Journal of Machine Learning Research 24 (2023) 1-49 Submitted 6/22; Revised 7/23; Published 11/23

Topological Hidden Markov Models

Adam B Kashlak KASHLAK@UALBERTA.CA
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB, T6G 2G1 Canada

Prachi Loliencar LOLIENCA@UALBERTA.CA
School of Dentistry
University of Alberta
Edmonton, AB, T6G 1C9 Canada

Giseon Heo GHEO@UALBERTA.CA
School of Dentistry
University of Alberta
Edmonton, AB, T6G 1C9 Canada

Editor: David Blei

Abstract

The Hidden Markov Model (HMM) is a classic modelling tool with a wide swath of applications. Its inception considered observations restricted to a finite alphabet, but it was quickly extended to multivariate continuous distributions. In this article, we further extend the HMM from mixtures of normal distributions in d-dimensional Euclidean space to general Gaussian measure mixtures in locally convex topological spaces, and hence, we christen this method the Topological Hidden Markov Model (THMM). The main innovation is the use of the Onsager-Machlup functional as a proxy for the probability density function in infinite dimensional spaces. This allows for choice of a Cameron-Martin space suitable for a given application. We demonstrate the versatility of this methodology by applying it to simulated diffusion processes such as Brownian and fractional Brownian sample paths as well as the Ornstein-Uhlenbeck process. Our methodology is applied to the identification of sleep states from overnight polysomnography time series data with the aim of diagnosing Obstructive Sleep Apnea in pediatric patients. It is also applied to a series of annual cumulative snowfall curves from 1940 to 1990 in the city of Edmonton, Alberta.

Keywords: functional data, Gaussian measures, locally convex spaces, Onsager-Machlup functional, stochastic processes.

1. Introduction

The Hidden Markov Model (HMM) was and still is a powerful tool for modelling diverse data sets. Its inception dates to the work of Leonard Baum and his colleagues at the Institute for Defense Analysis (Baum and Petrie, 1966; Baum et al., 1970) predating the advent of the expectation-maximization (EM) algorithm (Dempster et al., 1977; Wu, 1983). Originally developed for observations taking values within a finite alphabet, its applications included speech recognition and genome sequences (Ferguson, 1980; Poritz and Richter, 1986; Juang and Rabiner, 1985, 1991). Since then, it has been extended to many types of data including

multivariate elliptically symmetric distributions (Liporace, 1982) and skewed distributions (Chatzis, 2010).

When considering an ordered sequence of data points, the HMM assumes that the sequence of observed data is independent conditional on a discrete state variable driven by a Markov chain. For example, the daily number of people taking public transit to work (observation) may depend on the weather (state). If we consider two weather states, snowy and sunny, we can model the day-to-day changes in the weather as a Markov chain with a 2×2 transition matrix. In this example, the weather is obviously observable. However, the true power of the HMM is to learn hidden states that may not be immediately obvious. In this way, the HMM aims to cluster observations, but assuming only conditional independence (as opposed to full independence) based on the states of a Markov chain. An excellent tutorial on HMMs can be found in Rabiner (1989) with many references to early work on this model therein.

The following work considers extending the observation space of the Hidden Markov Model from Euclidean space to general locally convex topological vector spaces (LCTVS) specifically where the states of the Markov chain correspond to mean-shifted Gaussian measures. As a result of this general formulation, our so-called Topological Hidden Markov Model (THMM) can be adapted to many settings of interest from finite dimensional data to functional data and stochastic process data. In Section 5, we show the method's applicability to a variety of simulated sample paths from stochastic processes such as fractional Brownian motion and the Ornstein-Uhlenbeck process. In Section 6, we apply these algorithms to the task of identifying sleep states from an electroencephalogram (EEG) time series collected during overnight polysomnography. While this data is considered as a proof-of-concept for the methodology, the ultimate goal for subsequent research is to be able to quickly identify sleep disorders such as Pediatric Obstructive Sleep Apnea. Section 7 applies both the classic HMM and the new THMM algorithms to cumulative snowfall curves from the city of Edmonton, Alberta with an aim of identifying patterns in snowfall across a 50 year time span.

In infinite dimensional spaces, there is no analogue of Lebesgue measure and thus no probability densities and likelihood function to maximize. This is the central challenge when working with functional and stochastic process data. Our main innovation in this work is use the Onsager-Machlup functional for a general Gaussian measure (Bogachev, 1998) within the HMM emission function. This allows for fitting HMMs to data living in a wide variety of spaces such as $L^2[0,1]$ and the Sobolev space $W_0^{2,1}[0,1]$ among others. There have been many past works on deriving the Onsager-Machlup functional for various Gaussian processes that we will make use of (Takahashi and Watanabe, 1981; Zeitouni, 1989; Shepp and Zeitouni, 1992; Capitaine, 1995; Ikeda and Watanabe, 2014).

There have been a few recent ventures taking the classic HMM into the realm of functional data. In Martino et al. (2020), they focus on multivariate functional data and choose the emission function to be the inverse squared L^2 distance. The work of Sidrow et al. (2021) proposed a more complicated hierarchical HMM model that can be used to partition high frequency functional data with complicated dependence structures so that the pieces can in turn be modelled via time series or functional data methods. For analyzing longitudinal data, Altman (2007) proposes a mixed HMM that incorporates covariates and random effects into the model. Apart from functional data, there has been much recent work into

HMM fitting with nonparametric density estimation (De Castro et al., 2016; Gassiat and Rousseau, 2016; Gassiat et al., 2016; Lehéricy, 2018).

There have been many papers written on clustering of functional data without any regards to temporal ordering. In this work, we consider the `kmeans.fd` function from the `fda.usc` R package (Febrero-Bande and Oviedo de la Fuente, 2012). More sophisticated clustering and alignment methods can be found in `fdacluster` (Sangalli et al., 2010, 2014; Stamm, 2023). However, our data sets of interest are already aligned with regularly spaced observations, and thus we do not make use of `fdacluster` in this work. Nevertheless, alignment of functional data within an HMM framework would be a challenging future research topic. Other works on functional clustering methods include Bouveyron (2021) and chapter 9 of Ferraty and Vieu (2006).

The classic Hidden Markov Model is introduced below in Section 2. The two main algorithms, Baum-Welch and Viterbi, are outlined in Section 3. Section 4 details many specific models of interest including parametric models like Brownian motion with linear drift and non-parametric curve fitting. Section 5 demonstrates the power of the THMM applied to a variety of simulated data sets. Section 6 further demonstrates the power of the THMM by showing its efficacy in identifying sleep states from noisy pediatric EEG data streams. Section 7 considers climatological data. Future extensions to this work are briefly discussed in Section 8.

All theory contributions to justify this paper's methodology can be found in Appendix A. Appendix A.1 introduces notation for locally convex spaces, semi-norms, and Cameron-Martin spaces. Appendix A.2 defines the Onsager-Machlup functional. Theorems justifying the validity of the THMM are stated and proved in Appendices A.3 and A.4. We first prove that each step of the algorithm does, in fact, improve the analogue of the likelihood function which is based on the Onsager-Machlup functional. We secondly prove that the sequence of reestimated parameters produced by the Baum-Welch algorithm has at least one limit point and that all limit points of the sequence are critical points of the likelihood function. Lastly, we show that in the dual sense, the corresponding sequence of mixtures of Gaussian measures has a weak limit point as the number of iterations of the algorithm tends to infinity. Lastly, Appendix A.6 contains a theorem and proof and short discussion regarding model identifiability.

2. The Classic Hidden Markov Model

The classic HMM model is a type of dynamic Bayesian network (Sucar, 2015) with a long history of development. A standard diagram of such a model can be found in Figure 1 whereas Figure 5.5 in Sucar (2015) displays more complicated extensions of the HMM. For the classic HMM, we begin with an observation sequence $O_1, \ldots, O_T$ that lives in some space X, which could be a finite space $\{1, 2, \ldots, d\}$ or Euclidean space $\mathbb{R}^d$ or otherwise. In tandem, there is a hidden state sequence $s = (s_1, \ldots, s_T)$ where $s_t \in S = \{1, 2, \ldots, p\}$, which evolves as a p-state Markov chain with initial state probabilities $\eta_j = \mathrm{P}(s_1 = j)$ and $p \times p$ transition matrix A with ijth entry $a_{ij} = \mathrm{P}(s_{t+1} = j \,|\, s_t = i)$, which is assumed to be invariant to choice of $t = 1, \ldots, T-1$. However, nonhomogeneous Markov models with time varying state transition probabilities have also been developed (Hughes et al., 1999). Furthermore, there exist state dependent *emission* functions $b_j : X \to \mathbb{R}^+$ for $j = 1, \ldots, p$,

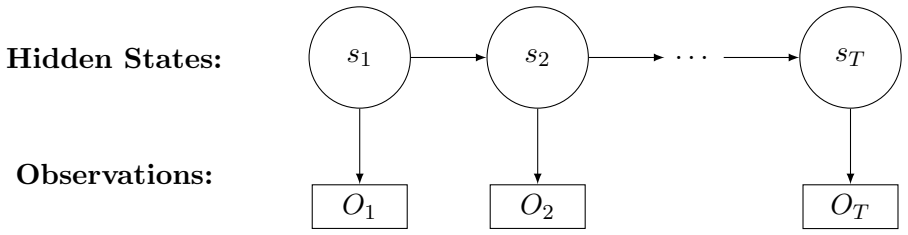

Figure 1: The classic Hidden Markov Model has observations O_t that are known, but has states s_t that are unknown. Conditional on state s_t, the observation O_t is independent of the other observations.

which assign a value to each observation O_t based on being emitted from each potential state $s_t = j$. In the case of, say, multivariate Gaussian data in $\mathbb{R}^p$, b_j is simply the d-dimensional probability density function with state dependent mean vector and covariance matrix. In the classic HMM, it is assumed that the observation sequence is comprised of T elements that are independent conditionally on the state sequence. This assumption is removed in more complex variants of the HMM such as the autoregressive HMM discussed in Rabiner (1989) and others.

The Baum-Welch algorithm offers an efficient way to estimate the unknown parameters in the HMM that maximize the likelihood function

$$L(O_1, \ldots, O_T \,|\, \lambda) = \sum_{s \in S^T} \eta_{s_1} \prod_{t=1}^{T} a_{s_t s_{t+1}} b_{s_t}(O_t)$$

where the summation is taken over $S^T = \{1, \ldots, p\}^T$, the space of all p^T state sequences, and λ represents the collection of model parameters. The crux of the Baum-Welch algorithm are the forward and backward probabilities

$$\begin{aligned} \alpha_t(j) &= \mathrm{P}\left(O_1, \ldots, O_t, s_t = j \,|\, \lambda\right) \\ \beta_t(j) &= \mathrm{P}\left(O_{t+1}, \ldots, O_T \,|\, s_t = j, \lambda\right), \end{aligned}$$

which can be computed recursively as outlined below in Algorithm 1. Estimation of the state means is achieved as a weighted sum of the observations where the weights come directly from the α and β probabilities. Namely, we wish to find the state means that maximize the following sum $\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)b_j(O_t)$ for each state j.

Our implementation of Baum-Welch discussed in the next section is very similar to the classic version. The main innovation is usage and justification of the Onsager-Machlup functional for the emission functions b_j as we do not have a probability density function to use for the b_j. Thus, the state dependent means live in the Cameron-Martin space $H(\gamma)$. More technical details can be found in the appendix.

3. The Topological HMM

Borrowing the notation from classic works on HMMs (Liporace, 1982; Rabiner, 1989), the goal for fitting an HMM to an observation sequence $O = (O_1, \ldots, O_T)$ is to find the model

Algorithm 1 The THMM Baum-Welch Algorithm

Initialize model parameters:

$\eta_j = \mathrm{P}(s_1 = j)$, for $j = 1, \ldots, p$ (Initial probabilities)

$a_{ij} = \mathrm{P}(s_{t+1} = j \mid s_t = i)$, for $i, j = 1, \ldots, p$ (Transition probabilities)

$h_j \in H(\gamma)$ for $j = 1, \ldots, p$ (Center element for each state)

Iterate until convergence:

Forward Pass:

$\alpha_1(j) = \eta_j b_j(O_1)$

For $t = 2, \ldots, T$:

$\alpha_t(j) = \mathrm{P}(O_1, \ldots, O_t, s_t = j \mid \eta, A, h) = \sum_{i=1}^{p} \alpha_{t-1}(i) a_{ij} b_j(O_t)$

Backward Pass:

$\beta_T(j) = 1$

For $t = (T-1), \ldots, 1$:

$\beta_t(j) = \mathrm{P}(O_{t+1}, \ldots, O_T \mid s_t = j, \eta, A, h) = \sum_{i=1}^{p} \beta_{t+1}(i) a_{ji} b_i(O_{t+1})$

Reestimation:

$\gamma_t(i) = \mathrm{P}(s_t = i \mid O, \eta, A, h) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{p} \alpha_t(j)\beta_t(j)}$

$\xi_t(i,j) = \mathrm{P}(s_t = i, s_{t+1} = j \mid O, \eta, A, h) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i',j'=1}^{p} \alpha_t(i') a_{i'j'} b_{j'}(O_{t+1}) \beta_{t+1}(j')}$

$\tilde{\eta}_j = \gamma_1(j)$

$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$

$h_j = \underset{h \in H(\gamma)}{\arg\max} \sum_{t=1}^{T} \alpha_t(j)\beta_t(j) b_j(O_t)$ where $b_j(O_t)$ depends on h.

parameters and the state sequence s that maximize the likelihood $L(O \mid s) = \prod_{t=1}^{T} b_{s_t}(O_t)$. The task of choosing the best parameters is achieved via the Baum-Welch algorithm. Determining the best state sequence is done by the Viterbi algorithm. These are detailed in Algorithms 1 and 2, respectively, which differ only from their original instantiations in the choice of emission function b_j and method of reestimation for the state means $h_j \in H(\gamma)$ for $j = 1, \ldots, p$. The space $H(\gamma)$ where means are selected from is a user-specified Cameron-Martin space. In the THMM setting, the emission functions b_j are Onsager-Machlup functionals, which is limit of the ratio of the Gaussian measures of two balls centred at some element h_j and at the origin element 0.

Many more technical details on this setup can be found in Appendices A.1 and A.2. Theorems and proofs regarding the convergence properties of the Baum-Welch algorithm in this setting are detailed in Appendices A.3 and A.4, which extend from the classic works of Liporace (1982) and Wu (1983). Lastly, identifiability of this model is discussed in detail in Appendix A.6, which extend the recent works of Gassiat and Rousseau (2016) and Gassiat et al. (2016).

3.1 Baum-Welch

The Baum-Welch Algorithm (Baum and Petrie, 1966), detailed in Algorithm 1, takes a form similar to that of an Expectation-Maximization algorithm. However, the Baum-Welch

Algorithm 2 The Viterbi Algorithm

Initialize $\delta_1(j) = \eta_j b_j(O_1)$ and $\phi_1(j) = 0$ for all $i = 1, \ldots, p$.
For $t = 2, \ldots, T$, compute the following for all $j = 1, \ldots, p$
$\delta_t(j) = \max_{i=1,\ldots,p} \{\delta_{t-1}(i) a_{ij}\} b_j(O_t)$
$\phi_t(j) = \arg\max_{i=1,\ldots,p} \{\delta_{t-1}(i) a_{ij}\}$
The final state is $\hat{s}_T = \arg\max_{i=1,\ldots,p} \delta_T(i)$.
For $t = (T-1), \ldots, 1$, compute
$\hat{s}_t = \phi_{t+1}(\hat{s}_{t+1})$.

algorithm predates the EM algorithm (Dempster et al., 1977) by about a decade. For more classic references on HMMs, see those within Rabiner (1989). The THMM version of Baum-Welch maximizes a replacement for the likelihood based on the Onsager-Machlup functionals. Such models can be fitted with respect to other criterion such as Viterbi Training (Lember and Koloydenko, 2008), for example.

The Baum-Welch algorithm works by computing so-called forward (alpha) and backward (beta) probabilities given the model parameters. It then reestimates the model parameters based on these probabilities. At each iteration, the likelihood increases and the algorithm is run until the relative change in the likelihood becomes minuscule. Given the alpha probabilities, the log likelihood is simply computed as $\ell(O) = \sum_{j=1}^{p} \log \alpha_T(j)$. Denoting the log likelihood at iteration r to be ℓ_r, we stop the algorithm when the relative change in the log likelihood, $(\ell_{r+1} - \ell_r)/\ell_{r+1}$, is less than a user specified tolerance, say, 10^{-6}. In practice, all of the terms in Algorithm 1 are computed on the log-scale to avoid numerical stability issues.

Reestimation of the means is performed by

$$h_j = \arg\max_{h \in H(\gamma)} \sum_{t=1}^{T} \alpha_t(j) \beta_t(j) b_j(O_t)$$

where the emission probabilities, $b_j(O_t)$, depend on choice of state mean h_j. How this equation is used depends on the type of model being fitted. Many specific examples are considered below in Section 4.

As with both the classic HMM and EM-style algorithms, the choice of initialization parameters can drastically affect the performance. In particular, we require each state to have a mean $h_j \in H(\gamma)$, which furthermore lies within the convex hull of the observations $O_1, \ldots, O_T$. This will be discussed more in Appendix A. For parametric settings, we estimate the parameters for each O_t and then pick random starting values roughly spread out in this convex set. Alternatively, in the non-parametric setting, we can either randomly select a single O_t for each state to start with or we can run a quick k-means clustering to automatically choose the starting state means.

3.2 Viterbi

Given an observation sequence $O_1, \ldots, O_T$, a state space, initial probabilities η_j, transition probabilities a_{ij}, and emission probabilities $b_j(O_t)$, the Viterbi algorithm (Viterbi, 1967)

finds the most probable state sequence; see Rabiner (1989) and references therein for more details.

Let $\delta_t(j) = \max_{(s_1,\ldots,s_{t-1})\in S^{t-1}} P(s_1, ..., s_{t-1}, s_t = j)$ be the highest probability of any state sequence from 1 to t such that the state at time t is j. The probability of the most likely state sequence ending in state $s_t = j$ can be computed in a recursive fashion by maximizing over s_{t-1} at each time step. Let $\phi_t(j)$ be the state at time $t-1$ that maximizes $\delta_t(j)$. The goal of this algorithm is to compute the *best* state sequence denoted $\hat{s}_1, \ldots, \hat{s}_T$. The Viterbi algorithm is detailed in Algorithm 2. We also consider soft clustering by considering the posterior probabilities of being in each given state at each given time. While we only make use of the classic Viterbi algorithm in this research, a comprehensive analysis of inference on Markov paths can be found in Lember and Koloydenko (2014) and the references therein. This may be of future interest especially when applied to complex data sets of interest.

4. Spaces of Interest

The following subsections contain specific settings of interest where the equations for the Onsager-Machlup functional have been worked out explicitly. The first is the classic Euclidean space setting, which coincides with the standard multivariate Gaussian HMM. Of more novel interest are the various types of Wiener processes such as Brownian motion and the Ornstein-Uhlenbeck process. Fractional Brownian motion with a fixed Hurst parameter and non-parametric state mean estimation are also considered.

4.1 Euclidean Space

For the simplest setting, we can consider $O_t \in \mathbb{R}^d$ and each state $s_t = j$ corresponding to a multivariate Gaussian distribution with mean μ_j and common covariance Σ. In this case, the Cameron-Martin norm is $|h|_{H(\gamma)} = \sup\{v^{\mathrm{T}}h \,:\, v^{\mathrm{T}}\Sigma v \leq 1\}$ where $\mathrm{E}(v) = 0$ and $\mathrm{Var}\,(v) = v^{\mathrm{T}}\Sigma v$. This leads to $|h|_{H(\gamma)} = \sqrt{h^{\mathrm{T}}\Sigma^{-1}h}$. The corresponding Cameron-Martin inner product is $\langle h, k\rangle_{H(\gamma)} = h^{\mathrm{T}}\Sigma^{-1}k$. Thus, the log-emission function is $\log b_j(O_t) = -\frac{1}{2}(O_t - h)^{\mathrm{T}}\Sigma^{-1}(O_t - h)$ and the reestimated mean vector is

$$\tilde{m}_j = \underset{h\in\mathbb{R}^d}{\arg\min} \sum_{t=1}^{T} \alpha_t(j)\beta_t(j)(O_t - h)^{\mathrm{T}}\Sigma^{-1}(O_t - h),$$

which can be simply solved via vector calculus to get

$$\tilde{m}_j = \frac{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)O_t}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}.$$

This formulation coincides with the classic HMM for multivariate Gaussian data when the covariance matrix is fixed. Our current formulation of the HMM using the Onsager-Machlup functional requires a fixed covariance structure across all system states. However, as we will show below, this still allows the THMM to model many diverse types of data.

Theoretical justification for the reestimated means can be found in Lemma 5. In this and the following specific cases, the reestimated means take the form of weighted averages of the observed data points.

4.2 Wiener Space with Smooth Norms

In a collection of articles and texts (Takahashi and Watanabe, 1981; Zeitouni, 1989; Shepp and Zeitouni, 1992; Capitaine, 1995; Ikeda and Watanabe, 2014), the Onsager-Machlup functional is derived for the diffusion process

$$dY_\tau = r(Y_\tau)d\tau + dW_\tau, \quad Y_0 = y, \quad Y_\tau \in \mathbb{R}^d, \quad \tau \in [0,1],$$

where $r : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth function and W_τ is d-dimensional Brownian motion. That is,

$$\log\left[\lim_{\varepsilon\to 0}\frac{\mathrm{P}\left(\|Y-\Phi\|<\varepsilon\right)}{\mathrm{P}\left(\|W\|<\varepsilon\right)}\right] = -\frac{1}{2}\sum_{k=1}^{d}\int_0^1\left|\dot{\Phi}_{k,\tau} - r_k(\Phi_\tau)\right|^2 d\tau - \frac{1}{2}\sum_{k=1}^{d}\int_0^1 \frac{\partial r_k}{\partial y_k}(\Phi_\tau)d\tau$$

where $\dot{\Phi}_\tau = d\Phi_\tau/d\tau$. Ikeda and Watanabe (2014) prove the above for the sup-norm and $\Phi \in C^2$, the space of twice differentiable functions. Shepp and Zeitouni (1992) extended this to all Φ such that $\Phi - y \in H(\gamma)$ and L^p norms for $p \geq 4$ and Hölder norms with $0 < \alpha < 1/3$. Capitaine (1995) further shows that this result holds for a wide class of smooth norms on Wiener space including Hölder norms with $0 < \alpha < 1/2$, Besov norms, and Sobolev norms.

Brownian Motion with Drift

Many common stochastic processes arise from choices of r (see Pavliotis (2014) Section 5.3). For example, in the case of one-dimensional Brownian motion with state j drift coefficient c_j, the diffusion equation is $dY_\tau = c_j d\tau + dW_\tau$, and the Onsager-Machlup functional / log-emission function is

$$\log b_j(O_{t,\tau}) = \log\left[\lim_{\varepsilon\to 0}\frac{\mathrm{P}\left(\|W_\tau + c_j\tau - O_{t,\tau}\|<\varepsilon\right)}{\mathrm{P}\left(\|W_\tau\|<\varepsilon\right)}\right] = -\frac{1}{2}\int_0^1\left|\dot{O}_{t,\tau} - c_j\right|^2 d\tau.$$

The observations $O_{t,\tau}$ should be projected into the Cameron-Martin space $H(\gamma)$ allowing for differentiation. In practice, smoothing methods may be required. The drift terms can be reestimated within the Baum-Welch algorithm by

$$\tilde{c}_j = \underset{h\in\mathbb{R}}{\arg\min}\sum_{t=1}^{T}\alpha_t(j)\beta_t(j)\left\{\frac{1}{2}\int_0^1\left|\dot{O}_{t,\tau} - h\right|^2 d\tau\right\}.$$

This leads to the weighted least squares estimate

$$\tilde{c}_j = \frac{\sum_{t=1}^{T}\alpha_t(j)\beta_t(j)[O_{t,1} - O_{t,0}]}{\sum_{t=1}^{T}\alpha_t(j)\beta_t(j)}.$$

Restricting to a one-dimensional drift parameter is convenient for exposition, but not necessary in practice. We could instead consider

$$dY_\tau = r(t)d\tau + dW_\tau, \quad Y_0 = y, \quad Y_\tau \in \mathbb{R}^d, \quad \tau \in [0,1],$$

where $r(t) = \sum_{i=1}^{m} c_i\psi_i(t)$ for some functional basis $\psi_1, \ldots, \psi_m$. Thus, we would have m parameters to estimate.

Ornstein-Uhlenbeck Process

In the case of the one-dimensional Ornstein-Uhlenbeck (OU) process, $dY_\tau = c_j(\mu_j - Y_\tau)d\tau + dW_\tau$, the Onsager-Machlup functional is

$$\log b_j(O_{t,\tau}) = \log\left[\lim_{\varepsilon\to 0}\frac{\mathrm{P}\left(\|W_\tau + c_j(\mu_j - Y_\tau) - O_{t,\tau}\| < \varepsilon\right)}{\mathrm{P}\left(\|W\| < \varepsilon\right)}\right] = -\frac{1}{2}\int_0^1 \left|\dot{O}_{t,\tau} - c_j(\mu_j - O_{t,\tau})\right|^2 d\tau + \frac{c_j}{2},$$

which leads to the following parameter reestimation:

$$(\tilde{c}_j, \tilde{\mu}_j) = \underset{h\in\mathbb{R}^+, k\in\mathbb{R}}{\arg\min}\sum_{t=1}^T \alpha_t(j)\beta_t(j)\left\{\frac{1}{2}\int_0^1 \left|\dot{O}_{t,\tau} - h(k - O_{t,\tau})\right|^2 d\tau - \frac{h}{2}\right\}.$$

The derivative with respect to k inside the curly brackets gives similarly to the above Brownian motion that

$$\begin{aligned}\tilde{\mu}_j &= \frac{\sum_{t=1}^T \alpha_t(j)\beta_t(j)\int_0^1\{h\dot{O}_{t,\tau} + h^2 O_{t,\tau}\}d\tau}{\sum_{t=1}^T \alpha_t(j)\beta_t(j)h^2}\\ &= \frac{\sum_{t=1}^T \alpha_t(j)\beta_t(j)\{O_{t,1} - O_{t,0} + h\int_0^1 O_{t,\tau}d\tau\}}{\sum_{t=1}^T \alpha_t(j)\beta_t(j)h}.\end{aligned}$$

The derivative with respect to h inside the curly brackets gives

$$\int_0^1 \left[\dot{O}_{t,\tau} - h(k - O_{t,\tau})\right](O_{t,\tau} - k)d\tau - \frac{1}{2} = -k(O_{t,1} - O_{t,0}) + \int_0^1 \dot{O}_{t,\tau}O_{t,\tau}d\tau + h\int_0^1 (k - O_{t,\tau})^2 d\tau - \frac{1}{2}.$$

Thus,

$$\tilde{c}_j = \frac{\sum_{t=1}^T \alpha_t(j)\beta_t(j)\left[\frac{1}{2} + k(O_{t,1} - O_{t,0}) - \int_0^1 \dot{O}_{t,\tau}O_{t,\tau}d\tau\right]}{\sum_{t=1}^T \alpha_t(j)\beta_t(j)\left[\int_0^1 (k - O_{t,\tau})^2 d\tau\right]}.$$

In practice, we reparametrize the OU process to stabilize this nonlinear optimization over the two parameters, which is discussed in Section 5.2.

4.3 Fractional Brownian Motion

In Moret and Nualart (2002), the Onsager-Machlup functional is derived for fractional Brownian motion in the *Singular Case*, which is where the Hurst parameter is $\frac{1}{4} < \nu < \frac{1}{2}$, and the *Regular Case* where the Hurst parameter is $\nu > \frac{1}{2}$. Note that for $\nu = \frac{1}{2}$, we have standard Brownian motion. The process considered is

$$Y_\tau = y + W_\tau^\nu + \int_0^\tau r(Y_s)ds$$

where $r \in C_b^2(\mathbb{R})$, the space of bounded functions with two continuous derivatives.

The Onsager-Machlup functional for the singular case from Theorem 7 in Moret and Nualart (2002) is

$$\log\left[\lim_{\varepsilon\to 0}\frac{\mathrm{P}\left(\|Y-\Phi\|<\varepsilon\right)}{\mathrm{P}\left(\|W^\nu\|<\varepsilon\right)}\right] = -\frac{1}{2}\int_0^1\left\{\dot{\Phi}_\tau - \tau^{-\upsilon}I_{0^+}^\upsilon \tau^\upsilon r(\Phi_\tau)\right\}^2 d\tau - \frac{1}{2}d_\nu\int_0^1 r'(\Phi_\tau)d\tau$$

where $K^\nu\dot{\Phi} = \Phi - x$, $\upsilon = |\nu - 1/2|$, K^ν is the operator such that $dW_t^\nu = K^\nu(t,s)dW_s$,

$$d_\nu = \sqrt{\frac{2\nu\Gamma(3/2-\nu)\Gamma(\nu+1/2)}{\Gamma(2-2\nu)}},$$

and $I_{a^+}^\upsilon f(x) = \Gamma(\upsilon)^{-1}\int_a^x (x-y)^{\upsilon-1}f(y)dy$ is called the *left fractional Riemann-Liouville Integral*.

In the simplest non-trivial setting of $r = c \in \mathbb{R}$ and fixed $\upsilon \in (0, 1/4)$, we aim to solve for the drift term c such that

$$\tilde{c}_j = \underset{h\in\mathbb{R}}{\arg\min}\sum_{t=1}^T \alpha_t(j)\beta_t(j)\left\{\frac{1}{2}\int_0^1\left\{\dot{\Phi}_\tau - h\tau^{-\upsilon}I_{0^+}^\upsilon\tau^\upsilon\right\}^2 d\tau\right\}.$$

In this case, the integral $I_{0^+}^\upsilon\tau^\upsilon$ is a scaled Beta function and $\tau^{-\upsilon}I_{0^+}^\upsilon\tau^\upsilon = \tau^\upsilon\Gamma(\upsilon+1)/\Gamma(2\upsilon+1)$. Thus, some simple calculus results in

$$\tilde{c}_j = \frac{\Gamma(2\upsilon+2)}{\Gamma(\upsilon+1)}\frac{\sum_{t=1}^T\alpha_t(j)\beta_t(j)\int_0^1\tau^\upsilon\dot{O}_{t,\tau}d\tau}{\sum_{t=1}^T\alpha_t(j)\beta_t(j)}.$$

Setting $\upsilon = 0$ in the above returns us to the formula for $\tilde{c}_j$ derived from Brownian motion with drift.

Similarly, the Onsager-Machlup functional for the regular case, $\nu > 1/2$, from Theorem 8 in Moret and Nualart (2002) is

$$\log\left[\lim_{\varepsilon\to 0}\frac{\mathrm{P}\left(\|Y-\Phi\|<\varepsilon\right)}{\mathrm{P}\left(\|W^\nu\|<\varepsilon\right)}\right] = -\frac{1}{2}\int_0^1\left\{\dot{\Phi}_\tau - \tau^{\omega}D_{0^+}^\omega\tau^{-\omega} r(\Phi_\tau)\right\}^2 d\tau - \frac{1}{2}d_\nu\int_0^1 r'(\Phi_\tau)d\tau$$

for $\omega = \nu - 1/2$ and $D_{0^+}^\omega$ is the left-sided Riemann-Liouville derivative defined as

$$D_{a^+}^\omega f(x) = \frac{1}{\Gamma(1-\omega)}\frac{d}{dx}\int_a^x\frac{f(y)}{(x-y)^\omega}dy.$$

If we consider the linear drift setting of $r = c \in \mathbb{R}$, then $\tau^\omega D_{0^+}^\omega\tau^{-\omega} = (1-2\omega)\tau^{-\omega}\Gamma(1-\omega)/\Gamma(2-2\omega)$. A little more calculus gives us

$$\tilde{c}_j = \frac{\Gamma(2-2\omega)}{\Gamma(1-\omega)}\frac{\sum_{t=1}^T\alpha_t(j)\beta_t(j)\int_0^1\tau^{-\omega}\dot{O}_{t,\tau}d\tau}{\sum_{t=1}^T\alpha_t(j)\beta_t(j)},$$

which coincides nicely with the singular case as $-\omega = \upsilon$.

Applying our THMM algorithm to fractional Brownian motion extends the range of possible stochastic processes we can consider. When the Hurst parameter $\nu > 1/2$, the process has positively correlated increments and thus appears smoother than standard Brownian motion. In comparison, processes with $\nu < 1/2$ exhibit negatively correlated increments and thus appear rougher than standard Brownian motion.

4.4 Non-Parametric State Means

The previous sections consider estimation of specific real valued parameters under different stochastic models. However, a more flexible approach is to treat estimation of the means non-parametrically. This is achieved by using the formulae derived from the Onsager-Machlup functional at the beginning of Section A.2 directly.

Indeed, the emission function and mean for state j can be shown to be

$$b_j(O_t) = \exp\left(-\frac{1}{2}|O_t - h_j|_H^2\right) \quad \text{and} \quad h_j = \frac{\sum_{t=1}^T \alpha_t(j)\beta_t(j)O_t}{\sum_{t=1}^T \alpha_t(j)\beta_t(j)},$$

respectively. To see the latter equation, let R_γ be the covariance operator for Gaussian measure γ. We note that

$$\begin{aligned}\sum_{t=1}^T \alpha_t(j)\beta_t(j)|O_t - h_j|_H^2 &= \sum_{t=1}^T \alpha_t(j)\beta_t(j)R_\gamma(O_t^* - h_j^*)(O_t^* - h_j^*) \\ &= \sum_{t=1}^T \alpha_t(j)\beta_t(j)\int_X (O_t^* - h_j^*)^2\gamma(dx),\end{aligned}$$

which is minimized by $h_j^* = \sum_{t=1}^T \alpha_t(j)\beta_t(j)O_t^*/\sum_{t=1}^T \alpha_t(j)\beta_t(j)$. Applying the operator R_γ to each side recovers the optimal h_j.

In practice, one must select a suitable Gaussian measure / Cameron-Martin space for the problem at hand. For example, in Section 6, we analyze a sequence of Electroencephalogram (EEG) signals under the standard Wiener measure. In this setting, $H = W_0^{2,1}[0,1]$, the Sobolev space of absolutely continuous functions h such that $\dot{h} := \partial h(\tau)/d\tau \in L^2[0,1]$ with $h(0) = 0$. As a consequence, the emission function becomes

$$b_j(O_t) = \exp\left(-\frac{1}{2}\int_0^1 \left|\dot{O}_t(\tau) - \dot{h}_j(\tau)\right|^2 d\tau\right).$$

However, it is worth emphasizing that other Cameron-Martin norms can be considered and may improve performance of algorithm.

5. Simulated Data Analysis

In the following sections, we test our THMM algorithm on a variety of simulated data sets from three different settings: Brownian motion with linear drift, the Ornstein-Uhlenbeck process, and fractional Brownian motion with linear drift. To evaluate its performance accuracy, we use the adjusted Rand index (ARI) as our performance metric. The ARI is a popular method of measuring the agreement between two sets of labels and is computed in R via the `adjustedRandIndex()` function in the `mclust` package (Scrucca et al., 2016). An ARI value of 1 indicates a perfect match whereas an ARI value of 0 indicates random guessing. We secondly compare clustering accuracy using cross entropy as an alternative measurement.

For each of the following simulations, it is possible to concoct an algorithm specifically designed to perform well on that specific data set; e.g. via feature selection. The true power

of the THMM approach is that it is generally applicable and adaptable to all of these settings of interest as well as others not considered in this work. R code to recreate these simulations can be found at `https://github.com/cachelack/Topological-Hidden-Markov-Model.git`. This also includes the THMM variants of the Baum-Welch and Viterbi algorithms.

Nevertheless, for the sake of comparison, we use functional Principal Components Analysis (fPCA) to project the $T = 200$ samples onto the first k principal components. Then, a classic multivariate Gaussian HMM is fit to this $\mathbb{R}^k$-valued data using the R package `mhsmm` (O'Connell and Højsgaard, 2011).

5.1 Brownian Motion with Drift

For a simple setting to test the THMM algorithm, we simulate a sequence of $T = 200$ Brownian sample paths with 5 different states corresponding to different drift parameters. For the "low separation" simulation, the drift parameters are $-4, -2, 0, 2, 4$. For the "medium separation" simulation, the drift parameters are $-8, -4, 0, 4, 8$. We consider two different transition matrices,

$$A_1 = \begin{pmatrix} 0.64 & 0.09 & 0.09 & 0.09 & 0.09 \\ 0.09 & 0.64 & 0.09 & 0.09 & 0.09 \\ 0.09 & 0.09 & 0.64 & 0.09 & 0.09 \\ 0.09 & 0.09 & 0.09 & 0.64 & 0.09 \\ 0.09 & 0.09 & 0.09 & 0.09 & 0.64 \end{pmatrix} \text{ and } A_2 = \begin{pmatrix} 0.04 & 0.44 & 0.04 & 0.04 & 0.44 \\ 0.44 & 0.04 & 0.44 & 0.04 & 0.04 \\ 0.04 & 0.44 & 0.04 & 0.44 & 0.04 \\ 0.04 & 0.04 & 0.44 & 0.04 & 0.44 \\ 0.44 & 0.04 & 0.04 & 0.44 & 0.04 \end{pmatrix},$$

where A_1 makes the Markov chain remain in the same state with high probability whereas A_2 results in more state switching behaviour. In latter simulations, we only consider A_1 for generating data as the performance of all methods considered does not appear to vary for transition matrices A_1 and A_2. Examples of this data are displayed in Figure 2.

The results of running the Baum-Welch and Viterbi algorithms making use of the Onsager-Machlup functional for Brownian motion with drift, see Section 4.2, are displayed in Tables 1 and 2. The performance of the THMM algorithm outperforms the combination of fPCA and the classic HMM in each of the four cases in both ARI and cross entropy. In each test, two principal components were used for the fPCA HMM approach, which gave the strongest performance after considering $1, \ldots, 5$ principal components.

We do note that our algorithm run in this setting is specifically designed to detect Brownian motion with drift and, in fact, estimates the drift parameters with high accuracy. If we did not wish to restrict ourselves to a such parametric model, we could fit a THMM model non-parametrically using the L^2 norm distance. For the data generated by matrix A_1, we have ARI values of 0.452 and 0.806 for low and medium separation, respectively. Meanwhile, Table 1 reports ARIs of 0.528 and 0.894 for the parametric setting. Hence, only a little accuracy is lost for not a priori knowing the correct model to fit to the data. For completeness, the same behaviour is seen for transition matrix A_2 where the ARIs for non-parametric fitting are 0.453 and 0.889 compared to the parametric values of 0.471 and 0.952 for low and medium separation, respectively.

In the simulations to follow, we only report findings for transition matrix A_1. This is because the performances of the methods under comparison are agnostic to this choice of transition matrix. Also, matrix A_1 models data that likely remains in a given state for more

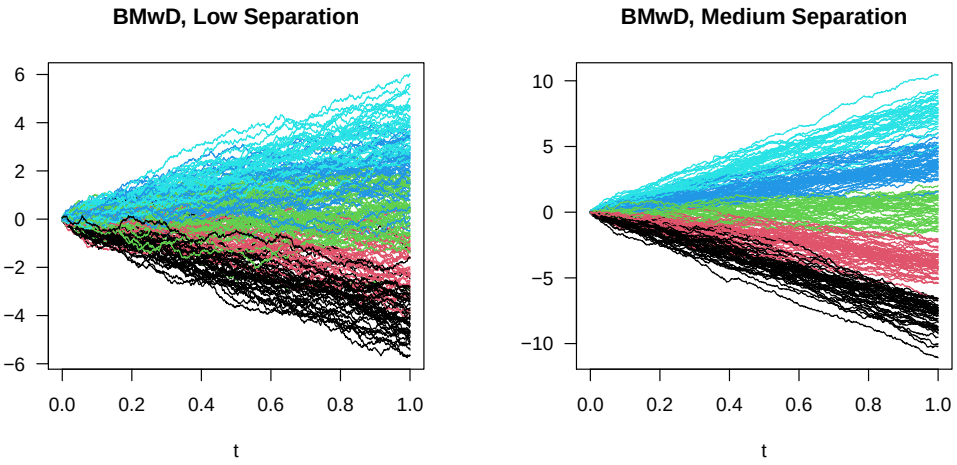

Figure 2: Simulated Brownian motion with five states with drift parameters $(-4, -2, 0, 2, 4)$ on the left and $(-8, -4, 0, 4, 8)$ on the right.

timesteps before switching to a new state, which is what we expect from our motivating pediatric obstructive sleep apnea data set.

5.2 Ornstein-Uhlenbeck Process

The one-dimensional Ornstein-Uhlenbeck process has the form

$$dY_\tau = c(\mu - Y_\tau)d\tau + dW_\tau$$

with two parameters. The mean parameter μ is where the process tends to in the long run, and the concentration parameter c determines how tightly the process fluctuates around its mean. However, this form is numerically challenging to optimize. Instead, the THMM estimates the transformed variables $b_0 = c\mu$ and $b_1 = c$. This is implemented in `R` via the `optim` function with the `L-BFGS-B` method. The function to minimize is

$$u(b_0, b_1) = \frac{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j) \left\{ \frac{1}{2} \int_0^1 \{\dot{O}_{t,\tau} - (b_0 - b_1 O_{t,\tau})\}^2 d\tau - \frac{b_1}{2} \right\}}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}$$

with derivatives

$$\frac{\partial u}{\partial b_0} = -\frac{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j) \left\{ \int_0^1 \{\dot{O}_{t,\tau} - (b_0 - b_1 O_{t,\tau})\} d\tau \right\}}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}$$

$$\frac{\partial u}{\partial b_1} = \frac{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j) \left\{ \int_0^1 \{\dot{O}_{t,\tau} - (b_0 - b_1 O_{t,\tau})\} O_{t,\tau} d\tau - \frac{1}{2} \right\}}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}$$

where we divide everything by $\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)$ for numerical stability reasons.

THMM with BMWD						
Low Separation						
	True States					Est
	A	B	C	D	E	drift
a	26	6	.	.	.	-3.98
b	10	27	10	.	.	-1.77
c	.	2	8	10	.	0.85
d	.	.	10	41	2	1.89
e	.	1	.	4	43	3.81
drift	-4	-2	0	2	4	
ARI: 0.528			**Ent: 2.859**			

THMM with BMWD						
Medium Separation						
	True States					Est
	A	B	C	D	E	drift
a	36	2	.	.	.	-7.99
b	.	33	4	.	.	-3.57
c	.	1	23	.	.	-0.03
d	.	.	1	53	.	3.72
e	.	.	.	2	45	7.94
drift	-8	-4	0	4	8	
ARI: 0.894			**Ent: 0.463**			

HMM with fPCA					
Low Separation					
True States					
	A	B	C	D	E
a	26	9	.	.	.
b	.	16	6	.	.
c	6	6	1	.	.
d	4	3	3	.	.
e	.	2	18	55	45
ARI: 0.338			**Ent: 2.277**		

HMM with fPCA					
Medium Separation					
True States					
	A	B	C	D	E
a	18	17	5	.	.
b	14	13	2	.	.
c	4	6	5	2	.
d	.	.	16	52	3
e	.	.	.	1	42
ARI: 0.534			**Ent: 3.070**		

Table 1: Confusion matrices showing the alignment of true states (A-E) with estimated states (a-e) for transition matrix A_1 for both the THMM (top) and fPCA HMM (bottom). The left is low separation, and the right is medium separation. The ARI and cross entropy are listed at the bottom of each table.

THMM with BMWD						
Low Separation						
	True States					Est
	A	B	C	D	E	drift
a	38	8	.	.	.	-3.57
b	2	30	20	.	.	-1.47
c	.	3	17	9	2	0.51
d	.	.	7	27	8	2.31
e	.	.	.	1	28	4.00
drift	-4	-2	0	2	4	
ARI: 0.471			**Ent: 2.56**			

THMM with BMWD						
Medium Separation						
	True States					Est
	A	B	C	D	E	drift
a	40	1	.	.	.	-7.79
b	.	40	.	.	.	-3.80
c	.	.	44	1	.	-0.16
d	.	.	.	35	1	4.19
e	.	.	.	1	37	7.87
drift	-8	-4	0	4	8	
ARI: 0.952			**Ent: 0.154**			

HMM with fPCA					
Low Separation					
True States					
	A	B	C	D	E
a	33	11	1	.	.
b	.	18	19	3	.
c	.	3	21	29	10
d	7	9	2	1	.
e	.	.	1	4	28
ARI: 0.347			**Ent: 3.32**		

HMM with fPCA					
Medium Separation					
True States					
	A	B	C	D	E
a	39	1	.	.	.
b	1	38	.	.	.
c	.	2	44	1	.
d	.	.	.	35	.
e	.	.	.	1	38
ARI: 0.925			**Ent: 0.226**		

Table 2: Confusion matrices showing the alignment of true states (A-E) with estimated states (a-e) for transition matrix A_2 for both the THMM (top) and fPCA HMM (bottom). The left is low separation, and the right is medium separation. The ARI and cross entropy are listed at the bottom of each table.

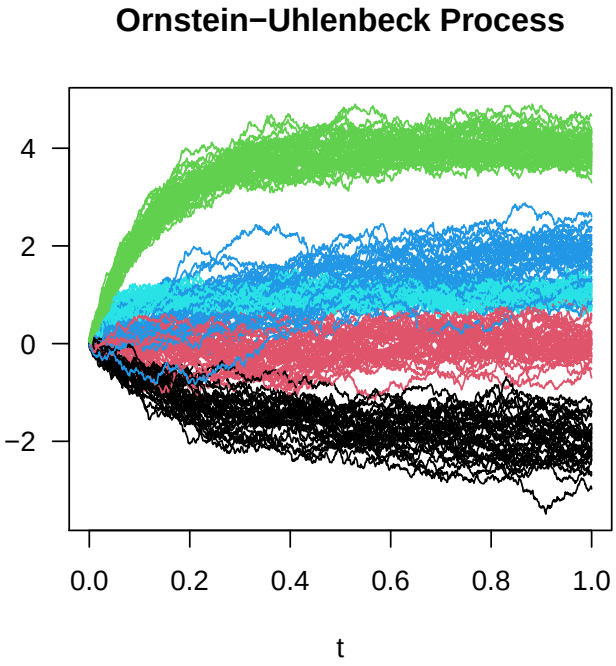

Figure 3: Simulated Ornstein-Uhlenbeck Processes with 5 states, mean parameters $\mu = (-2, 0, 4, 2, 1)$ and concentration parameters $c = (4, 4, 8, 2, 20)$.

In this simulation, five states were once again used to generate data with means $\mu = (-2, 0, 4, 2, 1)$ and $c = (4, 4, 8, 2, 20)$. The transition matrix is A_1 from above for Brownian motion with drift, and $T = 200$ again. The OU sample paths coloured by their state are displayed in Figure 3.

Table 3 displays the results of the Baum-Welch and Viterbi algorithms on this simulated data set as well as applying fPCA with 2 principal components and then fitting a classic HMM model on the data projected onto $\mathbb{R}^2$. In this setting, the THMM had slightly higher ARI and cross entropy values when compared to the fPCA HMM approach. Thus, their performances in this setting are comparable. Our algorithm had a hard time differentiating between states D and E with parameters $(2, 2)$ and $(1, 20)$. Of note, the THMM algorithm seems to recover the mean parameters with high accuracy, but does not perform as well with estimation of the concentration parameters for the OU process.

When we collapse states D and E into a single state and fit a 4-state THMM model and fPCA-HMM model to this data, we get almost perfect performance from the THMM whereas the fPCA HMM performs worse than before. These results are in Table 4.

OU Process via THMM								**OU Process via fPCA HMM**						
	True States					Estimated			**True States**					
	A	B	C	D	E	mean	conc		A	B	C	D	E	
a	40	.	.	.	.	-1.96	3.95	a	30	25	.	.	.	
b	.	38	.	.	.	-0.33	6.39	b	6	11	.	.	.	
c	.	.	44	.	.	4.10	5.15	c	.	.	28	.	.	
d	.	.	.	20	.	2.70	5.30	d	.	.	.	54	.	
e	.	3	.	17	38	1.22	5.54	e	.	.	.	1	45	
mean	-2	0	4	2	1									
conc	4	4	8	2	20									
ARI: 0.806			**Ent: 0.997**					**ARI: 0.797**			**Ent: 0.816**			

Table 3: Confusion matrix showing the alignment of true states (A-E) with estimated states (a-e) for the OU process data using transition matrix A_1.

OU Process via THMM							**OU Process via fPCA HMM**					
	True States				Estimated			**True States**				
	A	B	C	D	mean	conc		A	B	C	D	
a	36	0	0	0	-1.95	3.69	a	36	0	0	0	
b	0	36	0	1	-0.08	5.67	b	0	36	0	44	
c	0	0	28	0	3.13	4.35	c	0	0	28	0	
d	0	0	0	99	1.78	4.85	d	0	0	0	56	
ARI: 0.985			**Ent: 0.014**				**ARI: 0.525**			**Ent: 0.837**		

Table 4: Confusion matrix for the same data as from Table 3 but with states D and E merged into a single state D.

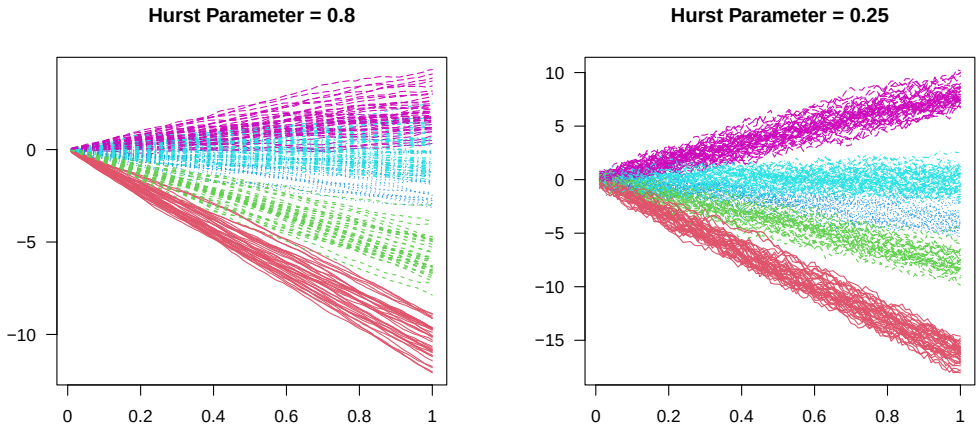

Figure 4: Example sample paths of fractional Brownian motion with Hurst parameter of 0.8 (left) and 0.25 (right).

5.3 Fractional Brownian Motion

To test the THMM algorithm applied to fractional Brownian motion, we first simulate 200 sample paths with a Hurst parameter of $\nu = 0.8$, which gives a Gaussian process with positively correlated increments, i.e. it is *smoother* than the standard Wiener process. Simulation was achieved by first simulating Gaussian white noise and transforming it based on the covariance function

$$\operatorname{cov}\left(Y_{\tau_1}, Y_{\tau_2}\right) = \frac{1}{2}\left(\tau_1^{2\nu} + \tau_2^{2\nu} - |\tau_1 - \tau_2|^{2\nu}\right).$$

This data is displayed in Figure 4. The same transition matrix A_1 was used and the 5-long vector of drift terms is $c = (-10, -6, -2, 0, 2)$. In this work, we treat the Hurst parameter as a tunable input to the THMM algorithm rather than a parameter to be estimated from the data. Hence, we run the THMM algorithm for $\nu = 0.25, 0.5, 0.8$ to compare performance.

Table 5 shows the results of this simulation. Choosing the Hurst parameter to be $\nu = 0.5$ gave the best ARI and cross entropy values. However, these values were extremely close to choosing $\nu = 0.8$ being the "right" choice in the sense of coinciding with how the data was generated. The choice of $\nu = 0.25$ performed much worse; this is expected as the algorithm is expecting much rougher sample paths than what it is given. Lastly, the fPCA approach with, once again, 2 principal components has worse performance than any of the THMM runs.

We repeat the same experiment but for fractional Brownian motion with a Hurst parameter of 0.25, which gives negatively correlated increments and *rougher* looking paths. The true drift coefficients are set to be $(-16, -8, -4, 0, 8)$ as the paths are harder to distinguish than in the previous simulation due to the added roughness of the sample paths. The results from Table 6 show that setting the Hurst parameter in the THMM algorithm

	Drift Vector					ARI	Ent
Truth	-10.	-6.	-2.	0.	2.		
$\nu = 0.25$	-13.42	-13.93	-10.65	-2.89	1.56	0.399	1.70
$\nu = 0.5$	-10.21	-5.72	-1.42	0.06	1.98	**0.692**	**1.19**
$\nu = 0.8$	-9.77	-5.43	-1.16	0.15	1.95	0.676	1.23
fPCA						0.342	1.98

Table 5: A comparison of the THMM algorithm run on fractional Brownian motion with Hurst parameter of 0.8 for different choices of ν as an input to the algorithm.

	Drift Vector					ARI	Ent
Truth	-16.	-8.	-4.	0.	8.		
$\nu = 0.25$	-19.22	-18.63	-7.82	-0.78	8.98	0.677	1.44
$\nu = 0.5$	-15.65	-7.58	-4.78	-0.17	7.72	**0.741**	**1.23**
$\nu = 0.8$	-14.62	-6.13	-2.66	0.15	7.62	0.662	1.70
fPCA						0.540	2.52

Table 6: A comparison of the THMM algorithm run on fractional Brownian motion with Hurst parameter of 0.25 for different choices of ν as an input to the algorithm.

to the $\nu = 1/2$ once again gave the best ARI and cross entropy values. The other THMM runs with $\nu = 0.25, 0.8$ gave comparable performance whereas fPCA, once again, displayed the worst performance.

Of note, in both Table's 5 and 6, the THMM algorithm with $\nu = 0.25$ results in the first two states coinciding and thus harming the performance. Furthermore, superior performance is shown to occur when we keep $\nu = 1/2$, which assumes standard Brownian motion. However, future investigations into fractional Brownian motion in the context of the THMM algorithm may yield other results. Certainly, data driven state-based estimation of the Hurst parameter would be of future interest.

5.4 Nonparametric Fitting and k-means

Often, we do not wish to impose a parametric model like Brownian motion with drift and instead want only for the THMM algorithm to find the most representative mean curve for each Markov state. To test the THMM's ability to fit such models, we generate $t = 1, \ldots, 200$ phase-shifted sinusoidal curves of the form

$$O_t(\tau) = \sin(2\pi(\tau + 0.2(j-1))) + \epsilon_t(\tau)$$

for $\tau \in [0, 1]$, state $j = 1, \ldots, 5$, and error term

$$\epsilon_t(\tau) = \sqrt{2} \sum_{k=1}^{16} Z_k \frac{\sin(k\pi\tau)}{k\pi}$$

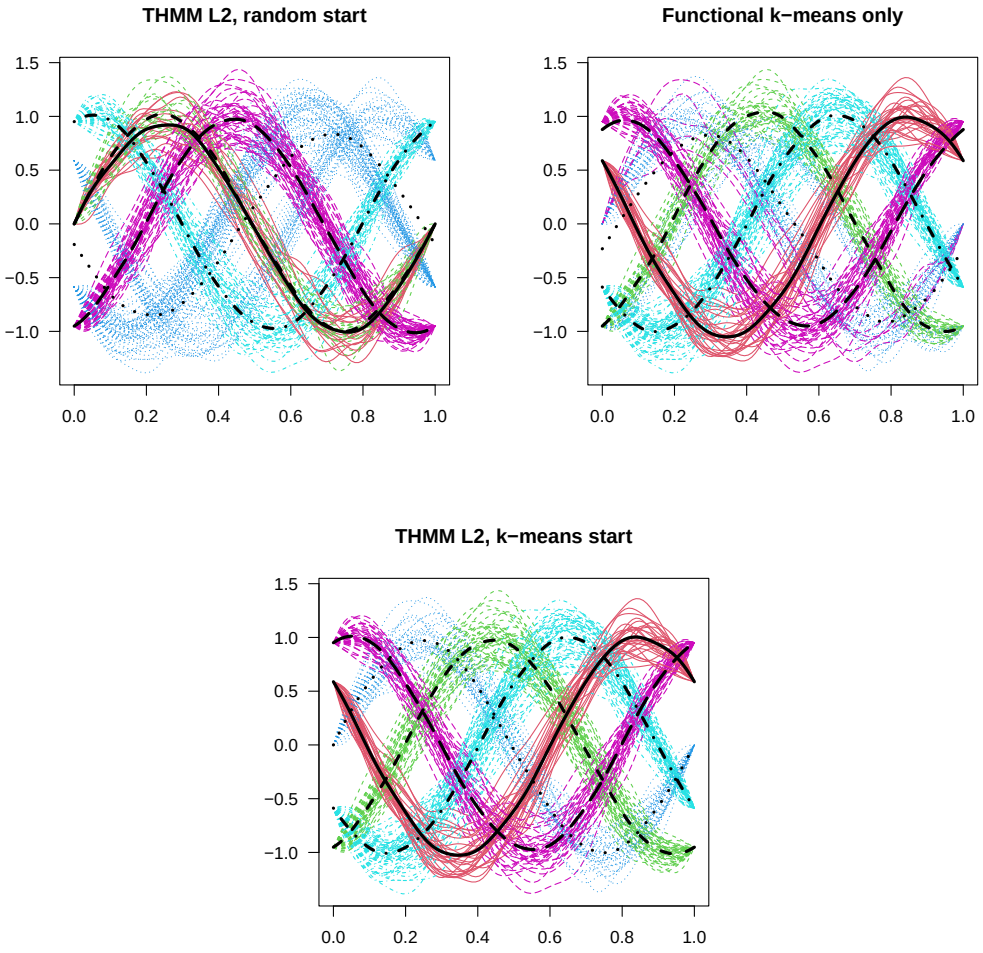

Figure 5: The generated sinusoidal data plotted along with the cluster mean curves.

where $Z_k \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.16)$ making ϵ_t the truncated Karhunen-Loeve expansion for a Brownian Bridge process. The five Markov states evolve with respect to the same transition matrix A_1 from above. The data so generated is displayed in Figure 5 along with the state mean curves fitted by three different methods. The results of all seven different methods applied to this simulated data set are displayed in Table 7

The choice of Cameron-Martin norm will drastically affect the performance of the THMM algorithm, and thus care must be shown when choosing the norm to use. In this example, imposing the L^2 norm,

$$d_{L^2}(O_t, h_j) = \int_0^1 |O_t - h_j|^2 d\tau,$$

results is a very quick convergence of the THMM algorithm in 264 iterations of the Baum-Welch algorithm. However, this version of the algorithm typically gets stuck in a sub-

Method	ARI	Ent	Iterations
Functional k-means	0.823	0.708	
THMM L^2	0.743	0.922	264
THMM $W_0^{2,1}$	0.000	4.173	10,000
fPCA & HMM	0.549	1.067	
THMM L^2 with k-means	**1.000**	**0.000**	11
THMM $W_0^{2,1}$ with k-means	0.001	0.119	10,000
fPCA & HMM with k-means	0.808	0.322	

Table 7: A comparison of methods for clustering the simulated sinusoidal data. The THMM under the L^2-norm starting from the output of functional k-means achieves perfect clustering.

optimal critical point with little hope of escaping. The problem of getting stuck in a local optimum—a common occurrence with EM-type algorithms—can be alleviated by first running functional k-means clustering and then implementing the THMM algorithm with the k-means output as starting vectors. Using k-means as a preprocessing step before a more complex EM-type algorithm is common in the clustering literature. In this case, the `kmeans.fd` function from the `fda.usc` R package (Febrero-Bande and Oviedo de la Fuente, 2012) was first applied and achieved an ARI of 0.823. Then, the k-means output was inputted into the THMM algorithm under the L^2 norm, which increased the ARI to a perfect 1.0 in only 11 Baum-Welch iterations.

In contrast to L^2, we can use the Sobolev $W_0^{2,1}$ norm imposing the metric

$$d_{W_0^{2,1}}(O_t, h_j) = \int_0^1 |\dot{O}_t - \dot{h}_j|^2 d\tau$$

on the THMM algorithm. Unlike in the L^2 case, under this norm, the algorithm was run for 10,000 iterations both with a random start and with starting state-means determined by functional k-means. In both cases, it completely failed to cluster this simulated data.

Lastly, our THMM was once again compared to the classic multivariate HMM with fPCA. In this simulation, the best results were achieved by setting the number of principal components to 4. The HMM was fit with both random starting mean vectors (ARI of 0.549) and with k-means determined mean vectors (ARI 0.808). The drastic increase in the ARI coincides with what occurred for the THMM algorithm under the L^2 norm, and demonstrates the utility in using an algorithm like k-means to find a good starting point for the Baum-Welch or other EM-type algorithms.

6. Pediatric Obstructive Sleep Apnea Data

This section demonstrates the performance of the THMM on a real data set of critical importance to the health of pediatric patients. Obstructive sleep apnea is a chronic condition characterized by frequent episodes of upper airway collapse during sleep, which over time can be detrimental to one's health.

6.1 Data Overview

The gold standard for diagnosis of obstructive sleep apnea in children is by overnight polysomnography in a hospital or sleep clinic. Polysomnography provides multi-channel time series including an electroencephalogram (EEG), electrocardiography (ECG), electrooculography (EOG), and electromyography (EMG). Even for a single patient, this data is vast and should eventually be considered jointly to label sleep states and identify sleep disorders. However to illustrate application of the THMM algorithms in this work, we chose one channel of EEG from one patient labelled as CF050 to be used as a proof-of-concept. This patient was selected from a group of seventy four pediatric patients with potential obstructive sleep apnea in a clinical study, Pro00057638, approved by University of Alberta. The sampling rate for EEG is 512 samples per second. Each signal was split into a sequence of epochs, i.e. 30 second intervals. These signals were transformed into power spectral densities (PSD) using Welch's method (Welch, 1967) for each epoch. The reason for dividing time series into 30 second intervals is to group them with respect to five sleep stages: wake, rapid eye-movement (REM), and non-rapid eye-movement (NREM). The NREM category is further divided into three states: NREM1, NREM2, NREM3. The sleep stages are labelled per epoch by a sleep technician. For every sleep stage not labelled as *wake*, the patient is considered to be asleep. The spectral densities for each epoch may then be used to identify possible underlying states such as the sleep stages. There were 948 epochs in patient CF050.

Markov models and HMMs have a long history of being applied to the modelling of sleep states (Zung et al., 1965; Yang and Hursch, 1973; Kemp and Kamphuisen, 1986). The work of Penny and Roberts (1998) considered an HMM with Gaussian observations for EEG analysis. This was followed by similar analyses in Flexer et al. (2002, 2005). More recent work includes Doroshenkov et al. (2007), Pan et al. (2012), and Chen et al. (2015). Typically, these approaches are based on various discretization and feature selection methods whose output is then fed into the classic HMM. Our THMM takes the entire power spectral density curve into account and hence obviates the need for such feature selection steps. Note that the following data analysis is meant as a proof-of-concept for the proposed methodology; a comprehensive analysis of the full multichannel data set is left to future work.

6.2 Raw Data Analysis

The EEG PSD curves for patient CF050 was run through the THMM variant of the Baum-Welch algorithm under Wiener measure and the $H = W_0^{2,1}[0,1]$ norm. The first experiment combined REM and the NREM states into one category "asleep" to contrast with the "awake" category. Figure 6 shows the data and the predicted vs the true state means both for the fully-functional THMM model and the fPCA + HMM model. The estimated mean curves are very similar to the true mean curves. The estimated Markov transition matrix and the "true" transition probabilities are, respectively,

$$\hat{A}_{W_0^{2,1}} = \begin{pmatrix} 0.910 & 0.090 \\ 0.176 & 0.824 \end{pmatrix} \quad \text{and } A = \begin{pmatrix} 0.933 & 0.067 \\ 0.025 & 0.975 \end{pmatrix}$$

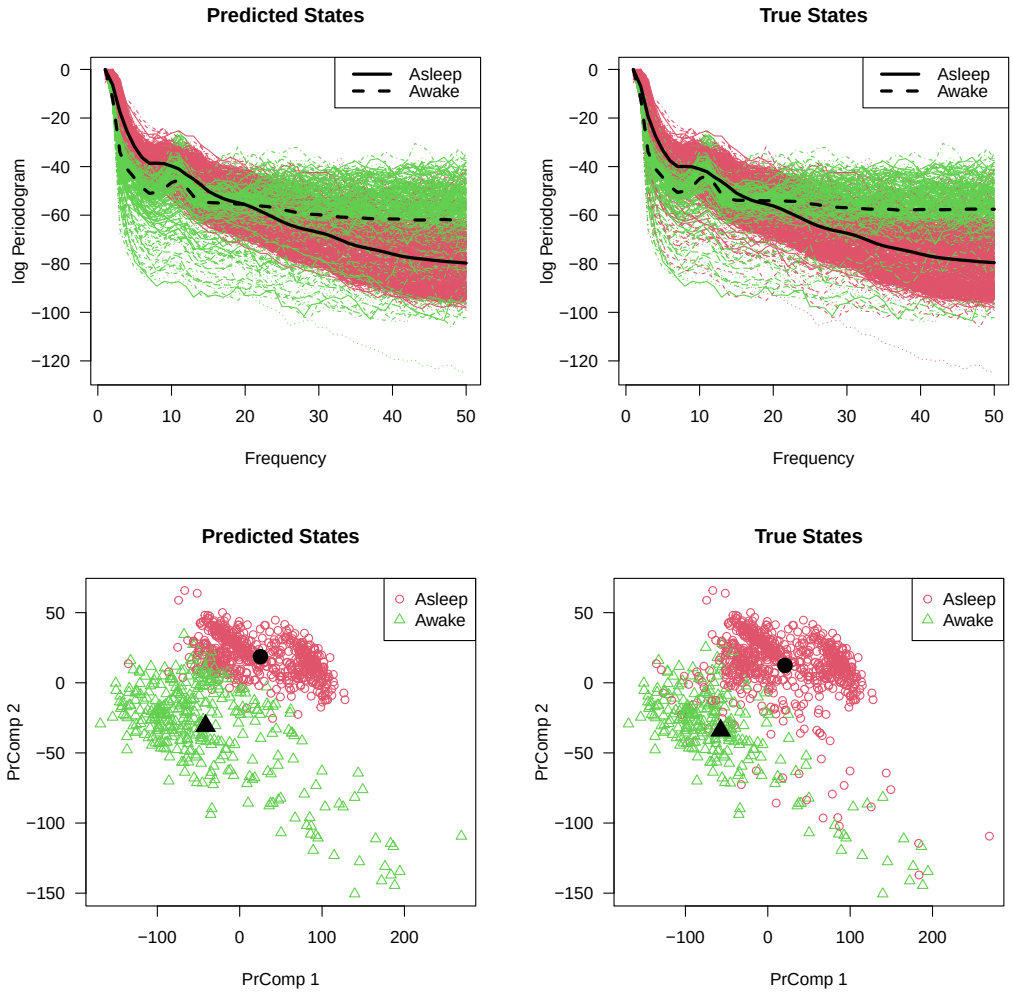

Figure 6: A comparison of the predicted and true states and means for the EEG PSD data set over the first 50 frequencies using the THMM model under the $W_0^{2,1}$ norm (top row) and using fPCA with 2 principal components and the multivariate HMM (bottom row).

where A was computed by counting the number of transitions between states as labelled by the sleep technician. It is worth noting that patients typically remain asleep or awake for many sequential epochs, i.e. 30-second time segments.

The leftside columns of Tables 8 and 9 display the results of two-state clustering of this raw OSA data under a variety of methods. The best performer with respect to ARI is the THMM algorithm under the $W_0^{2,1}$ norm. The best performer with respect to cross entropy is fPCA with the multivariate HMM when two principal components are used. The two functional k-means methods have poor performance, but these do not take into account temporal dependence in the data. Hence, this reinforces the necessity of using models that

		True States			
		Raw Data		**Smoothed Data**	
Method	Predicted	Awake	Asleep	Awake	Asleep
k-means L^2	Awake	214	233	212	232
	Asleep	39	462	41	463
k-means $W_0^{2,1}$	Awake	223	686	228	683
	Asleep	30	9	25	12
fPCA HMM	Awake	243	115	243	118
	Asleep	10	580	10	577
THMM L^2	Awake	223	367	223	366
	Asleep	30	328	30	329
THMM $W_0^{2,1}$	Awake	247	75	241	46
	Asleep	6	620	12	649
THMM $W_0^{2,2}$	Awake	250	313	247	46
	Asleep	3	382	6	649

Table 8: Confusion matrices for the results of fitting a two-state THMM to raw and kernel smoothed EEG PSD curves, left and right, respectively.

take temporal ordering into account. For the THMM model, the choice of norm makes a large difference in performance. Unlike in the simulated data, functional k-means was not used to seed the THMM algorithm as their performances were quite poor on this data.

As an additional experiment under the $W_0^{2,1}$, a particularly poor starting point for the Baum-Welch algorithm was chosen so that the progress of the algorithm could be tracked with respect to both the likelihood and the ARI. The left plot in Figure 7 shows how the likelihood grows over the initial iterations only to plateau around iteration 10. However, it begins to climb to a new plateau after 30-40 iterations. In unison, the right plot shows improvement in the ARI from below zero to its final value of 0.68. The final confusion matrix is displayed in Table 8. Most alternative starting points for the algorithm converged to the same final parameters. In some cases, the ARI rose above 0.680 only to fall back to it. Hence, it is worth emphasizing that each iteration of the Baum-Welch algorithm will increase the likelihood but may result in an increase or a decrease in the ARI.

A comparison of the predicted state sequence with the true state sequence is featured in Figure 8 where the states are predicted via the Viterbi algorithm from the THMM with the $W_0^{2,1}$ norm. The figure highlights the fact that the expected duration that the HMM remains in a given state follows a geometric distribution. Hence, the classic HMM will switch states more frequently than what may be observed in practice. Variable state duration or hidden semi-Markov models allow for longer sojourns in a given state and may improve the predicted state sequence for such EEG data as well as others. It is also possible to soft-cluster the data rather than making a hard decision on to which state the observation belongs. However, in this case, only 6 of the 948 observations had a probability between 10% and 90% of being in the awake state. That is, most of the predicted a posteriori state probabilities were numerically equal to either 1 or 0.

	Raw Data			Smoothed Data		
Method	ARI	Ent	Iterations	ARI	Ent	Iterations
k-means L^2	0.180	1.057		0.178	1.059	
k-means $W_0^{2,1}$	0.095	0.785		0.072	0.981	
fPCA HMM	0.536	**0.319**		0.527	0.327	
THMM L^2	0.013	0.811	12	0.014	0.812	12
THMM $W_0^{2,1}$	**0.680**	0.396	30	0.762	0.278	12
THMM $W_0^{2,2}$	0.104	0.858	19	**0.786**	**0.260**	15

Table 9: ARI, cross entropy, and number of iterations for comparison of two-state clustering methods for the pediatric OSA data.

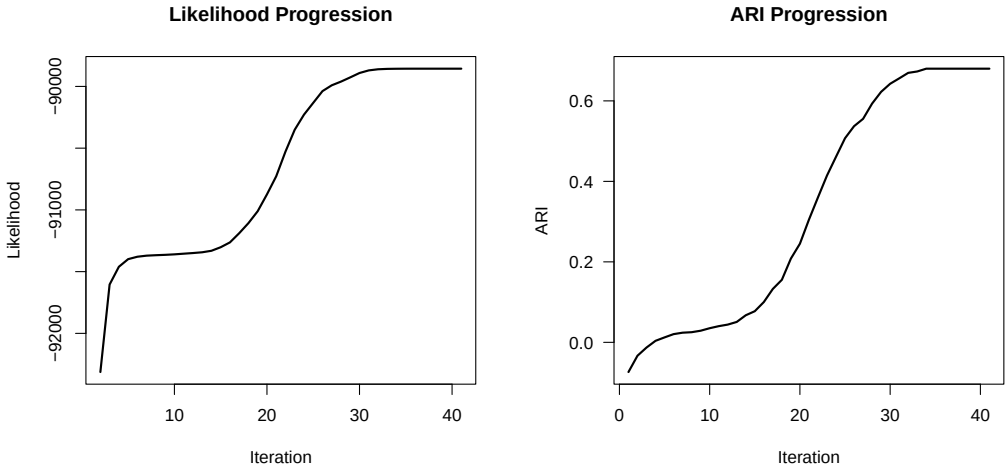

Figure 7: Tracking the increase in likelihood (left) and ARI (right) as the THMM algorithm runs.

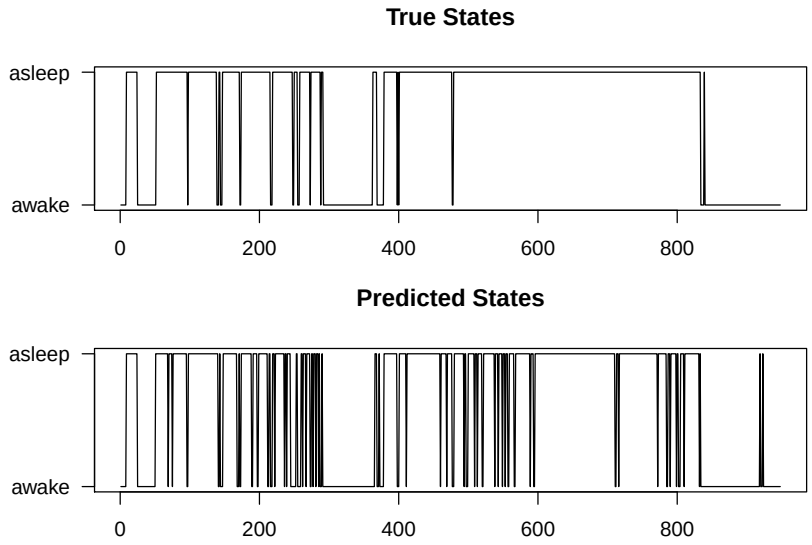

Figure 8: Comparison of the predicted state sequence via the THMM and the true state sequence based on the raw EEG PSD curves.

6.3 Kernel Smoothed Data Analysis

A subtle point from the above theory is that the observations O_t do not remain as elements of the LCTVS X, but are instead considered within the Cameron-Martin space H. In the case of Wiener measure and $W_0^{2,1}[0,1]$, that implies that we want a smooth analogue of O_t. Thus, we can use, for example, Nadaraya-Watson kernel regression to create a smoothed version of each O_t via the function `ksmooth` in the base `stats` package in R. Repeating the above experiment for a smoothed EEG PSD curves with two states improves the ARI from 0.680 to 0.762. However, the strongest performance with respect to both ARI and cross entropy is the THMM model under the $W_0^{2,2}$ norm. The confusion matrices and the ARI and cross entropy values are displayed on the right side of Tables 8 and 9, respectively. Of course, elements of $W_0^{2,2}$ are smoother than those of $W_0^{2,1}$, which in turn are smoother than those of L^2. Hence, it is intuitively reasonable to see a big improvement of performance under the $W_0^{2,2}$ norm after the data is preprocesses with a Gaussian kernel smoother.

6.4 Modelling Five Sleep States

The clustering task becomes much more challenging when considering all five sleeps states: awake, REM, NREM1, NREM2, and NREM3. Nevertheless, the THMM model can discern some patterns in the EEG PSD data set. For this analysis, we only consider the PSD curves smoothed as discussed above via Nadaraya-Watson kernel regression. This is because the THMM model and the fPCA HMM model both had stronger clustering performance after application of a Gaussian kernel smoother. In what follows, the fPCA HMM model was fit

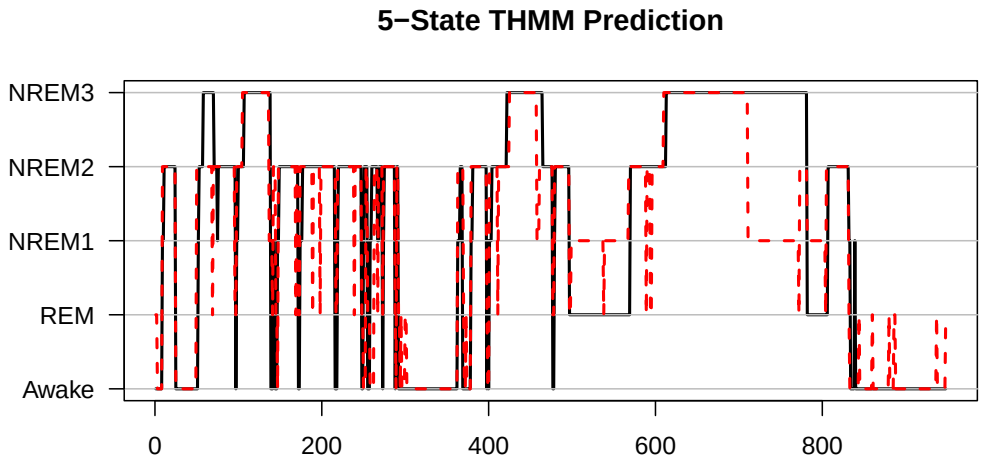

Figure 9: Comparison of the predicted state sequence (red dashed line) via the THMM under the $W_0^{2,1}$ norm and the true state sequence (black solid line) based on the kernel smoothed EEG PSD curves.

with five principal components that was chosen purposefully as it resulted in the best ARI and cross entropy values.

A comparison of the predicted sleep states and true sleep states can be seen in Figure 9. The red dashed line represents the states predicted by the THMM algorithm fitted to the kernel smoothed data using the $W_0^{2,1}$ norm. The ARI is only 0.632, but the algorithm still is able to track sleep states to some extent. Confusion matrices for the THMM with the $W_0^{2,1}$ norm, THMM with the $W_0^{2,2}$ norm, and fPCA HMM are displayed in Table 10, and ARI and cross entropy values can be found in Table 11. The THMM with $W_0^{2,1}$ and with $W_0^{2,2}$ both perform similarly and perform better than fPCA HMM even after fine tuning the number of principal components to maximize the ARI.

A close look at the confusion matrices in Table 10 show some interesting patterns. All three methods assign a significant number of *awake* epochs to the NREM1 state. Then, many of the true NREM1 states are confounded with the NREM2 states. This is a reasonable error to make as NREM1 corresponds to light sleep with NREM2 being deeper and NREM3 being the deepest. Such model interpretability will allow for future research involving more OSA data to achieve even stronger clustering accuracies.

7. Cumulative Snowfall Curves

An alternative application of the THMM is to model climate data. In this section, we consider 50 years (winters) of cumulative snowfall growth curves from the city of Edmonton, Alberta as recorded by the Meteorological Service of Canada (see `https://climate.`

		True States				
Method	Predicted	Awake	REM	NREM1	NREM2	NREM3
fPCA HMM	Awake	180	.	1	.	.
	REM	.	84	.	57	8
	NREM1	70	10	11	27	15
	NREM2	3	4	22	215	31
	NREM3	.	.	.	7	203
THMM $W_0^{2,1}$	Awake	207	2	2	8	3
	REM	2	88	2	16	.
	NREM1	33	5	2	7	3
	NREM2	11	3	28	270	43
	NREM3	.	.	.	5	208
THMM $W_0^{2,2}$	Awake	212	.	8	5	.
	REM	2	90	.	.	66
	NREM1	34	4	2	9	2
	NREM2	5	4	24	287	28
	NREM3	.	.	.	5	161

Table 10: Confusion matrices for the results of fitting a five-state THMM and five-state fPCA HMM to kernel smoothed EEG PSD curves.

Method	ARI	Ent	Iterations
fPCA HMM	0.535	2.191	
THMM $W_0^{2,1}$	0.632	**1.425**	62
THMM $W_0^{2,2}$	**0.636**	1.805	64

Table 11: ARI, cross entropy, and number of iterations for comparison of five-state clustering methods for the pediatric OSA data.

weather.gc.ca/). The data was collected at station number 3012208 at latitude 53°34'24 N longitude 113°31'06 W, which was the location of the now-closed City Centre Airport. The data considered spans from the winter of 1940/1941 until the winter of 1989/1990 and was collected daily. A Gaussian kernel smoother was used to preprocess these growth curves prior to analysis.

7.1 Two-State Models

If only the total snowfall is considered as a univariate time series, a classic Gaussian HMM model can be fit using the `mhsmm` package in R (O'Connell and Højsgaard, 2011). This naturally splits winters into high snowfall (mean = 192.5 cm) and low snowfall (mean = 111.9 cm) years. Similarly, two-state THMMs under the $W_0^{2,1}$ and the L^2 norms also split the winters into heavier and lighter snowfalls. However, the heavy snowfall category contains those years with consistently higher snowfall over the entire winter; i.e. the heavy snowfall curves are shifted up and left. We also fit the fPCA HMM to this data using four principal components. In this case, we cannot select the number of components based on purposefully maximizing the ARI as there is no ground truth to conform to. Instead, four was chosen to get the explained variation in the data above 99%. Lastly, the THMM with the $W_0^{2,2}$ norm was also fit to the data, but appeared to be assigning curves to states at random.

For each of the five methods considered, 20 models were fit and the fitted model that returned the highest likelihood was kept. Table 12 computes the ARI for the predicted state sequence for each pairing of fitted models. The THMM under the $W_0^{2,2}$ norm is seen to diverge from the other models. In contrast, the other four models all reasonably agree on which years should be put into which state. The states in these four models correspond to higher and lower snowfall years as can be seen visually in Figure 10. The fitted transition matrices indicates that back-to-back heavy snowfall winters are unlikely:

$$\tilde{A}_{L^2} = \begin{pmatrix} 0.154 & 0.846 \\ 0.306 & 0.694 \end{pmatrix}, \qquad \tilde{A}_{W_0^{2,1}} = \begin{pmatrix} 0 & 1 \\ 0.321 & 0.679 \end{pmatrix},$$
$$\tilde{A}_{\text{fPCA}} = \begin{pmatrix} 0.285 & 0.715 \\ 0.216 & 0.784 \end{pmatrix}, \qquad \tilde{A}_{\text{Uni}} = \begin{pmatrix} 0 & 1 \\ 0.255 & 0.745 \end{pmatrix}.$$

Furthermore, these models predict that heavier snowfall years occur at a rate of once every 3.1 to 4.6 years.

7.2 Four-State THMM

A four-state THMM model applied to these smoothed cumulative snowfall growth curves results in more interesting findings. Mainly, it identifies three Markov states corresponding to low, medium, and high snowfall. The fourth state is reserved for the winter of 1954/1955 alone, which had little snowfall until 18 April when 47.5 cm of snow fell during a three-day storm. This event was the largest recorded snowfall in Edmonton's recorded history. The

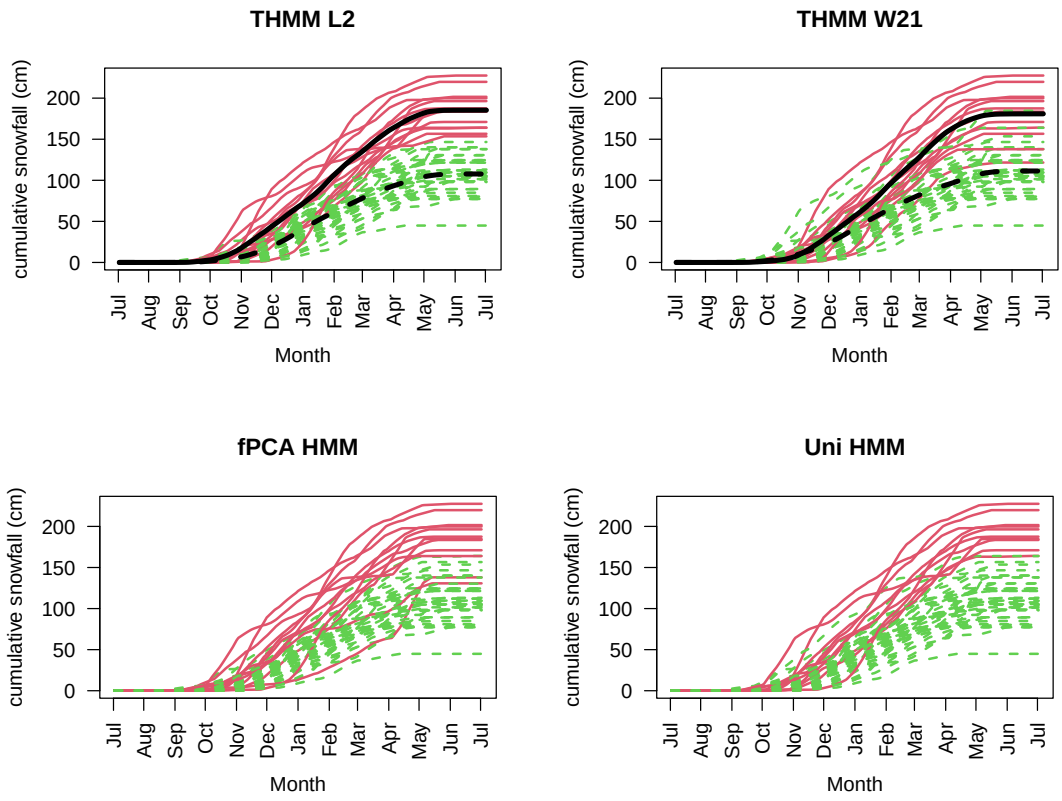

Figure 10: Fitted two-state THMM's (top row) under the L^2 and $W_0^{2,1}$ norms for cumulative snowfall curves and (bottom row) fPCA HMM for cumulative curves and a univariate HMM fit to total snowfall in the city of Edmonton, Alberta.

		THMM		HMM	
	L^2	$W_0^{2,1}$	$W_0^{2,2}$	fPCA	Uni
THMM L^2		0.611	-0.014	0.611	0.751
THMM $W_0^{2,1}$	0.611		-0.009	0.413	0.670
THMM $W_0^{2,2}$	-0.014	-0.009		-0.015	-0.007
fPCA HMM	0.611	0.413	-0.015		0.670
Uni HMM	0.751	0.670	-0.007	0.670	

Table 12: A comparison based on ARI of how well two-state HMM models fit to the snowfall data coincide with respect to their Viterbi paths.

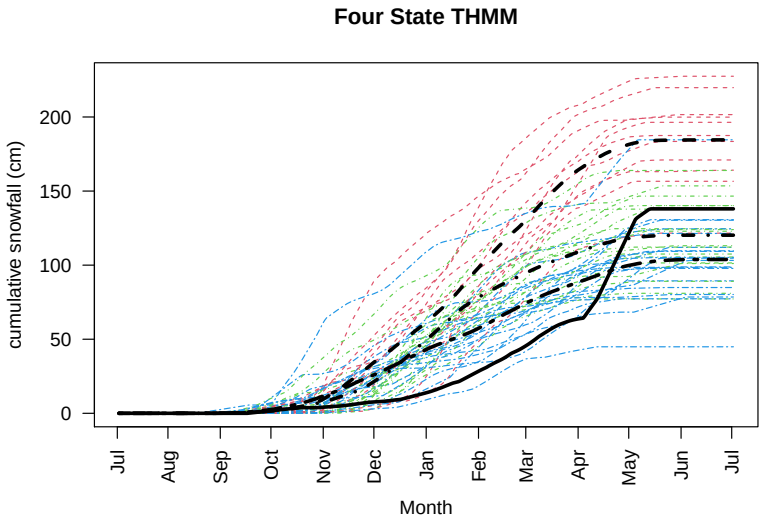

Figure 11: Fitted four state THMM for cumulative and total snowfall in the city of Edmonton Alberta, respectively.

fitted transition matrix is

$$\tilde{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0.43 & 0.57 \\ 0 & 0.40 & 0.52 & 0.08 \\ 0.05 & 0.15 & 0.18 & 0.62 \end{pmatrix}.$$

Note that $\tilde{A}_{2,2} = 0$ indicating that back-to-back heavy snowfall years do not occur in this 50 year data set. This is consistent with the two-state models discussed above. A plot of the data with the four mean curves superimposed on top is displayed in Figure 11.

8. Future Investigations

The following subsections propose areas of research for future investigation with respect to the THMM methodology.

8.1 Estimation of the Hurst Parameter

In this work, we only consider fractional Brownian motion with a fixed Hurst parameter chosen by the analyst. However much past research has gone into estimation of the Hurst parameter from data; see Berzin and León (2008); Kubilius and Mishura (2012); Hu et al. (2019) and others. Future work could incorporate estimation of the Hurst parameter within the THMM algorithms. This would allow for models to be fit where each Markov state will have its own estimated Hurst parameter and thus its own Gaussian Measure / Cameron-Martin space.

8.2 Regularized Learning

The Onsager-Machlup functional for a Gaussian measure as first presented in Bogachev (1998) is

$$\lim_{\varepsilon \to 0} \frac{\gamma(V_\varepsilon + h)}{\gamma(V_\varepsilon + k)} = \exp\left(\frac{1}{2}|\pi_q k|_H^2 - \frac{1}{2}|\pi_q h|_H^2 \right).$$

In this work, we set $k = 0$ and $h = O_t - h_j$. However, retaining k in the above equation results in the following optimization problem:

$$\arg\min_{h_j \in H} \sum_{t=1}^{T} \alpha_t(j)\beta_t(j) \left\{ |O_t - h_j|_H^2 - |k|_H^2 \right\}.$$

Thus, choosing a nonzero k to be a function of h_j will affect the final fit of the THMM.

8.3 Extensions Beyond the HMM

The Hidden Markov Model is an eminently useful modelling tool. However, there are many models that extend and complicate the beautiful simplicity of the original HMM. Future work can consider extending our proposed THMM from a finite number of discrete states to a countably infinite state space using ideas from the infinite HMM (Beal et al., 2001). Continuous state space HMMs have also been considered with respect to the Kalman filter. Other ways of adding dependency exist as well including autoregressive HMMs that are discussed in Juang and Rabiner (1985) and Rabiner (1989) as well as in the recent works of Lawler et al. (2019) and Sidrow et al. (2021) who use this tool to model animal behaviour. Lastly, a Topological Hidden Markov Random Field could be implemented to model spatial time series of climate data.

Acknowledgments

We would like to acknowledge support for this project from the Natural Sciences and Engineering Research Council of Canada and the McIntyre Memorial fund from the School of Dentistry at the University of Alberta.

Appendix A. Theoretical Guarantees

In what follows, detailed proofs are presented to demonstrate the theoretical soundness of considering an HMM and running the Baum-Welch algorithm in locally convex spaces.

A.1 Definitions and Notation

A locally convex topological vector space (LCTVS) generalizes normed spaces and can be constructed in a few equivalent ways. In this work, we define our LCTVS's via a family of seminorms.

Definition 1 (Seminorm) *Let X be a real vector space, then $q : X \to \mathbb{R}$ is a seminorm if it satisfies the following properties.*

1. *$q(x) \geq 0$ for all $x \in X$.*
2. *$q(cx) = |c|q(x)$ for all $x \in X$ and $c \in \mathbb{R}$.*
3. *$q(x_1 + x_2) \leq q(x_1) + q(x_2)$.*

Of note, unlike a norm, $q(x) = 0$ does not necessarily imply that $x = 0$.

Definition 2 (Locally Convex Space) *A real vector space X with $\mathcal{Q}$, a collection of seminorms, is said to be a locally convex topological vector space. The seminorms in $\mathcal{Q}$ induce a topology on X, which is the coarsest topology such that the $q \in \mathcal{Q}$ are continuous.*

An example of a LCTVS is $X = C_0([0,1], \mathbb{R})$, the space of continuous functions $x : [0,1] \to \mathbb{R}$, $x(0) = 0$ with $q_\tau(x) = |x(\tau)|$ for $\tau \in [0,1]$. More generally, we can consider X^*, the space of linear functionals on X, and take $q_f(x) = |f(x)|$ for any $x \in X$ and $f \in X^*$. Thus, we have a TVS X with the weak topology.

The focus of this work is on Gaussian measures on a LCTVS X. We say that γ is a Gaussian measure defined on the cylindrical σ-field $\mathcal{E}(X)$ if the induced measure $\gamma \circ f^{-1}$ on $\mathbb{R}$ is Gaussian for all $f \in X^*$. We furthermore take γ to be a Radon Gaussian measure, which is still sufficiently general for most applications of interest.

Definition 3 (Radon Measure) *A measure μ defined on the Borel σ-field $\mathcal{B}(X)$ for a topological space X is Radon if for every $B \in \mathcal{B}(X)$ and every $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset B$ such that $\mu(B \setminus K_\varepsilon) < \varepsilon$.*

In the case that X is a separable Fréchet space, the cylindrical σ-field coincides with the Borel σ-field and furthermore every Borel measure is Radon; see Bogachev (1998), Theorems A.3.7 and A.3.11.

We define similarly to Bogachev (1998) the following terms. For a locally convex space X, X^* is the topological dual space consisting of continuous linear functionals. The mean of γ is $a_\gamma(f) = \int_X f(x)\gamma(dx)$ with $a_\gamma \in X^{**}$. The covariance operator is $R_\gamma : X^* \to X^{**}$ defined by

$$R_\gamma(f)(g) = \int_X [f(x) - a_\gamma(f)]\,[g(x) - a_\gamma(g)]\,\gamma(dx).$$

X_γ^* is the closure of $\{f - a_\gamma(f) : f \in X^*\}$ embedded into $L^2(\gamma)$. The Cameron-Martin space is

$$H(\gamma) = \left\{h \in X : |h|_{H(\gamma)} < \infty\right\}$$

where the norm is $|h|_{H(\gamma)} = \sup\{l(h) : l \in X^*, R_\gamma(l)(l) \leq 1\}$. In what follows, we typically omit γ from the notation and just write the Cameron-Martin norm as $|h|_H$. Lemma 2.4.1 of Bogachev (1998) proves that if some $h \in X$ is in $H(\gamma)$, then there is a $h^* \in X_\gamma^*$ such that $h = R_\gamma(h^*)$ with $|h|_{H(\gamma)} = \|h^*\|_{L^2(\gamma)}$. Lastly, if $R_\gamma(X_\gamma^*) \subset X$, then $H(\gamma) = R_\gamma(X_\gamma^*)$ and $|R_\gamma(f)|_{H(\gamma)} = \sqrt{R_\gamma(f)(f)}$ where the operator R_γ is extended to X_γ^* as follows:

$$R_\gamma : X_\gamma^* \to X^{**},\ R_\gamma(f)(g) = \int_X f(x)\left[g(x) - a_\gamma(g)\right]\gamma(dx).$$

Thus, $\langle h, k\rangle_{H(\gamma)} = \langle h^*, k^*\rangle_{L^2(\gamma)}$, and the Cameron-Martin space $H(\gamma) \subset X$ gains the Hilbert space structure of X_γ^*. In this case, the mapping R_γ is an isomorphism between X_γ^* and $H(\gamma)$. This is necessarily true for Radon Gaussian measures (Bogachev, 1998, Theorem 3.2.3).

A.2 Onsager-Machlup Functional

For a Gaussian measure γ on a metric space X, we can consider the *Onsager-Machlup functional*, which is

$$I(a, b) = \lim_{\varepsilon \to 0} \frac{\gamma(K(a, \varepsilon))}{\gamma(K(b, \varepsilon))}, \quad a, b \in X$$

where $K(a, \varepsilon)$ is a closed ball of radius $\varepsilon > 0$ centred at $a \in X$. Hence, we consider the limit as the radii of the two balls shrink to zero. For a locally convex space, Bogachev (1998) introduces the notation for an epsilon ball $V_\varepsilon = \{x \in X : q(x) \leq \varepsilon\}$ with h-shift $V_\varepsilon + h = \{x + h \in X : q(x) \leq \varepsilon\}$ where q is a seminorm. We are interested in the ratio

$$\frac{\gamma(V_\varepsilon + h)}{\gamma(V_\varepsilon)} = \frac{\mathrm{e}^{-|h|_H^2/2}}{\gamma(V_\varepsilon)} \int_{V_\varepsilon} \mathrm{e}^{h^*(x)}\gamma(dx)$$

where $h^* \in X_\gamma^*$ is such that $h = R_\gamma h^*$. Furthermore, we denote the integral

$$J_\varepsilon(f) = \frac{1}{\gamma(V_\varepsilon)} \int_{V_\varepsilon} \mathrm{e}^{f(x)}\gamma(dx)$$

and $F_q = \{f \in X_\gamma^* : \lim_{\varepsilon \to 0} J_\varepsilon(f) = 1\}$ is a closed linear subspace of X_γ^* (Bogachev, 1998, Lemma 4.7.2). Via Lemma 4.7.4 (Bogachev, 1998), we can also define F_q as $R_\gamma^{-1} Z^\perp$ where $Z = \{a \in H(\gamma) : q(a) = 0\}$. Lastly, $P_q : X_\gamma^* \to F_q$ is an orthogonal projection.

Given this setup, the Onsager-Machlup function is (Bogachev, 1998, Corollary 4.7.8)

$$\lim_{\varepsilon \to 0} \frac{\gamma(V_\varepsilon + h)}{\gamma(V_\varepsilon + k)} = \exp\left(\frac{1}{2}|\pi_q k|_H^2 - \frac{1}{2}|\pi_q h|_H^2\right)$$

for $h, k \in H(\gamma)$ where π_q is the orthogonal projection onto $Z^\perp$, which can be written as $\pi_q = R_\gamma P_q R_\gamma^{-1}$. We will use this for the emission function within the HMM framework. Namely, we choose

$$b_j(O_t) = \lim_{\varepsilon \to 0} \frac{\gamma(V_\varepsilon + \{O_t - h_j\})}{\gamma(V_\varepsilon)} = \exp\left(-\frac{1}{2}|\pi_q\{O_t - h_j\}|_H^2\right).$$

Here, we require $O_t - h_j \in H(\gamma)$, which may necessitate a modification of O_t such as application of a kernel smoother.

A.3 Reestimation and Maximum Likelihood

In this section, we extend proofs from past work (Liporace, 1982; Wu, 1983) to show that the Baum-Welch and EM algorithms satisfy nice convergence properties. Given a finite sequence of observations $O = \{O_t\}_{t=1}^T$, initial state probabilities $\{\eta_j\}_{j\in S}$, a $p \times p$ Markov transition matrix A with ijth entry a_{ij}, state means $\{m_j\}_{j\in S}$ and a state sequence $s = (s_1, \ldots, s_T) \in S^T$, we can define our analogue of the likelihood function (refer to Section 2 for the classic HMM setup) to be

$$L_\lambda(O) = \sum_{s\in S^T} \left(\eta_{s_1} \prod_{t=1}^T a_{s_{t-1}s_t} \right) \exp\left\{ -\frac{1}{2} \sum_{t=1}^T |\pi_q(O_t - m_{s_t})|_H^2 \right\}$$

where the sum is taken over all state sequences and $\lambda = (\{\eta_j\}_{j\in S}, \{a_{ij}\}_{i,j\in S}, \{m_j\}_{j\in S})$ denotes the selection of parameters. Let Λ be the space of all possible parameters λ. The initial probabilities $\{\eta_j\}_{j\in S}$ and each row of $\{a_{ij}\}_{i,j\in S}$ lie in the $p-1$ simplex, i.e. the closed convex hull of the unit vectors in $\mathbb{R}^p$, which is compact. Furthermore, each m_j for $j \in S$ lies in H_0, a convex subset of the Cameron-Martin space $H(\gamma)$. Thus, Λ is a closed convex subset of $\mathbb{R}^p \times \mathbb{R}^{p\times p} \times H(\gamma)^p$. For a specific state sequence $s \in S^T$, we write $L_\lambda(O, s)$ to be the summand of $L_\lambda(O)$ for s. The reestimation transformation $Q : \Lambda \times \Lambda \to \mathbb{R}$ is a bivariate function given by

$$\begin{aligned} Q(\lambda, \tilde{\lambda}) &= \sum_{s\in S^T} L_\lambda(O, s) \log L_{\tilde{\lambda}}(O, s) \\ &= \sum_{s\in S^T} \left[L_\lambda(O, s) \left\{ \log \tilde{\eta}_{s_1} + \sum_{t=1}^T \log \tilde{a}_{s_{t-1}s_t} - \frac{1}{2} \sum_{t=1}^T |\pi_q(O_t - \tilde{m}_{s_t})|_H^2 \right\} \right]. \end{aligned}$$

The Baum-Welch algorithm is part of the family of majorize-minimization algorithms that optimize this function in an iterative fashion in order to obtain parameter estimates, instead of directly optimizing the likelihood. Our goal in this section is to prove that maximizing $Q(\lambda, \tilde{\lambda})$ over all $\tilde{\lambda}$ increases the likelihood, i.e. $L_\lambda(O) \le L_{\tilde{\lambda}}(O)$, and that the the reestimation procedure stabilizes only at critical points of the likelihood. When talking about differentiability in this paper, we will always be referring to the Fréchet derivative.

Lemma 4 *The likelihood function $L(\lambda) = L_\lambda(O)$ and the reestimation function Q are differentiable with respect to the state means.*

Proof

Firstly, we note that the squared norm on real Hilbert spaces is differentiable, since

$$\|x + h\|^2 = \langle x + h, x + h\rangle = \|x\|^2 + \|h\|^2 + 2\langle x, h\rangle = \|x\|^2 + 2\langle x, h\rangle + O(h)$$

is linear in h. Differentiability of both functions in $\{m_j\}_{j\in S}$ then follows from this and the smoothness of projections (Coleman, 2012, Corollary 6.2). ■

Lemma 5 *The function* $\psi(\{\tilde{m}_j\}_{j\in S}) = \sum_{s\in S^T} L_\lambda(O,s)\sum_{t=1}^T |\pi_q(O_t - \tilde{m}_{s_t})|_H^2$ *has global minima for each*

$$\tilde{m}_j \in \left\{ z + \frac{\sum_{t=1}^T \alpha_t(j)\beta_t(j)\pi_q(O_t)}{\sum_{t=1}^T \alpha_t(j)\beta_t(j)} \,:\, z \in Z \right\}$$

where $Z = \{a \in H \,:\, q(a) = 0\}$. *Furthermore, these are the only critical points of the function.*

Proof Let us denote $T_j(s) = \{t \,:\, s_t = j\} \subseteq \{1,\ldots,T\}$, $S_j(t) = \{s \,:\, s_t = j\} \subset S^T$, and assume the forward and backward probabilities are known at the current step, denoted by α_t and β_t respectively. We may then rewrite our function ψ as follows.

$$\begin{aligned}
\psi(\{\tilde{m}_j\}_{j\in S}) &= \sum_{s\in S^T} L_\lambda(O,s) \sum_{j=1}^p \sum_{t\in T_j(s)} |\pi_q(O_t - \tilde{m}_j)|_H^2 \\
&= \sum_{j=1}^p \sum_{t=1}^T \sum_{s\in S_j(t)} L_\lambda(O,s)|\pi_q(O_t - \tilde{m}_j)|_H^2 \\
&= \sum_{j=1}^p \sum_{t=1}^T \alpha_t(j)\beta_t(j)|\pi_q(O_t - \tilde{m}_j)|_H^2.
\end{aligned}$$

Each of the p terms in the outer sum is a non-negative function of only one of the parameters $\tilde{m}_j$ to be optimized. We may therefore separately optimize each of the terms. Abusing notation and letting $\tilde{m}_j$ represent the set of minimizers for ψ, we have

$$\tilde{m}_j = \operatorname*{arg\,min}_{m\in H} \sum_{t=1}^T \alpha_t(j)\beta_t(j)|\pi_q(O_t - m)|_H^2.$$

We note here that the arg min is invariant under translation by elements in Z. For $B_1(0) \subset X$, the ball of radius 1 about the origin,

$$\begin{aligned}
\pi_q(\tilde{m}_j) &= \operatorname*{arg\,min}_{h\in Z^\perp} \sum_{t=1}^T \alpha_t(j)\beta_t(j)|\pi_q(O_t) - h|_H^2 \\
&= \operatorname*{arg\,min}_{\substack{h_0\in Z^\perp \cap B_1(0) \\ c>0}} \sum_{t=1}^T \alpha_t(j)\beta_t(j)|\pi_q(O_t) - ch_0|_H^2 \\
&= \operatorname*{arg\,min}_{\substack{h_0\in Z^\perp \cap B_1(0) \\ c>0}} \sum_{t=1}^T \alpha_t(j)\beta_t(j) \left\{ |\pi_q(O_t)|_H^2 + c^2|h_0|_H^2 - 2c\left\langle \pi_q(O_t), h_0 \right\rangle_H \right\} \\
&= \operatorname*{arg\,min}_{\substack{h_0\in Z^\perp \cap B_1(0) \\ c>0}} \sum_{t=1}^T \alpha_t(j)\beta_t(j) \left\{ c^2 - 2c\left\langle \pi_q(O_t), h_0 \right\rangle_H \right\} \\
&= \operatorname*{arg\,min}_{\substack{h_0\in Z^\perp \cap B_1(0) \\ c>0}} \left\{ c^2 \sum_{t=1}^T \alpha_t(j)\beta_t(j) - 2c\left\langle \sum_{t=1}^T \alpha_t(j)\beta_t(j)\pi_q(O_t), h_0 \right\rangle_H \right\}.
\end{aligned}$$

Now we note that the variable h_0 minimizing the above equation is independent of the choice of c and that the second term is minimized exactly when h_0 is parallel to the term in the inner product, i.e. $h_0 = \sum_{t=1}^{T} \alpha_t(j)\beta_t(j)\pi_q(O_t)|\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)\pi_q(O_t)|_H^{-1}$. We therefore have

$$\begin{aligned}\underset{\substack{h_0 \in Z^{\perp} \cap B_1(0) \\ c>0}}{\arg\min} &\left\{ c^2 \sum_{t=1}^{T} \alpha_t(j)\beta_t(j) - 2c \left\langle \sum_{t=1}^{T} \alpha_t(j)\beta_t(j)\pi_q(O_t), h_0 \right\rangle_H \right\} \\ &= \underset{c>0}{\arg\min} \left\{ c^2 \sum_{t=1}^{T} \alpha_t(j)\beta_t(j) - 2c \left| \sum_{t=1}^{T} \alpha_t(j)\beta_t(j)\pi_q(O_t) \right|_H \right\}.\end{aligned}$$

Thus we have

$$c = \frac{|\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)\pi_q(O_t)|_H}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} \quad \text{and} \quad \pi_q(\tilde{m}_j) = ch_0 = \frac{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)\pi_q(O_t)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}.$$

It follows that the members of the product of sets $\{\pi_q(\tilde{m}_j) + Z\}$ over $j \in S$ are the global minima of ψ over the open convex set H_0, and thus, they are local minima, and hence, they are also critical points (Coleman, 2012, Corollary 2.5). The lack of other critical points follows from the convexity of ψ over H_0^p owing to the convexity of the square norm and linearity of π_q. See Theorem 7.4 (c) and Proposition 7.4 from Coleman (2012). ■

Theorem 6 *Every critical point of the reestimation function $Q(\lambda, \cdot)$ is a global maximum and at least one such point exists. Additionally, if q is a norm, this maximum is unique.*

Proof

The reestimation function $Q(\lambda, \tilde{\lambda})$ can be written as

$$\begin{aligned}Q(\lambda, \tilde{\lambda}) &= \sum_{s \in S^T} \left[L_\lambda(O, s) \left\{ \log \tilde{\eta}_{s_1} + \sum_{t=1}^{T} \log \tilde{a}_{s_{t-1}s_t} - \frac{1}{2} \sum_{t=1}^{T} |\pi_{q,s_t}(O_t - \tilde{m}_{s_t})|_H^2 \right\} \right] \\ &= \underbrace{\sum_{s \in S^T} L_\lambda(O, s) \log \tilde{\eta}_{s_1}}_{(1)} + \underbrace{\sum_{s \in S^T} L_\lambda(O, s) \sum_{t=1}^{T} \log \tilde{a}_{s_{t-1}s_t}}_{(2)} \\ &\quad \underbrace{- \frac{1}{2} \sum_{s \in S^T} \sum_{t=1}^{T} L_\lambda(O, s) |\pi_q(O_t - \tilde{m}_{s_t})|_H^2}_{(3)}.\end{aligned}$$

The arg max of the above in $\tilde{\lambda} = (\{\tilde{\eta}_j\}_{j \in S}, \{\tilde{a}_{ij}\}_{i,j \in S}, \{\tilde{m}_j\}_{j \in S})$ can be broken down into taking the arg max of (1) and (2) and the arg min of (3), in $\{\tilde{\eta}_j\}_{j \in S}$, $\{\tilde{a}_{ij}\}_{i,j \in S}$ and $\{\tilde{m}_j\}_{j \in S}$, respectively, in an independent manner. The max of $Q(\lambda, \cdot)$ occurs exactly at the Cartesian

product set of these points. For (1) and (2), we note that since $\sum_{i=1}^{p} \tilde{\eta}_i = 1$ and $\sum_{j=1}^{p} \tilde{a}_{ij} = 1$ for each i, the boundary of their respective optimizations over Λ (using Lagrange multipliers for example) occur when at least one term in the sum is 0. However, in this case, the logs approach $-\infty$ so that (1) and (2) cannot be maximized at the boundaries. It follows that any global maxima for these terms must occur at critical points. The reestimation formulae from Algorithm 1 give these points (see Theorem 2 in Liporace (1982) for derivation). Lastly, term (3) is maximized over the Hilbert space $H(\gamma)$, so that any extrema are necessarily critical points (Coleman, 2012, Corollary 2.5) and at least one such point exists by Lemma 5. The product of these sets then gives the existence of critical points for $Q(\lambda, \cdot)$.

Additionally, since the log function is strictly concave, terms (1) and (2) are strictly concave in their respective arguments. Due to the strict convexity of squared norms, term (3) is also concave in its argument, becoming strictly concave when q is a norm and π_q is the identity map on $H(\gamma)$. It follows that every critical point of $Q(\lambda, \cdot)$ is a global maximum and uniqueness holds due to strict concavity when q is a norm (Coleman, 2012, Theorem 7.4, Proposition 7.3, Proposition 7.4). ■

Theorem 7 *A point λ in the parameter space is a critical point of $L_\lambda(O)$ if and only if it is a fixed point of the reestimation function, i.e. $Q(\lambda, \lambda) = \max_{\tilde{\lambda}} Q(\lambda, \tilde{\lambda})$. Furthermore*

$$Q(\lambda, \tilde{\lambda}) > Q(\lambda, \lambda) \Rightarrow L_{\tilde{\lambda}}(O) > L_\lambda(O_t).$$

Hence increasing $Q(\lambda, \cdot)$ improves the likelihood.

Proof This closely follows the proof in Liporace (1982). First, we note that for any real valued function ψ on a normed space, $\psi(x)\frac{d\log(\psi(x))}{dx} = \frac{d\psi(x)}{dx}$. Suppose λ is a critical point of $L_\lambda(O)$. Then,

$$\begin{aligned} 0 = \nabla_{\tilde{\lambda}} L_{\tilde{\lambda}}(O)|_{\tilde{\lambda}=\lambda} &= \sum_{s \in S^T} \nabla_{\tilde{\lambda}} L_{\tilde{\lambda}}(O, s)|_{\tilde{\lambda}=\lambda} \\ &= \sum_{s \in S^T} L_\lambda(O, s) \nabla_{\tilde{\lambda}} \log(L_{\tilde{\lambda}}(O, s))|_{\tilde{\lambda}=\lambda} = \nabla_{\tilde{\lambda}} Q(\lambda, \tilde{\lambda})|_{\tilde{\lambda}=\lambda} \end{aligned}$$

implying that λ is a critical point of $Q(\lambda, \cdot)$, so that by Theorem 6, it is a fixed point of the reestimation. Furthermore, as $\log x \leq x - 1$ with equality holding if and only if $x = 1$,

$$\begin{aligned} Q(\lambda, \tilde{\lambda}) - Q(\lambda, \lambda) &= \sum_{s \in S^T} L_\lambda(O, s) \log \left\{ \frac{L_{\tilde{\lambda}}(O, s)}{L_\lambda(O, s)} \right\} \\ &\leq \sum_{s \in S^T} L_\lambda(O, s) \left(\frac{L_{\tilde{\lambda}}(O, s)}{L_\lambda(O, s)} - 1 \right) \\ &= L_{\tilde{\lambda}}(O) - L_\lambda(O). \end{aligned}$$

Thus, $Q(\lambda, \tilde{\lambda}) > Q(\lambda, \lambda)$ implies $L_{\tilde{\lambda}}(O) > L_\lambda(O)$ as inequality in the above equation is an equality if and only if $L_{\tilde{\lambda}}(O, s) = L_\lambda(O, s)$ for each $s \in S$ thereby implying $L_{\tilde{\lambda}}(O) = L_\lambda(O)$. ■

A.4 Limits for Model Parameters

In this section, we show that that the parameter sequence produced by the Baum-Welch algorithm satisfies nice convergence properties where the parameters have a limit point that is a critical point of the likelihood. A similar technique to the classical proofs in Wu (1983) is used. This requires two theorems, Berge's Maximum Theorem and Zangwill's Convergence Theorem, and some preliminary definitions and notation. Our main result is Theorem 10 below.

We recall that the parameter space Λ for $\lambda = (\{\eta_j\}_{j\in S}, \{a_{ij}\}_{i,j\in S}, \{m_j\}_{j\in S})$ is a closed convex subset of $\mathbb{R}^p \times \mathbb{R}^{p\times p} \times H(\gamma)^p$, which is additionally compact when projected onto $\mathbb{R}^p \times \mathbb{R}^{p\times p}$. Now by Lemma 5, when considering the optimization problem for the likelihood, each m_j can be considered to lie in $\overline{\text{conv}}(\{O_t\}) + Z$ where $\overline{\text{conv}}(\{O_t\}) \subset H(\gamma)$ is the closed convex hull of the observation sequence. Without loss of generality, we may take $m_j = \sum_{t=1}^T \alpha_t(j)\beta_t(j)\pi_q(O_t)/\sum_{t=1}^T \alpha_t(j)\beta_t(j)$ in our reestimation formulas by assigning the component in Z to be 0. Then, each state mean m_j lies in $\overline{\text{conv}}(\{O_t\})$, which is compact being the closed convex hull of a finite set of points (Charalambos and Border, 2013, Corollary 5.30) (Rudin, 1991, Theorem 3.20). Therefore, the arg max of our objective function over λ can be restricted to a compact convex subset $\mathcal{C} \subset \Lambda$. We note that in the case q is a norm, $Z = \{0\}$ and the reestimation formula automatically restricts to $\mathcal{C}$ (see Theorem 6).

Given metric spaces X, Y, a function $f : X \to P(Y)$, where $P(Y)$ is the power set of Y, is called a *correspondence* on X. Such a map is said to be closed or *upper hemicontinuous* if given sequences $\{x_n\} \subset X$ and $\{y_n\} \subset Y$ such that $y_n \in f(x_n)$ for each $n \in \mathbb{N}$, $x_n \to x$ and $y_n \to y$ imply $y \in f(x)$. On the other hand, it is said to be *lower hemicontinuous* if for any sequence $\{x_n\} \subset X$, $x_n \to x$ implies that for each $y \in f(x)$, there exists a subnet x_{k_n} of $\{x_n\}$ and a sequence $\{y_n\} \subset Y$ with $y_n \subset f(x_{k_n})$ such that $y_n \to y$. A correspondence that is both upper and lower hemicontinuous is said to be continuous.

Theorem 8 (Berge's Maximum Theorem) *(Charalambos and Border, 2013, Theorem 17.31) Let X and Y be Hausdorff topological spaces and let $\phi : X \to P(Y)$ be a continuous correspondence such that $\phi(x)$ is non-empty and compact for all $x \in X$. Additionally suppose $f : X \times Y \to \mathbb{R}$ is continuous. Then the correspondence $\mu : X \to P(Y)$ given by $\mu(x) = \arg\max_{y\in\phi(x)} f(x, y)$ has non-empty compact values and is upper hemicontinuous.*

Let $X = \Lambda$, $Y = \Lambda$ in the above theorem, and let $\phi : \Lambda \to P(\Lambda)$ be given by the constant map $\phi(x) = \mathcal{C}$ where we recall that $\mathcal{C}$ is non-empty, compact, and convex. This is continuous. By choosing $f : \Lambda \times \Lambda \to \mathbb{R}$ with $f(x, y) = Q(x, y)$, the map $\lambda \mapsto \mu(\lambda) = \arg\max_{y\in\mathcal{C}} Q(\lambda, y)$ is upper hemicontinuous by Theorem 8.

Theorem 9 (Zangwill's Convergence Theorem) *(Zangwill, 1969, page 91) Suppose M is a correspondence $M : X \to P(X)$ that generates a sequences $\{x_n\}$ with $x_{n+1} \in M(x_n)$ that is initiated with $x_0 \in X$. Suppose a "solution set" $\Gamma \subset X$ is given and suppose that*

1. *$\{x_n\} \subset S$ for a compact set $S \subset X$*
2. *There is continuous function f on X satisfying*

(a) if $x \in \Gamma$, then $f(x) \leq f(y)$ for all $y \in M(x)$

(b) if $x \notin \Gamma$, then $f(x) < f(y)$ for all $y \in M(x)$

3. M is upper hemicontinuous on $X \setminus \Gamma$

Then, every limit point of $\{x_n\}$ lies in Γ.

Theorem 10 *Suppose $\{\lambda_n\}$ is the sequence generated by Algorithm 1 with $M : \Lambda \to P(\Lambda)$ defined as $\lambda \mapsto \arg\max_{\tilde{\lambda} \in C} Q(\lambda, \tilde{\lambda})$ and $\lambda_{n+1} \in M(\lambda_n)$. Then, all the limit points of $\{\lambda_n\}$ are critical points of the likelihood, achieving the same likelihood, and the sequence $L_{\lambda_n}(O)$ converges to this value. Furthermore, at least one such point exists.*

Proof Let us denote $\mathcal{M} = \arg\max_{\lambda \in \Lambda} L_\lambda(O)$, the set of critical points of the likelihood function of $L(\lambda) = L_\lambda(O)$ on Λ and consider M as given. Note that the points $\{\lambda_n\}$ generated by Algorithm 1 satisfy the following conditions

1. $\{\lambda_n\} \subset \mathcal{C} \subset \Lambda$ where $\mathcal{C}$ is compact and convex;

2. $L : \Lambda \to \mathbb{R}$, $\lambda \mapsto L_\lambda(O)$ is continuous and by Theorem 7

   (a) If $\lambda \notin \mathcal{M}$, then, it is not a fixed point of the reestimation i.e. for all $\tilde{\lambda} \in M(\lambda)$, $Q(\lambda, \tilde{\lambda}) > Q(\lambda, \lambda)$ which implies $L(\tilde{\lambda}) > L(\lambda)$.

   (b) However if $\lambda \in \mathcal{M}$, then by the definition of $M(\lambda)$, $Q(\lambda, \tilde{\lambda}) \geq Q(\lambda, \lambda)$ which implies $L(\tilde{\lambda}) \geq L(\lambda)$.

3. Lastly, take $\phi : \Lambda \to P(\Lambda)$ to be the constant correspondence given by $\lambda \mapsto \mathcal{C}$. It is easily checked that this is hemicontinuous and clear that $\phi(\lambda)$ is compact and non-empty for each $\lambda \in \Lambda$. Further, since $Q : \Lambda \times \Lambda \to \mathbb{R}$, $(\lambda, \tilde{\lambda}) \mapsto Q(\lambda, \tilde{\lambda})$ is continuous, by Berge's maximum theorem, M is upper hemicontinuous on all of Λ, whence on $\Lambda \setminus \Gamma$.

As all of the requirements for the Zangwill's Convergence Theorem are satisfied, every limit point of $\{\lambda_n\}$ lies in $\mathcal{M}$ so that the first part of the theorem holds. Note that by Theorem 7 and the monotone convergence theorem $\{L(\lambda_n)\}$ converges to $\sup_n L(\lambda_n)$. If λ is any limit point of λ_n, then by the uniqueness of the limit, $L(\lambda) = \sup_n L(\lambda_n)$ whereby all such limit points give the same likelihood. That such a limit point always exists follows from $\mathcal{C}$ being compact. ■

We conclude this section by noting that if a stronger version of Theorem 7 holds where every fixed point of the reestimation is a local maximum of $L_\lambda(O)$, then Theorem 10 can be extended to the likelihood function converging to a local maximum. In the special case where either the likelihood or the log likelihood functions are concave, convergence to the global maximum follows due to the properties of concave functions.

A.5 Limits for Mixture Distributions

The previous theorems show that the parameter sequence $\{\lambda_i\}_{i=1}^{\infty}$ produced by the Algorithm 1 has a unique limit point when q is a norm. In this section, we show that this quickly implies the existence of a weak limit point for the sequence of corresponding measures $\{\mu_i\}_{i=1}^{\infty}$ where

$$\mu_i = \textstyle\sum_{s\in S^T} \omega_s^{(i)} \gamma^{(i)}(s)$$

for weights $\omega_s^{(i)} \propto \prod_{t=1}^{T} \alpha_t^{(i)}(s_t)\beta_t^{(i)}(s_t)$. These μ_i are finite mixtures of Gaussian measures.

Let $\lambda_0 \in \Lambda$ be the set of initial THMM parameters and γ_0 be a centred Gaussian measure on a locally convex space X. Each iteration of the Baum-Welch algorithm produces updated parameters λ_i, $i = 1, \ldots, \infty$, which consist of a p-long vector of initial state probabilities $\eta^{(i)}$, a $p \times p$ Markov transition matrix $A^{(i)}$, and means $m_j^{(i)} \in H(\gamma_0)$ for $j = 1, \ldots, p$. Conditioned on a fixed state sequence $s \in S^T = \{1, \ldots, p\}^T$ at iteration i, we have a Gaussian measure $\gamma^{(i)}(s) = \bigotimes_{t=1}^{T} \gamma_t^{(i)}(s_t)$ on the product space $X^{\otimes T}$ where $\gamma_t^{(i)}(s_t) = \gamma_0(\cdot - m_{s_t}^{(i)})$ is the Gaussian measure γ_0 on X shifted by $m_{s_t}^{(i)}$. Thus, the Baum-Welch algorithm produces a sequence of measures $\{\mu_i\}_{i=1}^{\infty}$ on $X^{\otimes T}$, which are mixtures of $|S^T| = p^T$ Gaussian measures.

Corollary 11 *The sequence of measures μ_i has a weak limit point.*

Proof The reestimated means at algorithm iterate i and state j, $m_j^{(i)}$, lie in the compact convex hull $K = \overline{\text{conv}}\{O_t\}_{t=1}^{T}$. It follows that the sequence $\{m^{(i)}\}_{i\in\mathbb{N}}$ lying in K^p has a limit point $m = (m_1, \ldots, m_p)$. Similarly, the Markovian parameters $\eta^{(i)}$ and $A^{(i)}$ lie in a compact set and thus have a limit point. Let $\omega_s^{(i)}$ be weights corresponding to each state sequence $s \in S^T$ where $\omega_s^{(i)} \propto \prod_{t=1}^{T} \alpha_t^{(i)}(s_t)\beta_t^{(i)}(s_t)$ such that $\sum_{s\in S^T} \omega_s^{(i)} = 1$. The mapping

$$\lambda_i = (\eta^{(i)}, A^{(i)}, m^{(i)}) \to \mu_i = \textstyle\sum_{s\in S^T} \omega_s^{(i)} \gamma^{(i)}(s)$$

is continuous when the latter probability space on $X^{\otimes T}$ is equipped with the weak topology. Thus, given a weak limit point λ of $\{\lambda_i\}_{i=1}^{\infty}$, μ, the corresponding image of the point under the above map, is a weak limit point for μ_i. ■

A.6 Identifiability

In this section, the THMM is shown to be identifiable by adapting the theory developed in Gassiat and Rousseau (2016); Gassiat et al. (2016) to the setting of Gaussian measures in locally convex spaces (Bogachev, 1998; Rudin, 1987, 1991). For such a model and a centred Gaussian measure γ_0, let the collection of model parameters be denoted as $\lambda = (\{\eta_j\}_{j\in S}, \{a_{ij}\}_{i,j\in S}, \{m_j\}_{j\in S})$ with state space $S = \{1, \ldots, p\}$. Of course, each m_j is assumed to be distinct. Then, the mixture of shifted Gaussian measures associated with O_1, the first observation, is

$$\mu_{\lambda,\gamma_0,1}(B) = \sum_{j=1}^{p} \eta_j \gamma_0(B - m_j)$$

for a Borel set $B \in \mathcal{B}(X)$. Similarly, the product measure from the first two observations is

$$\mu_{\lambda,\gamma_0,1,2}(B_1 \times B_2) = \sum_{j=1}^{p}\sum_{l=1}^{p} \eta_j a_{j,l} \gamma_0(B_1 - m_j)\gamma_0(B_2 - m_l)$$

for Borel sets $B_1, B_2 \in \mathcal{B}(X)$.

Theorem 12 *Let γ_0 and $\tilde{\gamma}_0$ be two centred Radon Gaussian measures on a locally convex space X, and let $\lambda = (\{\eta_j\}_{j\in S}, \{a_{jl}\}_{j,l\in S}, \{m_j\}_{j\in S})$ and $\tilde{\lambda} = (\{\tilde{\eta}_j\}_{j\in \tilde{S}}, \{\tilde{a}_{jl}\}_{j,l\in \tilde{S}}, \{\tilde{m}_j\}_{j\in \tilde{S}})$ denote two sets of THMM parameters with state spaces of sizes $p, \tilde{p} \in \mathbb{N}$, respectively, and such that $\det[(a)_{j,l}] \neq 0$ and $\det[(\tilde{a})_{j,l}] \neq 0$. If the two product measures $\mu_{\lambda,\gamma_0,1,2}$ and $\tilde{\mu}_{\tilde{\lambda},\tilde{\gamma}_0,1,2}$ coincide, then γ_0 and $\tilde{\gamma}_0$ coincide and $p = \tilde{p}$ and $\lambda = \tilde{\lambda}$.*

Proof Let X^* be the topological dual of X, and recall that for a Gaussian measure γ that X^*_γ is the closure of $\{f - a_\gamma(f) : f \in X^*\}$ embedded into $L^2(\gamma)$ where $a_\gamma \in X^{**}$ is the mean. By Theorem 2.7.2 of Bogachev (1998), two Gaussian measures γ and $\tilde{\gamma}$ on X are either equivalent or mutually singular. Since $\gamma(\cdot)$ is equivalent to $\gamma(\cdot - m)$ for any $m \in H(\gamma)$, the assumption that $\mu_{\lambda,\gamma_0,1,2}$ and $\tilde{\mu}_{\tilde{\lambda},\tilde{\gamma}_0,1,2}$ coincide implies that γ_0 and $\tilde{\gamma}_0$ are equivalent measures. Hence, $X^*_{\gamma_0}$ and $X^*_{\tilde{\gamma}_0}$ coincide as do $H(\gamma_0)$ and $H(\tilde{\gamma}_0)$; see Theorem 2.4.5 and Proposition 2.7.3 in Bogachev (1998). Hence, we only use the notation $X^*_{\gamma_0}$ and $H(\gamma_0)$ in what follows.

A centred Gaussian measure γ_0 has characteristic function $\phi_{\gamma_0} : X^*_{\gamma_0} \to \mathbb{C}$ defined to be

$$\phi_{\gamma_0}(f) = \exp\left(-\frac{1}{2}\sigma^2(f)\right)$$

for $f \in X^*_{\gamma_0}$ and $\sigma^2(f) = R_{\gamma_0}(f)(f)$ where R_{γ_0} is the covariance operator defined above in Appendix A.1; see Theorem 2.2.4 of Bogachev (1998) for this characterization of Gaussian measures. For a discrete random variable $m_{S_1} \in H(\gamma)$ that takes values $m_1, \ldots, m_p$ with probabilities $\eta_1, \ldots, \eta_p$, the characteristic function is

$$\phi_{\lambda,1}(f) = \sum_{j=1}^{p} \eta_j \exp\left(if(m_j)\right)$$

for $f \in X^*_{\gamma_0}$.

Let ϕ_{γ_0} and $\phi_{\tilde{\gamma}_0}$ be the characteristic functions of γ_0 and $\tilde{\gamma}_0$, respectively. Let $\phi_{\lambda,1}$ and $\phi_{\tilde{\lambda},1}$ be the characteristic functions of m_{S_1} and $\tilde{m}_{\tilde{S}_1}$ for discrete state space random variables $S_1 \in \{1, \ldots, p\}$ and $\tilde{S}_1 = \{1, \ldots, \tilde{p}\}$. Similarly, let $\phi_{\lambda,2}$ and $\phi_{\tilde{\lambda},2}$ be characteristic functions for m_{S_2} and $\tilde{m}_{\tilde{S}_2}$, respectively. Lastly, let $\phi_{\lambda,1,2}$ and $\phi_{\tilde{\lambda},1,2}$ be the characteristic functions for the joint distribution of (m_{S_1}, m_{S_2}) and $(\tilde{m}_{\tilde{S}_1}, \tilde{m}_{\tilde{S}_2})$, respectively.

By assumption, the measures $\mu_{\lambda,\gamma_0,1,2}$ and $\tilde{\mu}_{\tilde{\lambda},\tilde{\gamma}_0,1,2}$ coincide. Hence, the laws for observation O_1 coincide under the two models and thus for any $f \in X^*_{\gamma_0}$,

$$\phi_{\gamma_0}(f)\phi_{\lambda,1}(f) = \phi_{\tilde{\gamma}_0}(f)\phi_{\tilde{\lambda},1}(f).$$

Similarly, for O_2 and any $f \in X^*_{\gamma_0}$,

$$\phi_{\gamma_0}(f)\phi_{\lambda,2}(f) = \phi_{\tilde{\gamma}_0}(f)\phi_{\tilde{\lambda},2}(f).$$

And similarly, the joint law of (O_1, O_2) and any $f, g \in X^*_{\gamma_0}$,

$$\phi_{\gamma_0}(f)\phi_{\gamma_0}(g)\phi_{\lambda,1,2}(f,g) = \phi_{\tilde{\gamma}_0}(f)\phi_{\tilde{\gamma}_0}(g)\phi_{\tilde{\lambda},1,2}(f,g).$$

For fixed non-zero $f, g \in X^*_{\gamma_0}$ and any $t_1, t_2 \in \mathbb{R}$, the characteristic function of $f(m_{S_1})$ is $\varphi_{f,\lambda,1}(t_1) = \phi_{\lambda,1}(t_1 f)$. We similarly introduce the corresponding characteristic functions $\varphi_{f,\tilde{\lambda},1}(t_1)$, $\varphi_{g,\lambda,2}(t_2)$, $\varphi_{g,\lambda,2}(t_2)$, $\varphi_{f,g,\lambda,1,2}(t_1,t_2)$, $\varphi_{f,g,\tilde{\lambda},1,2}(t_1,t_2)$, $\varphi_{f,\gamma_0}(t_1)$, and $\varphi_{f,\tilde{\gamma}_0}(t_1)$. There exists $U \subset \mathbb{R}$, a neighbourhood of 0, such that $\varphi_{f,\gamma_0}(t) \neq 0$ and $\varphi_{f,\tilde{\gamma}_0}(t) \neq 0$ for $t \in U$. Thus, for $(t_1, t_2) \in U \times U$,

$$\varphi_{f,g,\lambda,1,2}(t_1,t_2)\varphi_{f,\tilde{\lambda},1}(t_1)\varphi_{g,\tilde{\lambda},2}(t_2) = \varphi_{f,g,\tilde{\lambda},1,2}(t_1,t_2)\varphi_{f,\lambda,1}(t_1)\varphi_{g,\lambda,2}(t_2).$$

For a fixed $t_1 \in \mathbb{R}$, the functions of $t_2 \in \mathbb{R}$ above have analytic continuations for all $z_2 \in \mathbb{C}$. Hence, $\varphi_{f,g,\lambda,1,2}(t_1, z_2)$, $\varphi_{g,\tilde{\lambda},2}(z_2)$, $\varphi_{f,g,\tilde{\lambda},1,2}(t_1, z_2)$, and $\varphi_{g,\lambda,2}(z_2)$ are entire functions. Similarly, for a fixed $z_2 \in \mathbb{C}$, the above functions of t_1 have analytic continuations for all $z_1 \in \mathbb{C}$. Hence, the above equality extends to

$$\varphi_{f,g,\lambda,1,2}(z_1,z_2)\varphi_{f,\tilde{\lambda},1}(z_1)\varphi_{g,\tilde{\lambda},2}(z_2) = \varphi_{f,g,\tilde{\lambda},1,2}(z_1,z_2)\varphi_{f,\lambda,1}(z_1)\varphi_{g,\lambda,2}(z_2).$$

Next, let $\mathcal{Z} \subset \mathbb{C}$ be the zero set for $\varphi_{f,\lambda,1}(z_1)$, and let $\tilde{\mathcal{Z}} \subset \mathbb{C}$ be the zero set for $\varphi_{f,\tilde{\lambda},1}(z_1)$. The aim of what follows is to show that $\mathcal{Z}$ and $\tilde{\mathcal{Z}}$ coincide. For $z_1 \in \mathcal{Z}$, the previous equality becomes

$$\varphi_{f,g,\lambda,1,2}(z_1,z_2)\varphi_{f,\tilde{\lambda},1}(z_1)\varphi_{g,\tilde{\lambda},2}(z_2) = 0$$

for all $z_2 \in \mathbb{C}$. Thus, the function $z \to \varphi_{f,g,\lambda,1,2}(z_1, z)$ is

$$\varphi_{f,g,\lambda,1,2}(z_1,z) = \sum_{l=1}^{p}\left[\sum_{j=1}^{p}\eta_j a_{j,l}\exp(iz_1 f(m_j))\right]\exp(izg(m_l)).$$

We have that $\varphi_{f,g,\lambda,1,2}(z_1, z) = 0$ for all $z \in \mathbb{C}$ if and only if $\sum_{j=1}^{p}\eta_j a_{j,l}\exp(iz_1 f(m_j)) = 0$ for all $l = 1, \ldots, p$. But the latter only occurs if $\det[(a)_{j,l}] = 0$, which contradicts our assumption. Thus, both $\varphi_{f,g,\lambda,1,2}(z_1, \cdot)$ and $\varphi_{g,\tilde{\lambda},2}(\cdot)$ are entire functions with isolated zeros. This and the above necessitates that $\varphi_{f,\tilde{\lambda},1}(z_1) = 0$, and thus $\mathcal{Z} \subseteq \tilde{\mathcal{Z}}$. Applying this same argument for some $z_1 \in \tilde{\mathcal{Z}}$ results in the reverse inclusion, and hence $\mathcal{Z} = \tilde{\mathcal{Z}}$. By Hadamard's / Weierstrass's factorization theorem (Rudin, 1987, Theorem 15.10), these two functions must coincide up to a factor of $\mathrm{e}^{q(z)}$ for some polynomial q of degree 0 or 1 given that $\varphi_{f,\lambda,1}(z)$ and $\varphi_{f,\tilde{\lambda},1}(z)$ have growth order 1. That is, $\varphi_{f,\lambda,1}(z) = \mathrm{e}^{q(z)}\varphi_{f,\tilde{\lambda},1}(z)$ for all $z \in \mathbb{C}$. As $\varphi_{f,\lambda,1}(0) = \varphi_{f,\tilde{\lambda},1}(0) = 1$, $q(z) = \omega z$ for some $\omega \in \mathbb{C}$. Furthermore, for z restricted to $\mathbb{R}$, $\bar{\varphi}_{f,\lambda,1}(z) = \varphi_{f,\lambda,1}(-z)$, which in turn implies that $\bar{\omega} = -\omega$. Thus, $\omega = ir$ for some $r \in \mathbb{R}$, and $p = \tilde{p}$.

As a consequence of Lemma 13 via the Hahn-Banach theorem on locally convex spaces (Rudin, 1991, Chapter 3), we can pick a linear functional $f^\star$ such that $f^\star(m_{j_1}) \neq f^\star(m_{j_2})$ for all $j_1 \neq j_2$. Hence, after some relabelling, let $f^\star(m_1) < f^\star(m_2) < \ldots < f^\star(m_p)$ and $f^\star(\tilde{m}_1) < f^\star(\tilde{m}_2) < \ldots < f^\star(\tilde{m}_p)$. By subtracting m_1 and $\tilde{m}_1$ from each m_i and $\tilde{m}_i$, respectively, without loss of generality, let $f^\star(m_1) = 0 < f^\star(m_j) < f^\star(m_{j+1})$ and $f^\star(\tilde{m}_1) =$

$0 < f^\star(\tilde{m}_j) < f^\star(\tilde{m}_{j+1})$ for all $1 < j < p$. Therefore, since $\varphi_{f^\star,\lambda,1}(z) = \mathrm{e}^{ir}\varphi_{f^\star,\tilde{\lambda},1}(z)$, we necessarily have that $f^\star(m_j) = f^\star(\tilde{m}_j) + r$, and setting $j = 1$ results in $r = 0$. And thus, $\varphi_{f^\star,\lambda,1}(z) = \varphi_{f^\star,\tilde{\lambda},1}(z)$. The same argument can show that there is a $g^\star$ such that $\varphi_{g^\star,\lambda,2}(z) = \varphi_{g^\star,\tilde{\lambda},2}(z)$ and in turn that $\varphi_{f^\star,g^\star,\lambda,1,2}(z_1,z_2) = \varphi_{f^\star,g^\star,\tilde{\lambda},1,2}(z_1,z_2)$ and that $(\{\eta_j\}_{j\in S}, \{a_{jl}\}_{j,l\in S}, \{f^\star(m_j)\}_{j\in S}) = (\{\tilde{\eta}_j\}_{j\in\tilde{S}}, \{\tilde{a}_{jl}\}_{j,l\in\tilde{S}}, \{f^\star(\tilde{m}_j)\}_{j\in\tilde{S}})$. Lastly, $\varphi_{f^\star,\gamma_0}(t) = \varphi_{f^\star,\tilde{\gamma}_0}(t)$ when $\varphi_{f^\star,\lambda,1}$ is not zero. However, since the zeros are isolated, $\varphi_{f^\star,\gamma_0}(t) = \varphi_{f^\star,\tilde{\gamma}_0}(t)$ for all t by continuity. Ergo, the one dimensional image of the THMM under application of the functional $f^\star$ is identifiable.

Furthermore, we have for any arbitrary $f \in X_\gamma^*$ that there exists an $r_f \in \mathbb{R}$ and permutation π_f on $\{1,\ldots,p\}$ possibly depending on choice of f such that $f(m_j) = f(\tilde{m}_{\pi_f(j)}) + r_f$. Thus, we claim that for some fixed π and $m_0 \in H(\gamma)$ that $m_j = \tilde{m}_{\pi(j)} + m_0$. Indeed, without loss of generality, we take π to be the identity, and assume this is not the case. Then, there are at least two indices $j \neq l$ such that $m_j = \tilde{m}_j + m_{0,j}$ and $m_l = \tilde{m}_l + m_{0,l}$ for $m_{0,j} \neq m_{0,l}$. However, by Lemma 13 and Hahn-Banach again, there exists a linear functional $f_0 \in X_\gamma^*$ such that $f_0(m_{0,j}) \neq f_0(m_{0,l})$. This contradicts $f_0(m_j) = f_0(\tilde{m}_j) + r_f$ for r_f independent of j. Thus, $m_j = \tilde{m}_{\pi(j)} + m_0$ for all $j = 1,\ldots,p$, Furthermore, $m_0 = 0$ as a result of $f^\star$ and $m_1 = \tilde{m}_1 = 0$ from the previous paragraph. Thus, parameters sets must be equal: $\lambda = \tilde{\lambda}$.

In conclusion, we have that $\phi_{\lambda,1}(f) = \phi_{\tilde{\lambda},1}(f)$ and $\phi_{\lambda,2}(g) = \phi_{\tilde{\lambda},2}(g)$ and $\phi_{\lambda,1,2}(f,g) = \phi_{\tilde{\lambda},1,2}(f,g)$ and, consequently, that $\phi_{\gamma_0}(f) = \phi_{\tilde{\gamma}_0}(f)$ for any $f, g \in X_\gamma^*$. Thus, γ_0 and $\tilde{\gamma}_0$ coincide on the cylindrical σ-field. As they are both Radon measures, they necessarily also coincide on the Borel sets. ■

Lemma 13 *Let X be a normed space with $m_1,\ldots,m_p \in X$ such that $m_i \neq m_j$ for all $i \neq j$. Then, there exists a linear functional $f \in X^*$ such that $f(m_i) \neq f(m_j)$ for all $i \neq j$.*

Proof First, let $p = 2$. Let $M = \{c_1m_1 + c_2m_2 : c_1, c_2 \in \mathbb{R}\}$ be a (one or two dimensional) linear subspace of X, and we assume without loss of generality that $m_1 \neq 0$. If there exists a $c_0 \in \mathbb{R}$, $c_0 \neq 1$, such that $m_2 = c_0m_1$ (note that c_0 could equal 0 in the case that $m_2 = 0$), then let $f \in M^*$ be the linear functional such that $f(m_1) = \|m_1\|$ and $f(m_2) = c_0\|m_1\|$. Thus, by Hahn-Banach (Rudin, 1991, Theorem 3.3), we extend f to an element of X^* that separates m_1 and m_2. Otherwise, if m_1 and m_2 are linearly independent, let $f(c_1m_1 + c_2m_2) = c_1d_1 + c_2d_2$ for any choice of distinct nonzero $d_1 \neq d_2 \in \mathbb{R}$, which is a linear functional on the two dimensional subspace M. Furthermore, let $q(x)$ be a seminorm on X such that $q(c_1m_1 + c_2m_2) = |c_1||d_1| + |c_2||d_2|$. As $|f(x)| \leq q(x)$ for $x \in M$, we can extend f to a linear functional on X once again by Hahn-Banach (Rudin, 1991, Theorem 3.3).

To proceed via induction, we next assume for distinct $m_1,\ldots,m_{p-1} \in X$ that there exists an $f_0 \in X^*$ such that $f_0(m_i) \neq f_0(m_j)$ for all $i,j = 1,\ldots,p-1$, $i \neq j$. Let $M_0 = \{\sum_{i=1}^{p-1} c_im_i : c_1,\ldots,c_{p-1} \in \mathbb{R}\}$ be a linear subspace of X. Let $m_p \in X$ be distinct from the other $m_1,\ldots,m_{p-1}$. We once again consider two cases. If $m_p \in M_0$, then there exist $b_1,\ldots,b_{p-1} \in \mathbb{R}$ such that $m_p = \sum_{i=1}^{p-1} b_im_i$. Thus, we can pick an $f \in M_0^*$ such that $f(m_i) = d_i \in \mathbb{R}$ for $i = 1,\ldots,p-1$ by the induction hypothesis where the $\{d_1,\ldots,d_{p-1}\}$ are a distinct set of real numbers. Furthermore, we can choose f such that these d_i satisfy $d_p := f(m_p) = \sum_{i=1}^{p-1} b_id_i \neq d_j$ for any $j = 1,\ldots,p-1$. If $m_p \notin M_0$, then

let $M_1 = \{c_p m_p : c_p \in \mathbb{R}\}$ be the one dimensional subspace of X spanned by m_p, and let $M = \{\sum_{i=1}^{p} c_i m_i : c_1, \ldots, c_{p-1} \in \mathbb{R}\}$ be the subspace spanned by all of the $m_1, \ldots, m_p$. Let $f_0 \in M^*$ be such that $f_0(m_i) = d_i$ for $i = 1, \ldots, p-1$ and $f_0(m_p) = 0$. Let $f_1 \in M^*$ be such that $f_1(m_p) = d_p \notin \{d_1, \ldots, d_{p-1}\}$ and $f_1(m) = 0$ for all $m \in M_0$. Defining $f := f_0 + f_1$ on M^* and extending to X^* via Hahn-Banach completes the proof. ■

References

Rachel MacKay Altman. Mixed Hidden Markov Models: an extension of the Hidden Markov Model to the longitudinal data setting. Journal of the American Statistical Association, 102(477):201–210, 2007.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. The annals of mathematical statistics, 37(6):1554–1563, 1966.

Leonard E Baum, Ted Petrie, George Soules, and Norman 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.

Matthew Beal, Zoubin Ghahramani, and Carl Rasmussen. The infinite Hidden Markov Model. Advances in neural information processing systems, 14, 2001.

Corinne Berzin and José R León. Estimation in models driven by fractional Brownian motion. In Annales de l'IHP Probabilités et statistiques, volume 44, pages 191–213, 2008.

Vladimir I Bogachev. Gaussian measures, volume 62. American Mathematical Society Providence, 1998.

Charles Bouveyron. funFEM: Clustering in the Discriminative Functional Subspace, 2021. URL `https://CRAN.R-project.org/package=funFEM`. R package version 1.2.

Mireille Capitaine. Onsager-Machlup functional for some smooth norms on Wiener space. Probability theory and related fields, 102(2):189–201, 1995.

Aliprantis D Charalambos and Kim C Border. Infinite Dimensional Analysis: A Hitchhiker's Guide. Springer-Verlag Berlin and Heidelberg GmbH & Company KG, 2013.

Sotirios P Chatzis. Hidden Markov Models with nonelliptically contoured state densities. IEEE transactions on pattern analysis and machine intelligence, 32(12):2297–2304, 2010.

Ying Chen, Xin Zhu, and Wenxi Chen. Automatic sleep staging based on ECG signals using Hidden Markov Models. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 530–533, 2015. doi: 10.1109/EMBC.2015.7318416.

Rodney Coleman. Calculus on normed vector spaces. Springer Science & Business Media, 2012.

Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. Minimax adaptive estimation of nonparametric Hidden Markov Models. The Journal of Machine Learning Research, 17 (1):3842–3884, 2016.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

L Doroshenkov, VA Konyshev, and Sergey Selishchev. Classification of human sleep stages based on EEG processing using Hidden Markov Models. Meditsinskaia tekhnika, 41:24–8, 01 2007. doi: 10.1007/s10527-007-0006-5.

Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software, 51(4):1–28, 2012. URL `https://www.jstatsoft.org/v51/i04/`.

John D Ferguson. Hidden Markov analysis: an introduction. Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech, 1980.

Frédéric Ferraty and Philippe Vieu. Nonparametric functional data analysis: theory and practice. Springer Science & Business Media, 2006.

Arthur Flexer, Georg Dorffner, Peter Sykacekand, and Iead Rezek. An automatic, continuous and probabilistic sleep stager based on a Hidden Markov Model. Applied Artificial Intelligence, 16(3):199–207, 2002. doi: 10.1080/088395102753559271. URL `https://doi.org/10.1080/088395102753559271`.

Arthur Flexer, Georg Gruber, and Georg Dorffner. A reliable probabilistic sleep stager based on a single EEG signal. Artificial intelligence in medicine, 33:199–207, 03 2005. doi: 10.1016/j.artmed.2004.04.004.

Elisabeth Gassiat and Judith Rousseau. Nonparametric finite translation Hidden Markov Models and extensions. Bernoulli, 22(1):193–212, 2016. ISSN 13507265. URL `http://www.jstor.org/stable/43863921`.

Élisabeth Gassiat, Alice Cleynen, and Stephane Robin. Inference in finite state space non parametric Hidden Markov Models and applications. Statistics and Computing, 26:61–71, 2016.

Yaozhong Hu, David Nualart, and Hongjuan Zhou. Parameter estimation for fractional Ornstein–Uhlenbeck processes of general Hurst parameter. Statistical Inference for Stochastic Processes, 22(1):111–142, 2019.

James P Hughes, Peter Guttorp, and Stephen P Charles. A non-homogeneous Hidden Markov Model for precipitation occurrence. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48(1):15–30, 1999.

Nobuyuki Ikeda and Shinzo Watanabe. Stochastic differential equations and diffusion processes. Elsevier, 2014.

Biing Hwang Juang and Laurence R Rabiner. Hidden Markov Models for speech recognition. Technometrics, 33(3):251–272, 1991.

Biing-Hwang Juang and Lawrence Rabiner. Mixture autoregressive Hidden Markov Models for speech signals. IEEE Transactions on Acoustics, Speech, and Signal Processing, 33 (6):1404–1413, 1985.

Bob Kemp and Hilbert A. C. Kamphuisen. Simulation of human hypnograms using a Markov chain model. Sleep, 9(3):405–14, 1986.

Kestutis Kubilius and Yuliya S Mishura. The rate of convergence of Hurst index estimate for the stochastic differential equation. Stochastic processes and their applications, 122 (11):3718–3739, 2012.

Ethan Lawler, Kim Whoriskey, William H Aeberhard, Chris Field, and Joanna Mills Flemming. The conditionally autoregressive Hidden Markov Model (carhmm): Inferring behavioural states from animal tracking data exhibiting conditional autocorrelation. Journal of Agricultural, Biological and Environmental Statistics, 24(4):651–668, 2019.

Luc Lehéricy. State-by-state minimax adaptive estimation for nonparametric Hidden Markov Models. The Journal of Machine Learning Research, 19(1):1432–1477, 2018.

Jüri Lember and Alexey Koloydenko. The adjusted Viterbi training for Hidden Markov Models. Bernoulli, 14(1):180 – 206, 2008. doi: 10.3150/07-BEJ105. URL `https://doi.org/10.3150/07-BEJ105`.

Jüri Lember and Alexey A Koloydenko. Bridging Viterbi and posterior decoding: a generalized risk approach to hidden path inference based on Hidden Markov Models. The Journal of Machine Learning Research, 15(1):1–58, 2014.

Louis A Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. IEEE Transactions on Information Theory, 28(5):729–734, 1982.

Andrea Martino, Giuseppina Guatteri, and Anna Maria Paganoni. Hidden Markov Models for multivariate functional data. Statistics & Probability Letters, 167:108917, 2020.

Sílvia Moret and David Nualart. Onsager-Machlup functional for the fractional Brownian motion. Probability theory and related fields, 124(2):227–260, 2002.

Jared O'Connell and Søren Højsgaard. Hidden Semi Markov Models for multiple observation sequences: The mhsmm package for R. Journal of Statistical Software, 39(4):1–22, 2011. URL `http://www.jstatsoft.org/v39/i04/`.

Shing-Tai Pan, Chih-En Kuo, Jian-Hong Zeng, and Sheng-Fu Liang. A transition-constrained discrete Hidden Markov Model for automatic sleep staging. Biomedical engineering online, 11(1):1–19, 2012.

Grigorios A Pavliotis. Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations, volume 60. Springer, 2014.

William D Penny and Stephen J Roberts. Gaussian observation Hidden Markov Models for EEG analysis. Imperial College TR-98-12, London, UK, Tech. Rep, 1998.

Alan B Poritz and A G Richter. On Hidden Markov Models in isolated word recognition. In ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 11, pages 705–708. IEEE, 1986.

Lawrence R Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286, 1989.

Walter Rudin. Real and complex analysis. Tata McGraw-Hill Education, 1987.

Walter Rudin. Functional Analysis. McGraw-Hill Science/Engineering/Math, 1991.

Laura M Sangalli, Piercesare Secchi, Simone Vantini, and Valeria Vitelli. K-mean alignment for curve clustering. Computational Statistics & Data Analysis, 54(5):1219–1233, 2010.

Laura M Sangalli, Piercesare Secchi, Simone Vantini, et al. Analysis of aneurisk65 data: k-mean alignment. Electronic Journal of Statistics, 8(2):1891–1904, 2014.

Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal, 8(1):289–317, 2016. URL `https://doi.org/10.32614/RJ-2016-021`.

Larry A Shepp and Ofer Zeitouni. A note on conditional exponential moments and Onsager-Machlup functionals. The Annals of Probability, pages 652–654, 1992.

Evan Sidrow, Nancy Heckman, Sarah ME Fortune, Andrew W Trites, Ian Murphy, and Marie Auger-Méthé. Modelling multi-scale, state-switching functional data with Hidden Markov Models. Canadian Journal of Statistics, 2021.

Aymeric Stamm. fdacluster: Joint Clustering and Alignment of Functional Data, 2023. URL `https://CRAN.R-project.org/package=fdacluster`. R package version 0.2.1.

Luis Enrique Sucar. Probabilistic graphical models. Advances in Computer Vision and Pattern Recognition. London: Springer London. doi, 10(978):1, 2015.

Y Takahashi and S Watanabe. The probability functionals (Onsager-Machlup functions) of diffusion processes. In Stochastic Integrals, pages 433–463. Springer, 1981.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE transactions on Information Theory, 13(2):260–269, 1967.

Peter Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Transactions on Audio and Electroacoustics, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

CF Jeff Wu. On the convergence properties of the EM algorithm. The Annals of statistics, pages 95–103, 1983.

Michelle C. Yang and Carolyn J. Hursch. The use of a semi-Markov model for describing sleep patterns. Biometrics, 29 4:667–76, 1973.

Willard I Zangwill. Nonlinear programming: a unified approach, volume 52. Prentice-Hall Englewood Cliffs, NJ, 1969.

Ofer Zeitouni. On the Onsager-Machlup functional of diffusion processes around non C2 curves. The Annals of Probability, pages 1037–1054, 1989.

William W. K. Zung, Thomas H. Naylor, Daniel T. Gianturco, and W. P. Wilson. Computer simulation of sleep EEG patterns with a Markov chain model. Recent advances in biological psychiatry, 8:335–55, 1965.