ORIGINAL PAPER

Variational Markov chain mixtures with automatic component selection

Christopher E. Miles[1] · **Robert J. Webber**[2]

Abstract

Markov state modeling has gained popularity in various scientific fields since it reduces complex time-series data sets into transitions between a few states. Yet common Markov state modeling frameworks assume a single Markov chain describes the data, so they suffer from an inability to discern heterogeneities. As an alternative, this paper models time-series data using a *mixture* of Markov chains, and it automatically determines the number of mixture components using the variational expectation-maximization algorithm. Variational EM simultaneously identifies the number of Markov chains and the dynamics of each chain without expensive model comparisons or posterior sampling. As a theoretical contribution, this paper identifies the natural limits of Markov chain mixture modeling by proving a lower bound on the classification error. This paper then presents numerical experiments where variational EM achieves performance consistent with the theoretically optimal error scaling. The experiments are based on synthetic and observational data sets including `Last.fm` music listening, ultramarathon running, and gene expression. In each of the three data sets, variational EM leads to the identification of meaningful heterogeneities.

Keywords Markov chain mixtures · Variational Bayes · EM algorithm · Markov state models

1 Introduction

This paper considers data emerging from observations, simulations, and experiments, which come in the form of dynamical trajectories. The trajectories can be modeled by a finite-state Markov chain that jumps between different regions ("states") of phase space. This modeling approach is known as *Markov state modeling* (Husic and Pande 2018), and it has become increasingly popular over the past two decades, including applications in chemistry (Noé et al. 2013; Röblitz and Weber 2013; Köhs et al. 2022), biology (Chu et al. 2017; Tse et al. 2018; Tan et al. 2019; Fang et al. 2018), and climate science (Finkel et al. 2023; Souza 2024; Springer et al. 2024).

Markov state modeling has roots in classical statistical modeling, including clustering-based discretizations and finite-state Markov chains. Yet the data sets are changing. Observations, simulations, and experiments are now generating large collections of short, noisy, and heterogeneous trajectories that cannot be described by a single Markov chain. To address the contemporary modeling challenge, this work combines the discretization step in Markov state modeling with a Markov chain mixture approach. As specific contributions, the paper proposes an efficient modeling algorithm and it derives a non-asymptotic information-theoretic lower bound on the classification error achievable from finite trajectory lengths.

1.1 Goals and contributions

This paper argues that Markov state modeling, which is traditionally applied to one Markov chain at a time, can be applied to learn several Markov chains simultaneously. Markov chain mixture modeling was previously proposed and developed in the works (Smyth 1996; Ramoni et al. 2002; Batu et al. 2004; Melnykov 2016; Zhou et al. 2021; Das et al. 2023a, b; Haan-Rietdijk et al. 2017; Frydman 2005). The paper builds on

✉ Christopher E. Miles
  chris.miles@uci.edu

✉ Robert J. Webber
  rwebber@ucsd.edu

[1] Department of Mathematics, University of California Irvine, 419 Rowland Hall, Irvine, CA 92697, United States

[2] Department of Mathematics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, United States

the existing research by asking and answering the following questions:

(a) What Markov chain mixture modeling approach can be easily adapted to high dimensions?
(b) What efficient algorithm can optimize the parameters in the model?
(c) What theoretical accuracy can the Markov chain mixture model attain?

For part (a), the paper applies a simple model (Smyth 1996) based on a mixture of Markov chains in a finite state space. Traditionally, the finite states were determined using expert knowledge of the predominant configurations of the system. However, the modern approach automatically identifies the states with a machine learning method such as spectral clustering.

For part (b), the paper adapts the variational expectation-maximization method (Bishop 2006, Sec. 10) to the setting of Markov chain mixture modeling. Variational EM has a major advantage over the classical EM algorithm, since classical EM requires fitting several models and then performing a post-hoc model comparison to determine the number of mixture components (Keribin 2000; Smyth 1996; Melnykov 2016). In contrast, variational EM determines the number of clusters automatically (Corduneanu and Bishop 2001; MacKay 2001; Rousseau and Mengersen 2011). Here the variational EM approach is developed for finite-state Markov chains. Extending this framework to continuous-state or non-Markovian dynamics is a natural direction for future work.

For part (c), the paper quantifies the optimal theoretical accuracy of any Markov chain mixture model in terms of the Kullback-Leibler divergence between mixture components. The main theoretical contribution is the following theorem, with the proof appearing in Sec. 4.

Theorem 1 (Classification error bound) *Consider any statistical estimator $\hat{Z}$ for the label Z in the Markov chain mixture model*

$$\mathbb{P}\{Z = i\} = \mu(i),$$
$$\mathbb{P}\{Y_0 = \alpha \mid Z = i\} = \nu_i(\alpha),$$
$$\mathbb{P}\{Y_t = \beta \mid Z = i, \ Y_{t-1} = \alpha\} = P_i(\alpha, \beta).$$

The estimator can depend on the observed data Y and the model parameters but not on the unknown label Z. Then, the classification error is at least

$$\mathbb{P}\{\hat{Z} \neq Z\} \geq \frac{1}{2} \sum_{i=1}^{k} \max_{j \neq i} \frac{e^{-D_{KL}(\mathbb{P}_i \| \mathbb{P}_j)}}{\mu(i)^{-1} + \mu(j)^{-1}}.$$

In this expression, the Kullback-Leibler divergence is defined as

$$D_{KL}(\mathbb{P}_i \| \mathbb{P}_j) = \sum_{Y} \mathbb{P}_i(Y) \log \left(\frac{\mathbb{P}_i(Y)}{\mathbb{P}_j(Y)} \right).$$

The summation runs over all possible trajectories of length T, and

$$\mathbb{P}_i(\alpha_0, \ldots, \alpha_T) = \nu_i(\alpha_0) \prod_{t=0}^{T-1} P_i(\alpha_t, \alpha_{t+1}).$$

is the probability of observing the trajectory $Y = (\alpha_0, \ldots, \alpha_T)$ given that $Z = i$.

The theorem lower bounds the classification error in terms of the Kullback–Leibler (KL) divergence between the trajectory distributions in the mixture model. Conceptually, it is a hypothesis-testing theorem: if the mixture includes two similar trajectory distributions, no estimator can reliably assign trajectories to the correct distributions. The proof lower bounds the classification error by comparing pairs of components and quantifying how well they can be distinguished using $D_{KL}(\mathbb{P}_i \| \mathbb{P}_j)$. The resulting bound exhibits the $e^{-D_{KL}}$ scaling familiar from the Bretagnolle–Huber inequality (Bretagnolle and Huber 1979); see Sec. 4 for further discussion.

In the context of Markov chain mixture modeling, the KL divergence in Theorem 1 is the divergence between *length-T trajectory distributions*. Under mild regularity conditions (e.g., ergodicity), this divergence grows linearly with T at a rate determined by the stationary transition dynamics. Consequently, the lower bound in Theorem 1 decays exponentially in T, highlighting trajectory length as a dominant driver of identifiability.

1.2 Plan for the Paper

The rest of the paper is organized as follows. Section 2 introduces Markov state modeling and Markov chain mixture modeling, Section 3 reviews previous related approaches, Section 3.3 introduces the variational EM algorithm, Section 4 provides theoretical analysis, Section 5 presents numerical experiments, and Section 6 concludes.

1.3 Notation

Scalars are in regular typeface: k, s, t, T, N, D. Lowercase letters k, s, t indicate scalar quantities that can vary from line to line. Uppercase letters T, N, D indicate scalar quantities that are fixed by the data set. Finite and infinite sets are in uppercase calligraphic letters, e.g., $\mathcal{X}, \mathcal{Y}$. Elements of finite

sets are often in lowercase Greek letters, e.g., $\alpha, \beta \in \mathcal{Y}$. Vectors are typically written in bold lowercase typeface: $\boldsymbol{\mu}, \boldsymbol{v}, \boldsymbol{q}$. Matrices are written in bold uppercase typeface: $\boldsymbol{K}, \boldsymbol{P}$. The entries of vectors and matrices are indicated with parentheses and regular typeface: $\mu(i), D(i, i)$.

Trajectories are a special class of vectors indexed by time $t = 0, 1, \ldots, T$ and written in uppercase bold typeface: $\boldsymbol{X}, \boldsymbol{Y}$. The states of trajectories are written in regular typeface with subscripts: X_t, Y_t. Superscripts denote trajectories in a collection: $\boldsymbol{Y}^1, \boldsymbol{Y}^2, \ldots, \boldsymbol{Y}^N$.

The probability measure $\mathbb{P}$ represents a ground-truth model, while the probability measure $\hat{\mathbb{P}}$ is a data-driven estimate.

2 Modeling: Critiquing the Markov State Model

A Markov state model is a classic approach for time-series analysis with a long history (Husic and Pande 2018). It has recently become popular in chemistry (Noé et al. 2013; Röblitz and Weber 2013; Köhs et al. 2022), biology (Chu et al. 2017; Tse et al. 2018; Tan et al. 2019; Fang et al. 2018), and climate science (Finkel et al. 2023; Souza 2024; Springer et al. 2024). A Markov state model approximates time-series data as a finite-state Markov chain that jumps between suitably defined "states" of the system.

A traditional Markov state model relies on three assumptions that are described in the following sections: the finite-state assumption (Sec. 2.1), the Markovian assumption (Sec. 2.2), and the one-chain assumption (Sec. 2.3). However, the one-chain assumption may be inappropriate, and Sec. 2.4 will relax the one-chain assumption and generalize the Markov state model to a Markov chain mixture model.

2.1 The Finite-state Assumption

In a Markov state model, a dynamical system on a general state space $\mathcal{X}$ is reduced to a dynamical system on a finite state space $\mathcal{Y} = \{1, 2, \ldots, s\}$. The finite states are assumed to be internally homogeneous and meaningfully different from one another. Clearly delineated states are crucial to the interpretability of the Markov state model.

There are various approaches for defining an appropriate set of finite states. As one approach, the states can be defined based on domain expertise. For example, user activity on `MSNBC.com` ("clickstreams") can be reduced to a state space $\mathcal{Y}$ consisting of different types of webpages, such as the homepage (state 1), the news pages (state 2), and the sports pages (state 3) (Melnykov 2016). These state definitions are based on the `MSNBC.com` site map and the conventional division of articles into "sports" versus "news".

As a more systematic alternative, the state space $\mathcal{X}$ can be decomposed using a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a set of data centers $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_s\}$. This leads to the division into Voronoi cells

$$\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_s, \quad \text{where}$$
$$\mathcal{X}_i = \{\boldsymbol{x} \in \mathcal{X} : d(\boldsymbol{x}, \boldsymbol{c}_i) \leq d(\boldsymbol{x}, \boldsymbol{c}_j) \text{ for } j \neq i\}.$$

Any trajectory $\boldsymbol{X}$ on the state space $\mathcal{X}$ can be approximated by a trajectory $\boldsymbol{Y}$ on the reduced state space $\mathcal{Y} = \{1, \ldots, s\}$ by setting $Y_t = i$ if $X_t \in \mathcal{X}_i$ with arbitrary tie-breaking at the boundaries. The rest of the paper will work with these discretized trajectories.

Identifying a suitable distance function d and a set of data centers $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_s\}$ can be tricky. When the data lies in $\mathcal{X} = \mathbb{R}^D$, one popular method combines all the trajectory locations into a data set $\{\boldsymbol{x}_i\}_{1 \leq i \leq M}$ and selects the centers $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_s\}$ that minimize the sum of square distances

$$\sum_{i=1}^{M} \min_{1 \leq j \leq s} d(\boldsymbol{x}_i, \boldsymbol{c}_j)^2.$$

where $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ is the Euclidean distance. This method is called k-means clustering (Lloyd 1982).

k-means clustering is common but not always effective at identifying finite states. Spectral clustering (Coifman et al. 2005) is an alternative method that adapts to the nonlinear manifold structure of the input data. Spectral clustering can be applied to any state space (not just $\mathcal{X} = \mathbb{R}^D$), as long as there is a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which quantifies the similarity between data points. **Fig.** 1 illustrates spectral clustering on a spiral data set using the Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{y}\|^2)$ with bandwidth $\sigma = 1$. Spectral clustering is now standard practice in Markov state modeling, so the algorithmic details are omitted; see (Coifman et al. 2005) for the classical formulation and Chen et al. (2025) for a scalable implementation based on randomized linear algebra. Later, Sec. 5.4 will apply spectral clustering to construct Markov states for gene expression data.

2.2 The Markovian Assumption

In a Markov state model, each trajectory $\boldsymbol{Y}$ is assumed to be *Markovian*. The Markovian assumption means that the position Y_t depends only on the immediate past position Y_{t-1}, regardless of all the earlier positions $Y_{t-2}, Y_{t-3}, \ldots$. Thus, all the transition probabilities can be written as

$$\mathbb{P}\{Y_t = \beta \mid Y_0 = \alpha_0, \ldots, Y_{t-1} = \alpha_{t-1}\}$$
$$= P(\alpha_{t-1}, \beta),$$

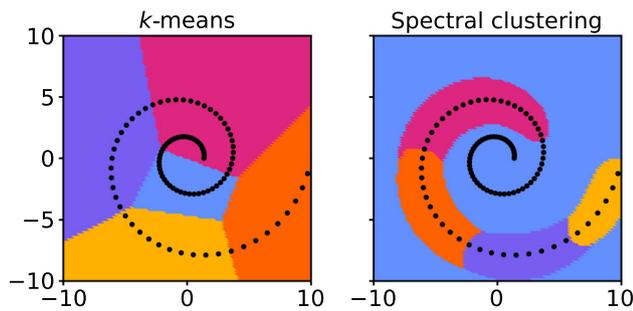for a suitable transition matrix $\boldsymbol{P} \in \mathbb{R}^{s \times s}$.

Fig. 1 Comparison of *k*-means (left) versus spectral clustering (right) applied to $M = 100$ data points in a spiral. Spectral clustering adapts to the nonlinear manifold structure of the data, *k*-means does not

Like all model assumptions, the Markovian assumption is an idealization of the truth. Many dynamical systems are not precisely Markovian because they exhibit prolonged history effects (Chierichetti et al. 2012). As an example of history dependence, consider the travel patterns of adults in Nanjing, China (Zhou et al. 2021). On a given day, a parent is unlikely to commute to their child's school more than twice (first taking the child to school, second picking the child up from school). Because of this history dependence, the location of the adult during the day does not precisely form a Markov chain. Nonetheless, a rich and detailed description of the daily lives of residents is possible using the Markov chain assumption (Zhou et al. 2021). In general, a Markov state model assumes that history effects can be disregarded while still providing a meaningful analysis.

When strict Markovianity fails due to finite memory, a standard remedy is *delay embedding*: redefine the state at time t as a tuple of lagged observations, e.g., $(Y_t, Y_{t-1}, \ldots, Y_{t-m+1})$. Delay embedding converts an mth-order Markov chain into a first-order Markov chain on an enlarged state space. Markov state modeling can be applied directly to the delay-embedded process, but the trade-off is an increased state-space size and hence more parameters to estimate.

2.3 The One-Chain Assumption

The final assumption of a Markov state model is rarely stated yet fundamental. The standard presentations of Markov state modeling (Pande et al. 2010; Chodera and Noé 2014; Husic and Pande 2018) all assume that a single Markov chain generates the trajectories. If the data set consists of multiple trajectories $Y^1, Y^2, \ldots, Y^N$, the trajectories are all assumed to be generated from the same transition matrix P. The one-chain assumption is well-justified if the trajectories are computer simulations from a fixed model. Yet, the one-chain assumption is challenged by the heterogeneities of real-world observations and experiments. For example, is there really one type of web user on MSNBC.com? Is there really one type of commuter in Nanjing, China? Several papers in the

statistics literature go beyond the one-chain assumption and instead they analyze MSNBC.com (Melnykov 2016) users and Nanjing commuters (Zhou et al. 2021) using a *mixture* of Markov chains with distinct transition matrices $P_1, \ldots, P_k$.

The resulting Markov chain mixture model will be described in the next section, and it will be the focus of the remainder of the paper. Although joint inference of states and Markov chain mixture modeling parameters is possible (Safinianaini et al. 2024), the paper assumes the state sequences have already been defined using domain knowledge or data-driven clustering. This design choice decouples state assignment from mixture inference to support scalable, domain-adaptable modeling.

2.4 The Markov chain mixture model

In a Markov chain mixture model, each trajectory Y is generated as follows:

$$\mathbb{P}\{Z = i\} = \mu(i),$$
$$\mathbb{P}\{Y_0 = \alpha \mid Z = i\} = v_i(\alpha), \qquad (1)$$
$$\mathbb{P}\{Y_t = \beta \mid Z = i, Y_{t-1} = \alpha\} = P_i(\alpha, \beta).$$

The variable $Z \in \{1, \ldots, k\}$ is a latent variable indicating the Markov chain label. $\mu(i)$ describes the probability of observing each label $Z = i$. The vectors $v_i \in [0, 1]^s$ and matrices $P_i \in [0, 1]^{s \times s}$ give the initialization probabilities and transition probabilities for the Markov chain with label $Z = i$. The total number of free parameters in the model is $ks^2 - 1$, and typical parameter values are $k = 10^0$–10^2 and $s = 10^0$–10^2. The Markov chain mixture model was presented in the papers (Smyth 1996; Ramoni et al. 2002; Batu et al. 2004), and it has been applied in areas including consumer behavior (Melnykov 2016; Zhou et al. 2021), biology (Das et al. 2023a, b), psychology (Haan-Rietdijk et al. 2017), and economics (Frydman 2005).

3 Algorithms: Comparing Approaches for Markov Chain Mixtures

This section discusses past algorithms for fitting the parameters in a Markov chain mixture model, and it emphasizes three main themes. First, many researchers rely on the expectation-maximization (EM) algorithm for parameter inference (Sec. 3.1). Second, researchers have identified the key importance of the trajectory length T for ensuring classification accuracy (Sec. 3.2). Third, this paper adapts the variational EM algorithm (Bishop 2006,Sec. 10) instead of traditional EM since variational EM learns the number of mixture components automatically (Sec. 3.3).

3.1 The Expectation-Maximization Algorithm

In several applications of Markov chain mixture modeling, the expectation-maximization (EM) algorithm gives fast and reliable parameter estimates (Smyth 1996; Ramoni et al. 2002; Melnykov 2016; Zhou et al. 2021), especially when it is initialized carefully (Kausik et al. 2023; Gupta et al. 2016; Spaeh and Tsourakakis 2023). In EM, each iteration alternates between (i) estimating conditional probabilities for the latent variables $\mathbb{P}\{Z^n = i \mid Y^n\}$ and (ii) updating parameters $(\boldsymbol{\mu}, \boldsymbol{v}_i, \boldsymbol{P}_i)$ using the expected counts of initial states and transitions under the latent variable distribution. Because EM for finite-state Markov-chain mixtures is standard (see, e.g., Dempster et al. (2018); Kausik et al. (2023)), this paper omits the algorithmic details and focuses instead on the detailed exposition of variational EM in Sec. 3.3.

As a limitation, EM is a nonconvex optimization algorithm that converges to local maxima in the likelihood landscape. Therefore, the accuracy of the converged parameters depends on the initialization, which can be chosen either randomly (Melnykov 2016; Zhou et al. 2021) or by a more structured approach (Smyth 1996; Gupta et al. 2016; Spaeh and Tsourakakis 2023; Kausik et al. 2023). In a random initialization, the user can generate probability estimates $\hat{\mathbb{P}}\{Z^n = i\}$ for $i = 1, \ldots, k$ that are uniformly distributed on the probability simplex. A convenient procedure for generating uniformly distributed vectors on the simplex is to first produce k independent exponential random variables and then normalize them to sum to one (Onn and Weissman 2009).

As another limitation, EM does not automatically determine the number of mixture components. Rather, the number of components must be specified as input. Many users fit several models with different numbers of components and then perform a comparison using AIC, BIC, or cross-validation (Keribin 2000; Smyth 1996; Melnykov 2016). When the user searches over a space of k different models, this procedure can increase the runtime by a factor of k or more. Sec. 3.3 will present an improvement of the EM algorithm called variational EM (VEM) that *does* automatically determine the number of components.

3.2 The Trajectory Length T

The second theme in the Markov chain mixture modeling literature is the close link between the classification accuracy and the trajectory length T. Put simply, classifying short trajectories with the correct labels is *hard*. For example, Ramoni and coauthors (Ramoni et al. 2002,Sec. 5.3.1) describe numerical experiments with increasing trajectory lengths that lead to steadily increasing accuracy.

Researchers have tackled the issue of trajectory length by focusing on the extremes of short trajectories ($T = 3$, Gupta et al. (2016); Spaeh and Tsourakakis (2023)) or long trajectories ($T \to \infty$, Khaleghi et al. (2016); Fitzpatrick and Stewart (2022); Kausik et al. (2023)). A couple papers propose structured initialization methods for EM based on short trajectories of length $T = 3$ (Gupta et al. 2016; Spaeh and Tsourakakis 2023), but these approaches require massive amounts of data. For example, $N = 10^8$ trajectories to achieve classification error $< 10\%$ (Gupta et al. 2016). At the other extreme, algorithms can ensure perfect parameter identification in the limit $T \to \infty$ (Khaleghi et al. 2016; Kausik et al. 2023).

Sec. 4 provides further insight into the intermediate regimes where T is neither short nor long, by bounding the classification error for any Markov chain mixture model with any trajectory length T. Since the KL divergence grows roughly linearly with T, the lower bound on the classification error decays roughly exponentially in T, indicating the importance of long trajectories.

3.3 Variational EM Algorithm for Markov Chain Mixtures

The current section describes the variational EM method for Markov chain mixtures in detail and outlines the steps for a user to apply the method to data.

As the first conceptual step, the user must reduce the data to a finite state space $\mathcal{Y} = \{1, \ldots, s\}$. The states can be chosen based on domain expertise or based on spectral clustering (Coifman et al. 2005), as is common in the Markov state modeling literature. The typical number of states is $s = 10^0$–10^2.

Next, the user must choose the maximum number of mixture components, typically $k = 10^1$–10^2. As long as $k \geq k_{\text{true}}$, the approach is capable of identifying the true number of components. However, there is motivation to not set k too large, as the computational cost is proportional to k.

The data set is then described using the Bayesian mixture model:

$$\begin{aligned}
\boldsymbol{\mu} &\sim \text{Dir}(\mathbf{1}_k/k), \\
\boldsymbol{v}_i &\sim \text{Dir}(\mathbf{1}_s), \\
\boldsymbol{P}_i(\alpha \,\cdot) &\sim \text{Dir}(\mathbf{1}_s), \\
\mathbb{P}\{Z = i\} &= \mu(i), \\
\mathbb{P}\{Y_0 = \alpha \mid Z = i\} &= v_i(\alpha), \\
\mathbb{P}\{Y_t = \beta \mid Z = i, \, Y_{t-1} = \alpha\} &= P_i(\alpha, \beta).
\end{aligned} \tag{2}$$

This Bayesian formulation combines standard Bayesian priors for Markov state modeling (Bacallado et al. 2009) and mixture modeling (Gormley et al. 2023). Specifically, the model places uninformative $\text{Dir}(\mathbf{1}_s)$ prior distributions on the parameters $\boldsymbol{v}_i$ and the rows of the matrix $\boldsymbol{P}_i(\alpha, \cdot)$, and it places a sparsity-promoting $\text{Dir}(\mathbf{1}_k/k)$ prior distribution on the mixture parameter $\boldsymbol{\mu}$. The $\text{Dir}(\mathbf{1}_k/k)$ prior leads to

a posterior distribution that automatically prunes extraneous mixture components by setting the corresponding mixture probabilities to $\mu(i) \approx 0$ (Hemant Ishwaran and Sun 2001; Rousseau and Mengersen 2011; Malsiner-Walli et al. 2014).

Given the model (2) and the observed trajectories, the Bayesian posterior distribution is hard to compute exactly. However, the posterior can be efficiently approximated by a factored distribution (see Bishop 2006,Sec. 10.2):

$$\mathbb{P}(\{Z^n, \boldsymbol{\mu}, \boldsymbol{v}_i, \boldsymbol{P}_i\}) \approx Q_1(\{Z^n\}) \, Q_2(\{\boldsymbol{\mu}, \boldsymbol{v}_i, \boldsymbol{P}_i\}).$$

This factorization assumes the latent variables $\{Z^n\}$ are independent from the mixture parameters $\{\boldsymbol{\mu}, \boldsymbol{v}_i, \boldsymbol{P}_i\}$. Given this factorization, the approximation minimizing the Kullback-Leibler divergence with respect to the posterior is a sequence of mutually independent random variables

$$\begin{aligned} \boldsymbol{\mu} &\sim \text{Dir}(\boldsymbol{N}), \\ \boldsymbol{v}_i &\sim \text{Dir}(\boldsymbol{N}_i), \\ \boldsymbol{P}_i(\alpha, \cdot) &\sim \text{Dir}(\boldsymbol{N}_{i,\alpha}). \end{aligned} \tag{3}$$

with parameters that can be optimized using the variational EM algorithm in Alg. 1.

Examining Alg. 1, there are a couple of key changes from the standard EM algorithm (Sec. 3.1). First, the parameters $\{\boldsymbol{\mu}, \boldsymbol{v}_i, \boldsymbol{P}_i\}$ are described by Dirichlet distributions, which encode point estimates and uncertainty intervals. The user can extract a point estimate for any parameter by using the expected value formula for the Dirichlet distribution, e.g.,

$$\hat{\mu}(i) = \mathbb{E}[\mu(i)] = \frac{N(i)}{\sum_{j=1}^{k} N(j)}. \tag{4}$$

The user can extract uncertainty intervals by using the variance formula for the Dirichlet distribution, e.g.,

$$\text{Var}[\mu(i)] = \frac{N(i) \sum_{j \neq i} N(j)}{\left(\sum_{j=1}^{k} N(j)\right)^2 \left(\sum_{j=1}^{k} N(j) + 1\right)}.$$

The automatic availability of uncertainty information is a helpful feature of the Bayesian approach.

As the second change, variational EM automatically determines the number of mixture components. Upon convergence, only a small number of mixture parameters satisfy $\hat{N}(i) \geq 1$. The optimization drives all the remaining mixture parameters to the lowest possible value $\hat{N}(i) = 1/k$. These components are not represented by any labels $\hat{Z}^n = i$, so they are effectively removed from the model. The tendency of variational EM to identify a parsimonious mixture model is called *automatic relevance determination*. It was observed in the papers (Corduneanu and Bishop 2001; MacKay 2001) and later justified based on asymptotic arguments in Rousseau and Mengersen (2011).

Algorithm 1 Variational EM Algorithm

Input: Trajectories Y^n, estimates $\hat{\mathbb{P}}\{Z^n = i\}$.
Output: Refined estimates $\hat{N}, \hat{N}_i, \hat{N}_{i,\alpha}, \hat{Z}^n$.
1: For each $n = 1, \ldots, N$, form

$$U^n(\alpha) = \mathbb{1}\{Y_0^n = \alpha\},$$

$$V^n(\alpha, \beta) = \sum_{t=0}^{T-1} \mathbb{1}\{Y_t^n = \alpha, \, Y_{t+1}^n = \beta\}.$$

2: **for** $t = 1, 2, \ldots$ **do**
3:    Compute Dirichlet parameters

$$\hat{N}(i) = \frac{1}{k} + \sum_{n=1}^{N} \hat{\mathbb{P}}\{Z^n = i\},$$

$$\hat{N}_i(\alpha) = 1 + \sum_{n=1}^{N} \hat{\mathbb{P}}\{Z^n = i\} \, U^n(\alpha),$$

$$\hat{N}_{i,\alpha}(\beta) = 1 + \sum_{n=1}^{N} \hat{\mathbb{P}}\{Z^n = i\} V^n(\alpha, \beta).$$

4:    Compute vectors $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{v}}_i, \hat{\boldsymbol{P}}_i(\alpha, \cdot)$ via

$$\log \hat{\mu}(i) = \psi(\hat{N}(i)) - \psi\left(\sum_{j=1}^{k} \hat{N}(j)\right),$$

$$\log \hat{v}_i(\alpha) = \psi(\hat{N}_i(\alpha)) - \psi\left(\sum_{\beta=1}^{s} \hat{N}_i(\beta)\right),$$

$$\log \hat{P}_i(\alpha, \beta) = \psi(\hat{N}_{i,\alpha}(\beta)) - \psi\left(\sum_{\gamma=1}^{s} \hat{N}_{i,\alpha}(\gamma)\right).$$

$\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function.
5:    Update probabilities $\hat{\mathbb{P}}\{Z^n = i\}$ via

$$\frac{\hat{\mu}(i)}{C_n} \prod_{\alpha=1}^{s} \hat{v}_i(\alpha)^{U^n(\alpha)} \prod_{\alpha,\beta=1}^{s} \hat{P}_i(\alpha, \beta)^{V^n(\alpha,\beta)},$$

and choose C_n so that $\sum_{1 \leq i \leq k} \hat{\mathbb{P}}\{Z^n = i\} = 1$.
6:    Calculate variational lower bound

$$\mathcal{L} = \sum_{n=1}^{N} \log C_n + \log\left[\frac{B(\mathbf{1}_k/k)}{B(\hat{N})} \prod_{i=1}^{k} \hat{\mu}_i^{\hat{N}(i)-1/k} \right]$$

$$+ \sum_{i=1}^{k} \log\left[\frac{B(\mathbf{1}_s)}{B(\hat{N}_i)} \prod_{\alpha=1}^{s} \hat{v}_i(\alpha)^{\hat{N}_i(\alpha)-1} \right]$$

$$+ \sum_{i=1}^{k} \sum_{\alpha=1}^{s} \log\left[\frac{B(\mathbf{1}_s)}{B(\hat{N}_{i,\alpha})} \prod_{\beta=1}^{s} \hat{P}_i(\alpha, \beta)^{\hat{N}_{i,\alpha}(\beta)-1} \right],$$

where $B(N) = \left(\prod_i \Gamma(N_i)\right) / \Gamma(\sum_i N_i)$.
7:    **Break** if $|\Delta\mathcal{L}| < 10^{-12} \cdot NT$.
8: **end for**
9: Return $\hat{N}, \hat{N}_i, \hat{N}_{i,\alpha}, \hat{Z}^n = \underset{1 \leq i \leq k}{\arg\max} \, \hat{\mathbb{P}}\{Z^n = i\}$.

To apply Alg. 1 in practice, the user needs some way to initialize the parameter estimates. Since there is limited theoretical understanding of the best way to initialize variational EM, this paper randomly initializes parameters $\hat{\mathbb{P}}\{Z^n = i\}$ for $i = 1, \ldots, k$ from the probability simplex. After running Alg. 1 with 10^2 independent initializations, the paper selects the optimal parameters $\{\hat{N}, \hat{N}_i, \hat{N}_{i,\alpha}, \hat{Z}^n\}$ that maximize the likelihood lower bound $\mathcal{L}$.

4 Theory: Limits on Markov Chain Mixture Modeling

Markov chain mixture modeling cannot succeed when the trajectories are too short. This section provides a theoretical explanation by proving and interpreting Theorem. 1.

Proof of Thm. 1 Given the observed data Y, the conditional distribution of the label Z is

$$\mathbb{P}\{Z = i \mid Y\} = \frac{\mu(i)\mathbb{P}_i(Y)}{\sum_{j=1}^k \mu(j)\mathbb{P}_j(Y)} \tag{5}$$

for each $i = 1, \ldots, k$. The optimal classification algorithm selects the labels according to

$$\hat{Z} \in \underset{1 \le i \le k}{\operatorname{argmax}} \, \mathbb{P}\{Z = i \mid Y\},$$

and the resulting statistical error is quantified by

$$\mathbb{P}\{Z \ne \hat{Z} \mid Y\} = 1 - \max_{1 \le i \le k} \mathbb{P}\{Z = i \mid Y\}.$$

Next, use the inequality $1 - \alpha \ge (1 - \alpha^2)/2$ to obtain the lower bound

$$
\begin{aligned}
&\mathbb{P}\{Z \ne \hat{Z} \mid Y\} \\
&= 1 - \max_{1 \le i \le k} \mathbb{P}\{Z = i \mid Y\} \\
&\ge \frac{1}{2} - \frac{1}{2} \max_{1 \le i \le k} \mathbb{P}\{Z = i \mid Y\}^2 \\
&\ge \frac{1}{2} - \frac{1}{2} \sum_{i=1}^k \mathbb{P}\{Z = i \mid Y\}^2 \\
&= \frac{1}{2} \sum_{i=1}^k \mathbb{P}\{Z = i \mid Y\}(1 - \mathbb{P}\{Z = i \mid Y\}).
\end{aligned}
$$

Multiply both sides by $\mathbb{P}(Y)$ and sum over all possible trajectories Y to obtain the formula

$$
\begin{aligned}
\mathbb{P}\{Z \ne \hat{Z}\} &= \sum_Y \mathbb{P}\{Z \ne \hat{Z} \mid Y\} \mathbb{P}(Y) \\
&\ge \frac{1}{2} \sum_{i,Y} \mathbb{P}\{Z = i, Y\}(1 - \mathbb{P}\{Z = i \mid Y\}) \\
&= \frac{1}{2} \sum_{i=1}^k \mu(i) \frac{\sum_Y \mathbb{P}\{Z = i, Y\}(1 - \mathbb{P}\{Z = i \mid Y\})}{\mu(i)}.
\end{aligned}
$$

The latter quantity is a conditional expectation, so it follows

$$
\begin{aligned}
&\mathbb{P}\{Z \ne \hat{Z}\} \\
&\ge \frac{1}{2} \sum_{i=1}^k \mu(i) \, \mathbb{E}\big[1 - \mathbb{P}\{Z = i \mid Y\} \mid Z = i\big].
\end{aligned}
\tag{6}
$$

The conditional expectation turns out to be analytically tractable and allows us to derive the bound directly.

The rest of the proof is based on bounding $\mathbb{E}\big[1 - \mathbb{P}\{Z = i \mid Y\} \mid Z = i\big]$ from below using the Kullback-Leibler divergence. To obtain a convenient lower bound, use eq. (5) to derive

$$
\begin{aligned}
&\mathbb{E}\big[1 - \mathbb{P}\{Z = i \mid Y\} \mid Z = i\big] \\
&= \mathbb{E}\left[\frac{\sum_{j \ne i} \mu(j)\mathbb{P}_j(Y)}{\sum_{j=1}^k \mu(j)\mathbb{P}_j(Y)} \,\bigg|\, Z = i\right] \\
&\ge \mathbb{E}\left[\frac{\mu(j)\mathbb{P}_j(Y)}{\mu(i)\mathbb{P}_i(Y) + \mu(j)\mathbb{P}_j(Y)} \,\bigg|\, Z = i\right],
\end{aligned}
$$

for each $j \ne i$. The inequality is due to the fact that $f(x) = x/(c + x)$ is an increasing function of $x > 0$, for any $c > 0$. Next, by an application of Jensen's inequality,

$$
\begin{aligned}
&\log\left(\mathbb{E}\left[\frac{\mu(j)\mathbb{P}_j(Y)}{\mu(i)\mathbb{P}_i(Y) + \mu(j)\mathbb{P}_j(Y)} \,\bigg|\, Z = i\right]\right) \\
&\ge \mathbb{E}\left[\log\left(\frac{\mu(j)\mathbb{P}_j(Y)}{\mu(i)\mathbb{P}_i(Y) + \mu(j)\mathbb{P}_j(Y)}\right) \,\bigg|\, Z = i\right] \\
&= \mathbb{E}\left[\log\left(\frac{\mathbb{P}_j(Y)}{\mathbb{P}_i(Y)}\right) \,\bigg|\, Z = i\right] \\
&\quad - \mathbb{E}\left[\log\left(\frac{\mu(i)}{\mu(j)} + \frac{\mathbb{P}_j(Y)}{\mathbb{P}_i(Y)}\right) \,\bigg|\, Z = i\right].
\end{aligned}
$$

The first term in the final expression is the negative Kullback-Leibler divergence

$$\mathbb{E}\left[\log\left(\frac{\mathbb{P}_j(Y)}{\mathbb{P}_i(Y)}\right) \,\bigg|\, Z = i\right] = -\, \mathrm{D}_{\mathrm{KL}}(\mathbb{P}_i \parallel \mathbb{P}_j).$$

The second term can be bounded by another application of Jensen's inequality:

$$-\mathbb{E}\left[\log\left(\frac{\mu(i)}{\mu(j)} + \frac{\mathbb{P}_j(\boldsymbol{Y})}{\mathbb{P}_i(\boldsymbol{Y})}\right)\middle|\, Z = i\right]$$
$$\geq -\log\left(\frac{\mu(i)}{\mu(j)} + \mathbb{E}\left[\frac{\mathbb{P}_j(\boldsymbol{Y})}{\mathbb{P}_i(\boldsymbol{Y})}\middle|\, Z = i\right]\right).$$

Moreover, observe that

$$\mathbb{E}\left[\frac{\mathbb{P}_j(\boldsymbol{Y})}{\mathbb{P}_i(\boldsymbol{Y})}\middle|\, Z = i\right] = \sum_{\boldsymbol{Y}} \mathbb{P}_i(\boldsymbol{Y})\frac{\mathbb{P}_j(\boldsymbol{Y})}{\mathbb{P}_i(\boldsymbol{Y})}$$
$$= \sum_{\boldsymbol{Y}} \mathbb{P}_j(\boldsymbol{Y}) = 1.$$

Tying together the above expressions, it follows

$$\log\left(\mathbb{E}\left[1 - \mathbb{P}\{Z = i \mid \boldsymbol{Y}\} \mid Z = i\right]\right)$$
$$\geq -\,\mathrm{D_{KL}}(\mathbb{P}_i \,\|\, \mathbb{P}_j) - \log\left(\frac{\mu(i)}{\mu(j)} + 1\right)$$

and therefore

$$\mathbb{E}\left[1 - \mathbb{P}\{Z = i \mid \boldsymbol{Y}\} \mid Z = i\right] \geq \max_{j \neq i} \frac{e^{-\mathrm{D_{KL}}(\mathbb{P}_i \,\|\, \mathbb{P}_j)}}{\frac{\mu(i)}{\mu(j)} + 1}.$$

Combine with eq. (6) to conclude that the classification error is at least

$$\mathbb{P}\{\hat{Z} \neq Z\} \geq \frac{1}{2}\sum_{i=1}^{k}\max_{j \neq i} \frac{e^{-\mathrm{D_{KL}}(\mathbb{P}_i \,\|\, \mathbb{P}_j)}}{\mu(i)^{-1} + \mu(j)^{-1}}.$$

This completes the proof.   □

Two points are worth emphasizing. First, Theorem 1 is non-asymptotic: it applies to any finite trajectory length T without requiring mixing or stationarity. It thus provides a baseline limit in the "many short trajectories" regime, isolating an intrinsic identifiability barrier.

Second, the $e^{-\,\mathrm{D_{KL}}}$ scaling in the error bound reflects a classical connection between classification error and information divergences. The optimal error for distinguishing $\mathbb{P}_i$ from $\mathbb{P}_j$ in a two-class, equally weighted mixture model is

$$\frac{1}{2}\sum_{\boldsymbol{Y}} \min\{\mathbb{P}_i(\boldsymbol{Y}), \mathbb{P}_j(\boldsymbol{Y})\} = \frac{1}{2} - \frac{1}{2}d_{\mathrm{TV}}(\mathbb{P}_i, \mathbb{P}_j),$$

where the total variation distance is

$$d_{\mathrm{TV}}(\mathbb{P}_i, \mathbb{P}_j) = \frac{1}{2}\sum_{\boldsymbol{Y}} |\mathbb{P}_i(\boldsymbol{Y}) - \mathbb{P}_j(\boldsymbol{Y})|.$$

The Bretagnolle–Huber inequality (Bretagnolle and Huber 1979) then bounds

$$1 - d_{\mathrm{TV}}(\mathbb{P}_i, \mathbb{P}_j) \geq \frac{1}{2}e^{-\,\mathrm{D_{KL}}(\mathbb{P}_i \,\|\, \mathbb{P}_j)}$$

Thm. 1 derives a comparable bound for the Markov chain mixture model, making explicit how the KL divergence controls the overall classification error. This result complements recent work by Lee et al. (2025), who prove high-probability error bounds for the Markov chain mixture model and also find that KL divergence governs the clustering error.

To interpret the theorem, it is instructive to consider what happens to the KL divergence in the specific case of a Markov chain mixture as the trajectory length T approaches infinity. If the Markov chain i has a unique stationary measure $\boldsymbol{\pi}_i \in [0, 1]^s$, then in the limit $T \to \infty$ the ergodic theorem for Markov chains (Norris 1997, Sec. 1.10) guarantees

$$\frac{1}{T}\,\mathrm{D_{KL}}(\mathbb{P}_i \,\|\, \mathbb{P}_j)$$
$$= \frac{1}{T}\sum_{\alpha_0,\ldots,\alpha_T} \mathbb{P}_i(\alpha_0,\ldots,\alpha_T)\log\left(\frac{\mathbb{P}_i(\alpha_0,\ldots,\alpha_T)}{\mathbb{P}_j(\alpha_0,\ldots,\alpha_T)}\right)$$
$$\to \sum_{\alpha,\beta=1}^{s}\pi_i(\alpha)P_i(\alpha,\beta)\log\left(\frac{P_i(\alpha,\beta)}{P_j(\alpha,\beta)}\right)$$
$$= \sum_{\alpha=1}^{s}\pi_i(\alpha)\,\mathrm{D_{KL}}(P_i(\alpha,\cdot) \,\|\, P_j(\alpha,\cdot)).$$

By this derivation, the KL divergence $\mathrm{D_{KL}}(\mathbb{P}_i \,\|\, \mathbb{P}_j)$ is approximately T times the averaged Kullback-Leibler divergence $\mathrm{D_{KL}}(P_i(\alpha,\cdot) \,\|\, P_j(\alpha,\cdot))$ over all the states. The linear dependence of the Kullback-Leibler divergence on T has a practical implication: longer trajectories help substantially. Doubling T can turn a hard problem into an easy one, for example reducing error from 10% to 1%.

5 Numerical Experiments

This section describes four numerical experiments with the variational EM algorithm. The first experiment (Sec. 5.1) uses synthetic Markov chain data. The middle two experiments (Secs. 5.2 and 5.3) use real-world observations. The last experiment (Sec. 5.4) uses synthetic genetics data. Code to run the experiments is available at https://github.com/chris-miles/MarkovChain-VEM.

5.1 Synthetic Markov Chain Data

The first experiment is based on simulations of the Markov chain mixture model (1) with uniformly random parameters $\boldsymbol{\mu}$, $\boldsymbol{v}_i$, and $\boldsymbol{P}_i$. The goal of the experiment is to test whether
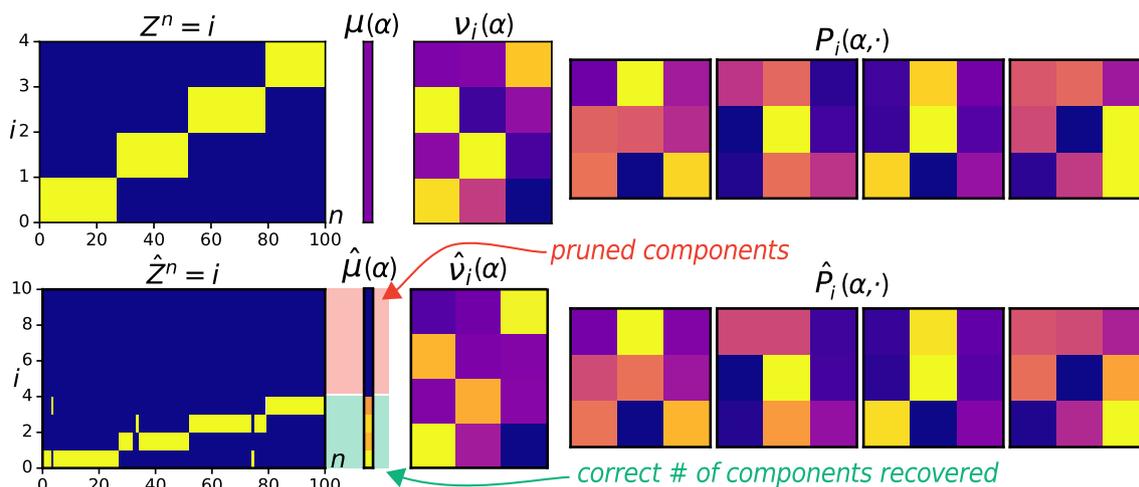
Fig. 2 True parameters and parameter estimates for synthetic trajectories with $k_{\text{true}} = 4$, $s = 3$, $N = 100$, and $T = 30$. Top row shows true parameters. Bottom row shows parameter estimates from variational EM with a maximum of $k = 10$ components. Even with limited data, VEM correctly identifies the number of components

variational EM correctly identifies the number of components and correctly classifies trajectories into components.

As the first result, variational EM accurately identifies the number of components even with a relatively small data set. **Fig. 2** shows the true parameters (top) and parameter estimates (bottom) when $k_{\text{true}} = 4$, $s = 3$, $N = 100$, and $T = 30$. Although variational EM is run with a maximum number of $k = 10$ components, it leads to a fitted model with just 4 components. The algorithm assigns no trajectories to 6 of the 10 components, thereby pruning them from the model. The automatic identification of the number of mixture components is a major advantage compared to previous Markov chain mixture modeling approaches (Keribin 2000; Smyth 1996; Melnykov 2016).

The reader may wonder, is there a cost to setting the value of k excessively high? While there is an obvious computational burden to setting $k > k_{\text{true}}$, the results in **Fig. 3** show that the classification accuracy, defined by

$$\text{acc} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{Z^n = \hat{Z}^n\}, \tag{7}$$

is not hindered. Indeed, the results with $k = 4 = k_{\text{true}}$ (top panel) are nearly identical to results with $k = 10 > k_{\text{true}}$ (bottom panel).

As the next result, **Fig. 3** shows the trajectory length T is a major factor determining the classification accuracy (7). The accuracy reaches a threshold for each T value as $N \to \infty$, and increasing T exponentially increases the threshold. These results are in line with the theoretical analysis in Sec. 4, which also suggests an exponential dependence on T.
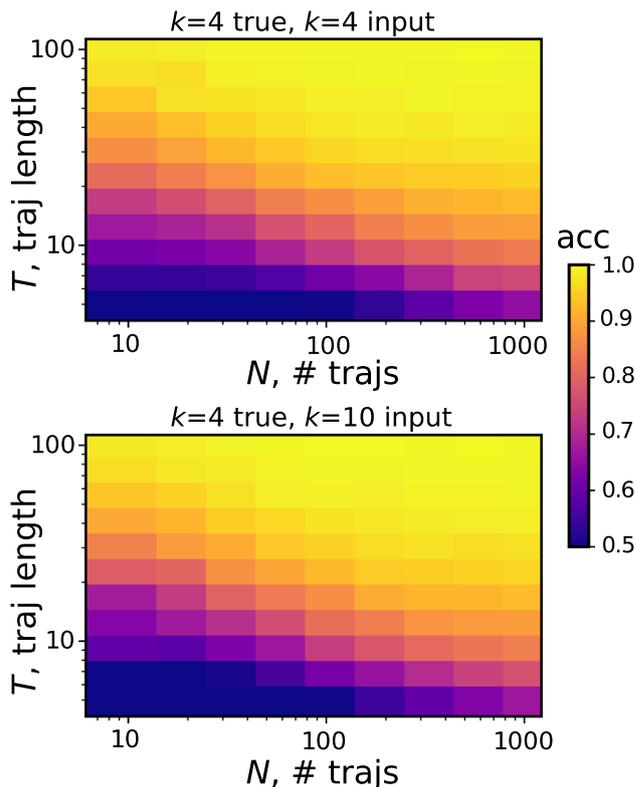

Fig. 3 Classification accuracy with $k_{\text{true}} = 4$, $s = 3$, and varying settings of N and T. The accuracy is similar using $k = 4$ (top) versus $k = 10$ (bottom). In both plots, the accuracy is averaged over 250 independent trials

EM and variational EM are known to terminate in locally optimal parameter values that can be far from the global optimum (Redner and Walker 1984). To safeguard against
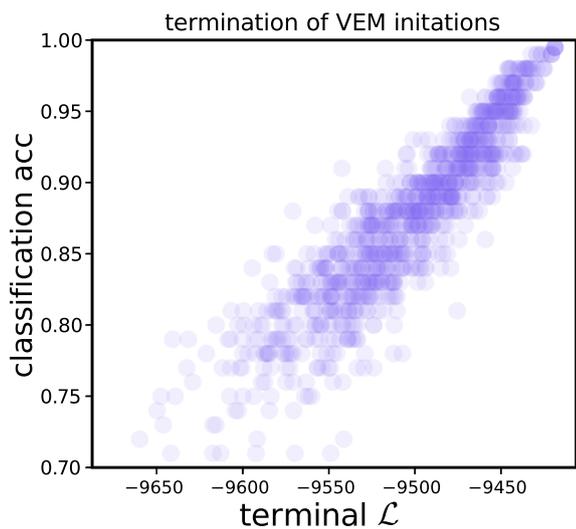
Fig. 4 Demonstration of the local optima over 1000 random variational EM initializations when $k = 15$, $k_{\text{true}} = 10$, $s = 7$, $N = 100$, and $T = 50$. Many variational EM runs end with nearly optimal assignments, some do not

misconvergence, this paper randomly initializes variational EM many times and accepts the results that maximize the likelihood bound $\mathcal{L}$. The outcomes of 10^3 random initialization are shown in **Fig. 4**. For the figure, the parameters $k = 15$, $k_{\text{true}} = 10$, $s = 7$, $N = 100$, and $T = 50$ were deliberately chosen to make the optimization challenging and produce a range of variational EM solutions. Perfect accuracy is achieved in some variational EM runs, but other runs result in low accuracy and low $\mathcal{L}$ values. While the sensitivity of variational EM to the initial parameters may appear prohibitive, given the speed of the algorithm and its ease of parallelization, running hundreds to thousands of independent initializations is feasible and sufficient to yield reliable results. Nonetheless, there remains significant future interest in identifying more principled initializations for variational EM or more robust stochastic variants (Celeux et al. 1996).

5.2 `Last.fm` User Data

The next experiment uses a publicly available data set that records the listening histories of `Last.fm` users in 2007. The works (Kausik et al. 2023; Gupta et al. 2016) report trajectory-level clustering results on this `Last.fm` Markov-chain-mixture benchmark, making them natural baselines.

Following the previous work (Kausik et al. 2023), the ground-truth model is specified as follows. The data contains $k_{\text{true}} = 10$ Markov chains, corresponding to the `Last.fm` users with the greatest number of song listens. The data is coded using $s = 100$ states based on assigning a discrete label to each song in a user's listening history by the predominant genre. If a user sequentially listens to songs from the same genre, these songs are collapsed into a single state.
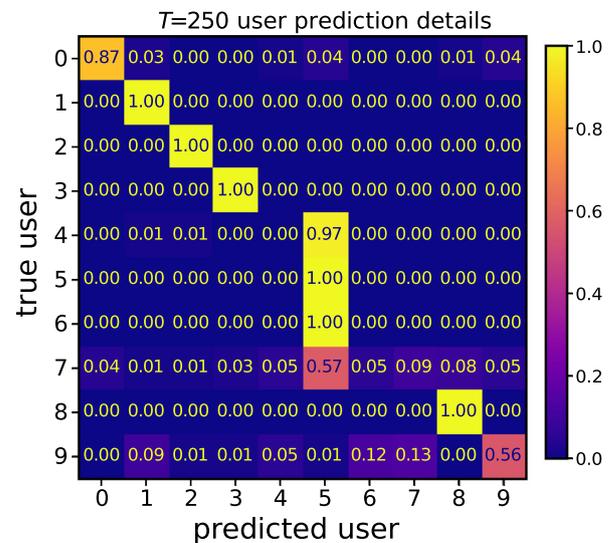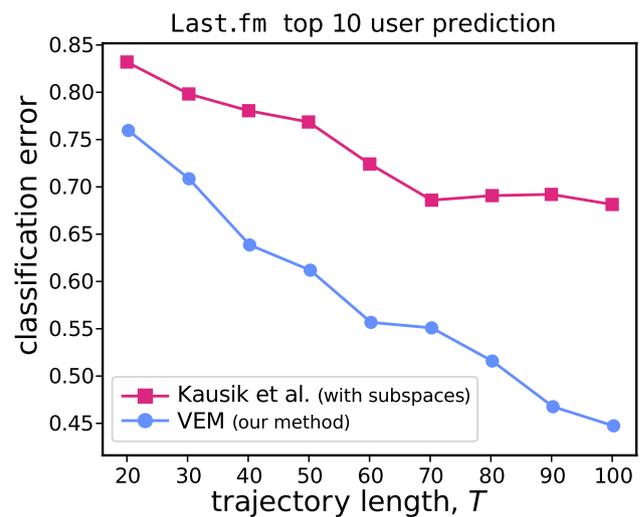

Fig. 5 Classification accuracy for the top 10 `Last.fm` users, given $N = 75$ trajectories and $s = 100$ possible genres for each user. Top: Variational EM is twice as accurate as the best method reported in Kausik et al. (2023). Bottom: Confusion matrix for identifying the top 10 users with $T = 250$

Hence, each Markov chain models the *genre transitions* for a `Last.fm` user. To make the learning problem difficult, each user's listening history is broken into 75 equal-length segments, and the segments are truncated to a short trajectory length $20 \leq T \leq 100$.

Fig. 5 shows that the Bayesian variational EM approach with $k = k_{\text{true}} = 10$ components achieves classification accuracy approximately $2\times$ higher than the experiments in Kausik et al. (2023). The error is also much lower than the 90% classification error in Gupta et al. (2016). These results are included for context, although the earliest benchmark (Gupta et al. 2016) used more numerous and shorter trajectories ($N = 30{,}000$, $T = 3$), so the numbers are not directly comparable.
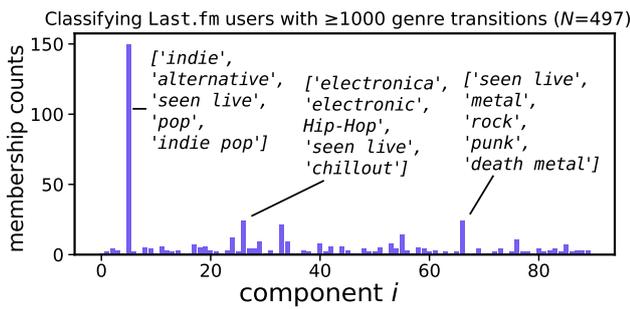
Fig. 6 Classification of users for a `Last.fm` data set with $N = 497$ users, $s = 100$ genres, and trajectory length $T = 1000$. Popular components are labeled with the most frequent genres

Fig. 7 Application of Markov chain mixture modeling to the 2012 International Association of Ultrarunners (IAU) World Championship data. With a maximum of $k = 10$ components, three running patterns emerged: (1) constant pace, (2) a fast start followed by subsequent resting, (3) erratic performance. Black lines show component means

While variational EM outperforms previous approaches, the misclassification rate for length $T = 100$ trajectories is still notably high, 45%. To further improve the accuracy, the trajectories were extended to length $T = 250$, resulting in a reduction to 30% error. The right panel of **Fig. 5** shows the confusion matrix, which indicates that most of the misclassifications come from 4 users who all had similar listening histories.

For further insight, a new data set was prepared with a length $T = 1000$ trajectory for each of the top $N = 497$ `Last.fm` users. Variational EM was applied to this data set with a maximum number of $k = 100$ components. **Fig. 6** shows the mixture modeling results, which indicate that a few components have substantially large memberships. The 3 largest components correspond to well-known genres of 'indie', 'electronic', and 'metal'. The overwhelming popularity of the indie component explains why distinguishing `Last.fm` users may be difficult. Many indie rock listeners had quite similar listening histories in 2007 (Eck et al. 2007).

5.3 Ultrarunners data set

The third experiment is based on data from the 2012 International Association of Ultrarunners (IAU) World Championship held in Katowice, Poland (Bartolucci and Murphy 2015; Roick et al. 2020). The goal of this experiment is to find the predominant pacing patterns in the data and infer which pacing patterns lead to the best overall performance. Conventional wisdom among ultrarunners contends that a slower pace to start is ideal (Berger et al. 2023), as many runners tend to overexert themselves during the start and tire later during a race.

Pacing is an individualized measure of performance, so each runner's data over the 24-hour event was normalized based on average speed. First, 12 runners that did not complete a single lap were removed from the data set, yielding $N = 248$ total trajectories. Then, each runner's average speed was calculated by dividing the total number of laps
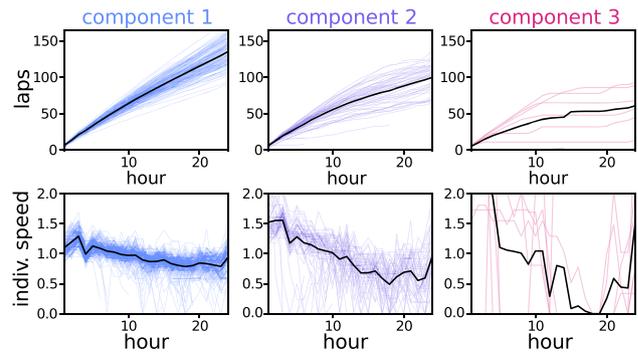
over the 2.314 km course by the total number of hours until finishing or dropping out of the event (typically 24 hours). Each runner's data was normalized by dividing the number of laps per hour by the average speed. Last, the data was categorized into $s = 4$ Markov states: "resting" when the hourly speed is less than 0.5 times the average speed, "normal" when the hourly speed is between 0.5 and 1.5 times the average speed, "strained" when the hourly speed is greater than 1.5 times the average speed, and a terminal "ended" state.

The variational EM algorithm identifies three distinct running patterns in the data, as displayed in **Fig. 7**. The predominant pattern (70% of runners) involves a roughly constant pace, with slightly higher relative speed at the start of the race. The second-most-common pattern (25% of runners) involves overexertion at the start of the race and a significant slowdown later. The third pattern involves erratic running with no apparent strategy. A previous analysis identified similar running patterns in the 2012 ultrarunner data (Bartolucci and Murphy 2015).

The Markov state mixture model is interesting in two ways. First, the runners in group 1 ran notably faster than those in group 2 even though the average speed was not included as input into the algorithm. Rather, a constant pace emerged organically as the top-performing pattern. Second, the ideal strategy of a slower opening pace was not observed in any group. While the slow start is considered optimal, it is also notably difficult to achieve in races (Berger et al. 2023).

5.4 Synthetic Gene Expression Data

The last experiment is based on synthetic data generated from the mutual-inhibition, self-activation (MISA) gene circuit. The MISA circuit has been observed in many biological systems (Graf and Enver 2009; Huang 2009; Zhou and Huang
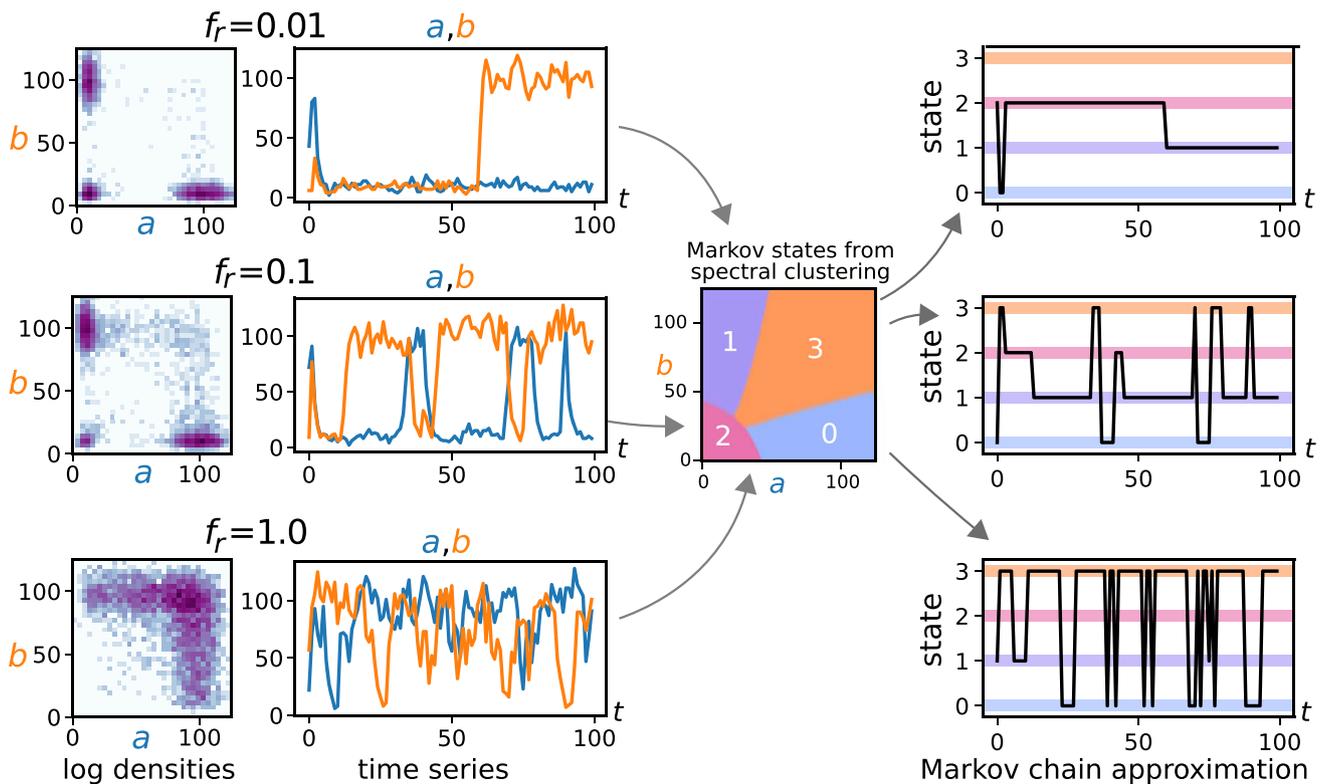
Fig. 8 Schematic of finite state approximation for the mutual-inhibition, self-activation (MISA) gene circuit. Counts of proteins a and b are visualized as scatter plots (column 1) and time series (column 2). (a, b) trajectories are classified into 4 states (column 3), and trajectories are projected based on the 4 states (column 4). Rows indicate different magnitudes of the rate parameter $f_r = 0.01, 0.1, 1.0$
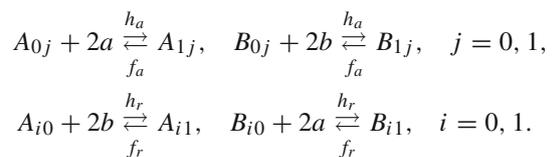
2011; Smith et al. 2016), and it has been studied extensively by theorists (Schultz et al. 2008; Morelli et al. 2008; Feng and Wang 2012; Gallivan et al. 2020). It consists of an A gene and a B gene that inhibit each other and activate themselves through the production of a and b proteins. Mathematically, it can be modeled as a chemical reaction network (Anderson and Kurtz 2015) containing the following species and reactions. There is one A gene and one B gene that each exhibit conditions $ij \in \{00, 01, 10, 11\}$, where $i = 1$ indicates the presence of an activator and $j = 1$ indicates the presence of a repressor. The A and B genes produce proteins at a rate g_{ij}:

$$A_{ij} \xrightarrow{g_{ij}} A_{ij} + a, \quad ij = 00, 01, 10, 11,$$

$$B_{ij} \xrightarrow{g_{ij}} B_{ij} + b, \quad ij = 00, 01, 10, 11,$$
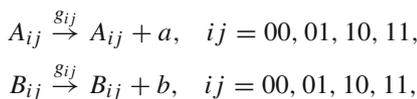
where $g_{00} = g_{01} = g_{11} = 10$ and $g_{10} = 100$. The proteins degrade at rate $d = 1$:

$$a \xrightarrow{d} \varnothing, \quad b \xrightarrow{d} \varnothing.$$
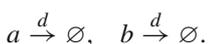
Last, the proteins influence the conditions of the A gene and B gene as follows:

$$A_{0j} + 2a \underset{f_a}{\overset{h_a}{\rightleftarrows}} A_{1j}, \quad B_{0j} + 2b \underset{f_a}{\overset{h_a}{\rightleftarrows}} B_{1j}, \quad j = 0, 1,$$

$$A_{i0} + 2b \underset{f_r}{\overset{h_r}{\rightleftarrows}} A_{i1}, \quad B_{i0} + 2a \underset{f_r}{\overset{h_r}{\rightleftarrows}} B_{i1}, \quad i = 0, 1.$$

Three of the rate parameters are fixed to $h_a = 10^{-1}$, $f_a = 1$, and $h_r = 10^{-3}$. However, f_r is a free parameter that controls the extent of protein activation versus repression in the system. In summary, the MISA gene circuit can be encoded as a vector $X_t \in \{00, 01, 10, 11\}^2 \times \mathbb{N}^2$. The first two coordinates indicate the conditions of the A gene and B gene, and the last two coordinates indicate the populations of the a and b proteins.

The MISA trajectories were simulated using the stochastic simulation algorithm (SSA) implemented in `PyGillespie` (Matthew et al. 2023) with various values of the f_r parameter. The SSA algorithm produces a continuous-time trajectory $(X_t)_{t \geq 0}$; however, the states were sampled at uniformly spaced times $t = 0, 1, 2, \ldots$ to produce a discrete-time data set.

Fig. 8 displays the stochastic switching in the populations of a and b proteins. The protein populations toggle between four metastable states, which are associated with each gene
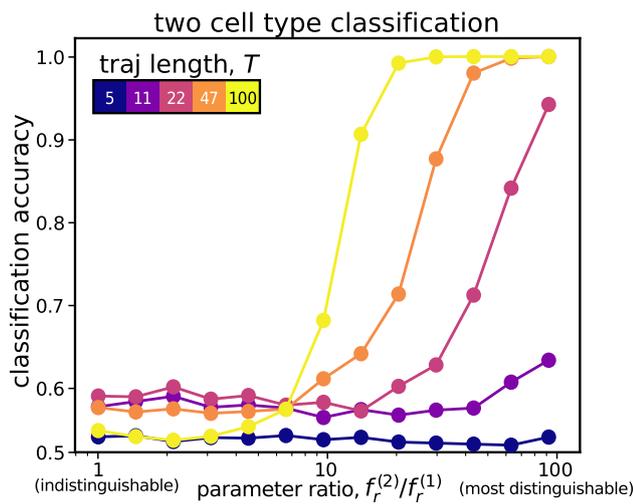
Fig. 9 Classification accuracy for synthetic MISA data containing 15 trajectories with $f_r^{(1)} = 0.01$ and 15 trajectories with $f_r^{(2)} \neq f_r^{(1)}$. Trajectories are reliably identified when $T \geq 25$ and $f_r^{(2)}/f_r^{(1)} \geq 25$

being "on" or "off". When the rate f_r is low (top panels), the genes are frequently off, leading to protein populations of just 0–20. When f_r is high (bottom panels), both genes are frequently on, leading to increased populations of 50–150 proteins. The most interesting behavior occurs for intermediate values of f_r (middle panels), because then one gene is typically on and the other is off. The genes actively compete to produce more proteins.

The MISA gene circuit is not Markovian in the protein numbers a and b. Nonetheless, a Markov state model can be constructed by applying spectral clustering (Coifman et al. 2005) to the (a, b) data. The middle panel of **Fig.** 8 shows the outcome of spectral clustering using a Gaussian kernel with bandwidth $\sigma = 50$. The four identified clusters are interpretable and they efficiently represent the metastable dynamics.

Variational EM can be used to untangle multiple Markov chains from a single MISA data set. This application area is increasingly important, as MISA models with variable parameters are now being engineered in labs (Li et al. 2018) with the goal of understanding long chains of cell differentiation (Zhou and Huang 2011). As an example application, researchers have obtained time-series data for a MISA gene circuit in the bacteriophage λ switch (Fang et al. 2018). They analyzed the time-series data assuming a single Markov chain, but the Markov chain mixture model provides the ability to discern heterogeneities within the samples.

To test the feasibility of the mixture modeling approach, spectral clustering and variational EM were applied to a range of synthetic MISA data sets containing 15 trajectories with an unbinding rate $f_r^{(1)} = 0.01$ and 15 trajectories with a different unbinding rate $f_r^{(2)} \neq f_r^{(1)}$. The algorithms were used to untangle the two populations of trajectories, leading

to the results in **Fig.** 9. The results indicate that the classification accuracy depends greatly on the trajectory length T and the parameter ratio $f_r^{(2)}/f_r^{(1)}$. When $f_r^{(2)}/f_r^{(1)} \geq 25$ and $T \geq 25$, the populations can be separated with near-perfect accuracy. However, short trajectories with $T \leq 5$ cannot be reliably separated for any $f_r^{(2)}$ parameter. These results support the theoretical analysis in Thm. 1, since the Kullback-Leibler divergence increases with $f_r^{(2)}/f_r^{(1)}$ and increases linearly with T.

6 Conclusion

This paper has proposed an extension of Markov state modeling that enables the study of heterogeneities within time-series data. Previous work employed Markov state modeling under the assumption that the observations come from a homogeneous population modeled by a single Markov chain. This paper develops the theory and practice of learning a mixture of different Markov chains simultaneously.

The paper has motivated and tested a variational EM algorithm that automatically identifies the number of chains and the dynamics of each chain. The proposed mixture modeling approach deliberately combines classical, well-understood components, based on the discretization step from Markov state modeling and the fitting step from variational EM. It contributes an interpretable and competitive baseline for heterogeneous dynamical data. The algorithm is computationally efficient: unlike past work, it identifies the number of chains organically, without relying on expensive model comparisons or posterior sampling.

The current bottleneck, shared among all the standard methods for fitting Markov chain mixture models (EM and variational EM), is the convergence to locally optimal parameters. Running these methods with many random initializations is the currently proposed solution, but it is the computationally limiting aspect of the approach. Future work should identify and rigorously justify an initialization strategy that works well for variational EM with any trajectory length T.

The paper also provides a theorem that lower bounds the classification error for any Markov chain mixture model, similar to classical information-theoretic bounds. The bound is stated in terms of the Kullback-Leibler divergence between the underlying Markov chains, and it suggests the optimal classification error decreases exponentially in T because the trajectory-level KL divergence grows linearly in T. This prediction is repeatedly supported in the numerical experiments (Secs. 5.1, 5.2, 5.4). The theory quantifies the established wisdom (Ramoni et al. 2002) that long trajectories are better than short ones. The bound requires no ergodicity or mixing-time assumptions, making it applicable to transient

or non-stationary dynamics common in short experimental trajectories.

Last, there is a natural question: when should scientists use Markov chain mixture models in the future? Three of the four experiments (Secs. 5.2, 5.3, 5.4) involved real data or real biological systems, and the mixture modeling approach led to accurate and interpretable results. One of the authors (C.E.M.) listened to indie rock on `Last.fm` in 2007 while the other author (R.J.W.) runs ultramarathons competitively, and they can attest to the qualitative accuracy of the Markov chain mixture models. More broadly, these results reinforce a simple point: careful probabilistic modeling and inference, even when built from classical ingredients, can remain competitive in modern high-throughput settings, serving as an interpretable baseline against which newer approaches are compared.

In conclusion, Markov chain mixture models have a track record of distinguishing meaningful heterogeneities in human behavior, for example, heterogeneous patterns of surfing the internet (Melnykov 2016), commuting between home, work, and school (Zhou et al. 2021), listening to music (Sec. 5.2), or running ultramarathons (Sec. 5.3). The variational EM algorithm is an efficient strategy for fitting these models while automatically selecting the number of components. Moving forward, Markov chain mixture models could be extended to other areas (chemistry, biology, climate science) where Markov state modeling is already popular. Indeed, the experiments in Sec. 5.4 demonstrate promising first steps toward applications in the rapidly evolving area of gene expression data analysis (Eisen et al. 1998; Ernst et al. 2005; McDowell et al. 2018; Mitra and MacLean 2021).

Author Contributions Both authors conceived of the project, created the main algorithm, and wrote the manuscript. R.J.W. developed the mathematical theory. C.E.M. performed numerical experiments.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Code availability Code to reproduce the numerical experiments can be found in the GitHub repository https://github.com/chris-miles/MarkovChain-VEM.

Competing interests The authors declare no competing interests.

References

Anderson, D.F., Kurtz, T.G.: Stochastic Analysis of Biochemical Systems. Springer, Switzerland. (2015). https://doi.org/10.1007/978-3-319-16895-1

Bacallado, S., Chodera, J.D., Pande, V.: Bayesian comparison of markov models of molecular dynamics with detailed balance constraint. J. Chem. Phys. **131**(4), 045106 (2009). https://doi.org/10.1063/1.3192309

Bartolucci, F., Murphy, T.B.: A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. Journal of Quantitative Analysis in Sports **11**(4), 193–203 (2015). https://doi.org/10.1515/jqas-2014-0060

Batu, T., Guha, S., Kannan, S.: Inferring mixtures of Markov chains. In: 17th Annual Conference on Learning Theory, 186–199 (2004). https://doi.org/10.1007/978-3-540-27819-1_13

Berger, N.J.A., Best, R., Best, A.W., Lane, A.M., Millet, G.Y., Barwood, M., Marcora, S., Wilson, P., Bearden, S.: Limits of ultra: Towards an interdisciplinary understanding of ultra-endurance running performance. Sports Med. **54**(1), 73–93 (2023). https://doi.org/10.1007/s40279-023-01936-8

Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York. (2006). https://doi.org/10.1007/978-0-387-45528-0

Bretagnolle, J., Huber, C.: Estimation des densités:risque minimax. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **47**(2), 119–137 (1979)

Celeux, G., Chauveau, D., Diebolt, J.: Stochastic versions of the EM algorithm: An experimental study in the mixture case. J. Stat. Comput. Simul. **55**(4), 287–314 (1996) https://doi.org/10.1080/00949659608811772

Chen, Y., Epperly, E.N., Tropp, J.A., Webber, R.J.: Randomly pivoted Cholesky: Practical approximation of a kernel matrix with few entry evaluations. Commun. Pure Appl. Math. **78**(5), 995–1041 (2025). https://doi.org/10.1002/cpa.22234

Chierichetti, F., Kumar, R., Raghavan, P., Sarlos, T.: Are web users really Markovian? In: Proceedings of the 21st International Conference on World Wide Web (2012). https://doi.org/10.1145/2187836.2187919

Chodera, J.D., Noé, F.: Markov state models of biomolecular conformational dynamics. Curr. Opin. Struct. Biol. **25**, 135–144 (2014). https://doi.org/10.1016/j.sbi.2014.04.002

Chu, B.K., Tse, M.J., Sato, R.R., Read, E.L.: Markov state models of gene regulatory networks. BMC Syst. Biol. (2017). https://doi.org/10.1186/s12918-017-0394-4

Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., Zucker, S.W.: Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proc. Natl. Acad. Sci. **102**(21), 7426–7431 (2005). https://doi.org/10.1073/pnas.0500334102

Corduneanu, A., Bishop, C.M.: Hyperparameters for soft Bayesian model selection. In: Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics (2001). https://proceedings.mlr.press/r3/corduneanu01a.html

Das, P., Sen, D., De, D., Hou, J., Abad, Z.S.H., Kim, N., Xia, Z., Cai, T.: Clustering sequence data with mixture Markov chains with covariates using multiple simplex constrained optimization routine (MSiCOR). Journal of Computational and Graphical Statistics 0(0), 1–14 (2023a) https://doi.org/10.1080/10618600.2023.2257258

Das, P., Weisenfeld, D., Dahal, K., De, D., Feathers, V., Coblyn, J.S., Weinblatt, M.E., Shadick, N.A., Cai, T., Liao, K.P.: Utilizing biologic disease-modifying anti-rheumatic treatment sequences to subphenotype rheumatoid arthritis. Arthritis Research & Therapy (2023b). https://doi.org/10.1186/s13075-023-03072-0

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) 39(1), 1–22 (2018). https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Eck, D., Lamere, P., Bertin-Mahieux, T., Green, S.: Automatic generation of social tags for music recommendation. In: Proceedings of the 20th International Conference on Neural Information Processing Systems (2007). https://doi.org/10.5555/2981562.2981611

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95(25), 14863–14868 (1998). https://doi.org/10.1073/pnas.95.25.14863

Ernst, J., Nau, G.J., Bar-Joseph, Z.: Clustering short time series gene expression data. Bioinformatics 21(suppl_1), 159–168 (2005) https://doi.org/10.1093/bioinformatics/bti1022

Fang, X., Liu, Q., Bohrer, C., Hensel, Z., Han, W., Wang, J., Xiao, J.: Cell fate potentials and switching kinetics uncovered in a classic bistable genetic switch. Nat. Commun. (2018). https://doi.org/10.1038/s41467-018-05071-1

Feng, H., Wang, J.: A new mechanism of stem cell differentiation through slow binding/unbinding of regulators to genes. Sci. Rep. 2(1), 550 (2012). https://doi.org/10.1038/srep00550

Finkel, J., Gerber, E.P., Abbot, D.S., Weare, J.: Revealing the statistics of extreme events hidden in short weather forecast data. AGU Advances 4(2), 2023–000881 (2023). https://doi.org/10.1029/2023AV000881

Fitzpatrick, M., Stewart, M.: Asymptotics for Markov chain mixture detection. Econometrics and Statistics 22, 56–66 (2022). https://doi.org/10.1016/j.ecosta.2021.11.004

Frydman, H.: Estimation in the mixture of Markov chains moving with different speeds. J. Am. Stat. Assoc. 100(471), 1046–1053 (2005). https://doi.org/10.1198/016214505000000024

Gallivan, C.P., Ren, H., Read, E.L.: Analysis of single-cell gene pair coexpression landscapes by stochastic kinetic modeling reveals gene-pair interactions in development. Front. Genet. 10, 1387 (2020). https://doi.org/10.3389/fgene.2019.01387

Gormley, I.C., Murphy, T.B., Raftery, A.E.: Model-based clustering. Annual Review of Statistics and Its Application 10, 573–595 (2023). https://doi.org/10.1146/annurev-statistics-033121-115326

Graf, T., Enver, T.: Forcing cells to change lineages. Nature 462(7273), 587–594 (2009). https://doi.org/10.1038/nature08533

Gupta, R., Kumar, R., Vassilvitskii, S.: On mixtures of Markov chains. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (2016). https://doi.org/10.5555/3157382.3157483

Haan-Rietdijk, S., Kuppens, P., Bergeman, C.S., Sheeber, L.B., Allen, N.B., Hamaker, E.L.: On the use of mixed markov models for intensive longitudinal data. Multivar. Behav. Res. 52(6), 747–767 (2017). https://doi.org/10.1080/00273171.2017.1370364

Hemant Ishwaran, L.F.J., Sun, J.: Bayesian model selection in finite mixtures by marginal density decompositions. J. Am. Stat. Assoc. 96(456), 1316–1332 (2001). https://doi.org/10.1198/016214501753382255

Huang, S.: Reprogramming cell fates: reconciling rarity with robustness. BioEssays 31(5), 546–560 (2009). https://doi.org/10.1002/bies.200800189

Husic, B.E., Pande, V.S.: Markov state models: From an art to a science. J. Am. Chem. Soc. 140(7), 2386–2396 (2018). https://doi.org/10.1021/jacs.7b12191

Kausik, C., Tan, K., Tewari, A.: Learning mixtures of Markov chains and MDPs. In: Proceedings of the 40th International Conference on Machine Learning (2023). https://proceedings.mlr.press/v202/kausik23a.html

Keribin, C.: Consistent estimation of the order of mixture models. Sankhyā: The Indian Journal of Statistics, Series A 62(1), 49–66 (2000)

Khaleghi, A., Ryabko, D., Mary, J., Preux, P.: Consistent algorithms for clustering time series. J. Mach. Learn. Res. 17(3), 1–32 (2016)

Köhs, L., Kukovetz, K., Rauh, O., Koeppl, H.: Nonparametric Bayesian inference for meta-stable conformational dynamics. Phys. Biol. 19(5), 056006 (2022). https://doi.org/10.1088/1478-3975/ac885e

Lee, J., Jedra, Y., Proutière, A., Yun, S.-Y.: Near-optimal clustering in mixture of markov chains. arXiv preprint arXiv:2506.01324 (2025)

Li, T., Dong, Y., Zhang, X., Ji, X., Luo, C., Lou, C., Zhang, H.M., Ouyang, Q.: Engineering of a genetic circuit with regulatable multistability. Integr. Biol. 10(8), 474–482 (2018). https://doi.org/10.1039/c8ib00030a

Lloyd, S.: Least squares quantization in pcm. IEEE Trans. Inf. Theory 28(2), 129–137 (1982). https://doi.org/10.1109/TIT.1982.1056489

MacKay, D.J.: Local Minima, Symmetry-breaking, and Model Pruning in Variational Free Energy Minimization. Department of Physics, University of Cambridge (2001). https://www.inference.org.uk/mackay/minima.pdf

Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B.: Model-based clustering based on sparse finite gaussian mixtures. Stat. Comput. 26(1–2), 303–324 (2014). https://doi.org/10.1007/s11222-014-9500-2

Matthew, S., Carter, F., Cooper, J., Dippel, M., Green, E., Hodges, S., Kidwell, M., Nickerson, D., Rumsey, B., Reeve, J., et al.: Gillespy2: A biochemical modeling framework for simulation driven biological discovery. Letters in Biomathematics 10(1), 87–103 (2023)

McDowell, I.C., Manandhar, D., Vockley, C.M., Schmid, A.K., Reddy, T.E., Engelhardt, B.E.: Clustering gene expression time series data using an infinite Gaussian process mixture model. PLoS Comput. Biol. 14(1), 1–27 (2018). https://doi.org/10.1371/journal.pcbi.1005896

Melnykov, V.: ClickClust: An R package for model-based clustering of categorical sequences. J. Stat. Softw. 74(9), 1–34 (2016). https://doi.org/10.18637/jss.v074.i09

Mitra, R., MacLean, A.L.: RVAgene: Generative modeling of gene expression time series data. Bioinformatics 37(19), 3252–3262 (2021). https://doi.org/10.1093/bioinformatics/btab260

Morelli, M.J., Tănase-Nicola, S., Allen, R.J., Ten Wolde, P.R.: Reaction coordinates for the flipping of genetic switches. Biophys. J. 94(9), 3413–3423 (2008). https://doi.org/10.1529/biophysj.107.116699

Noé, F., Wu, H., Prinz, J.-H., Plattner, N.: Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. J. Chem. Phys. 139(18), 184114 (2013). https://doi.org/10.1063/1.4828816

Norris, J.R.: Markov Chains. Cambridge University Press, Cambridge (1997). https://doi.org/10.1017/CBO9780511810633

Onn, S., Weissman, I.: Generating uniform random vectors over a simplex with implications to the volume of a certain polytope and to multivariate extremes. Ann. Oper. Res. **189**(1), 331–342 (2009). https://doi.org/10.1007/s10479-009-0567-7

Pande, V.S., Beauchamp, K., Bowman, G.R.: Everything you wanted to know about markov state models but were afraid to ask. Methods **52**(1), 99–105 (2010). https://doi.org/10.1016/j.ymeth.2010.06.002. (**Protein Folding**)

Ramoni, M., Sebastiani, P., Cohen, P.: Bayesian clustering by dynamics. Mach. Learn. **47**(1), 91–121 (2002). https://doi.org/10.1023/a:1013635829250

Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. **26**(2), 195–239 (1984). https://doi.org/10.1137/1026034

Röblitz, S., Weber, M.: Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. Adv. Data Anal. Classif. **7**(2), 147–179 (2013). https://doi.org/10.1007/s11634-013-0134-6

Roick, T., Karlis, D., McNicholas, P.D.: Clustering discrete-valued time series. Adv. Data Anal. Classif. **15**(1), 209–229 (2020). https://doi.org/10.1007/s11634-020-00395-7

Rousseau, J., Mengersen, K.: Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(5), 689–710 (2011). https://doi.org/10.1111/j.1467-9868.2011.00781.x

Safinianaini, N., De Souza, C.P., Roth, A., Koptagel, H., Toosi, H., Lagergren, J.: Copymix: mixture model based single-cell clustering and copy number profiling using variational inference. Comput. Biol. Chem. **113**, 108257 (2024)

Schultz, D., Walczak, A.M., Onuchic, J.N., Wolynes, P.G.: Extinction and resurrection in gene networks. Proc. Natl. Acad. Sci. **105**(49), 19165–19170 (2008). https://doi.org/10.1073/pnas.0810366105

Smith, T.D., Tse, M.J., Read, E.L., Liu, W.F.: Regulation of macrophage polarization and plasticity by complex activation signals. Integr. Biol. **8**(9), 946–955 (2016). https://doi.org/10.1039/c6ib00105j

Smyth, P.: Clustering sequences with hidden Markov models. In: Proceedings of the 9th International Conference on Neural Information Processing Systems (1996). https://doi.org/10.5555/2998981.2999073

Souza, A.N.: Representing turbulent statistics with partitions of state space. part 1. theory and methodology. Journal of Fluid Mechanics 997, 1 (2024)

Spaeh, F., Tsourakakis, C.: Learning mixtures of Markov chains with quality guarantees. In: Proceedings of the ACM Web Conference (2023). https://doi.org/10.1145/3543507.3583524

Springer, S., Laio, A., Galfi, V.M., Lucarini, V.: Unsupervised detection of large-scale weather patterns in the northern hemisphere via markov state modelling: from blockings to teleconnections. npj Climate and Atmospheric Science 7(1), 105 (2024)

Tan, Z.W., Guarnera, E., Berezovsky, I.N.: Exploring chromatin hierarchical organization via markov state modelling. PLoS Comput. Biol. **14**(12), 1–35 (2019). https://doi.org/10.1371/journal.pcbi.1006686

Tse, M.J., Chu, B.K., Gallivan, C.P., Read, E.L.: Rare-event sampling of epigenetic landscapes and phenotype transitions. PLoS Comput. Biol. **14**(8), 1–28 (2018). https://doi.org/10.1371/journal.pcbi.1006336

Zhou, J.X., Huang, S.: Understanding gene circuits at cell-fate branch points for rational cell reprogramming. Trends Genet. **27**(2), 55–62 (2011). https://doi.org/10.1016/j.tig.2010.11.002

Zhou, Y., Yuan, Q., Yang, C., Wang, Y.: Who you are determines how you travel: Clustering human activity patterns with a Markov-chain-based mixture model. Travel Behaviour and Society **24**, 102–112 (2021). https://doi.org/10.1016/j.tbs.2021.03.005