Hidden Markov Models

Learn Bayes Methods Week
at the Karolinska Institutet

Vianey Leos Barajas
Department of Statistical Sciences/School of the Environment
University of Toronto

Overview

- **First part:**

  - medical examples
  - finite mixture models
  - Markov chains
  - hidden Markov models (HMMs)
  - **simulating data in R**
  - *break*

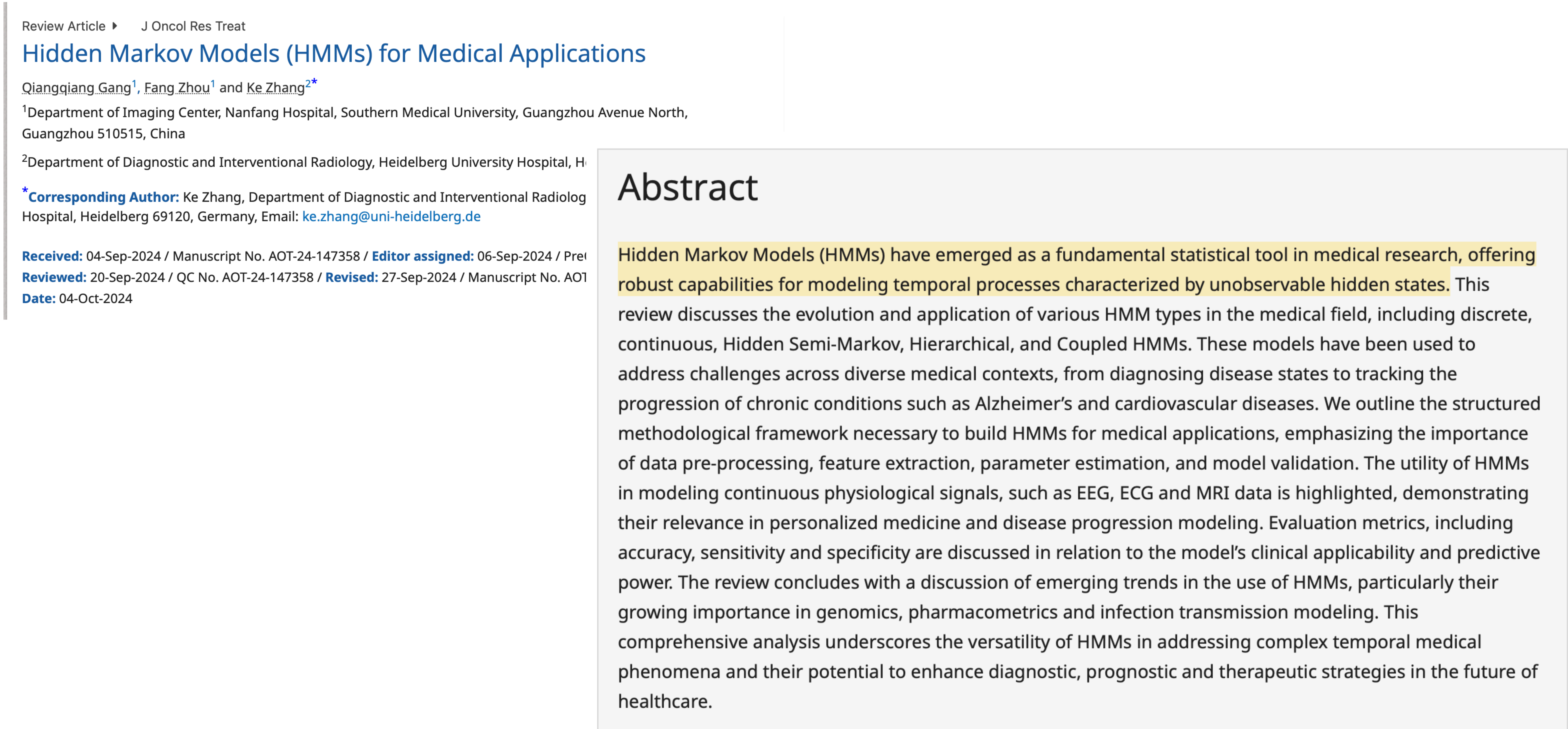
- **Second part:**

  - forward algorithm & likelihood evaluation
  - Bayesian inference
  - state decoding
  - **fitting HMMs in Stan**
  - *break*

- **Third part (discussion):**

  - missing values
  - covariates
  - mixed HMMs and random effects
  - multivariate observations
  - continuous-time HMMs + other extensions

HMMs in Medicine



Review Article ▸ J Oncol Res Treat

Hidden Markov Models (HMMs) for Medical Applications

Qiangqiang Gang[1], Fang Zhou[1] and Ke Zhang[2]*

[1]Department of Imaging Center, Nanfang Hospital, Southern Medical University, Guangzhou Avenue North, Guangzhou 510515, China

[2]Department of Diagnostic and Interventional Radiology, Heidelberg University Hospital, H

***Corresponding Author:** Ke Zhang, Department of Diagnostic and Interventional Radiolog Hospital, Heidelberg 69120, Germany, Email: ke.zhang@uni-heidelberg.de

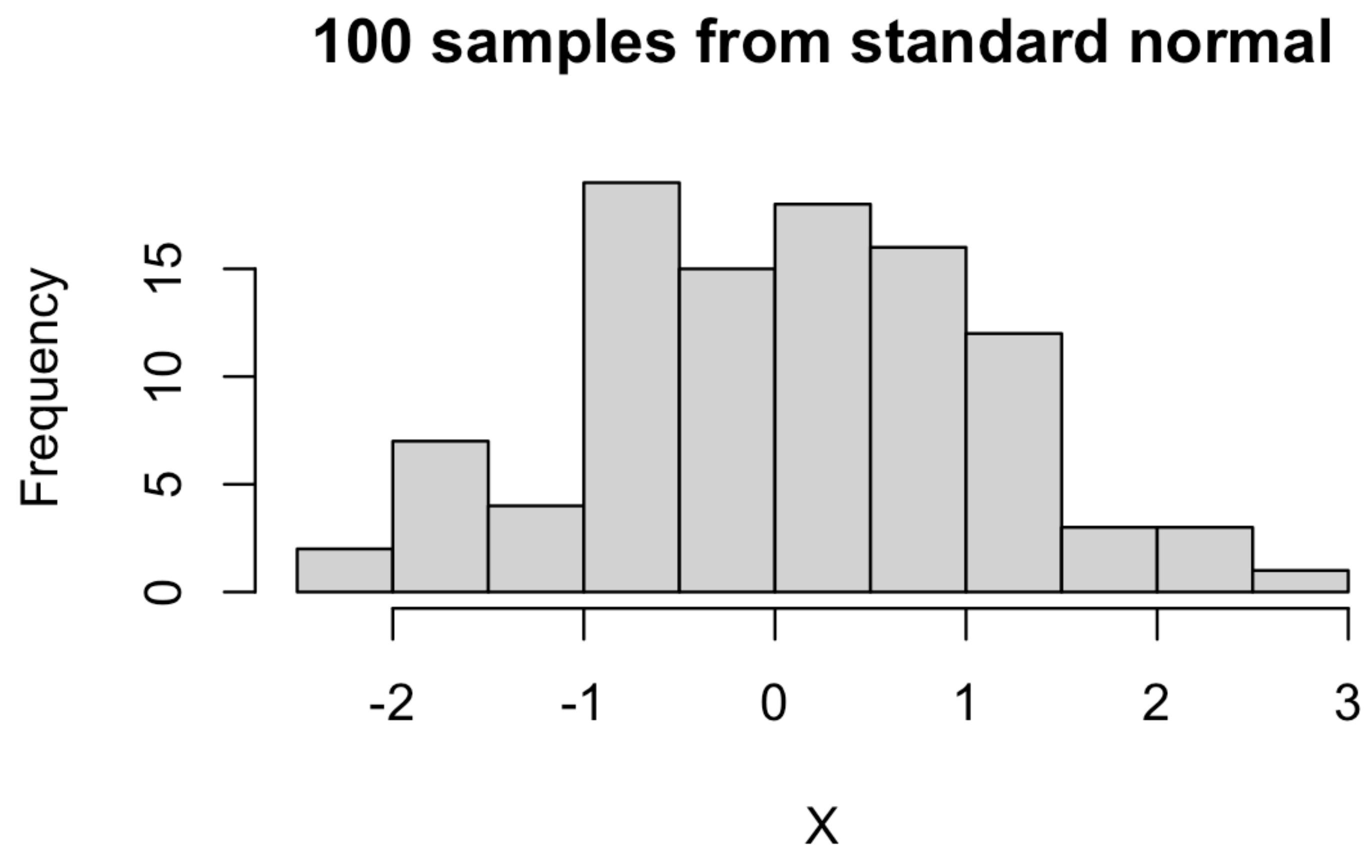
Received: 04-Sep-2024 / Manuscript No. AOT-24-147358 / **Editor assigned:** 06-Sep-2024 / Pre
Reviewed: 20-Sep-2024 / QC No. AOT-24-147358 / **Revised:** 27-Sep-2024 / Manuscript No. AOT
Date: 04-Oct-2024

Abstract

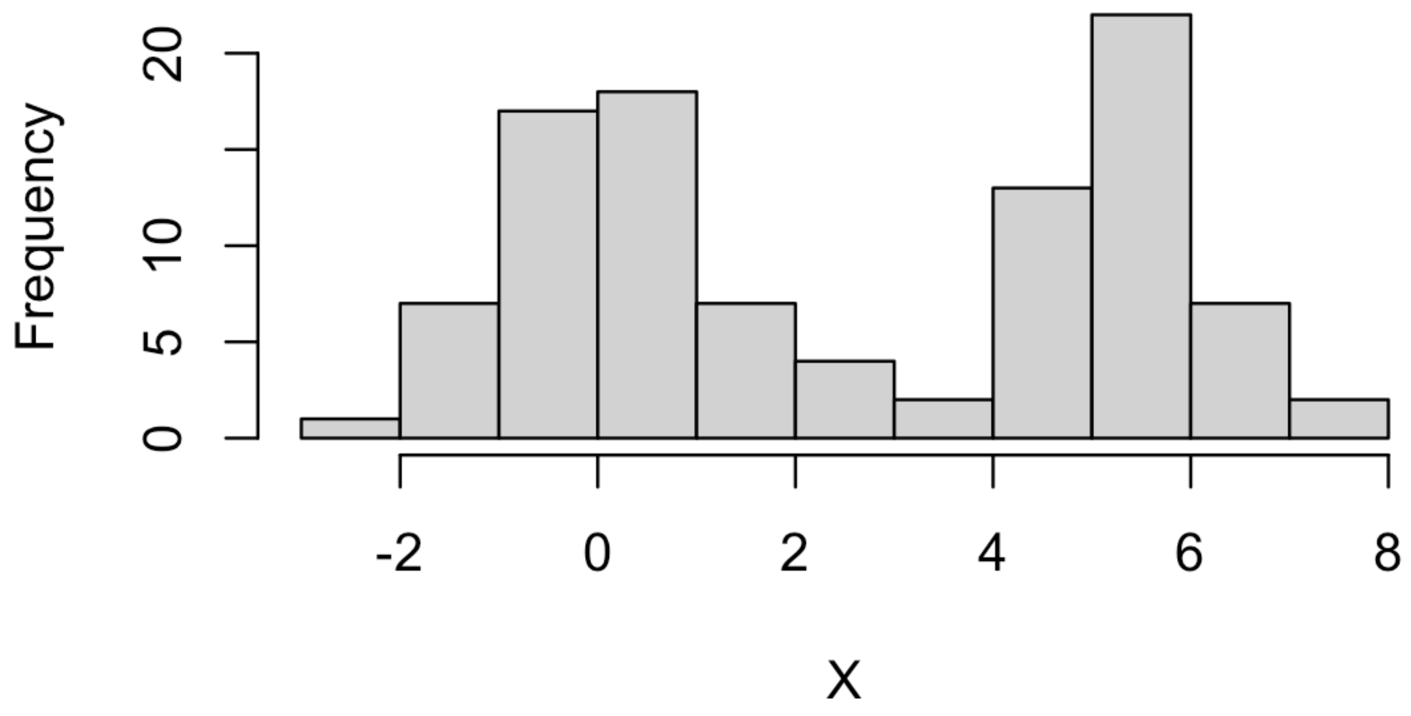
Hidden Markov Models (HMMs) have emerged as a fundamental statistical tool in medical research, offering robust capabilities for modeling temporal processes characterized by unobservable hidden states. This review discusses the evolution and application of various HMM types in the medical field, including discrete, continuous, Hidden Semi-Markov, Hierarchical, and Coupled HMMs. These models have been used to address challenges across diverse medical contexts, from diagnosing disease states to tracking the progression of chronic conditions such as Alzheimer’s and cardiovascular diseases. We outline the structured methodological framework necessary to build HMMs for medical applications, emphasizing the importance of data pre-processing, feature extraction, parameter estimation, and model validation. The utility of HMMs in modeling continuous physiological signals, such as EEG, ECG and MRI data is highlighted, demonstrating their relevance in personalized medicine and disease progression modeling. Evaluation metrics, including accuracy, sensitivity and specificity are discussed in relation to the model’s clinical applicability and predictive power. The review concludes with a discussion of emerging trends in the use of HMMs, particularly their growing importance in genomics, pharmacometrics and infection transmission modeling. This comprehensive analysis underscores the versatility of HMMs in addressing complex temporal medical phenomena and their potential to enhance diagnostic, prognostic and therapeutic strategies in the future of healthcare.

Finite Mixture Models

- Let's start with something simple, $Y \sim N(0,1)$. A histogram of a 100 samples of Y might look like:



- But what happens when our data look like:



Finite Mixture Models

- Finite mixture models (or independent mixture models) consist of a finite number of component distributions and a mechanism that 'mixes' them

- A finite mixture model with K components is given by:

$$f(Y) = \sum_{k=1}^{K} \pi_k f_k(Y)$$

- where $f_k(Y)$ corresponds to the k^{th} component distribution and π_k to the probability that the component is 'active'

- $\pi_k \in (0,1)$ and $\sum_{k=1}^{K} \pi_k = 1$

- Model is fully characterized by the parameters of each of the component distributions $\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\right)$ and $\left(\pi_1, \ldots, \pi_K\right)$

Finite Mixture Models

Likelihood evaluation + Clustering

- Given a sample of N observations, the **likelihood function** is given by:

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k f_k(y_n) \right\}$$

- **Label-switching** — a re-ordering of the labels leads to the same likelihood function/model.

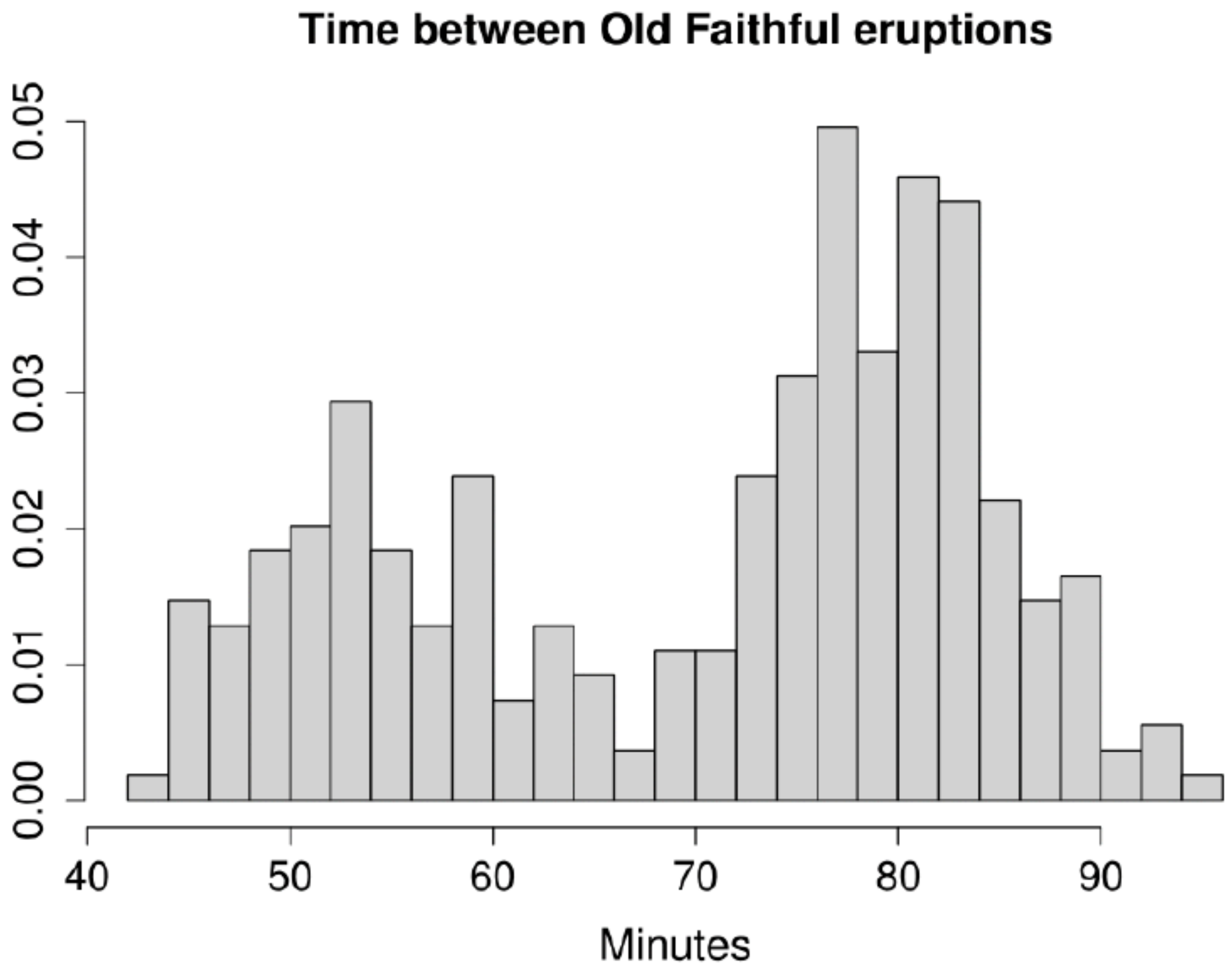
- **Clustering using Bayes' rule** — we can compute the probability that an observation was generated according to one of K possibilities by introducing random variables $\{Z_n\}_{n=1}^{N}$

$$Pr(Z_n = k \mid Y_n) = \frac{\pi_k f_k(Y_n)}{\sum_{k=1}^{K} \pi_k f_k(Y_n)}$$

Example: Time between Old Faithful eruptions

- Waiting times between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA

- The observations seem to exhibit two patterns, arising from one of two possible distributions

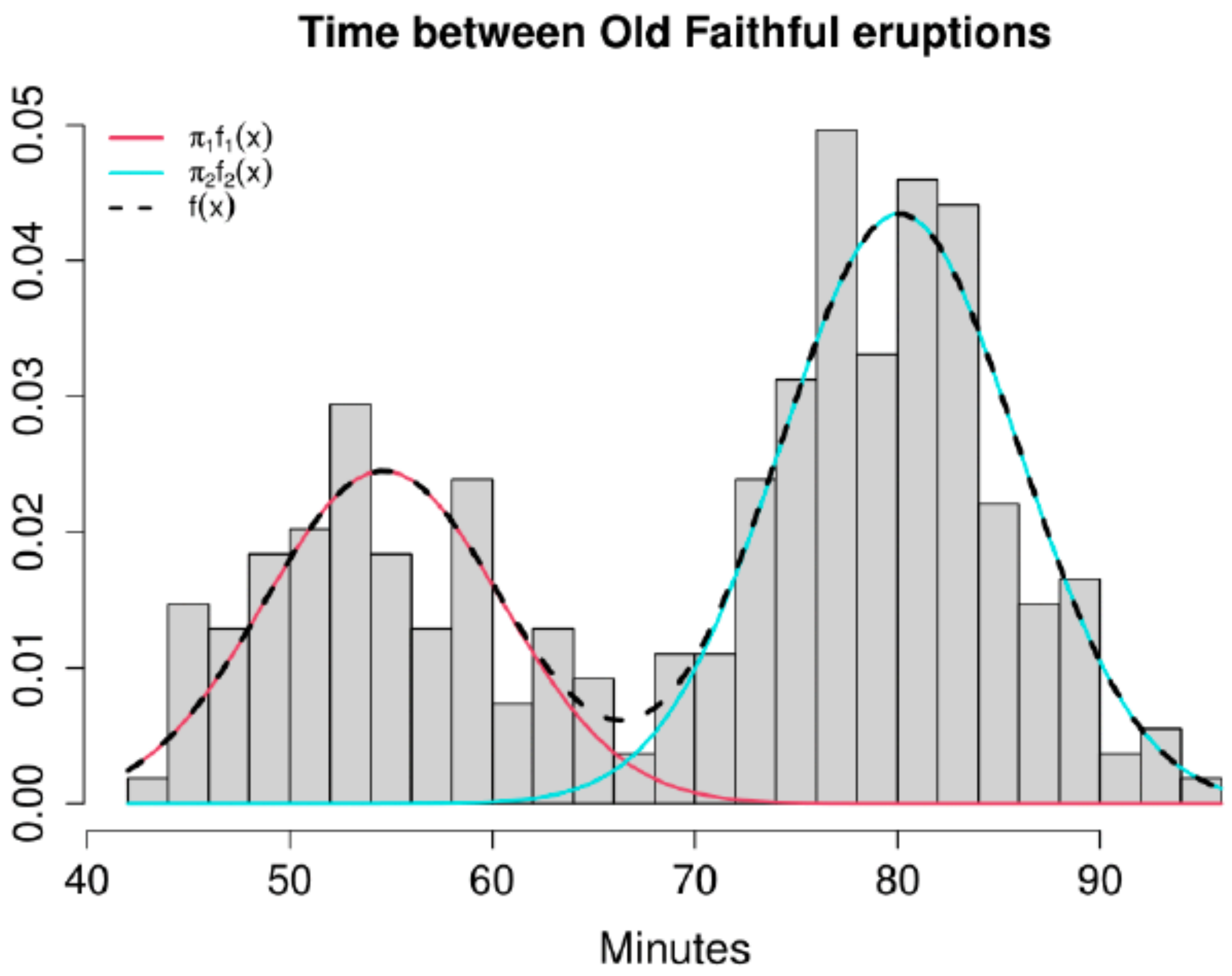


Back to Old Faithful

Point estimates

- Estimating a Gaussian mixture model for the old faithful data with 2 components:

  - $f_1(Y)$ was estimated to be $\mathcal{N}(54.6, 5.9^2)$
  - $f_2(Y)$ was estimated to be $\mathcal{N}(80.1, 5.9^2)$
  - π_1 was estimated to be 0.36
  - π_2 was estimated to be 0.64



Serial dependence

- Data collected over time or in sequence commonly show dependence between consecutive time steps

- **Dependent mixtures** better accommodate system dynamics arising from serial correlation

- In many cases it is reasonable to assume that the component distribution active at time t will more likely remain active at time $t + 1$

- **Markov chains** are a natural selection to model such dependence

Markov Chains

- A **discrete-time Markov chain** is a stochastic process $\{Z_t \in \{1,\ldots,K\}; t = 1,2,\ldots\}$ that satisfies the Markov property $Pr\left(Z_t \,|\, Z_t, Z_{t-1}, \ldots, Z_1\right) = Pr\left(Z_t \,|\, Z_{t-1}\right)$

- It is fully characterized by:

  - K, the number of components (which we'll denote as states from now on)

  - $\boldsymbol{\delta}$, the **initial state distribution**, with entries $Pr(Z_1 = k) = \delta_k$

  - the evolution of the states over time are governed by a transition probability matrix, $\boldsymbol{\Gamma}$, with entries $\gamma_{i,j}^{(t)} = Pr(Z_t = j \,|\, Z_{t-1} = i)$ for $i, j \in \{1,\ldots,K\}$

Markov Chains

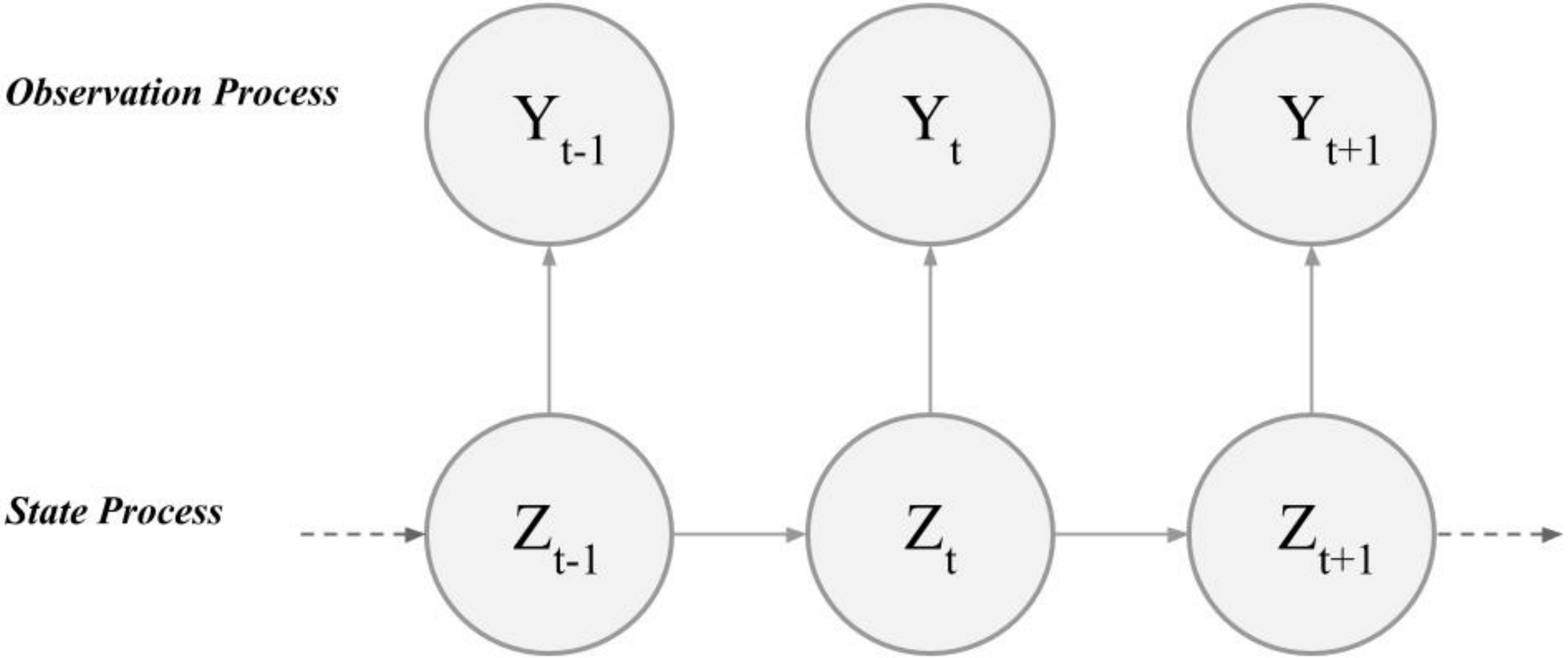
- $\gamma_{i,j}^{(t)}$ is the probability that the chain enters state j at time $t+1$ given that it is in state i at time t
- The chain is called homogeneous when $\gamma_{i,j}^{(t)} = \gamma_{i,j}$ for all $t \in \{1, \ldots, T\}$
- $\boldsymbol{\Gamma}$ is the transition probability matrix, given by

$$\begin{pmatrix} \gamma_{1,1} & \cdots & \gamma_{1,K} \\ \vdots & \ddots & \vdots \\ \gamma_{K,1} & \cdots & \gamma_{K,K} \end{pmatrix}$$

with $\gamma_{i,j} \in [0,1]$ for all $i, j \in \{1, \ldots, K\}$ and $\displaystyle\sum_{j=1}^{K} \gamma_{i,j} = 1$

Hidden Markov models

Combining finite mixture models + dependence via a Markov chain



Some hidden Markov model basics

(discrete-time finite-state)

- A hidden Markov model (HMM) is a doubly stochastic process composed of an **observation process** $\{Y_t\}_{t=1}^{T}$ and a **state process** $\{Z_t\}_{t=1}^{T}$

- A basic HMM assumes **conditional independence** of the observation process given the states, $Y_k \perp Y_h \mid \boldsymbol{Z}$ for $k \neq h$ and $k, h \in \{1, \ldots, T\}$

Some hidden Markov model basics

(discrete-time finite-state)

- The state process is assumed to be a **first-order Markov chain** evolving in discrete-time with transition probability matrix $\boldsymbol{\Gamma}$, $\Gamma_{ij} = \Pr(Z_t = j \mid Z_{t-1} = i)$, and initial state distribution $\boldsymbol{\delta}$, $\delta_i = \Pr(Z_1 = i)$

- The observations are generated according to a set of **state-dependent distributions**, $f(Y_t \mid Z_t = i)$, where $i \in \{1,\ldots,N\}$ and where $N \in \mathbb{N}$ is the # of states

Simulating from a 3-State HMM

- Parameter values

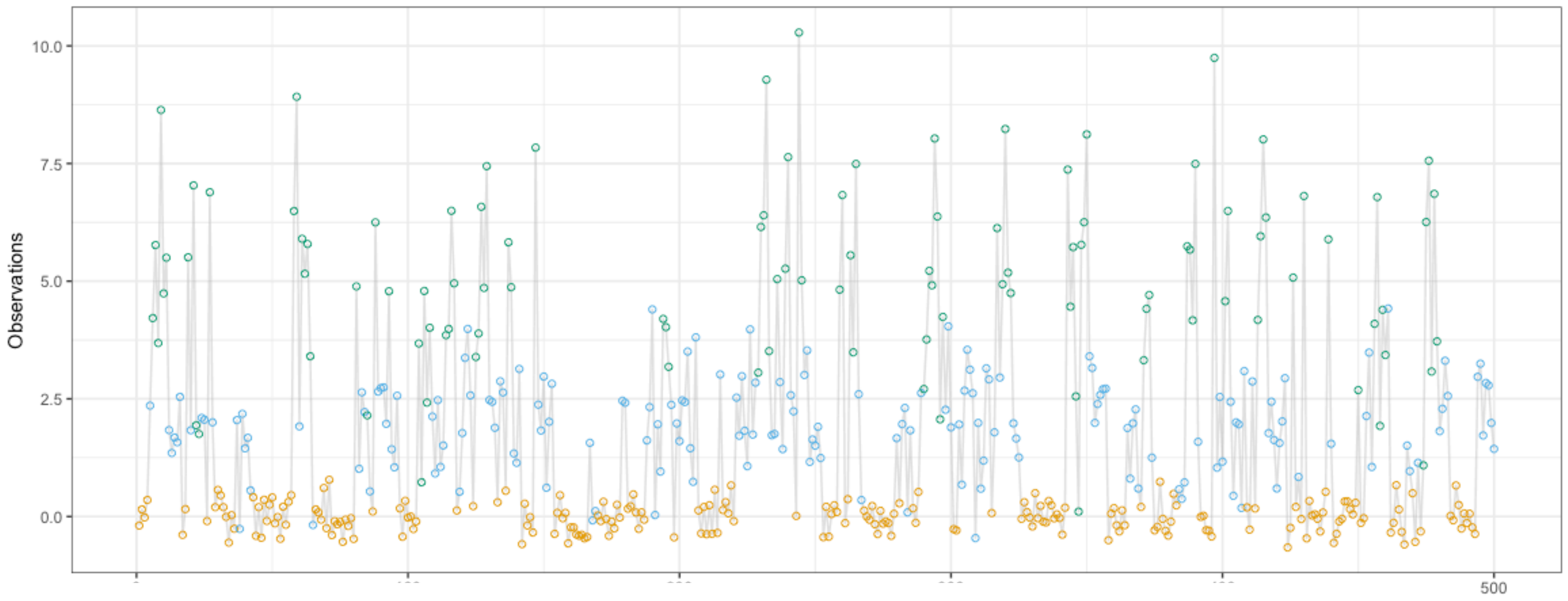
$$N = 3$$

$$f_n(Y_t) \sim N(\mu_n, \sigma_n) \;\; \textbf{for} \;\; n \in \{1,2,3\}$$

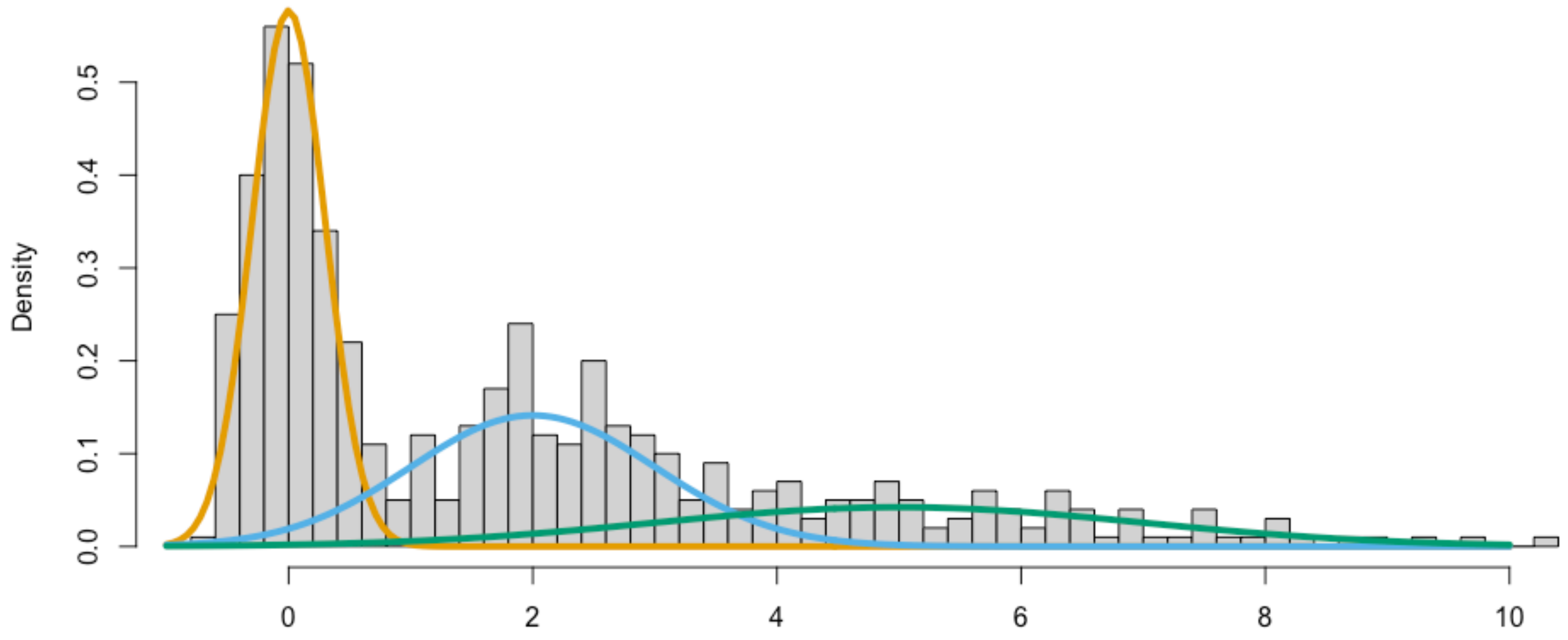
$$\mu \in \{0,2,5\} \quad \sigma \in \{0.3,1,2\}$$

$$\boldsymbol{\delta} = [1/3,1/3,1/3]$$

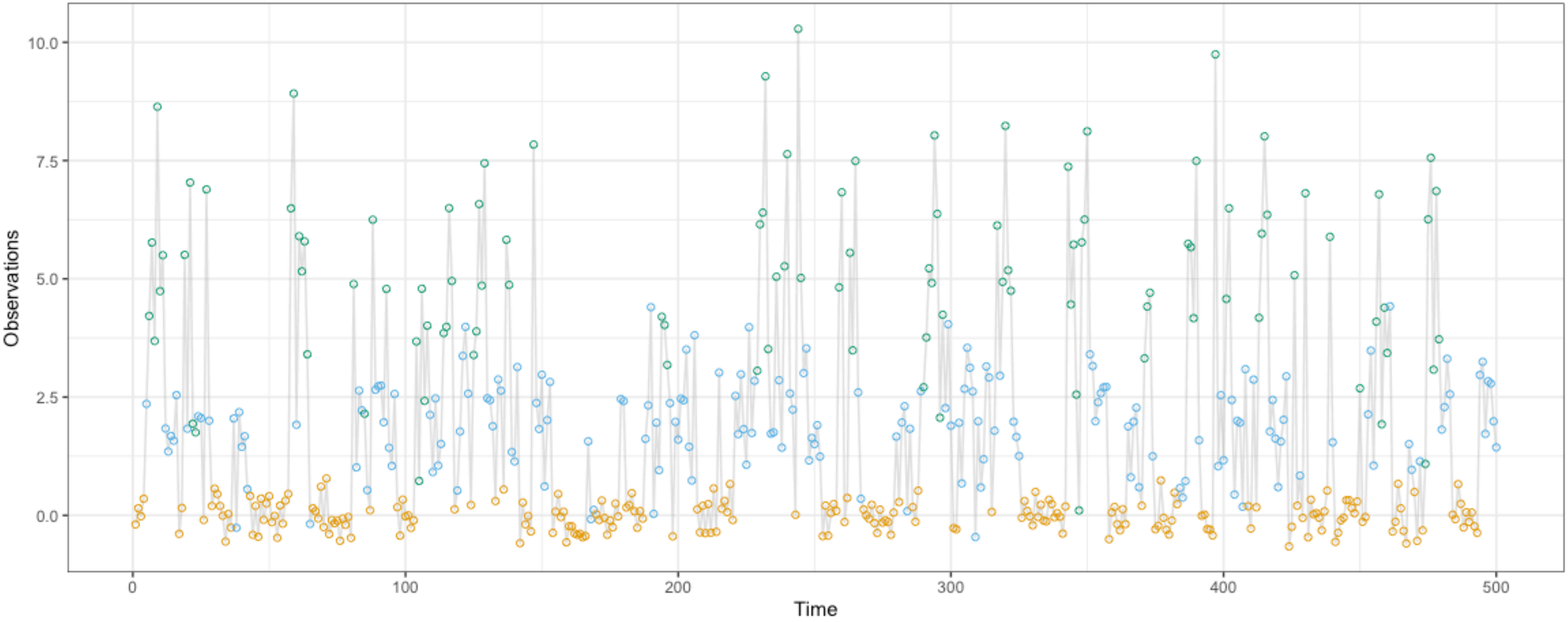
$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.05 & 0.3 & 0.65 \end{pmatrix}$$



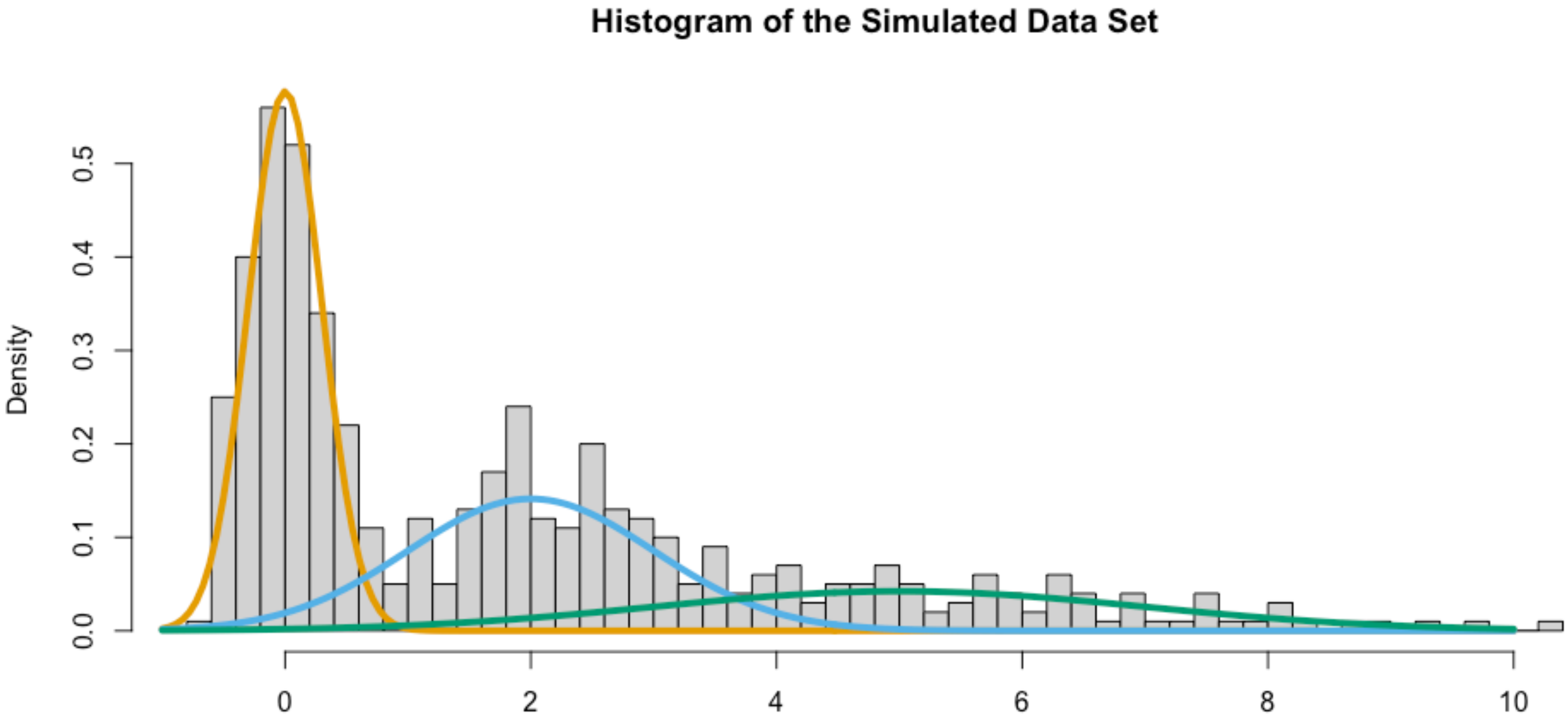
Histogram of the Simulated Data Set



Simulated data: first 500 of 2000 observations



Simulated Data: Marginal Distribution



Simulate Data from an HMM

HMM likelihoods

(Only one can be used in Stan)

- There are generally two likelihood functions used for inference in HMMs:

complete-data likelihood – $\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{Z} \,|\, \boldsymbol{Y})$

likelihood – $\mathscr{L}(\boldsymbol{\theta} \,|\, \boldsymbol{Y}) = \sum_{n=1}^{N} \cdots \sum_{n=1}^{N} f(\boldsymbol{Y}, Z_1 = n, Z_2 = n, \cdots, Z_T = n \,|\, \boldsymbol{\theta})$

- The likelihood can be concisely expressed in matrix form. Let $\boldsymbol{P}(Y_t)$ = diag $\{f(Y_t \,|\, Z_t = 1), \ldots, f(Y_t \,|\, Z_t = N)\}$ and $\mathbf{1}$ be a column vector of 1's of length N, then

$$\mathscr{L}(\boldsymbol{\theta} \,|\, \boldsymbol{Y}) = \boldsymbol{\delta}^T \boldsymbol{P}(Y_1) \boldsymbol{\Gamma} \boldsymbol{P}(Y_2) \cdots \boldsymbol{\Gamma} \boldsymbol{P}(Y_T) \mathbf{1}$$

with evaluation done via the ***forward algorithm.***

Forward algorithm

- The forward algorithm is an approach to efficiently computing the likelihood of an HMM.

- We can construct ‘forward variables’ $\{\boldsymbol{\alpha}_t\}_{t=1}^{T}$ where

$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}^\top \boldsymbol{P}(Y_1)$ and $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\boldsymbol{P}(y_t)$

such that $\displaystyle\sum_{k=1}^{K} \alpha_{T,k} = \mathscr{L}(\boldsymbol{\theta} \,|\, \boldsymbol{Y})$

- In this way, the likelihood function can be evaluated with $O(TK^2)$ operations

General HMM Application Categories

- **Unsupervised** - $\mathscr{L}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ - no states are known/observed

→ most common in ecological applications

- **Semi-supervised** - $\mathscr{L}(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z}_j = z_j)$ - some states are known/observed

- **Supervised** - $\mathscr{L}(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{Z})$ - all states are known/observed

- In all cases, simple modifications to the likelihood function evaluation and forward algorithm can be made to adapt for different applications.

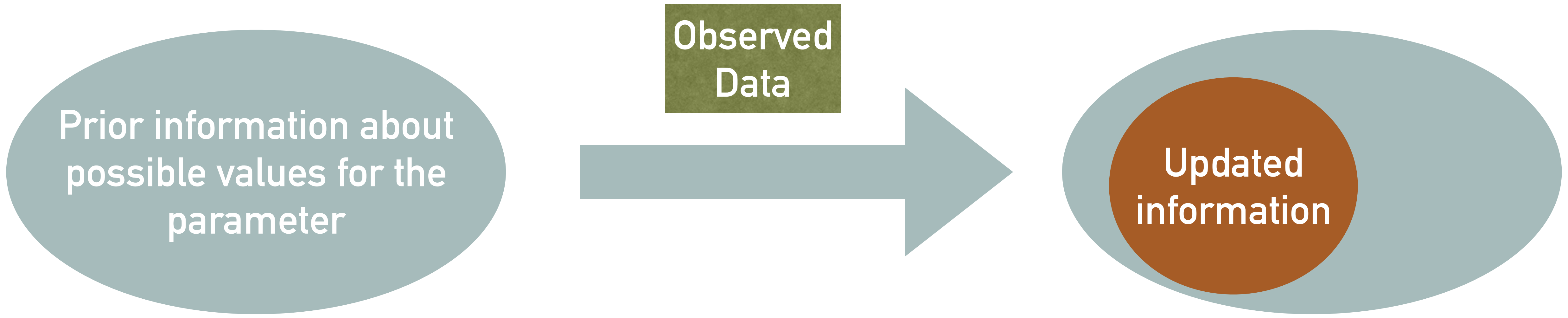
Scientific insights gained from using HMMs

- Can connect the observations to different patterns via state decoding
- We have distributions for the different patterns of interest
- Model state-switching behaviour + amount of time spent in a state before switching
- In a loose sense, we can estimate the ‘number of patterns seen’
- We can incorporate covariates to understand drivers of patterns (we’ll talk about this soon)

Bayesian inference for HMMs

- Set priors for HMMs:

  state-dependent distributions + rows of transition probability matrix



Prior information about possible values for the parameter

Observed Data

Updated information

$\Pr(\beta)$

$\Pr(y \mid \beta)$

$\Pr(\beta \mid y) = \dfrac{\Pr(y, \beta)}{\Pr(y)}$

Prior

Likelihood

Posterior

Calculated using ‘forward algorithm’

State Decoding

- In HMMs + extensions, assigning an observation to a state is known as state decoding

- Two common approaches:

  local state decoding — $\text{argmax}_{n \in \{1,\ldots,N\}} \Pr(Z_t = n \mid \boldsymbol{Y})$
  forward-backward algorithm

  global state decoding — $\text{argmax}_{\boldsymbol{n} \in \{1,\ldots,N\}^T} \Pr(\boldsymbol{Z} = \boldsymbol{n} \mid \boldsymbol{Y})$
  Viterbi algorithm

- Less common — taking posterior draws from the joint distribution of states via forward-filtering backward-sampling

Let’s fit an HMM to data

3-state HMM with Normal state-dependent distributions

- The choices we make when fitting an HMM in practice:

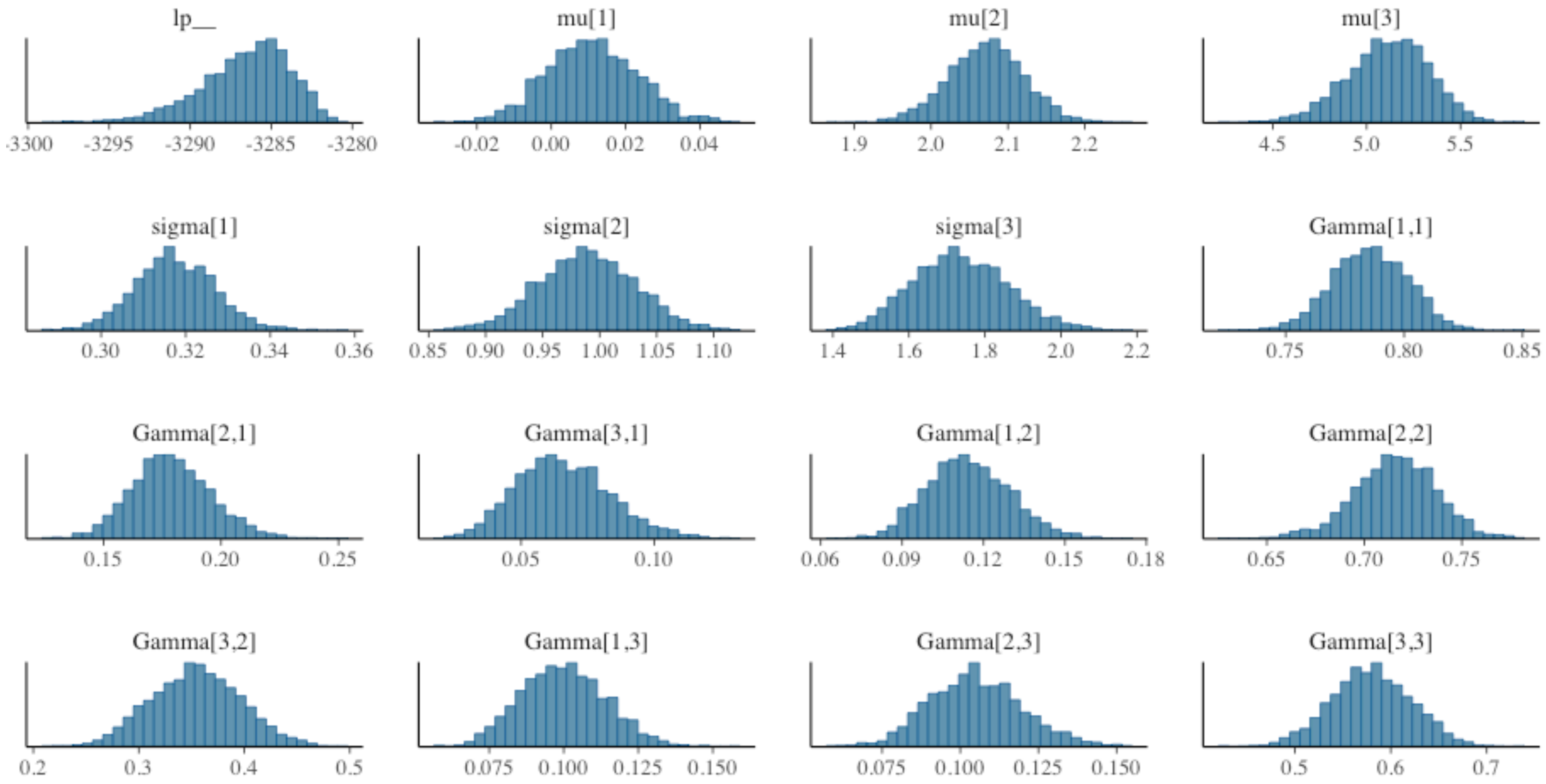
★ the number of states ⚠️

★ the forms of the state-dependent distributions⚠️

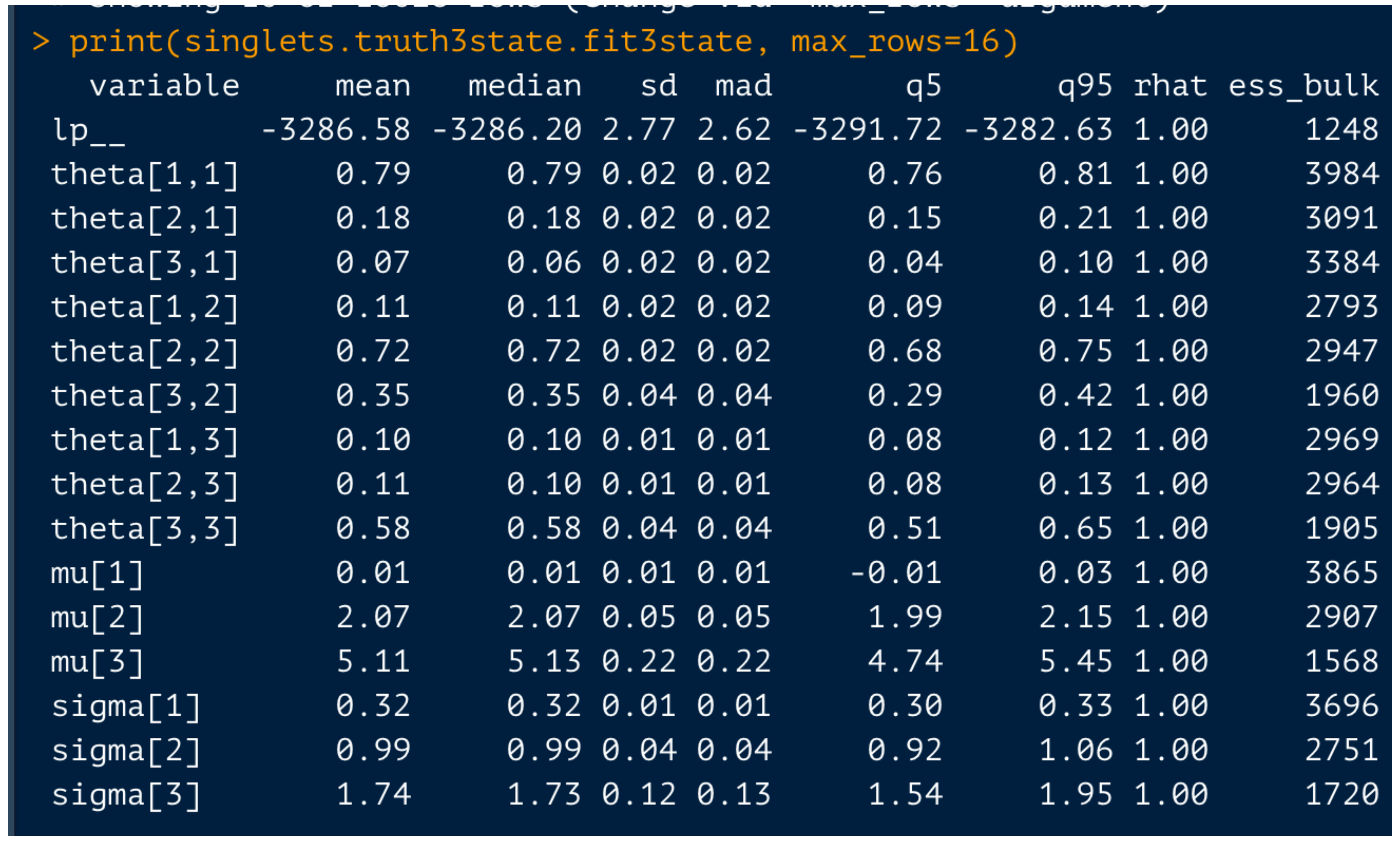
There are other assumptions but for now let’s consider that an HMM is the *true* data generating process.

When all goes well

Sampling 500 observations from true 3-state HMM



Stan output:

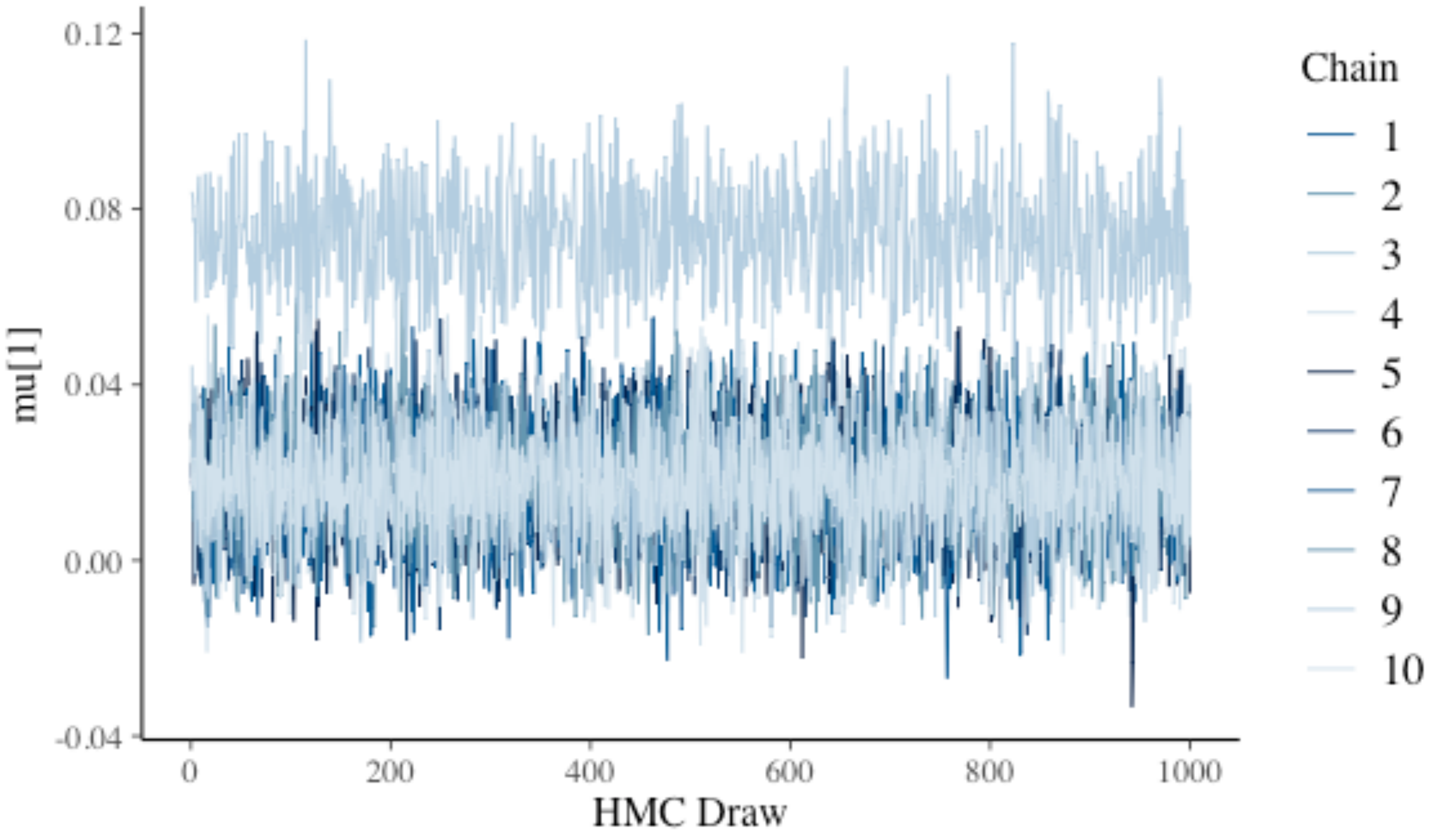


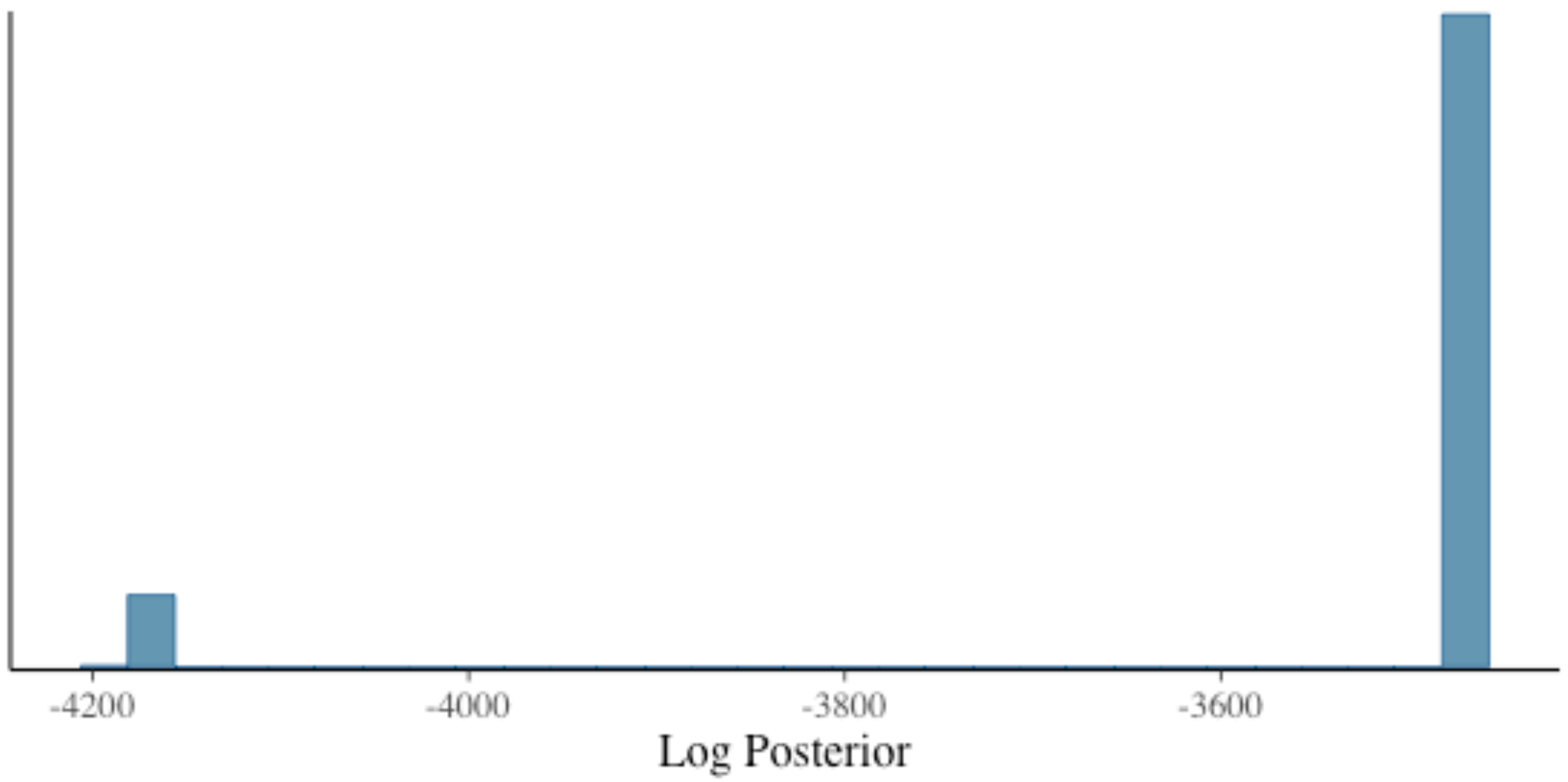
```
> print(singlets.truth3state.fit3state, max_rows=16)
   variable      mean    median   sd  mad       q5      q95 rhat ess_bulk
 lp__        -3286.58 -3286.20 2.77 2.62 -3291.72 -3282.63 1.00     1248
 theta[1,1]      0.79     0.79 0.02 0.02     0.76     0.81 1.00     3984
 theta[2,1]      0.18     0.18 0.02 0.02     0.15     0.21 1.00     3091
 theta[3,1]      0.07     0.06 0.02 0.02     0.04     0.10 1.00     3384
 theta[1,2]      0.11     0.11 0.02 0.02     0.09     0.14 1.00     2793
 theta[2,2]      0.72     0.72 0.02 0.02     0.68     0.75 1.00     2947
 theta[3,2]      0.35     0.35 0.04 0.04     0.29     0.42 1.00     1960
 theta[1,3]      0.10     0.10 0.01 0.01     0.08     0.12 1.00     2969
 theta[2,3]      0.11     0.10 0.01 0.01     0.08     0.13 1.00     2964
 theta[3,3]      0.58     0.58 0.04 0.04     0.51     0.65 1.00     1905
 mu[1]           0.01     0.01 0.01 0.01    -0.01     0.03 1.00     3865
 mu[2]           2.07     2.07 0.05 0.05     1.99     2.15 1.00     2907
 mu[3]           5.11     5.13 0.22 0.22     4.74     5.45 1.00     1568
 sigma[1]        0.32     0.32 0.01 0.01     0.30     0.33 1.00     3696
 sigma[2]        0.99     0.99 0.04 0.04     0.92     1.06 1.00     2751
 sigma[3]        1.74     1.73 0.12 0.13     1.54     1.95 1.00     1720
```

Challenges

People think I’m an expert in HMMs, I say

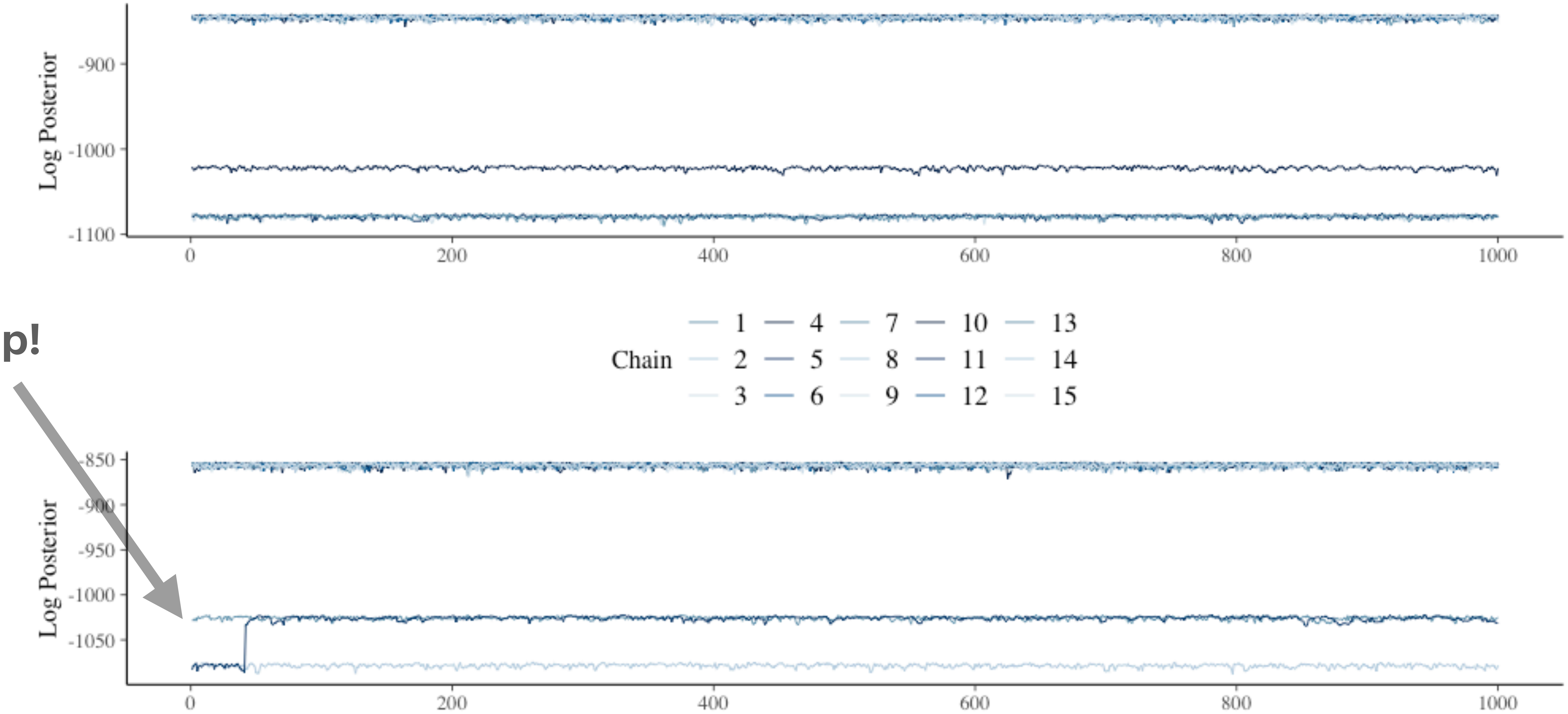
I’m an expert in failing to fit hidden Markov models





Fitting misspecified HMMs

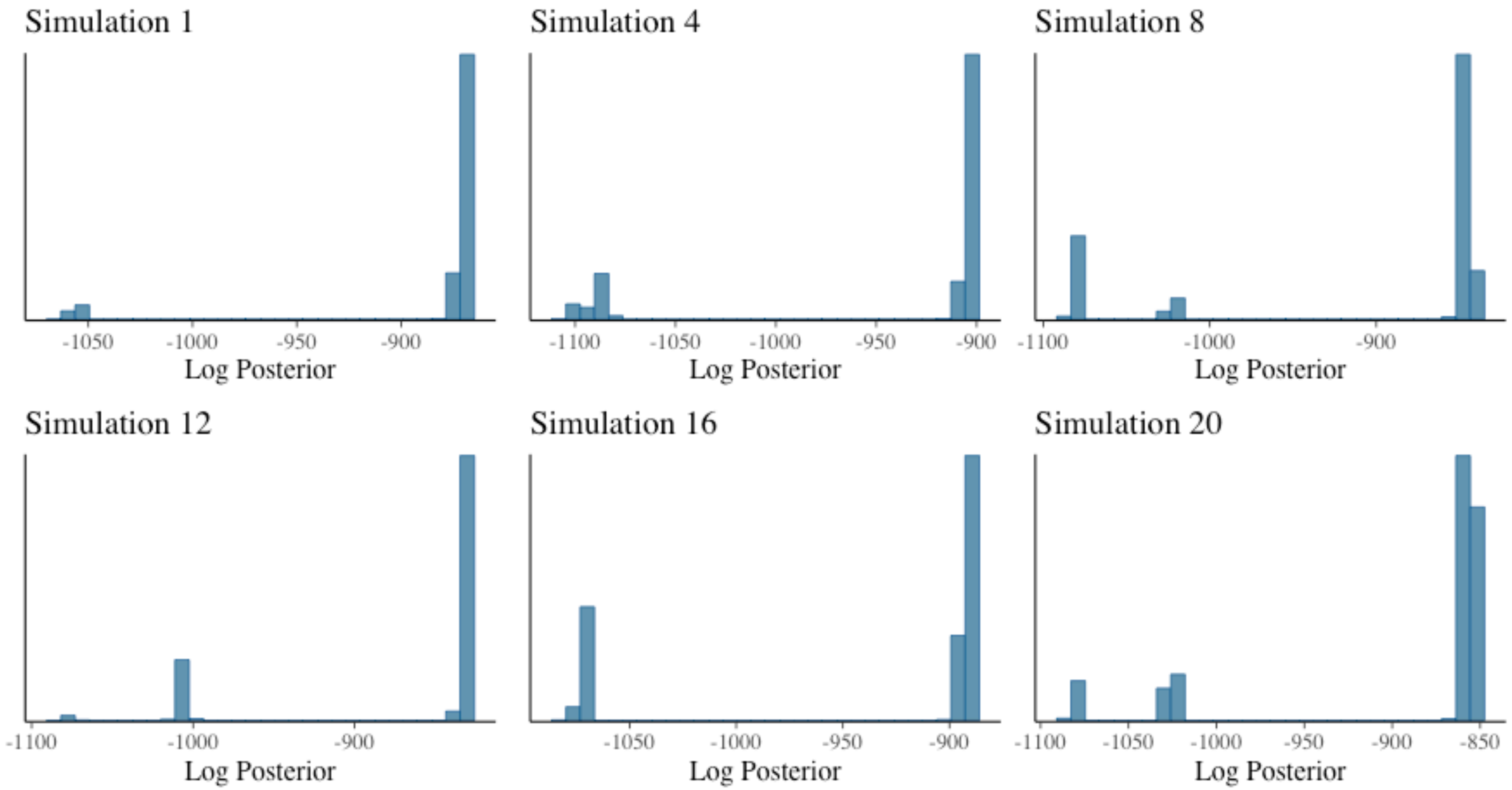
2-state HMM when truth is a 3-state HMM



Jump!

Fitting misspecified HMMs

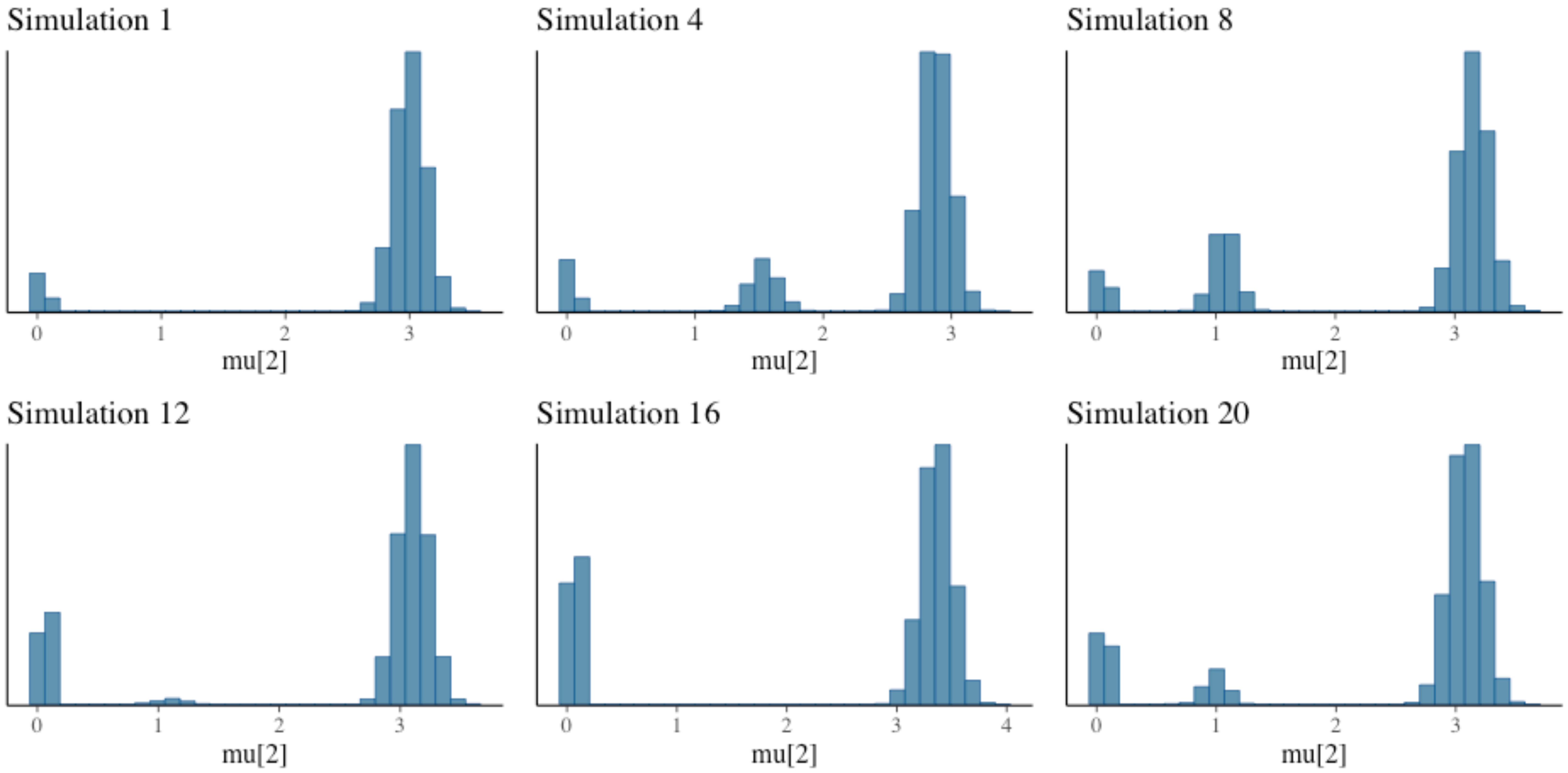
Log-posterior: $p(\boldsymbol{\theta} \mid \boldsymbol{y})$



- **Multi-modal log-posterior**
- **Each mode corresponds to different sets of parameter values that could have generated the data**
- **Not clear how much probability is associated with each mode**

Fitting misspecified HMMs

Posterior draws for μ_2 across 6 simulations:



- ◆ **These are NOT marginal distributions of μ_2**
- ◆ **These ARE values of μ_2 that are consistent with the data (and prior spec.)**

Building intuition behind misspecification

- Fitting (increasingly complex) HMM to complex data often leads to multimodal posteriors along the way

- Simulating data and fitting under misspecification helps build intuition of how to do model building when things are seemingly ‘going wrong’

- For me, I really wanted to understand — **how do different misspeficiations manifest themselves in terms of the behaviour of the likelihood/posterior distribution?** Stan is really helpful here.

Label-switching affects Bayesian inference

Fit HMMs using Bayesian inference in Stan

Extras

- Missing values

- Covariates

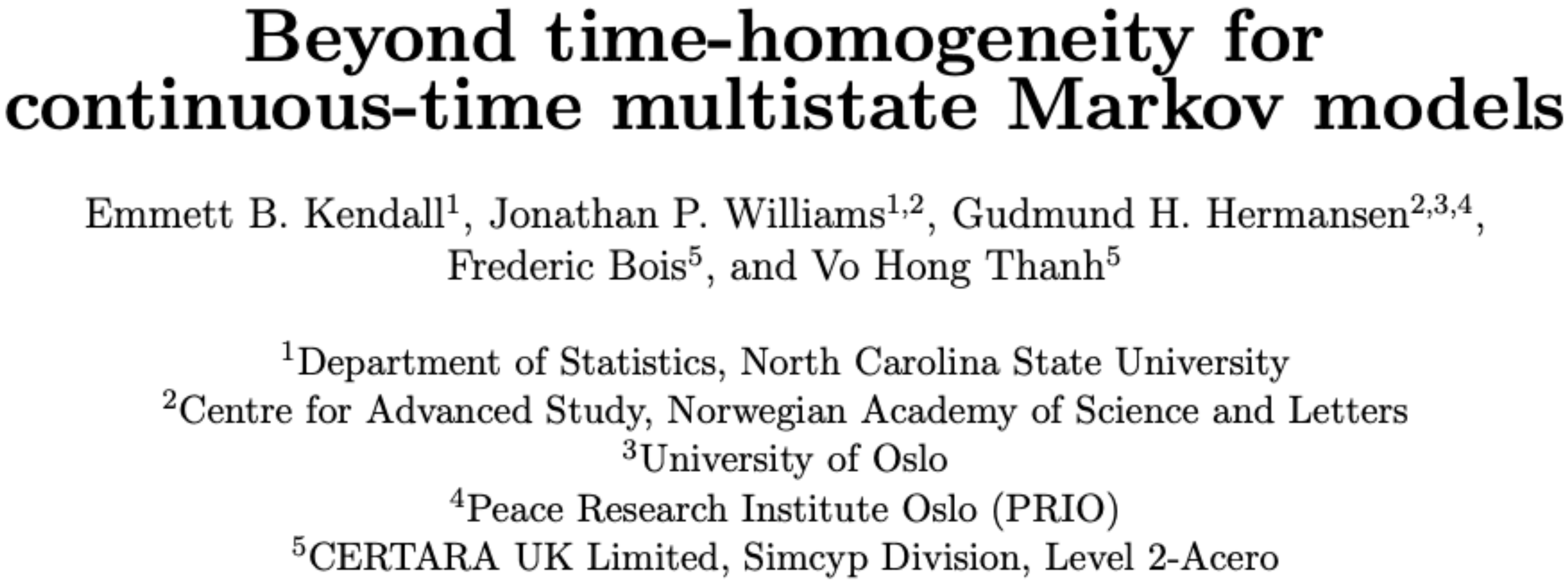
- Multivariate observations

Extras

- Mixed HMMs + random effects

- Continuous-time HMMs + more

Continuous-time HMMs



Beyond time-homogeneity for continuous-time multistate Markov models

Emmett B. Kendall[1], Jonathan P. Williams[1,2], Gudmund H. Hermansen[2,3,4],
Frederic Bois[5], and Vo Hong Thanh[5]

[1]Department of Statistics, North Carolina State University
[2]Centre for Advanced Study, Norwegian Academy of Science and Letters
[3]University of Oslo
[4]Peace Research Institute Oslo (PRIO)
[5]CERTARA UK Limited, Simcyp Division, Level 2-Acero