilde{W}_{i,j}}\right]_{k,l} \\
&= \operatorname{Tr}\left\{\left[\frac{\partial\|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial\tilde{Z}}\right]^{\mathrm{T}}\frac{\partial\tilde{Z}}{\partial\tilde{W}_{i,j}}\right\}.
\end{aligned}$$

To simplify the equations that follow, we use Einstein notation. We can write

$$\begin{aligned}
\|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}} &= \frac{1}{(T-1)^2}\operatorname{Tr}\left(K_{Y_1}K_{Y_2}\right) - 2\frac{1}{T-1}\operatorname{Tr}\left(K_{Y_1Y}\tilde{Z}K_{YY_2}\right) + \operatorname{Tr}(\tilde{Z}^{\mathrm{T}}K_Y\tilde{Z}K_Y), \\
&= \frac{1}{(T-1)^2}[K_{Y_1}]_{i,i_1}[K_{Y_2}]_{i_1,i} - \frac{2}{T-1}[K_{Y_1Y}]_{j,j_1}[\tilde{Z}]_{j_1,j_2}[K_{YY_2}]_{j_2,j} \\
&\quad + [\tilde{Z}^{\mathrm{T}}]_{k,k_1}[K_Y]_{k_1,k_2}[\tilde{Z}]_{k_2,k_3}[K_Y]_{k_3,k},
\end{aligned}$$

from which it follows that the partial derivative with respect to $\tilde{Z}_{m,n}$, for $m,n \in \{1,\ldots,n\}$, is

$$\begin{aligned}
\frac{\partial\|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial\tilde{Z}_{m,n}} &= -\frac{2}{T-1}[K_{Y_1Y}]_{j,m}[K_{YY_2}]_{n,j} + [\tilde{Z}^{\mathrm{T}}]_{k,k_1}[K_Y]_{k_1,m}[K_Y]_{n,k} \\
&\quad + [K_Y]_{n,k_2}[\tilde{Z}]_{k_2,k_3}[K_Y]_{k_3,m}, \\
&= -\frac{2}{T-1}[K_{YY_2}K_{Y_1Y}]_{n,m} + [K_Y(\tilde{Z}^{\mathrm{T}} + \tilde{Z})K_Y]_{n,m}.
\end{aligned}$$

The partial derivative with respect to $\tilde{Z}$ is therefore

$$\frac{\partial \|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}}}{\partial \tilde{Z}} = -\frac{2}{T-1}(K_{YY_2}K_{Y_1Y})^{\mathrm{T}} + (K_Y(\tilde{Z}^{\mathrm{T}} + \tilde{Z})K_Y)^{\mathrm{T}}$$
$$= -\frac{2}{T-1}K_{YY_1}K_{Y_2Y} + K_Y(\tilde{Z} + \tilde{Z}^{\mathrm{T}})K_Y.$$

The partial derivative of $\tilde{Z}$ with respect to $\tilde{W}_{i,j}$ can be computed as

$$\frac{\partial \tilde{Z}}{\partial \tilde{W}_{i,j}} = \frac{1}{T}WK_XK_XJ^{j,i}K_XW^{\mathrm{T}},$$

for $i,j \in \{1,\ldots,n\}$, where $J^{j,i}$ is the matrix with (j,i)-th element equal to 1, and 0 otherwise. Combining the two partial derivatives above, we have

$$\frac{\partial \|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}}}{\partial \tilde{W}_{i,j}} = \mathrm{Tr}\left\{\left[\frac{\partial \|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}}}{\partial \tilde{Z}}\right]^{\mathrm{T}} \frac{\partial \tilde{Z}}{\partial \tilde{W}_{i,j}}\right\}$$
$$= \mathrm{Tr}\left\{\left[-\frac{2}{T-1}K_{YY_1}K_{Y_2Y} + K_Y(\tilde{Z} + \tilde{Z}^{\mathrm{T}})K_Y\right]^{\mathrm{T}} \frac{1}{T}WK_XK_XJ^{j,i}K_XW^{\mathrm{T}}\right\}.$$

Note that if $C := AJ^{i,j}B$, then C is equal to the outer product of the i-th column of A and the j-th row of B. In particular, the diagonal of C is equal to the elementwise multiplication of the i-th columns and j-th row of A and B respectively. Thus, $\mathrm{Tr}(AJ^{i,j}B)$ is the sum of the i-th column of A multiplied by the j-th row of B. Let $D_{i,j} := \mathrm{Tr}(AJ^{i,j}B)$, then D can be expressed as $D = A^{\mathrm{T}}B^{\mathrm{T}} = (BA)^{\mathrm{T}}$. It therefore follows that

$$\frac{\partial \|\hat{\mu}_{Y_1Y_2} - \tilde{\mu}_{Y_1Y_2}\|^2_{\mathcal{H}_{\mathcal{Y}}^{\otimes 2}}}{\partial \tilde{W}} = K_XW^{\mathrm{T}}\left[-\frac{2}{T-1}K_{YY_1}K_{Y_2Y} + K_Y(\tilde{Z} + \tilde{Z}^{\mathrm{T}})K_Y\right]^{\mathrm{T}} \frac{1}{T}WK_XK_X.$$

The second term of the empirical surrogate loss function is independent of $\tilde{W}$. The third term $\|\hat{\mu}_X - \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathcal{H}_X}$ does depend on $\tilde{W}$ and its partial derivative with respect to $\tilde{W}$ is computed as follows. Recall that

$$\|\hat{\mu}_X - \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathcal{H}_{\mathcal{X}}} = \frac{1}{T^2}1^{\mathrm{T}}K_X1 - 2\frac{1}{T^2}1^{\mathrm{T}}K_X\tilde{W}K_X1 + \frac{1}{T^2}1^{\mathrm{T}}K_X\tilde{W}^{\mathrm{T}}K_X\tilde{W}K_X1.$$

The partial derivative follows readily as

$$\frac{\partial \|\hat{\mu}_X - \hat{\mathcal{U}}^{\tilde{W}}_{X_2|X_1}\hat{\mu}_X\|^2_{\mathcal{H}_{\mathcal{X}}}}{\partial \tilde{W}} = -\frac{2}{T^2}K_X11^{\mathrm{T}}K_X + \frac{2}{T^2}K_X\tilde{W}K_X11^{\mathrm{T}}K_X.$$

Combining the above, we have the following partial derivative of the empirical surrogate loss function

$$\frac{\hat{L}_S(W,\tilde{W})}{\partial \tilde{W}} = \frac{1}{T}K_XW^{\mathrm{T}}A^{\mathrm{T}}WK_XK_X + \lambda_2[D + \frac{2}{T^2}K_X\tilde{W}K_X11^{\mathrm{T}}K_X],$$

where we have defined $A := -\frac{2}{T-1}K_{YY_1}K_{Y_2Y} + K_Y(\tilde{Z}+\tilde{Z}^{\mathrm{T}})K_Y$, and $D := -\frac{2}{T^2}K_X 1 1^{\mathrm{T}} K_X$.

The partial derivative with respect to W can be derived in a similar manner. The partial derivative of the first term of the empirical surrogate loss is equal to

$$\frac{\partial \|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial W_{i,j}} = \mathrm{Tr}\left\{ \left[\frac{\partial \|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial \tilde{Z}} \right]^{\mathrm{T}} \frac{\partial \tilde{Z}}{\partial W_{i,j}} \right\}.$$

The partial derivative with respect to $\tilde{Z}$ was computed above. The partial derivative of $\tilde{Z}$ with respect to $W_{i,j}$ is

$$\frac{\partial \tilde{Z}}{\partial W_{i,j}} = \frac{1}{T} J^{i,j} K_X K_X \tilde{W}^{\mathrm{T}} K_X W^{\mathrm{T}} + \frac{1}{T} W K_X K_X \tilde{W}^{\mathrm{T}} K_X J^{j,i},$$

and it therefore follows that the partial derivative with respect to $W_{i,j}$, $i,j \in \{1,\ldots,n\}$ is

$$\begin{aligned}\frac{\partial \|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial W_{i,j}} &= \mathrm{Tr}\left\{ \left[-\frac{2}{T-1} K_{YY_1}K_{Y_2Y} + K_Y(\tilde{Z}+\tilde{Z}^{\mathrm{T}})K_Y \right]^{\mathrm{T}} \right. \\ &\quad \left. \times \left[\frac{1}{T} J^{i,j} K_X K_X \tilde{W}^{\mathrm{T}} K_X W^{\mathrm{T}} + \frac{1}{T} W K_X K_X \tilde{W}^{\mathrm{T}} K_X J^{j,i} \right] \right\}.\end{aligned}$$

Following the same reasoning as above, the partial derivative with respect to W is

$$\begin{aligned}\frac{\partial \|\hat{\mu}_{Y_1Y_2} - f(\tilde{Z}(W,\tilde{W}))\|^2_{\mathscr{H}_{\mathscr{Y}}^{\otimes 2}}}{\partial W} &= \left[-\frac{2}{T-1} K_{YY_1}K_{Y_2Y} + K_Y(\tilde{Z}+\tilde{Z}^{\mathrm{T}})K_Y \right] (\frac{1}{T} K_X K_X \tilde{W}^{\mathrm{T}} K_X W^{\mathrm{T}})^{\mathrm{T}} \\ &\quad + \left[-\frac{2}{T-1} K_{YY_1}K_{Y_2Y} + K_Y(\tilde{Z}+\tilde{Z}^{\mathrm{T}})K_Y \right]^{\mathrm{T}} \frac{1}{T} W K_X K_X \tilde{W}^{\mathrm{T}} K_X \\ &= (A + A^{\mathrm{T}}) W (B + B^{\mathrm{T}}),\end{aligned}$$

where $A := -\frac{2}{T-1}K_{YY_1}K_{Y_2Y} + K_Y(\tilde{Z}+\tilde{Z}^{\mathrm{T}})K_Y$, $B := \frac{1}{T}K_X K_X \tilde{W}^{\mathrm{T}} K_X$.

We also require the derivative of $\|\hat{\mu}_Y - \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_X\|^2_{\mathscr{H}_Y}$ with respect to W. Recall that

$$\|\hat{\mu}_Y - \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_X\|^2_{\mathscr{H}_Y} = \frac{1}{T^2} 1^{\mathrm{T}} K_Y 1 - 2\frac{1}{T^2} 1^{\mathrm{T}} K_Y W K_X 1 + \frac{1}{T^2} 1^{\mathrm{T}} K_X W^{\mathrm{T}} K_Y W K_X 1.$$

Differentiating with respect to W we obtain

$$\frac{\partial \|\hat{\mu}_Y - \hat{\mathscr{U}}^W_{Y|X}\hat{\mu}_X\|^2_{\mathscr{H}_Y}}{\partial W} = -\frac{2}{T^2} K_Y 1 1^{\mathrm{T}} K_X + \frac{2}{T^2} K_Y W K_X 1 1^{\mathrm{T}} K_X.$$

Hence, we have the following partial derivative of the empirical surrogate loss function

$$\frac{\partial \hat{L}_S(W,\tilde{W})}{\partial W} = (A + A^{\mathrm{T}}) W (B + B^{\mathrm{T}}) + \lambda_1 (C + \frac{2}{T^2} K_Y W K_X 1 1^{\mathrm{T}} K_X), \tag{4.12}$$

where we have defined $C := -\frac{2}{T^2} K_Y 1 1^{\mathrm{T}} K_X$, and A and B are as defined above. ■

Bibliography

A. Anandkumar, D. Hsu, and S. M. Kakade.
A method of moments for mixture models and hidden markov models.
In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 33.1–33.34, 2012.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky.
Tensor decompositions for learning latent variable models.
Journal of Machine Learning Research, 15:2773–2832, 2014.

N. Aronszajn.
Theory of reproducing kernels.
Transactions of the American Mathematical Society, 68(3):337–404, 1950.

C. R. Baker.
Joint measures and cross-covariance operators.
Transactions of the American Mathematical Society, 186:273–289, 1973.

L. E. Baum and T. Petrie.
Statistical inference for probabilistic functions of finite state markov chains.
The Annals of Mathematical Statistics, 37(6):1554–1563, 1966.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss.
A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains.
The Annals of Mathematical Statistics, 41(1):164–171, 1970.

A. Berlinet and C. Thomas-Agnan.
Reproducing kernel Hilbert spaces in probability and statistics.
Springer Science & Business Media, 2011.

R. Cao, A. Cuevas, and W. G. Manteiga.
A comparative study of several smoothing methods in density estimation.
Computational Statistics & Data Analysis, 17(2):153–176, 1994.

O. Cappé, E. Moulines, and T. Rydén.
Inference in hidden markov models.
In *Proceedings of EUSFLAT Conference*, pages 14–16, 2009.

J. T. Chang.
Full reconstruction of markov models on evolutionary trees: identifiability and consistency.
Mathematical Biosciences, 137(1):51–73, 1996.

Y. Chen, M. Welling, and A. Smola.
Super-samples from kernel herding.
In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 109–116, 2010.

N. Chopin, O. Papaspiliopoulos, et al.
An introduction to sequential Monte Carlo, volume 4.
Springer, 2020.

L. Couvreur and C. Couvreur.
Wavelet-based non-parametric hmm's: theory and applications.
In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 1, pages 604–607. IEEE, 2000.

N. Cristianini and J. Shawe-Taylor.
An introduction to support vector machines and other kernel-based learning methods.
Cambridge University Press, 2000.

G. Da Prato and J. Zabczyk.
Stochastic equations in infinite dimensions.
Cambridge university press, 2014.

Y. De Castro, E. Gassiat, and S. Le Corff.
Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden markov models.
IEEE Transactions on Information Theory, 63(8):4758–4777, 2017.

H. Dette and A. A. Zhigljavsky.
Reproducing kernel hilbert spaces, polynomials, and the classical moment problem.
SIAM/ASA Journal on Uncertainty Quantification, 9(4):1589–1614, 2021.

L. Devroye and G. Lugosi.
Combinatorial methods in density estimation.
Springer Science & Business Media, 2001.

J. Diestel and J. Uhl.
Vector measures, math, 1977.

N. Dinculeanu.
Vector integration and stochastic integration in Banach spaces, volume 48.
John Wiley & Sons, 2000.

R. Douc, E. Moulines, and T. Rydén.
Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime.
Annals of Statistics, 32(5):2254–2304, 2004.

A. Doucet, N. De Freitas, and N. Gordon.
An introduction to sequential monte carlo methods.
Sequential Monte Carlo Methods in Practice, pages 3–14, 2001.

S. Eleftheriadis, T. Nicholson, M. Deisenroth, and J. Hensman.
Identification of gaussian process state space models.
Advances in Neural Information Processing Systems, 30, 2017.

J. Fan and T. H. Yim.
A crossvalidation method for estimating conditional densities.
Biometrika, 91(4):819–834, 2004.

J. Fan, Q. Yao, and H. Tong.
Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems.
Biometrika, 83(1):189–206, 1996.

S. Fischer and I. Steinwart.
Sobolev norm learning rates for regularized least-squares algorithms.
The Journal of Machine Learning Research, 21(1):8464–8501, 2020.

R. Frigola, Y. Chen, and C. E. Rasmussen.
Variational gaussian process state-space models.
Advances in Neural Information Processing Systems, 27, 2014.

K. Fukumizu.
Nonparametric bayesian inference with kernel mean embedding.
Modern Methodology and Applications in Spatial-Temporal Modeling, pages 1–24, 2015.

K. Fukumizu, F. Bach, and M. Jordan.
Kernel dimensionality reduction for supervised learning.
Advances in Neural Information Processing Systems, 16, 2003.

K. Fukumizu, L. Song, and A. Gretton.
Kernel bayes' rule: Bayesian inference with positive definite kernels.
The Journal of Machine Learning Research, 14(1):3753–3783, 2013.

É. Gassiat, A. Cleynen, and S. Robin.
Inference in finite state space non parametric hidden markov models and applications.
Statistics and Computing, 26:61–71, 2016.

E. Gassiat, S. Le Corff, and L. Lehéricy.
Identifiability and consistent estimation of nonparametric translation hidden markov models with general state space.
The Journal of Machine Learning Research, 21(1):4589–4628, 2020.

E. Giné and A. Guillou.
Rates of strong uniform consistency for multivariate kernel density estimators.
In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921.
Elsevier, 2002.

I. Gohberg, S. Goldberg, and M. A. Kaashoek.
Classes of linear operators Vol. 1, volume 63.
Birkhäuser, 1990.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola.
A kernel method for the two-sample-problem.
Advances in Neural Information Processing Systems, 19, 2006.

S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil.
Conditional mean embeddings as regressors.
In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1803–1810, 2012.

J. Hoffmann-Jørgensen and G. Pisier.
The law of large numbers and the central limit theorem in banach spaces.
The Annals of Probability, pages 587–599, 1976.

M. P. Holmes, A. G. Gray, and C. L. Isbell Jr.
Fast nonparametric conditional density estimation.
In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 175–182, 2007.

D. Hsu, S. M. Kakade, and T. Zhang.
A spectral algorithm for learning hidden markov models.
Journal of Computer and System Sciences, 78(5):1460–1480, 2012.

R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald.
Estimating and visualizing conditional densities.
Journal of Computational and Graphical Statistics, 5(4):315–336, 1996.

H. Jaeger.
Observable operator models for discrete stochastic time series.
Neural Computation, 12(6):1371–1398, 2000.

M. Janzamin, R. Ge, J. Kossaifi, A. Anandkumar, et al.
Spectral learning on matrices and tensors.
Foundations and Trends® in Machine Learning, 12(5-6):393–536, 2019.

H. Jiang.
Uniform convergence rates for kernel density estimation.
In *International Conference on Machine Learning*, pages 1694–1703. PMLR, 2017.

M. C. Jones, J. S. Marron, and S. J. Sheather.
A brief survey of bandwidth selection for density estimation.
Journal of the American Statistical Association, 91(433):401–407, 1996.

M. Kanagawa and K. Fukumizu.
Recovering distributions from gaussian rkhs embeddings.
In *Artificial Intelligence and Statistics*, pages 457–465. PMLR, 2014.

M. Kanagawa, B. K. Sriperumbudur, and K. Fukumizu.
Convergence guarantees for kernel-based quadrature rules in misspecified settings.
Advances in Neural Information Processing Systems, 29, 2016.

D. P. Kingma and J. Ba.
Adam: A method for stochastic optimization.
arXiv preprint arXiv:1412.6980, 2014.

G. Kitagawa.
Non-gaussian state—space modeling of nonstationary time series.
Journal of the American statistical association, 82(400):1032–1041, 1987.

I. Klebanov, I. Schuster, and T. J. Sullivan.
A rigorous theory of conditional mean embeddings.
SIAM Journal on Mathematics of Data Science, 2(3):583–606, 2020.

J.-Y. Kwok and I.-H. Tsang.
The pre-image problem in kernel methods.
IEEE Transactions on Neural Networks, 15(6):1517–1525, 2004.

M. F. Lambert, J. P. Whiting, and A. V. Metcalfe.
A non-parametric hidden markov model for climate state identification.
Hydrology and Earth System Sciences, 7(5):652–667, 2003.

R. Langrock, T. Kneib, A. Sohn, and S. L. DeRuiter.
Nonparametric inference in hidden markov models using p-splines.
Biometrics, 71(2):520–528, 2015.

R. Langrock, T. Adam, V. Leos-Barajas, S. Mews, D. L. Miller, and Y. P. Papastamatiou.
Spline-based nonparametric inference in general state-switching models.
Statistica Neerlandica, 72(3):179–200, 2018.

Y. LeCun, C. Cortes, and C. Burges.
Mnist handwritten digit database.
ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2, 2010.

L. Lehéricy.
State-by-state minimax adaptive estimation for nonparametric hidden markov models.
The Journal of Machine Learning Research, 19(1):1432–1477, 2018.

L. Lehéricy.
Consistent order estimation for nonparametric hidden markov models.
Bernoulli, 25(1):464–498, 2019.

Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton.
Optimal rates for regularized conditional mean embedding learning.
Advances in Neural Information Processing Systems, 35:4433–4445, 2022.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin.
Towards a learning theory of cause-effect inference.
In *International Conference on Machine Learning*, pages 1452–1461. PMLR, 2015.

H. Q. Minh.
Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory.
Constructive Approximation, 32:307–338, 2010.

M. Mollenhauer, I. Schuster, S. Klus, and C. Schütte.
Singular value decomposition of operators on reproducing kernel hilbert spaces.
In *Advances in Dynamics, Optimization and Computation: A volume dedicated to Michael Dellnitz on the occasion of his 60th birthday*, pages 109–131. Springer, 2020.

K. Muandet, B. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf.

Kernel mean shrinkage estimators.
Journal of Machine Learning Research, 17, 2016.

K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al.
Kernel mean embedding of distributions: A review and beyond.
Foundations and Trends® in Machine Learning, 10(1-2):1–141, 2017.

É. Nadaraya.
On non-parametric estimates of density functions and regression curves.
Theory of Probability & Its Applications, 10(1):186–190, 1965.

Y. Nishiyama, M. Kanagawa, A. Gretton, and K. Fukumizu.
Model-based kernel sum rule: kernel bayesian inference with probabilistic models.
Machine Learning, 109(5):939–972, 2020.

J. Nocedal and S. J. Wright.
Numerical optimization.
Springer, 1999.

H. Owhadi and C. Scovel.
Separability of reproducing kernel spaces.
Proceedings of the American Mathematical Society, 145(5):2131–2138, 2017.

B. U. Park and J. S. Marron.
Comparison of data-driven bandwidth selectors.
Journal of the American Statistical Association, 85(409):66–72, 1990.

J. Park and K. Muandet.
A measure-theoretic approach to kernel conditional mean embeddings.
Advances in neural information processing systems, 33:21247–21259, 2020.

E. Parzen.
On estimation of a probability density function and mode.
The Annals of Mathematical Statistics, 33(3):1065–1076, 1962.

D. Paulin.
Concentration inequalities for markov chains by marton couplings and spectral methods.
Electron. J. Probab, 20(79):1–32, 2015.

V. I. Paulsen and M. Raghupathi.
An introduction to the theory of reproducing kernel Hilbert spaces, volume 152.
Cambridge University Press, 2016.

L. R. Rabiner.
A tutorial on hidden markov models and selected applications in speech recognition.
Proceedings of the IEEE, 77(2):257–286, 1989.

M. Rosenblatt.
Conditional probability density and regression estimators.
Multivariate Analysis II, 25:31, 1969.

V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan.
Finding a "kneedle" in a haystack: Detecting knee points in system behavior.
In *International Conference on Distributed Computing Systems Workshops*, pages 166–171. IEEE, 2011.

B. Schölkopf and A. Smola.
Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.
MIT Press, Cambridge, MA, USA, 2001.
ISBN 0262194759.

B. Schölkopf, A. Smola, and K.-R. Müller.
Kernel principal component analysis.
In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

I. Schuster, M. Mollenhauer, S. Klus, and K. Muandet.
Kernel conditional density operators.
In *International Conference on Artificial Intelligence and Statistics*, pages 993–1004. PMLR, 2020.

L. Shang and K.-P. Chan.
Nonparametric discriminant hmm and application to facial expression recognition.
In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2090–2096. IEEE, 2009.

B. W. Silverman.
Density estimation for statistics and data analysis, volume 26.
CRC press, 1986.

A. Smola, A. Gretton, L. Song, and B. Schölkopf.
A hilbert space embedding for distributions.
In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

L. Song.
Learning via hilbert space embedding of distributions.
University of Sydney (2008), 17, 2008.

L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf.
Tailoring density estimation via reproducing kernel moment matching.
In *Proceedings of the 25th International Conference on Machine learning*, pages 992–999, 2008.

L. Song, J. Huang, A. Smola, and K. Fukumizu.
Hilbert space embeddings of conditional distributions with applications to dynamical systems.
In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola.
Hilbert space embeddings of hidden markov models.
In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 991–998, 2010.

L. Song, K. Fukumizu, and A. Gretton.
Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models.
IEEE Signal Processing Magazine, 30(4):98–111, 2013.

L. Song, A. Anandkumar, B. Dai, and B. Xie.
Nonparametric estimation of multi-view latent variable models.
In *International Conference on Machine Learning*, pages 640–648. PMLR, 2014.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet.
Universality, characteristic kernels and rkhs embedding of measures.
Journal of Machine Learning Research, 12(7), 2011.

M. Stanke, S. Waack, et al.
Gene prediction with a hidden markov model and a new intron submodel.
Bioinformatics-Oxford, 19(2):215–225, 2003.

C. J. Stone.
Optimal rates of convergence for nonparametric estimators.
The Annals of Statistics, pages 1348–1360, 1980.

S. Taylor.
Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75.
Time Series Analysis: Theory and Practice, 1:203–226, 1982.

A. Tsybakov.
Introduction to Nonparametric Estimation.
Springer Series in Statistics. Springer New York, 2008.

S. Volant, C. Bérard, M.-L. Martin-Magniette, and S. Robin.
Hidden markov models with mixtures as emission distributions.
Statistics and Computing, 24(4):493–504, 2014.

V. Vovk.
Kernel ridge regression.
In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 105–116. Springer, 2013.

M. Welling.
Herding dynamical weights to learn.
In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128, 2009.

C. K. Williams and C. E. Rasmussen.
Gaussian processes for machine learning, volume 2.
MIT press Cambridge, MA, 2006.

C. Yau, O. Papaspiliopoulos, G. O. Roberts, and C. Holmes.
Bayesian non-parametric hidden markov models with applications in genomics.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 73(1):37–57, 2011.