Geometric Learning of Hidden Markov Models via a Method of Moments Algorithm

Berlin Chen, Cyrus Mostajeran, Salem Said

MDPI

Article

Geometric Learning of Hidden Markov Models via a Method of Moments Algorithm†

Berlin Chen [1], Cyrus Mostajeran [2,3], and Salem Said [4]

[1] Princeton Neuroscience Institute, Princeton University, USA
[2] School of Physical and Mathematical Sciences, Nanyang Technological University (NTU), Singapore
[3] Department of Engineering, University of Cambridge, United Kingdom
[4] CNRS, Laboratoire Jean Kuntzmann, Université Grenoble-Alpes, Grenoble, France
* Cyrus Mostajeran
† Submitted to International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, IHP, Paris, July 18-22, 2022.

Version November 4, 2023 submitted to Entropy

Abstract: We present a novel algorithm for learning the parameters of hidden Markov models (HMMs) in a geometric setting where the observations take values in Riemannian manifolds. In particular, we elevate a recent second-order method of moments algorithm that incorporates non-consecutive correlations to a more general setting where observations take place in a Riemannian symmetric space of non-positive curvature and the observation likelihoods are Riemannian Gaussians. The resulting algorithm decouples into a Riemannian Gaussian mixture model estimation algorithm followed by a sequence of convex optimization procedures. We demonstrate through examples that the learner can result in significantly improved speed and numerical accuracy compared to existing learners.

Keywords: hidden Markov models; method of moments; Riemannian geometry; Riemannian Gaussian mixtures; covariance matrices; geometric statistics

1. Introduction

Hidden Markov models (HMMs) describe states with Markovian dynamics that are hidden in the sense that they are only accessible via observations by a noisy sensor. Specifically, at every time-step k, an observation y_k is sampled from an observation space $\mathcal{Y}$ according to the HMM's *observation likelihoods*, which specify the probability of making a particular observation, conditioned on the system being in a certain state. Despite their structural simplicity, HMMs are capable of modeling complex signals and have indeed become a standard tool in the modeling of stochastic time-series [1] in recent decades and have found applications in a wide range of fields including computational biology [2,3], signal and image analysis [4], speech recognition [5,6], and financial modeling [7].

In order to apply an HMM, it is often necessary to estimate its parameters from data. The standard approach to estimating the parameters of an HMM is using a *maximum likelihood* (ML) criterion. Numerical algorithms for computing the ML estimate are dominated by iterative local-search procedures that aim to maximize the likelihood of observed data, such as the *expectation-maximization* (EM) algorithm [1,4]. Unfortunately, these schemes are only guaranteed to converge to local stationary points of the typically non-convex likelihood function and as a result often become trapped in local optima. Thus, to have a chance of converging to a global optimum, a good initialization is usually required. Another drawback of such methods is the significant computational cost associated with long runtimes due to costly iterations for large datasets.

In order to overcome such challenges, methods of moments have been introduced for HMMs [8–14]. Originally, these methods relied on empirical estimation of correlations between consecutive pair-

or triplet-wise observations to compute estimates of the HMM parameters. Although computationally attractive, such methods suffered from a loss of accuracy due to a focus on low order correlations in the data. In response, Mattila et al. [15,16] extended these methods to include non-consecutive correlations in the data, resulting in improved accuracy while retaining their attractive computational properties.

1.1. Hidden Markov models with manifold-valued observations

The development and analysis of statistical procedures and optimization algorithms on manifolds and nonlinear spaces more broadly have been the subject of intense and growing research interest in recent decades due to the ubiquity of manifold-valued data in a wide range of applications [17–23]. Since the application of Euclidean algorithms to such data often has a significantly negative impact on the accuracy and interpretability of the results, it is necessary to devise algorithms that respect the intrinsic geometry of the data. In this work, we turn our attention to HMMs with observations in a Riemannian manifold [24,25]. In particular, we restrict our attention to the class of models with observations in Riemannian symmetric spaces of non-positive curvature, which include hyperbolic spaces, as well as spaces of real, complex, and quaternionic positive definite matrices. We have three motivations for this restriction: (1) standard operations on such spaces have relatively favorable computational properties due to symmetries, (2) there exists a theory of Riemannian Gaussian distributions on such spaces together with associated algorithms such as Riemannian Gaussian mixture estimation [26,27], and (3) they are applicable to a substantial class of problems involving manifold-valued data, including applications with data in the form of covariance matrices [27].

1.2. Contributions and paper outline

Our main contribution in this paper is to extend the second-order method of moments algorithm with non-consecutive correlations developed by Mattila et al. [15,16] to the setting of HMMs with observations in a Riemannian symmetric space of non-positive curvature, where the observation likelihoods take the form of Riemannian Gaussians [27,28]. The paper is organized as follows. In Section 2, we describe HMMs with manifold-valued observations and review the necessary geometric background. In Section 3, we review the method of moments algorithms for HMMs and describe how they manifest in the geometric setting. In Section 4, we present a number of simulations based on these algorithms and conclude with a discussion in Section 5.

1.3. Notation

We denote the i-th entry of a vector by $[\cdot]_i$, and the element at row i and column j of a matrix by $[\cdot]_{ij}$. Vectors are assumed to be column vectors unless transposed. The vector of all ones is denoted $\mathbf{1}$. We interpret inequalities between vectors and matrices to hold elementwise. The operator diag acts on vectors and returns the matrix where the vector has been placed on the diagonal, and all other elements set to zero. The matrix Frobenius norm is denoted $\|\cdot\|_F$. The probability of an event A is denoted $\mathbb{P}(A)$.

2. Hidden Markov models on manifolds

We consider a discrete-time hidden Markov model with a finite-state Markov chain on the state space $\mathcal{X} = \{1, \dots, N\}$ with time-homogeneous $N \times N$ transition probability matrix P with elements

$$[P]_{ij} = \mathbb{P}[x_{k+1} = j | x_k = i]. \tag{1}$$

The initial and stationary distributions of the HMM exist under appropriate assumptions and are denoted by $\pi_0 \in \mathbb{R}^N$ and $\pi_\infty \in \mathbb{R}^N$, respectively. The HMM is said to be *stationary* if $\pi_0 = \pi_\infty$.

We assume that the states are hidden and can only be accessed through observations in a Riemannian symmetric space of non-positive curvature so that the Riemannian Gaussian distribution with probability density function

$$p(y|\bar{y},\sigma) = \frac{1}{Z(\sigma)} \exp\left[-\frac{d^2(y,\bar{y})}{2\sigma^2}\right] \tag{2}$$

with respect to the Riemannian volume measure $dv(y)$ on $\mathcal{Y}$ is well-defined for any $\bar{y} \in \mathcal{Y}$ and $\sigma > 0$, as outlined in [27]. $d(\cdot,\cdot)$ denotes the Riemannian distance function on $\mathcal{Y}$ and $Z(\sigma)$ denotes the normalization factor of the Riemannian Gaussian, whose efficient computation has been the subject of interest in recent years [28–31]. We assume that the observations are sampled from $\mathcal{Y}$ according to conditional probability densities

$$B(y_k = y|x_k = j) = p(y|\bar{y}_j,\sigma_j), \tag{3}$$

for $j = 1,\dots,N$ where $p(\cdot|\bar{y}_j,\sigma_j)$ is a Riemannian Gaussian density function of the form (2) with mean $\bar{y}_j \in \mathcal{Y}$ and dispersion $\sigma_j > 0$.

To use an HMM for applications such as filtering or prediction, its model parameters must be specified or estimated in advance. This task can be formulated as the following learning problem for HMMs:

Problem 1. *Given a sequence $y_1,\dots,y_D$ of observations in $\mathcal{Y}$ generated by an HMM of known state space $\mathcal{X} = \{1,\dots,N\}$, estimate the conditional probability densities B and the matrix of transition probabilities P.*

The learning problem is well-posed under the standard assumptions that the HMM is ergodic (irreducible and aperiodic) and identifiable [4,10,15,16]. A special case of the learning problem that is worth noting is that of the *known-sensor HMM*, in which the observation likelihoods B are assumed to be known. Known-sensor HMMs are motivated by applications in which the sensor is designed by the user, such as a target tracking system whose sensor specifications can be determined prior to deployment.

Various methods since the inception of HMMs have focused on maximizing the likelihood in terms of both B and P; however, recent efforts have demonstrated the potential of methods that decouple the problem [12,13] and estimate B and P sequentially. Specifically, in *parametric-output HMMs* (e.g., Gaussian HMMs), the observation likelihoods are estimated via a general mixture model learner as a first step, followed by identification of the transition matrix P as a second step [12]. In the first step, assuming that the underlying Markov chain behaves well (e.g. is recurrent) and mixes rapidly, in stationarity, each observation y_k from the HMM can be interpreted as having been sampled from the mixture distribution density

$$p(y) = \sum_{i=1}^{N} [\pi_\infty]_i B(y|\bar{y}_i,\sigma_i). \tag{4}$$

Since we are assuming that the observation likelihoods belong to the family of isotropic Riemannian Gaussians on $\mathcal{Y}$, the density (4) can be estimated using one of several algorithms for the estimation of mixtures of Riemannian Gaussian distributions including expectation-maximization (EM) [26,27], stochastic EM [32], and online variants [33]. The second step is then equivalent to the identification of a known-sensor HMM.

3. Method of moments algorithms for geometric learning of hidden Markov models

3.1. Method of moments for HMMs

We begin with a brief review of the method of moments algorithm for HMMs developed by Mattila et al. in [15]. The significance of this work is that it extends previous method of moments algorithms for HMMs that were based on correlations between consecutive pair- or triplet-wise observations to include non-consecutive correlations in the data. In doing so, the authors improve the accuracy of the approach by reducing the volume of neglected information inherent in the data while maintaining the computationally attractive properties of previous method of moments algorithms.

Before presenting the algorithm in the setting of HMMs with manifold-valued observations, we briefly review a summary of the key steps involved in the second-order algorithm of Mattila et al. [15] in the simplest setting where the observations take place in a finite observation alphabet $\{1, \dots, Y\}$ with a known $N \times Y$ observation matrix B:

$$[B]_{ij} = \mathbb{P}[y_k = j | x_k = i]. \tag{5}$$

Methods of moments for HMMs (e.g. [8–14]) involve the empirical estimation of low-order correlations in the data, such as pairs $\mathbb{P}[y_k, y_{k+1}]$ or triplets $\mathbb{P}[y_k, y_{k+1}, y_{k+2}]$, followed by computation of the HMM parameter estimates by minimizing the discrepancy between the empirical estimates and their analytical expressions via a series of convex optimization problems. In Mattila et al. [15], the authors extend such methods to include non-consecutive correlations of the form $\mathbb{P}[y_k, y_{k+\tau}]$ with $\tau = 1, 2, \dots, \bar{\tau}$ where the number $\bar{\tau}$ is a user-defined lag parameter.

The *lag-τ second-order moments* $M_2(k, \tau) \in \mathbb{R}^{Y \times Y}$ of the HMM are defined as the matrices

$$[M_2(k, \tau)]_{ij} = \mathbb{P}[y_k = i, y_{k+\tau} = j], \tag{6}$$

where $i, j = 1, \dots, Y$ and $\tau \geq 0$. The case $\tau = 0$ reduces to the first-order moments $[M_1(k)]_i = \mathbb{P}[y_k = i]$, where $M_1(k) \in \mathbb{R}^Y$, which for notational convenience is expressed as a special case of second-order moments by writing $M_2(k, 0) = \operatorname{diag}(M_1(k))$. For a stationary HMM (i.e., $\pi_0 = \pi_\infty$), it can be readily verified that the lag-τ second-order moments are related to the HMM parameters according to the equations

$$M_2(k, \tau) = B^T \operatorname{diag}(\pi_\infty) P^\tau B, \qquad M_2(k, 0) = \operatorname{diag}(B^T \pi_\infty), \tag{7}$$

for any $\tau > 0$.

The lag-τ second-order moments can be empirically estimated from data as $\hat{M}_2(\tau)$ according to the equation

$$[\hat{M}_2(\tau)]_{ij} = \frac{1}{D - \tau} \sum_{k=1}^{D-\tau} I\{y_k = i, y_{k+\tau} = j\}, \tag{8}$$

for $\tau = 0, 1, \dots, \bar{\tau}$, where D is the number of observations and I denotes the indicator function. The next step in the method is moment matching through the minimization of the discrepancy between the empirical estimate $\hat{M}(\tau)$ and its analytical expression by solving the following convex (quadratic) optimization problems:

1. Solve

$$\begin{aligned} \min_{\hat{\pi}_\infty \in \mathbb{R}^{N \times N}} \quad & \|\hat{M}_2(0) - \operatorname{diag}(B^T \hat{\pi}_\infty)\|_F^2 \\ \text{s.t.} \quad & \hat{\pi}_\infty \geq 0, \quad \mathbf{1}^T \hat{\pi}_\infty = 1, \end{aligned} \tag{9}$$

   and set $\hat{A}(0) = \operatorname{diag}(\hat{\pi}_\infty)$.

2. For $\tau = 1, \ldots, \bar{\tau}$, solve

$$\begin{aligned} \min_{\hat{P}(\tau) \in \mathbb{R}^{N \times N}} \quad & \|\hat{M}_2(\tau) - B^T \hat{A}(\tau - 1)\hat{P}(\tau)B\|_F^2 \\ \text{s.t.} \quad & \hat{P}(\tau) \geq 0, \quad \hat{P}(\tau)\mathbf{1} = \mathbf{1}, \end{aligned} \tag{10}$$

and set $\hat{A}(\tau) = \hat{A}(\tau - 1)\hat{P}(\tau)$.

The output of the above moment matching procedure is a sequence $\hat{A}(0), \ldots, \hat{A}(\bar{\tau})$. In the final step, we use this sequence to estimate the transition matrix P by solving the following least-squares problem, which incorporates information from every lag by construction.

$$\begin{aligned} \min_{\hat{P} \in \mathbb{R}^{N \times N}} \quad & \left\| \begin{bmatrix} \hat{A}(0) \\ \vdots \\ \hat{A}(\bar{\tau} - 1) \end{bmatrix} \hat{P} - \begin{bmatrix} \hat{A}(1) \\ \vdots \\ \hat{A}(\bar{\tau}) \end{bmatrix} \right\|_F^2 \\ \text{s.t.} \quad & \hat{P} \geq 0, \quad \hat{P}\mathbf{1} = \mathbf{1}. \end{aligned} \tag{11}$$

The dominant contribution to the computational cost of the above algorithm is independent of the data size D and scales linearly with the number of lags $\bar{\tau}$ included. In contrast, each iteration of the EM algorithm has a complexity of $\mathcal{O}(N^2 D)$. In addition to favorable computational properties, it is shown in [15,16] that the above algorithm is strongly consistent under reasonable assumptions. That is, as the number of samples grows, we expect the estimate of the transition matrix P to converge to its true value.

3.2. Geometric learning of HMMs using method of moments

We now return to the problem of estimating the parameters of an HMM with observations in a Riemannian manifold $\mathcal{Y}$ via an extension of the second-order method of moments presented earlier. We assume conditional probability densities to be given by Riemannian Gaussians of the form (2). The first stage of the process is to estimate the means and variances of the observation densities from data by employing a Riemannian Gaussian mixture learner [27,32,33]. In the case of a known-sensor HMM, this would be unnecessary as the observation densities are known a priori. In the next stage, we use a kernel trick outlined in [12,16] to extend the pairwise correlations between discrete-valued observations $M_2(\tau)$ to an analogous quantity $H(\tau) \in \mathbb{R}^{N \times N}$ applicable in the setting of continuous observation spaces. H is then related to the parameters of the HMM according to the equations

$$\begin{aligned} H(0) &= \operatorname{diag}(K\pi_\infty), \\ H(\tau) &= K^T \operatorname{diag}(\pi_\infty) P^\tau K, \end{aligned} \tag{12}$$

for $\tau = 1, \ldots, \bar{\tau} \in \mathbb{N}$, where π_∞ is the HMM stationary distribution which can be estimated from (4), and $K \in \mathbb{R}^{N \times N}$ is defined as

$$[K]_{ij} = \int_{\mathcal{Y}} B(y \mid x = i) B(y \mid x = j)\, dv(y). \tag{13}$$

The $N \times N$ matrix K in (13) is called the the *effective observation matrix* in [12,16] and replaces the $N \times Y$ observation matrix (5). We can compute K using Monte Carlo techniques based on sampling from Riemannian Gaussians [27].

The elements of the left-hand side of (12) can be interpreted as conditional expectations with respect to the joint probability distribution of y_k and $y_{k+\tau}$, which can be empirically estimated from HMM observations as

$$[\hat{H}(0)]_{ii} = \frac{1}{D}\sum_{k=1}^{D} B(y_k|x=i), \tag{14}$$

$$[\hat{H}(\tau)]_{ij} = \frac{1}{D-\tau}\sum_{k=1}^{D-\tau} B(y_k|x=i)B(y_{k+\tau}|x=j) \tag{15}$$

in analogy with empirical estimate (8) employed in the case of HMMs with a discrete observation space.

Following the estimation of $H(\tau)$ and the computation of K, the moment matching procedure now takes the form of minimizing the discrepancy between the empirical estimate $\hat{H}(\tau)$ and the corresponding analytical expressions in (12). Specifically, in the case of the known-sensor HMM, we solve the following sequence of convex (quadratic) optimization problems:

1. Solve
$$\begin{aligned} \min_{\hat{\pi}_\infty \in \mathbb{R}^{N\times N}} \quad & \|\hat{H}(0) - \mathrm{diag}(K^T\hat{\pi}_\infty)\|_F^2 \\ \text{s.t.} \quad & \hat{\pi}_\infty \geq 0, \quad \mathbf{1}^T\hat{\pi}_\infty = 1, \end{aligned} \tag{16}$$
and set $\hat{A}(0) = \mathrm{diag}(\hat{\pi}_\infty)$.

2. For $\tau = 1, \ldots, \bar{\tau}$, solve
$$\begin{aligned} \min_{\hat{P}(\tau)\in\mathbb{R}^{N\times N}} \quad & \|\hat{H}(\tau) - K^T\hat{A}(\tau-1)\hat{P}(\tau)K\|_F^2 \\ \text{s.t.} \quad & \hat{P}(\tau) \geq 0, \quad \hat{P}(\tau)\mathbf{1} = \mathbf{1}, \end{aligned} \tag{17}$$
and set $\hat{A}(\tau) = \hat{A}(\tau-1)\hat{P}(\tau)$.

The output is once again a sequence $\hat{A}(0), \ldots, \hat{A}(\bar{\tau})$, which is used to compute an estimate for the transition matrix P by solving (11).

To summarize, the algorithm follows a 2-stage procedure to learn the parameters of an HMM with observations in a Riemannian manifold admitting well-defined Gaussian densities of the form (2) from data. In stage 1, Riemannian Gaussian mixture estimation is employed to compute estimates for the conditional likelihoods B, which are then used in stage 2 to compute an estimate for the transition probabilities P by solving a series of convex optimization problems.

4. Simulations

We now present the results of several numerical experiments on learning HMMs with manifold-valued observations. In the first example, observations take place in the Poincaré disk model of hyperbolic 2-space. Poincaré models of hyperbolic spaces have been a subject of increasing interest in machine learning in recent years due to their ability to efficiently represent hierarchical data [34]. In the second example, we consider a model with observations in the manifold of 2×2 symmetric positive definite (SPD) matrices equipped with the standard affine-invariant Rao-Fisher metric [26].

4.1. Example 1: Observations in hyperbolic space

We consider the example of an HMM with $N = 3$ hidden states with initial distribution $\pi_0 = (1, 0, 0)^T$ and transition matrix

$$P = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \tag{18}$$

and observations generated from a Riemannian Gaussian model in the Poincaré disk $\mathcal{Y} = \{y \in \mathbb{C} : |y| < 1\}$ with associated means $\bar{y}_1 = 0$, $\bar{y}_2 = 0.29 + 0.82i$, $\bar{y}_3 = -0.29 + 0.82i$ and standard deviations $\sigma_1 = 0.1$, $\sigma_2 = 0.4$, $\sigma_3 = 0.4$ as studied in [24] in the context of estimation using the EM algorithm. The Riemannian distance function $d(\cdot, \cdot)$ and the Riemannian Gaussian normalization factor $Z(\sigma)$ are given by

$$d(y, z) = \operatorname{acosh}\left(1 + \frac{2|y - z|^2}{(1 - |y|^2)(1 - |z|^2)}\right), \qquad Z(\sigma) = 2\pi\sqrt{\frac{\pi}{2}}\sigma e^{\frac{\sigma^2}{2}} \operatorname{erf}\left(\frac{\sigma}{\sqrt{2}}\right), \tag{19}$$

respectively, where erf denotes the error function [35].

We employed the second-order method of moments algorithm of Section 3.2 to learn the parameters of this HMM from observations alone. The model was fitted on 20 HMM chains, each with 10,000 observations. In our implementation, we used the mixture estimation algorithm of [26] to estimate the density (4). The full results are reported in Table 1, where the true and estimated Gaussian means are denoted by $\bar{y}_i$ and $\hat{y}_i$, respectively. On repeating the experiment with varying $\bar{\tau}$ and the same random seed—and hence the same estimates for means and dispersions by construction—we observed that incorporating non-consecutive data (i.e., $\bar{\tau} > 1$) up to $\bar{\tau} = 3$ significantly improved our estimate for P and produced a more accurate estimate than alternative algorithms [24,25]. Comparing the empirical performance of our algorithm to the numerical results reported in [24], we observed that our algorithm performed competitively, while requiring only a fraction of the runtime with the same number of observations. In comparison to the online learning algorithm of [25], which we employed on the same learning problem, we observed improved performance for $\bar{\tau} > 1$, with the method of moments algorithm with $\bar{\tau} = 3$ producing the most accurate estimate of P out of all considered methods. Interestingly, the runtime of our algorithm was not noticeably affected by the choice of $\bar{\tau}$ in this example since the mixture estimation and computation of K (13) accounted for the dominant contribution to the computational cost.

Table 1. *Comparison of the performance of the method of moments algorithm proposed in this paper against previously published algorithms for estimating HMMs with observations in the Poincaré disk.*

	EM algorithm from [24]	**Online algorithm from [25]**	**Our proposed algorithm with (a) $\bar{\tau} = 1$, (b) $\bar{\tau} = 2$, (c) $\bar{\tau} = 3$**
Mean error, $\left(\sum_i d^2(\bar{y}_i, \hat{y}_i)\right)^{1/2}$	0.88	0.97	0.69
Dispersion error, $\left(\sum_i (\sigma_i - \hat{\sigma}_i)^2\right)^{1/2}$	0.42	0.37	0.34
Transition matrix error, $\|P - \hat{P}\|_F$	0.35	0.30	(a) 0.42, (b) 0.26, (c) 0.21
Average runtime	$\sim$ 1 hour	$\sim$ 190 sec	$\sim$ 20 sec

4.2. Example 2: Observations in the manifold of 2×2 SPD matrices with $N = 5$ hidden states

We now consider an HMM with $N = 5$ hidden states that are accessible through noisy observations in the manifold of 2×2 SPD matrices generated from a Riemannian Gaussian model with means $\bar{y}_i$ and standard deviations σ_i given in Table 2. Here the Riemannian distance function $d(\cdot, \cdot)$ and the Riemannian Gaussian normalization factor $Z(\sigma)$ are given by

$$d(y, z) = \|\log(y^{-1/2} z y^{-1/2})\|_F, \qquad Z(\sigma) = (2\pi)^{\frac{3}{2}} \sigma^2 e^{\frac{\sigma^2}{4}} \operatorname{erf}\left(\frac{\sigma}{2}\right). \tag{20}$$

While the expression for the Riemannian distance function holds true for higher dimensional SPD matrices, the analytical expression for $Z(\sigma)$ in (20) is only valid in the 2×2 case. Nonetheless, $Z(\sigma)$ can be directly computed or approximated for higher dimensional SPD matrices [26–31].

The transition matrix P of the underlying Markov chain is

$$P = \begin{bmatrix} 0.3 & 0.1 & 0.2 & 0.1 & 0.3 \\ 0.1 & 0.4 & 0.2 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.3 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.2 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.1 & 0.3 \end{bmatrix}. \tag{21}$$

We employed our proposed geometric second-order method of moments algorithm with $\bar{\tau} = 1$ to sequentially estimate the underlying Gaussian model and the probability transition matrix from 10,000 observations. The results of the Gaussian mixture estimation procedure are reported in Table 2 and demonstrate a high level of accuracy. The estimated Riemannian Gaussian model with means $\hat{y}_i$ and standard deviations $\hat{\sigma}_i$ as well as the observations used to learn the model are visualized in Figure 1.

Table 2. *True and estimated Riemannian Guassian mixture model parameters. $\hat{y}_i$ and $\hat{\sigma}_i$ denote the estimated Riemannian Gaussian means and standard deviations, respectively. π_∞ and $\hat{\pi}_\infty$ denote the true and estimated stationary distributions, respectively.*

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
$\bar{y}_i$	$\begin{bmatrix} 1.646 & 0.056 \\ 0.056 & 2.379 \end{bmatrix}$	$\begin{bmatrix} 2.294 & 0.744 \\ 0.744 & 1.415 \end{bmatrix}$	$\begin{bmatrix} 2.631 & -0.127 \\ -0.127 & 1.277 \end{bmatrix}$	$\begin{bmatrix} 0.674 & 0.454 \\ 0.454 & 2.056 \end{bmatrix}$	$\begin{bmatrix} 1.829 & -0.919 \\ -0.919 & 1.602 \end{bmatrix}$
$\hat{y}_i$	$\begin{bmatrix} 1.642 & 0.051 \\ 0.051 & 2.383 \end{bmatrix}$	$\begin{bmatrix} 2.300 & 0.743 \\ 0.743 & 1.412 \end{bmatrix}$	$\begin{bmatrix} 2.642 & -0.128 \\ -0.128 & 1.277 \end{bmatrix}$	$\begin{bmatrix} 0.672 & 0.454 \\ 0.454 & 2.057 \end{bmatrix}$	$\begin{bmatrix} 1.830 & -0.920 \\ -0.920 & 1.604 \end{bmatrix}$
σ_i	0.1	0.1	0.1	0.1	0.1
$\hat{\sigma}_i$	0.099	0.100	0.099	0.101	0.101
π_∞	0.227	0.171	0.199	0.195	0.207
$\hat{\pi}_\infty$	0.229	0.159	0.201	0.195	0.216

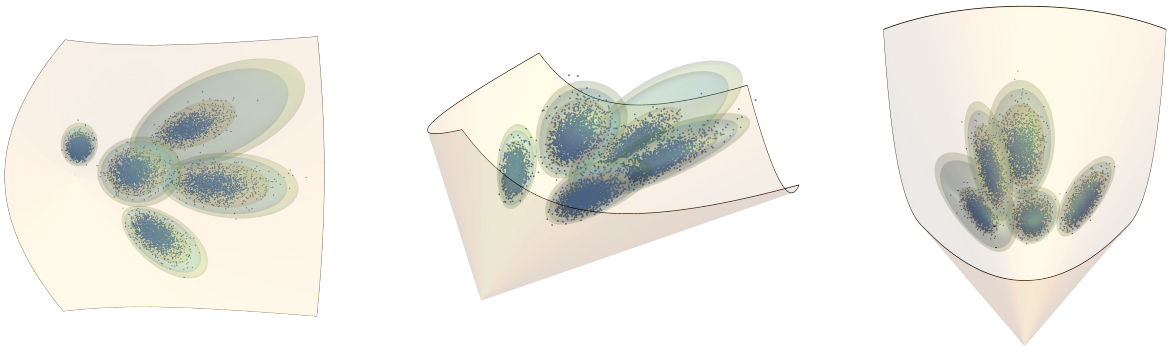

Figure 1. Visual representation of the Riemannian Gaussian model estimated from 10,000 observations from three vantage points: top view (left), side view (middle), and front view (right). Each 2×2 SPD-valued observation is plotted as a point in the interior of the pointed convex cone $\{(a, b, c) \in \mathbb{R}^3 : a \geq 0,\ ac - b^2 \geq 0\}$. The shaded compact regions within the cone are superlevel sets of the 5 estimated Riemannian Gaussian densities that represent the observation likelihoods.

The estimated transition matrix $\hat{P}$ is

$$\hat{P} = \begin{bmatrix} 0.291 & 0.088 & 0.195 & 0.092 & 0.334 \\ 0.104 & 0.409 & 0.185 & 0.188 & 0.114 \\ 0.199 & 0.206 & 0.297 & 0.098 & 0.200 \\ 0.091 & 0.113 & 0.202 & 0.482 & 0.112 \\ 0.407 & 0.105 & 0.106 & 0.083 & 0.299 \end{bmatrix}, \tag{22}$$

which yields a relative approximation error of

$$\frac{\|P - \hat{P}\|_F}{\|P\|_F} = 0.050 \tag{23}$$

with respect to the Frobenius norm. The mean error in the estimated transition probabilities is

$$\frac{1}{N^2} \sum_{i,j=1}^{N} |[P]_{ij} - [\hat{P}]_{ij}| \approx 0.01. \tag{24}$$

5. Conclusion

In this paper, we have shown that the recent method of moments algorithms for HMMs can be generalized to geometric settings in which observations take place in Riemannian manifolds. We observe through simple numerical simulations that the documented advantages of method of moments algorithms, including their competitive accuracy and attractive computational and statistical properties, may continue to hold in the geometric setting. Nonetheless, we expect unique computational challenges to arise in applications involving high-dimensional Riemannian manifolds. Specifically, using Markov chain Monte Carlo (MCMC) algorithms to compute the effective observation matrix K defined in (13) may become prohibitively expensive in high dimensions, which is not the case in the Euclidean setting as K admits a closed form analytic expression for multivariate Gaussian HMMs. Thus, a key technical challenge for the effective application of the proposed algorithm in problems involving high-dimensional manifolds is to devise algorithms for the efficient and scalable computation of K. Further developments of the approach may include extensions to models that incorporate third- or higher-order moments or more elaborate dynamics and control inputs.

Funding: B.C. acknowledges funding from the Faculty of Mathematics at the University of Cambridge as part of the Cambridge Mathematics Placements (CMP) Programme. C.M. was supported by an NTU Presidential Postdoctoral Fellowship and an Early Career Research Fellowship at Fitzwilliam College, Cambridge.

References

1. Krishnamurthy, V. *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*; Cambridge University Press, 2016. doi:10.1017/CBO9781316471104.
2. Durbin, R.; Eddy, S.R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press, 1998. doi:10.1017/CBO9780511790492.
3. Vidyasagar, M. *Hidden Markov Processes: Theory and Applications to Biology*; Princeton University Press, 2014.
4. Cappé, O.; Moulines, E.; Rydén, T. *Inference in hidden Markov models*; Springer Science & Business Media, 2005.
5. Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **1989**, *77*, 257–286. doi:10.1109/5.18626.
6. Gales, M.; Young, S.; others. The application of hidden Markov models in speech recognition. *Foundations and Trends® in Signal Processing* **2008**, *1*, 195–304.
7. Mamon, R.S.; Elliott, R.J. *Hidden Markov models in finance*; Springer, 2007.
8. Chang, J.T. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences* **1996**, *137*, 51–73. doi:https://doi.org/10.1016/S0025-5564(96)00075-2.

9. Mossel, E.; Roch, S. Learning nonsingular phylogenies and hidden Markov models. *The Annals of Applied Probability* **2006**, *16*, 583 – 614. doi:10.1214/105051606000000024.
10. Hsu, D.; Kakade, S.M.; Zhang, T. A spectral algorithm for learning Hidden Markov Models. *Journal of Computer and System Sciences* **2012**, *78*, 1460–1480. JCSS Special Issue: Cloud Computing 2011, doi:https://doi.org/10.1016/j.jcss.2011.12.025.
11. Anandkumar, A.; Hsu, D.; Kakade, S.M. A Method of Moments for Mixture Models and Hidden Markov Models. Proceedings of the 25th Annual Conference on Learning Theory; Mannor, S.; Srebro, N.; Williamson, R.C., Eds.; PMLR: Edinburgh, Scotland, 2012; Vol. 23, *Proceedings of Machine Learning Research*, pp. 33.1–33.34.
12. Kontorovich, A.; Nadler, B.; Weiss, R. On Learning Parametric-Output HMMs. Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. JMLR.org, 2013, ICML'13, p. III–702–III–710.
13. Mattila, R.; Rojas, C.R.; Krishnamurthy, V.; Wahlberg, B. Asymptotically Efficient Identification of Known-Sensor Hidden Markov Models. *IEEE Signal Processing Letters* **2017**, *24*, 1813–1817. doi:10.1109/LSP.2017.2759902.
14. Huang, K.; Fu, X.; Sidiropoulos, N. Learning Hidden Markov Models from Pairwise Co-occurrences with Application to Topic Modeling. Proceedings of the 35th International Conference on Machine Learning; Dy, J.; Krause, A., Eds. PMLR, 2018, Vol. 80, *Proceedings of Machine Learning Research*, pp. 2068–2077.
15. Mattila, R.; Rojas, C.; Moulines, E.; Krishnamurthy, V.; Wahlberg, B. Fast and Consistent Learning of Hidden Markov Models by Incorporating Non-Consecutive Correlations. Proceedings of the 37th International Conference on Machine Learning; III, H.D.; Singh, A., Eds. PMLR, 2020, Vol. 119, *Proceedings of Machine Learning Research*, pp. 6785–6796.
16. Mattila, R. Hidden Markov models: Identification, inverse filtering and applications. PhD thesis, KTH Royal Institute of Technology, 2020.
17. Absil, P.A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*; Princeton University Press, 2009. doi:doi:10.1515/9781400830244.
18. Barachant, A.; Bonnet, S.; Congedo, M.; Jutten, C. Multiclass Brain–Computer Interface Classification by Riemannian Geometry. *IEEE Transactions on Biomedical Engineering* **2012**, *59*, 920–928. doi:10.1109/TBME.2011.2172210.
19. Boumal, N.; Mishra, B.; Absil, P.A.; Sepulchre, R. Manopt, a Matlab Toolbox for Optimization on Manifolds. *J. Mach. Learn. Res.* **2014**, *15*, 1455–1459.
20. Pennec, X.; Sommer, S.; Fletcher, T. *Riemannian geometric statistics in medical image analysis*; Academic Press, 2020.
21. Miolane, N.; Guigui, N.; Brigant, A.L.; Mathe, J.; Hou, B.; Thanwerdas, Y.; Heyder, S.; Peltre, O.; Koep, N.; Zaatiti, H.; Hajri, H.; Cabanes, Y.; Gerald, T.; Chauchat, P.; Shewmake, C.; Brooks, D.; Kainz, B.; Donnat, C.; Holmes, S.; Pennec, X. Geomstats: A Python Package for Riemannian Geometry in Machine Learning. *Journal of Machine Learning Research* **2020**, *21*, 1–9.
22. Mostajeran, C.; Grussler, C.; Sepulchre, R. Geometric Matrix Midranges. *SIAM Journal on Matrix Analysis and Applications* **2020**, *41*, 1347–1368. doi:10.1137/19M1273475.
23. Van Goffrier, G.W.; Mostajeran, C.; Sepulchre, R. Inductive Geometric Matrix Midranges. *IFAC-PapersOnLine* **2021**, *54*, 584–589. 24th International Symposium on Mathematical Theory of Networks and Systems MTNS 2020, doi:https://doi.org/10.1016/j.ifacol.2021.06.120.
24. Said, S.; Le Bihan, N.; Manton, J. Hidden Markov chains and fields with observations in Riemannian manifolds. *IFAC-PapersOnLine* **2021**, *54*, 719–724. doi:10.1016/j.ifacol.2021.06.135.
25. Tupker, Q.; Said, S.; Mostajeran, C. Online Learning of Riemannian Hidden Markov Models in Homogeneous Hadamard Spaces. Geometric Science of Information; Nielsen, F.; Barbaresco, F., Eds.; Springer International Publishing: Cham, 2021; pp. 37–44.
26. Said, S.; Bombrun, L.; Berthoumieu, Y.; Manton, J.H. Riemannian Gaussian Distributions on the Space of Symmetric Positive Definite Matrices. *IEEE Transactions on Information Theory* **2017**, *63*, 2153–2170. doi:10.1109/TIT.2017.2653803.
27. Said, S.; Hajri, H.; Bombrun, L.; Vemuri, B.C. Gaussian Distributions on Riemannian Symmetric Spaces: Statistical Learning With Structured Covariance Matrices. *IEEE Transactions on Information Theory* **2018**, *64*, 752–772. doi:10.1109/TIT.2017.2713829.

28. Said, S.; Mostajeran, C.; Heuveline, S. Gaussian distributions on Riemannian symmetric spaces of nonpositive curvature; Handbook of Statistics, Elsevier, 2022. doi:https://doi.org/10.1016/bs.host.2022.03.004.
29. Heuveline, S.; Said, S.; Mostajeran, C. Gaussian Distributions on Riemannian Symmetric Spaces in the Large N Limit. Geometric Science of Information; Nielsen, F.; Barbaresco, F., Eds.; Springer International Publishing: Cham, 2021; pp. 20–28.
30. Santilli, L.; Tierz, M. Riemannian Gaussian distributions, random matrix ensembles and diffusion kernels. *Nuclear Physics B* **2021**, *973*, 115582. doi:https://doi.org/10.1016/j.nuclphysb.2021.115582.
31. Said, S.; Heuveline, S.; Mostajeran, C. Riemannian statistics meets random matrix theory : towards learning from high-dimensional covariance matrices. *IEEE Transactions on Information Theory* **2022**. doi:10.1109/TIT.2022.3199479.
32. Zanini, P.; Said, S.; Cavalcante, C.C.; Berthoumieu, Y. Stochastic EM algorithm for mixture estimation on manifolds. 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2017, pp. 1–5. doi:10.1109/CAMSAP.2017.8313158.
33. Zanini, P.; Said, S.; Berthoumieu, Y.; Congedo, M.; Jutten, C. Riemannian Online Algorithms for Estimating Mixture Model Parameters. Geometric Science of Information (GSI 2017); , 2017. doi:10.1007/978-3-319-68445-1_78.
34. Nickel, M.; Kiela, D. Poincaré Embeddings for Learning Hierarchical Representations. Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30.
35. Said, S.; Bombrun, L.; Berthoumieu, Y. New Riemannian Priors on the Univariate Normal Model. *Entropy* **2014**, *16*, 4015–4031. doi:10.3390/e16074015.