R User Group – Singapore (RUGS)

Hidden **M**arkov **M**odels & Applications Using **R**

Truc Viet 'Joe' Le

What is a **Model**?

- A mathematical formalization of the ***relationships*** between variables:
  - Independent variables (*X*) and dependent variables (*Y*)
  - $Y = f(X)$, e.g., regression models

- A model is a *statistical model* when the variables are ***probabilistically/stochastically related***
  - *Y* and *X* are related through a probability distribution function *f*
  - $Y = \Pr(X=x)$, e.g., Gaussian (normal) distribution

- **Hidden Markov Models** (HMMs) are **statistical models**

Descriptive & **Generative** Models

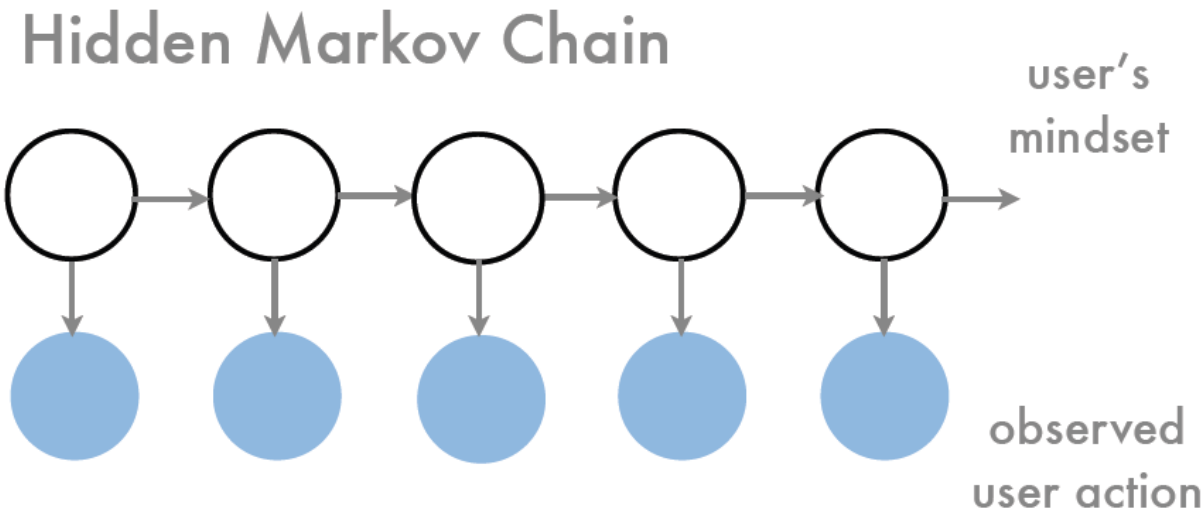
- *Y*: **observed**, measurable variables (e.g., symptoms); *X*: underlying, latent (hidden), **unobserved** variable (e.g., disease)

- A **descriptive model** is the ***conditional probability distribution*** **Pr(*X*|*Y*)** → dependence of the unobserved variable on the observed
  - E.g., logistic/linear regressions

- A **generative model** randomly **generates observable data** given the estimated parameters
  - Specifies the ***joint probability distribution*** between the observed and unobserved variables **Pr(*X*, *Y*)**
  - Used to simulate/generate values of any variables in the model → forecasting, testing hypotheses
  - E.g., **HMMs**, finite mixture models (special case of HMM)

Hidden **M**arkov **M**odels (**HMM**s)

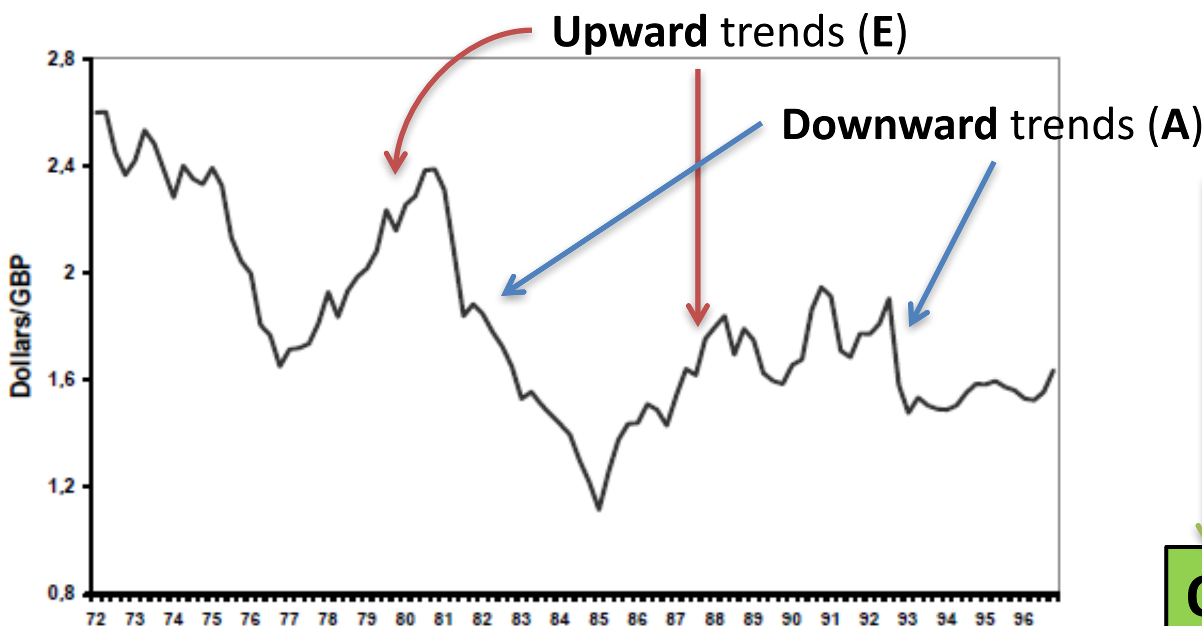
- Describe the **relationship** between **two** stochastic processes: the ***observed*** process & the ***unobserved*** (hidden/latent) underlying process

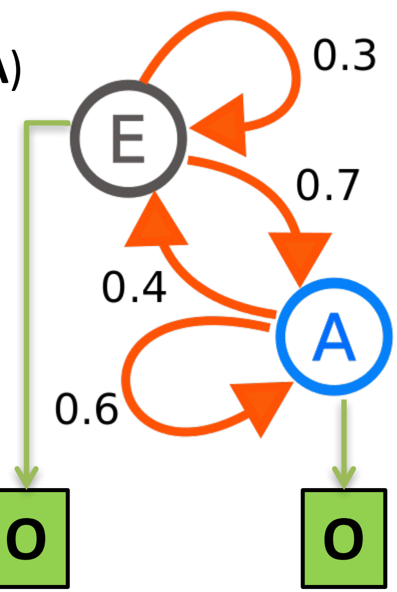


  – **Hidden process** follows a **Markov chain**
    - Describe the **hidden states** by random variable **X**
  – **Observations** are typically a sequence (e.g., time series) and are ***conditionally independent*** given the sequence of hidden states
    - Describe the **observations** by random variable **Y**

Example: **Regime Switching** Model

- Modeling the **hidden "regimes"** of financial markets – switches between periods of **high volatility** & **low volatility**, **bearish** & **bullish**, etc.

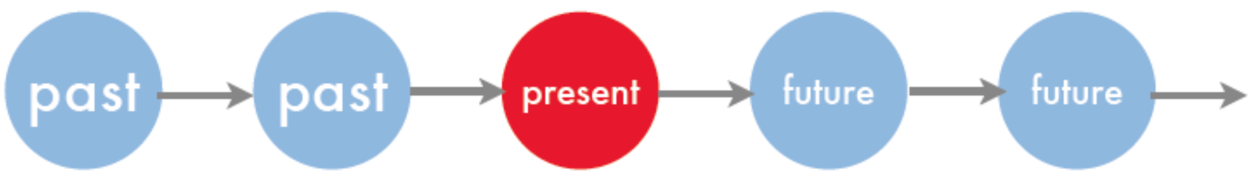




- Recently, **Markov Switching Multifractal** (MSM) asset pricing model

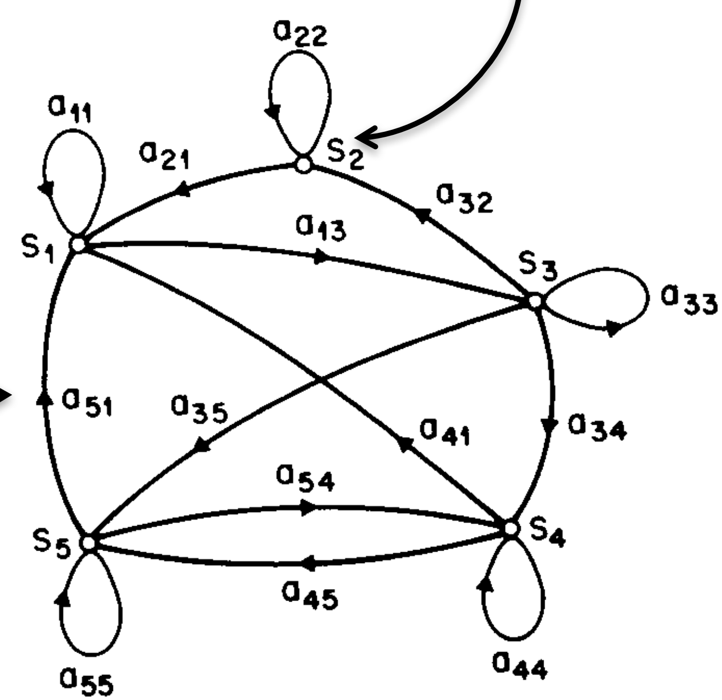
Markovian Property of the **Hidden States**

Markov Chain



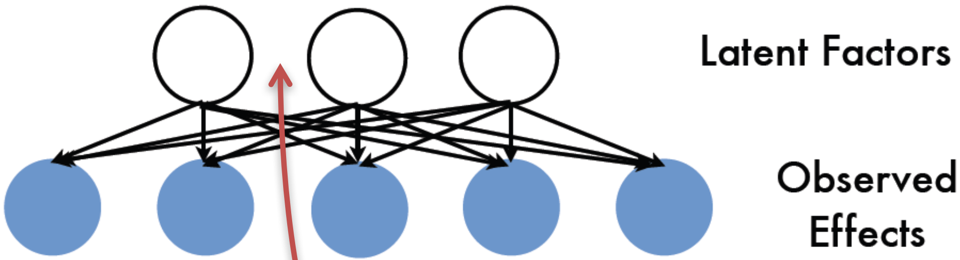
The **present state** depends on the **immediate past state** and nothing else!

State

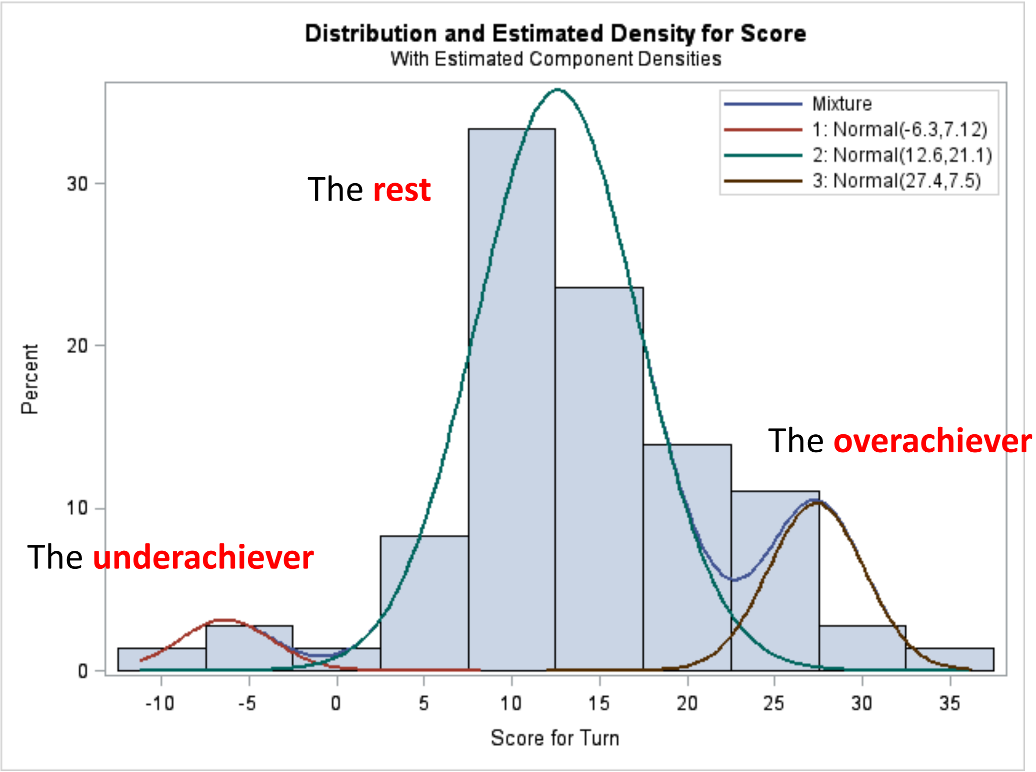


Transition probability

Example: **Factor** Analysis

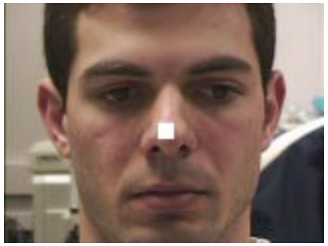


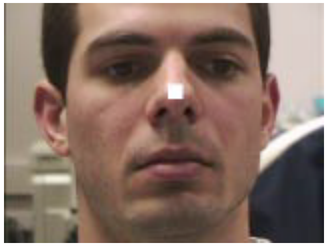
Finite mixture of **3 Gaussians**. Notice there are **no transitions** between the hidden states (aka ***latent factors***)

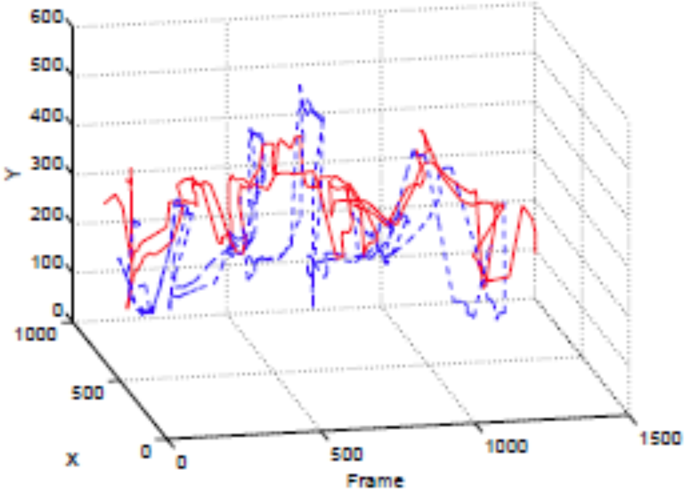


Example: **Facial** Recognition

- Learning of **moving facial images** over time
- Each **facial feature** (e.g., nose, eyes, etc.) is a **hidden state** of the HMM
- **Observed variables** are the (x, y) coords of the features on the images







Alon, J., Sclaroff, S., Kollios, G., & Pavlovic, V. (2003, June). Discovering clusters in motion time-series data. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* (Vol. 1, pp. I-375). IEEE.

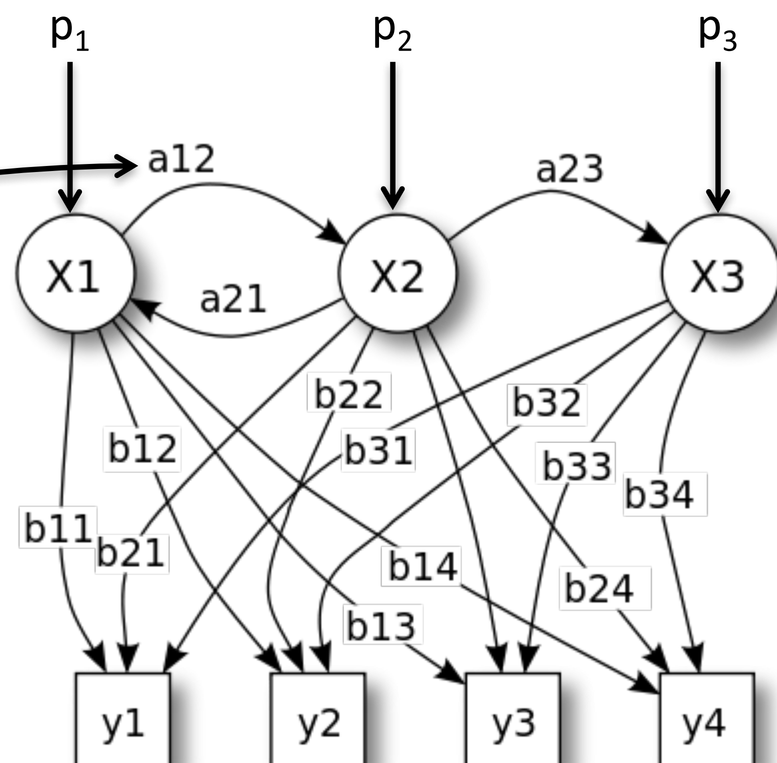
Parameters of an HMM

(1) Initial probabilities

$\mathbf{p_i}$ = Pr(S_1=p_i) for i = 1, … , 3

(2) Transition probabilities/ Transition matrix of the *underlying* Markov chain

e.g., $\mathbf{a_{12}}$ = Pr(S_t=X2|S_{t-1}=X1)



(3) Emission probabilities

$\mathbf{b_{34}}$ = Pr(Y_t=y4|X_t=X3)

Observations (**Y**)

Estimation of an HMM's **Parameters** (Model **Learning**)

- HMMs are typically learned using the **Expectation-Maximization** (EM) algorithm – *not* discussed here

- The **parameters** of an HMM are:
  - Set of hidden states **S** = $\{S_1, S_2, \ldots, S_N\}$ for an ***N***-state model
  - Vector of **initial** (state) probabilities **p** = $(p_1, p_2, \ldots, p_N)$
  - **Transition** probability matrix (NxN) **A** = $\{a_{ij}\}$, where
    - $a_{ij} = \Pr(X_t = S_j \mid X_{t-1} = S_i)$
  - **Emission** (response) probability distribution/density function ***f*** = $\Pr(Y_t = y \mid X_t = x)$
    - Could be discrete/continuous/categorical function
    - Could also be **multivariate**

Model **Selection**, **Validation** & **Inference** in HMMs

- Since HMM is a *generative model* → Validate it by testing **how well it reflects the reality**

- Generate random observations using the learned HMM (from the partial data) and compare those with the holdout (test) set (e.g., **cross validation**)

- Many **statistical tests** exist for this purpose (depending on the emission density function)

- How to select the **optimal # hidden states *N***? Typically using **AIC** or **BIC** (**B**ayesian **I**nformation **C**riterion)

- **Inference** (*not* discussed)
  - Joint probability of an observed sequence
  - Joint probability of a sequence of hidden states given the observations → **Viterbi algorithm**

Applying HMMs Using

WITH THE DEPMIXS4 PACKAGE

Real-world Application

MODELING VISITORS' TRAJECTORIES IN SENTOSA USING HMM'S

Background on Sentosa **Play (Day) Pass**

- Is an **attraction bundling scheme** marketed by Sentosa
- Play **one price** and redeem up to **17** participating attractions in Sentosa – **“Up to 70% Savings!”**
- **Price variability** depends on:
  - Adult/Child
  - Weekday/Weekend
  - 1Day/2Day Pass
- Pass valid from **9am** to **6pm** (10-hour period) daily (for one-day use only)

17 Participating Attractions

Spatio-temporal Trajectories

- Collected for **7 months** in 2012–2013
- Involves over **30K visitors** in total, each produces a trajectory

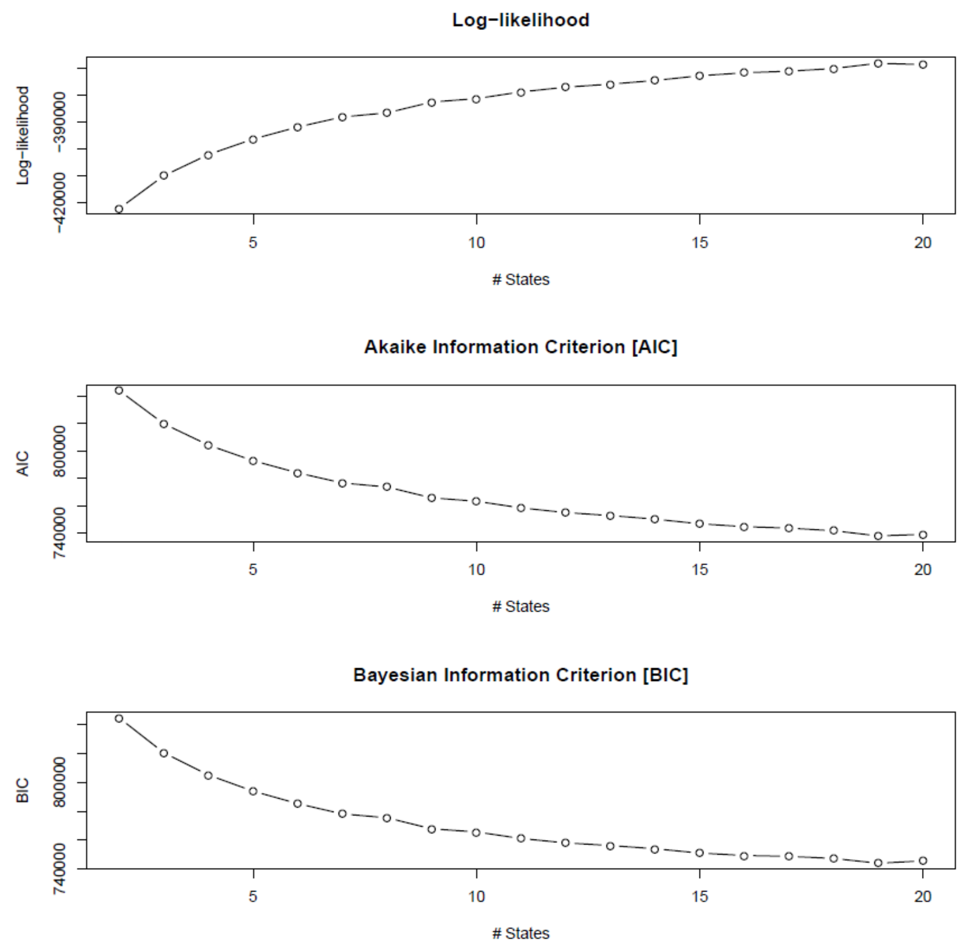
- Each trajectory is a **temporally-ordered sequence** of attraction visits with length varying from 1 to 17 (approx. normal w/ mean around 8-9)

- Each trajectory is a ***bivariate*** **spatio-temporal** sequence
  - Sequence of **attractions (events)**: *discrete* r.v.
  - Sequence of **times-to-event**: *continuous* r.v. (# mins from 9am until the visit)

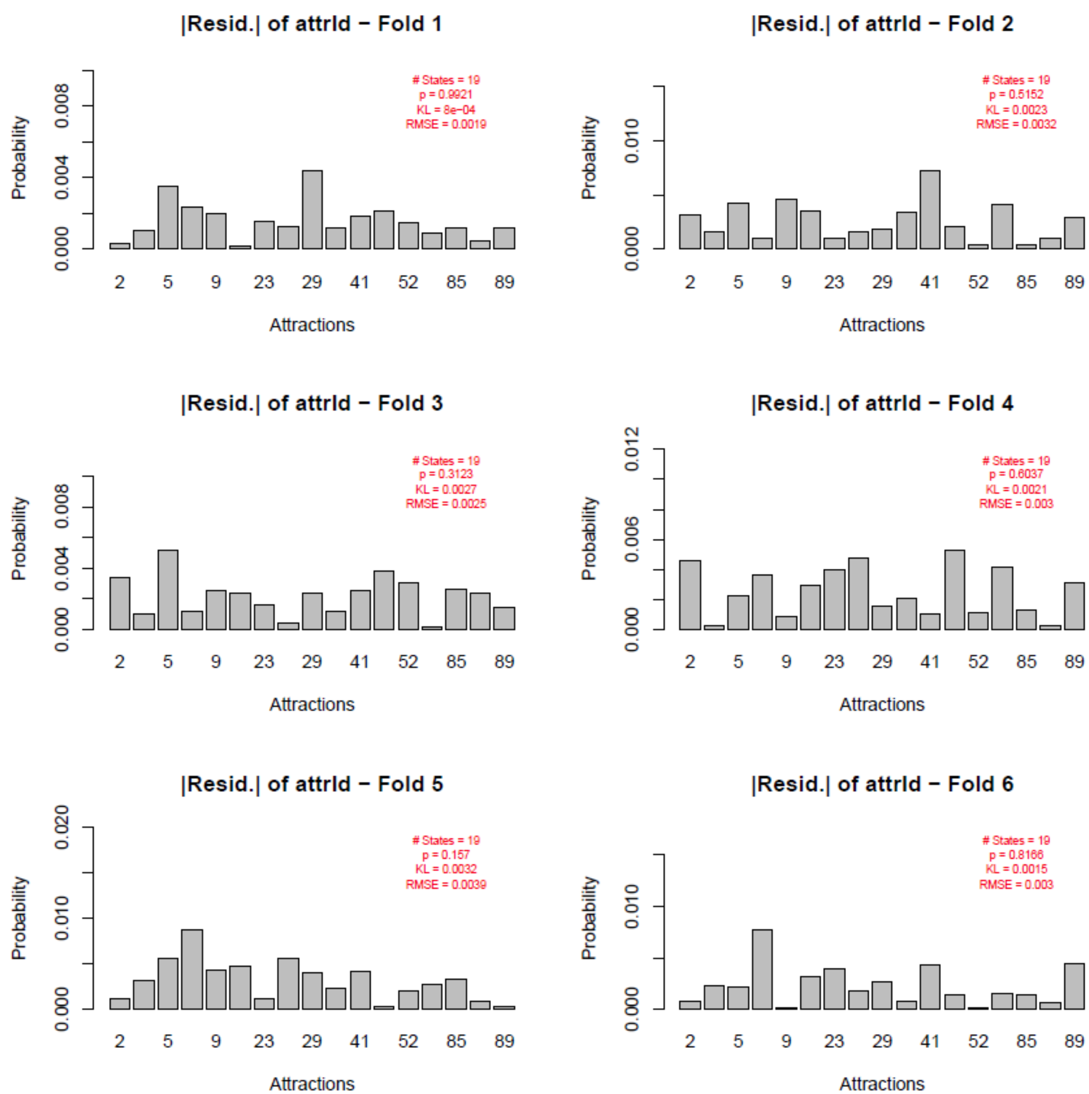
- Reflects the diverse **behaviors** (*observed*) and the **preferences/tastes** (*unobserved*) of visitors that are **confined** by the pass's T&C's and the physical clustering of the attractions + human activities (e.g., lunchtime)

Model **Specification** & **Fitting**

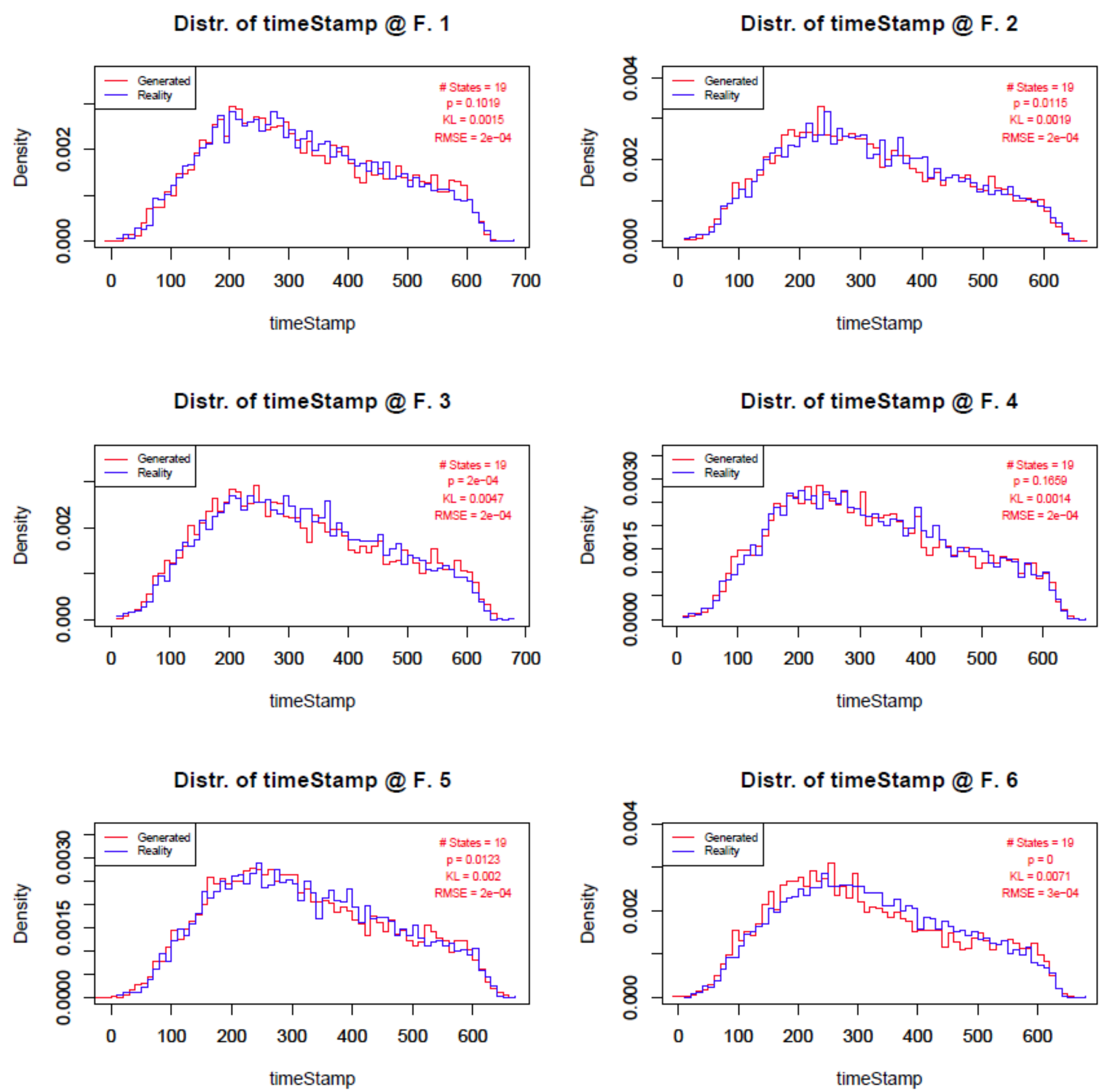
- Specify an HMM using **`depmixS4`** with:
  - **Bivariate response**: ***multinomial*** (discrete attractions/events) + ***Gaussian*** (continuous time-to-event)
  - Incremental fitting to determine the **optimal # states** using **BIC**
  - Important to **specify the independent sequences** for each of these individuals (**30K** reduced to **14K** through **groupings**)
  - Takes **very long time** due to **HUGE** dataset!



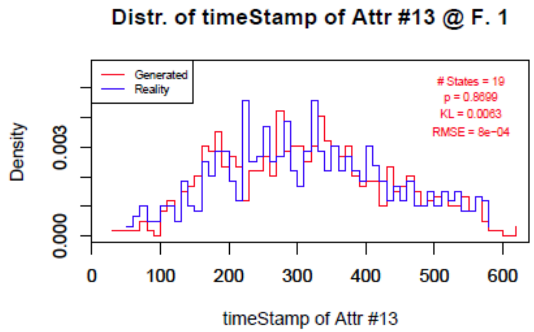
Model **Validation** w/ **6-fold** Cross Validation (1)

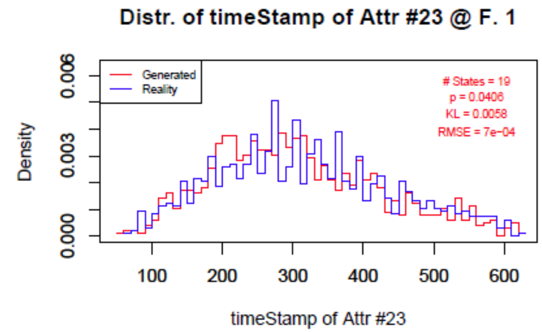


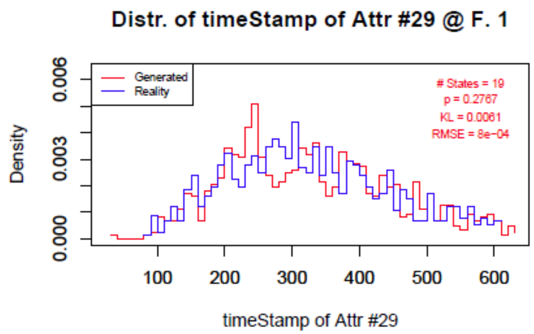
Model **Validation** w/ **6-fold** Cross Validation (2)

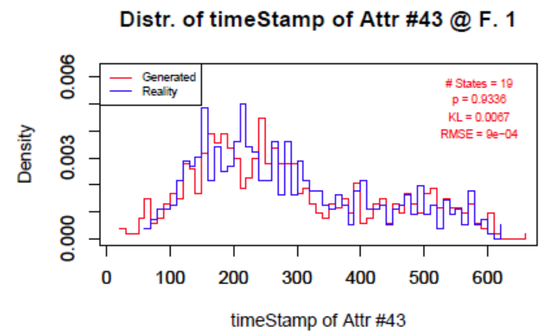


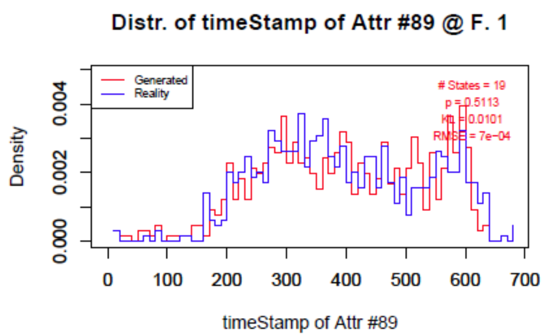
Distr. of **Time-to-Event** for *Each Attraction*

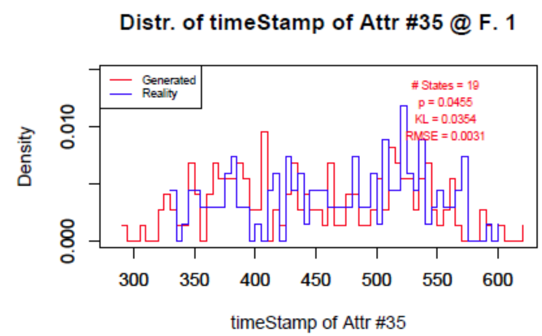




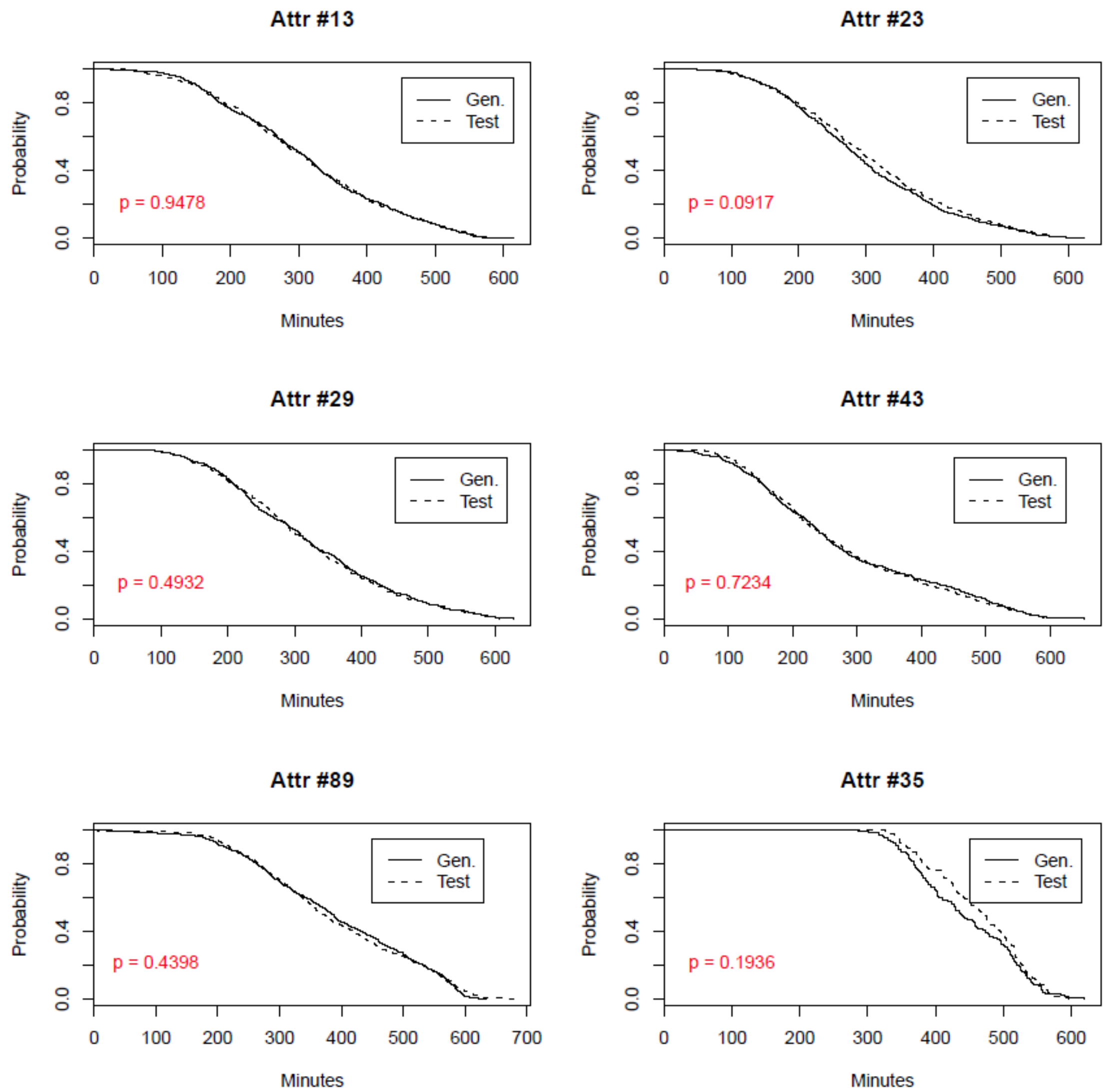




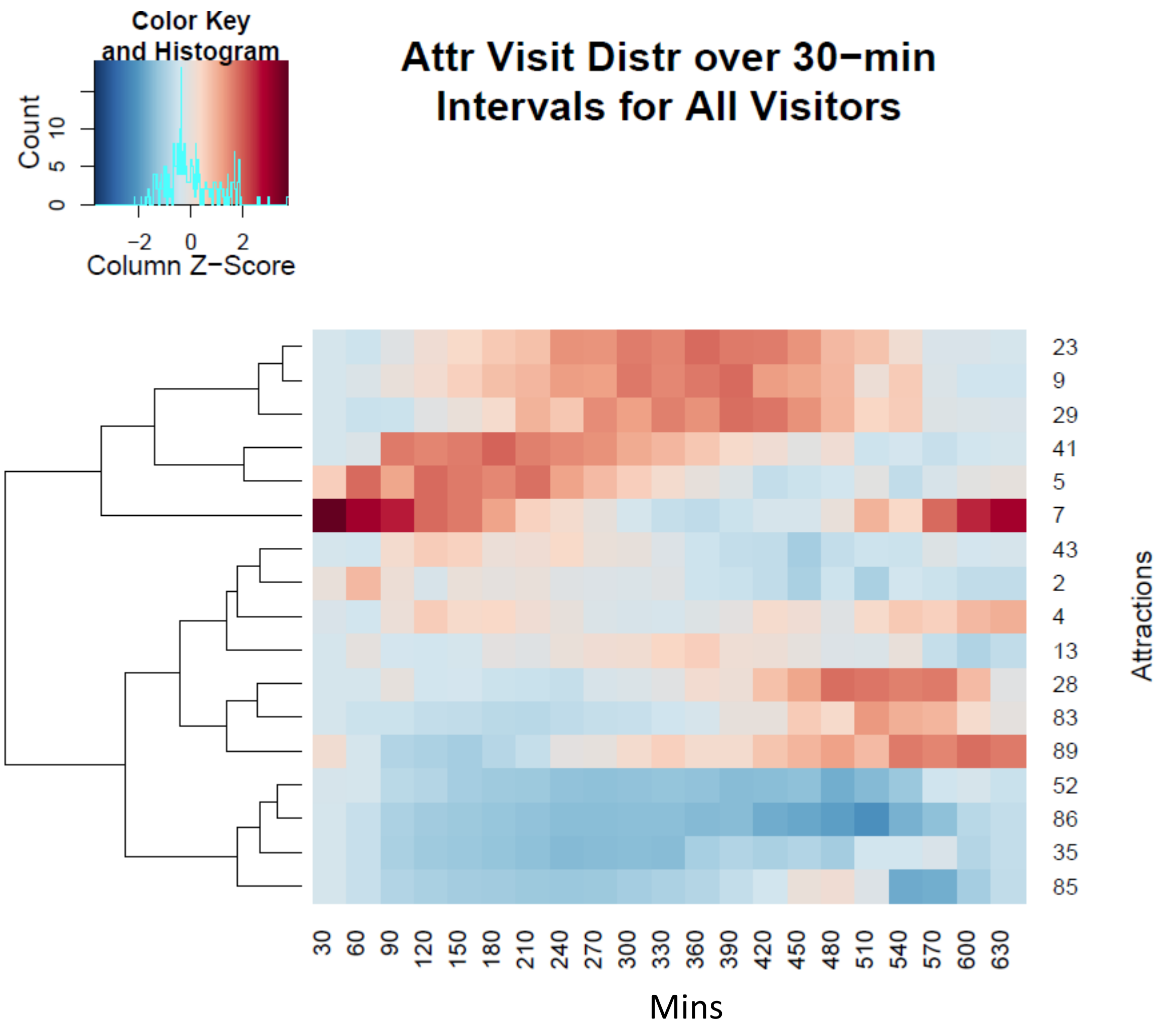




Logrank Tests of Distributions of **Time-to-Event** Using **Survival Analysis**



Physical Clustering of the Attractions



Learned Clusters through the HMM's **Emission Parameters**