Bayesian formulation and implementation for a non-stationary semi-Markov model with covariates (Work in progress)

Sébastien Coube [1]

INRAE, Toulouse University

1. Special mention for Nathalie Peyrard whose critical yet benevolent feedback is invaluable

1. Introduction + motivating examples

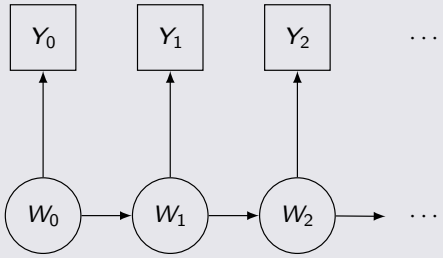
2. Model formulation

3. Inference

Semi-Markov ?

Hidden Markov Model : HMM

- Observations $Y_0, Y_1, Y_2, \ldots$, indexed in discrete time
- **Discrete and hidden latent states** $W_0, W_1, W_2, \ldots$
- Probability of jumping from one latent state to another at each time period

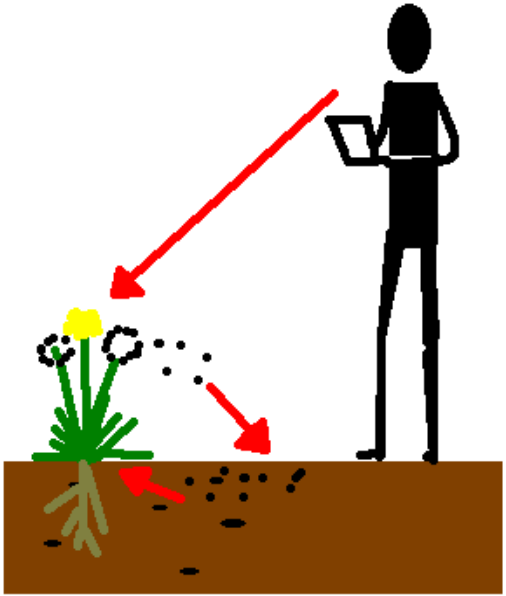


Hidden **Semi** *Markov Model* : HSMM

- In an HMM, we have **sojourn times in a state** that follow a **geometric distribution**
- Not always realistic, e.g., when you have the flu, you often do not recover in the first few days
- The time spent in a state can follow another distribution : semi-Markov

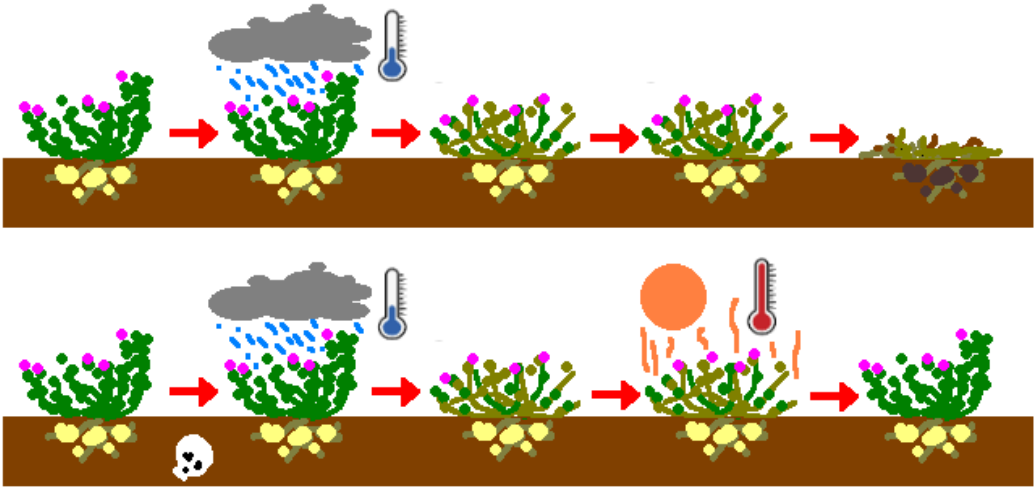
Weeds, from Hanna Bacave's work [1]

- Weeds are observed
- Seeds in the ground are hidden
- **Retroaction** of weeds to seeds



Mildew on my mother's potato patches

- Damp and cold weather induces mildew
- Treatment has a preventive action
- Heatwaves cure mildew
- After some time, mildew kills potato plants and bligts the tubers



What now ?

What's in common ?

- We can model the situation with a H(S)MM. That's a fiction of course, but still, not a delirious one
  - Presence (or not) of seeds in the ground, observed plants
  - Potato plant is either healthy, sick, or dead, mildew is directly observed
- There are **explanatory variables** that act upon the latent state
  - The production of seeds (who also happens to be the observed variable)
  - The weather, the treatment
- Those variables **change at each period of time**

What do we do from here ?

- Do what we usually do with H(S)MMs with a little more spice
  - Classifying the latent state
  - Clustering the observations (not exactly the same thing as classifying)
  - Predicting the outcome variable, the latent state, in future times
- Evaluate the impact of those explanatory variables upon the changes latent state

1. Introduction + motivating examples

2. Model formulation

3. Inference

Linking Covariates to Latent State Changes

Reformulating the Model to Incorporate Covariate Effects

- Problem : how can we account for the **effect of variables during the stay in a given state** ?
- Solution : reformulate as a Markov model over the pair "state" $\times$ "time already spent in the state"
- Idea stolen from Hanna Bacave [1] : assign a probability to the next latent state...
  - for each current latent state, ...
  - for each duration spent in that current latent state, ...
  - and for each possible value of the explanatory variables
- (The time spent is automatically updated)

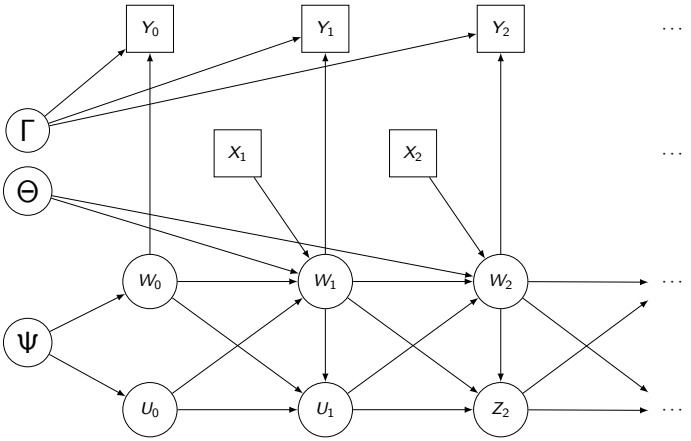
Important idea : the **global behavior emerges from a collection of local behaviors**

Generalizing the Formulation

- Problem : Hanna works with a binary variable
  - The weed produces seeds...
  - The weed does not produce seeds...

  ... I want to be able to work with other formats of variables as well
- Solution : to create an (interpretable) link with **explanatory variables in a general format**, use **categorical regression**
  - Softmax for now
  - Why not multi-level, if needed in the future ?
  - Why not hybrid methods (Peyhardi [2]), if needed in the future ?

A Picture Is Worth a Thousand Words



- Squares ☐ represent observations
  - $X_t \rightarrow$ explanatory variables at time t
  - $Y_t \rightarrow$ emissions at time t
- Circles ◯ represent parameters to be inferred by the model
- Greek letters are **high-level parameters**
  - $\Theta \rightarrow$ transition parameters
  - $\Gamma \rightarrow$ emission parameters
  - $\Psi \rightarrow$ initialization parameters
- Latin letters are **low-level parameters**
  - $W_t \rightarrow$ latent state at time t
  - $U_t \rightarrow$ time already spent in the state at time t (if just entered, $U_t = 1$)

Time Basis Functions for Semi-Markov Model

A quote from Håvard Rue, around June 2023

Sébastien, you are trying to model every single observation. This is doing statistics like in the 1970's. In modern statistics, first you make a structure, then the observations land on top of the structure.

Major Potential Problems of "Bunch Of Softmax"

- Overfitting
- Algorithmic issues
- Interpretability

Idea : impose a **low-rank structure** on the coefficients of multinomial regressions to lighten the model

Introducing Structure and Parsimony

- First step : build a **directed graph** $\rightarrow$ for a given state, indicate where we are allowed to go
- Second step : **zoom in on each node** $\rightarrow$ choose **time basis functions** to describe the evolution of transition probabilities as a function of time spent in the state
- Third step : **zoom in on each arrow** $\rightarrow$ combine the variables and basis functions for each legal output state. We are not required to make all combinations

An Example : Seasonal Illness, Day by Day. An SIRS model on one individual

Model Elements

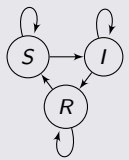
Latent states W_t

- Susceptible state.
- Infected state.
- Recovered, immunized state.

Observations :

- Explanatory variable X_t : average temperature
- Emitted variable Y_t : individual's temperature

Step 1 : Build the Transition Graph



Step 2 : Zoom in on a Node

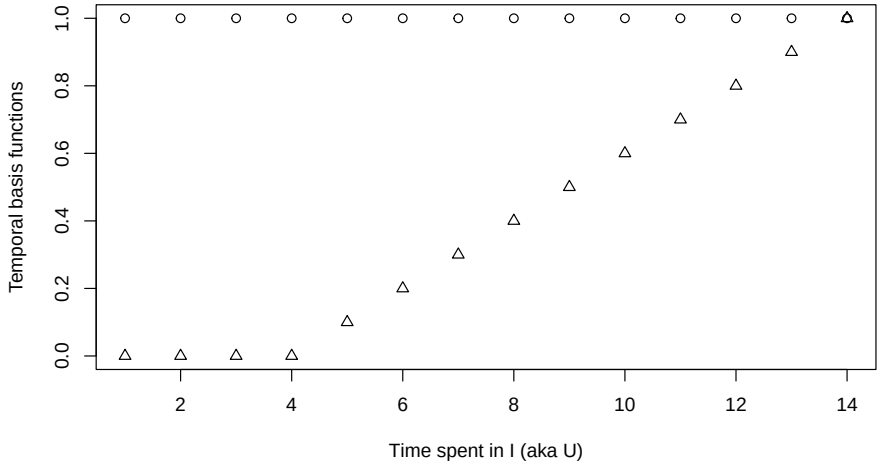
- S : Markov model. Time basis function : (1)
- I : Semi-Markov model, with constant transition probabilities after $U = 14$ days. Time basis functions :
  - $(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$
  - $(0, 0, 0, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)$
- R : Semi-Markov model, with constant transition probabilities after $U = 6$ months. Time basis functions :
  - $(1, 1 \leftarrow 6 \text{ months} \rightarrow 1, 1)$
  - $(0, 0 \leftarrow 80\text{d} \rightarrow 0, 0, 0.01, 0.02 \leftarrow 100\text{d} \rightarrow 0.99, 1)$

Step 3 : Zoom in on an Edge

- $S \rightarrow I$: (1)$\times$ (Intercept & temperature)
- $I \rightarrow R$: 2 functions $\times$ Intercept
- $R \rightarrow S$: 2 functions $\times$ Intercept

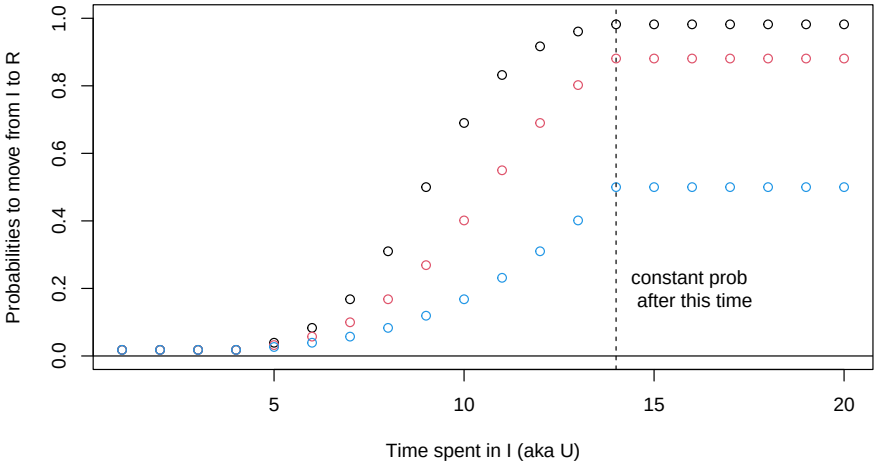
$I \rightarrow R$: Time Basis Functions

They are the "alphabet" who describes how the transition probabilities evolve along the time spent in the state "I"



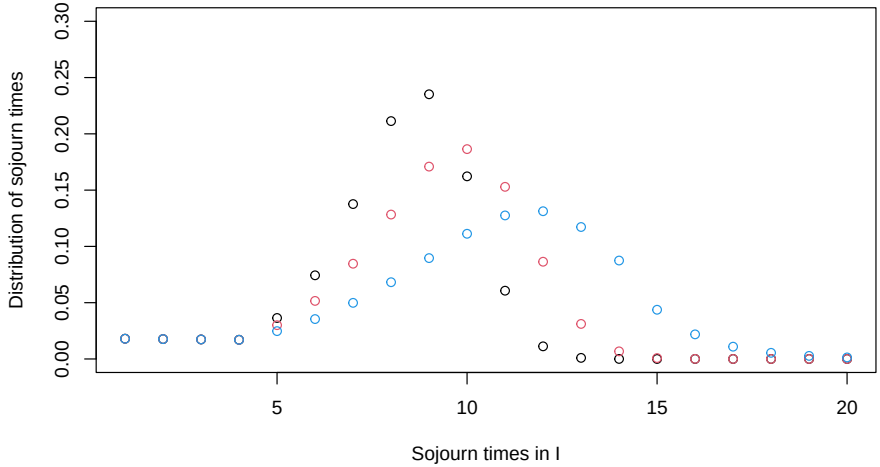
$I \to R$: Transition Probabilities, Different Scenarios

Each color corresponds to certain values of the parameters

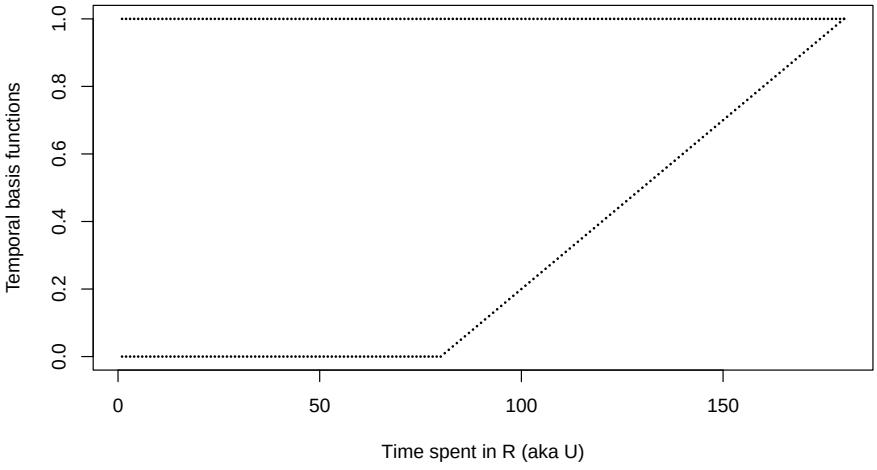


$I \to R$: Sojourn Time Distributions, Different Scenarios

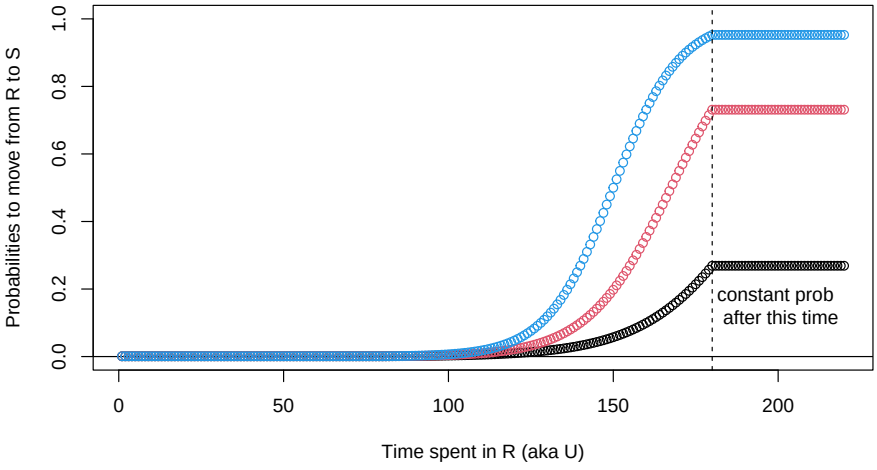
A global behavior for the sojourn times arises from the local behavior of the transition probabilities



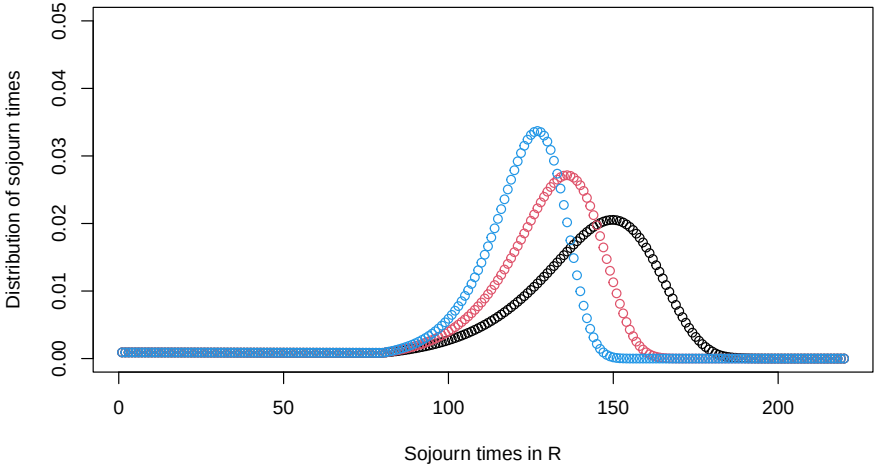
$R \rightarrow S$: Time Basis Functions



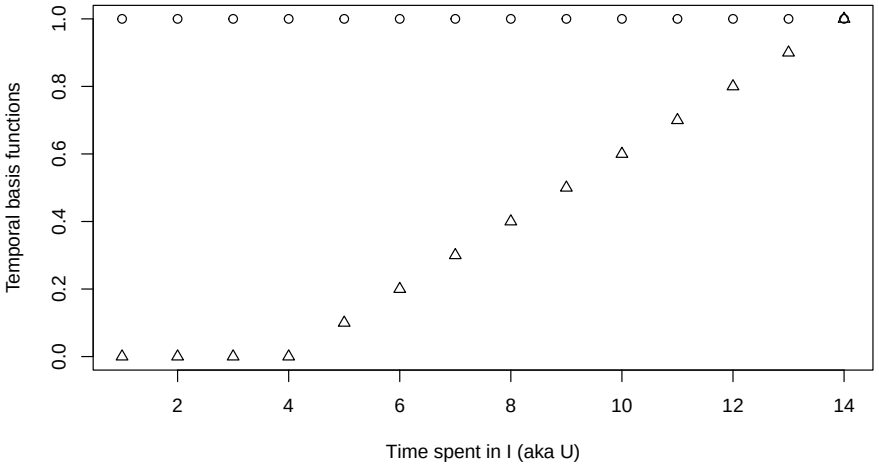
$R \to S$: Transition Probabilities, Different Scenarios



$R \to S$: Sojourn Time Distributions, Different Scenarios

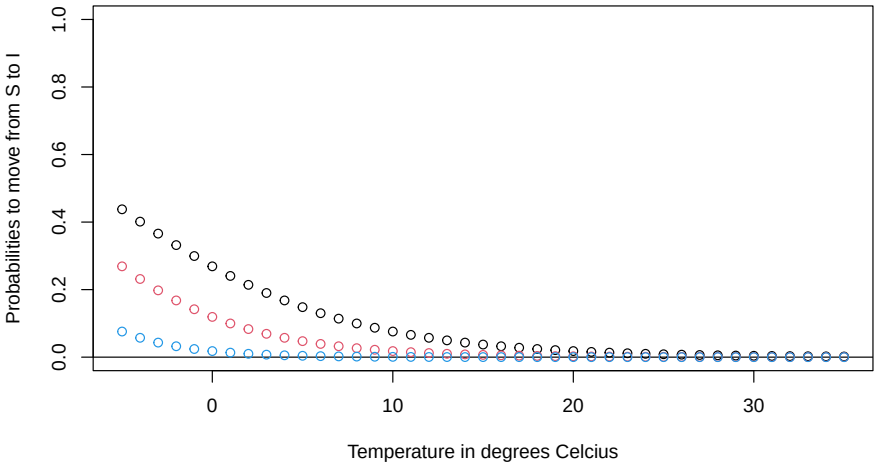


$S \rightarrow I$: Time Basis Functions

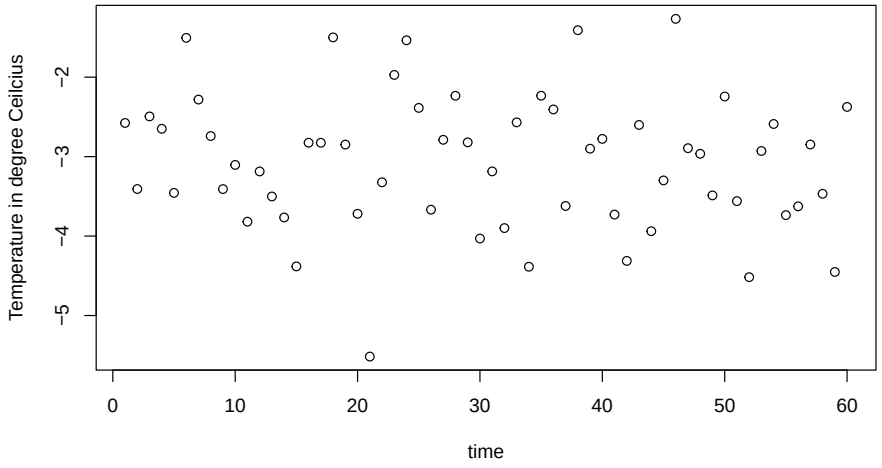


$S \to I$: Transition Probabilities, Different Scenarios

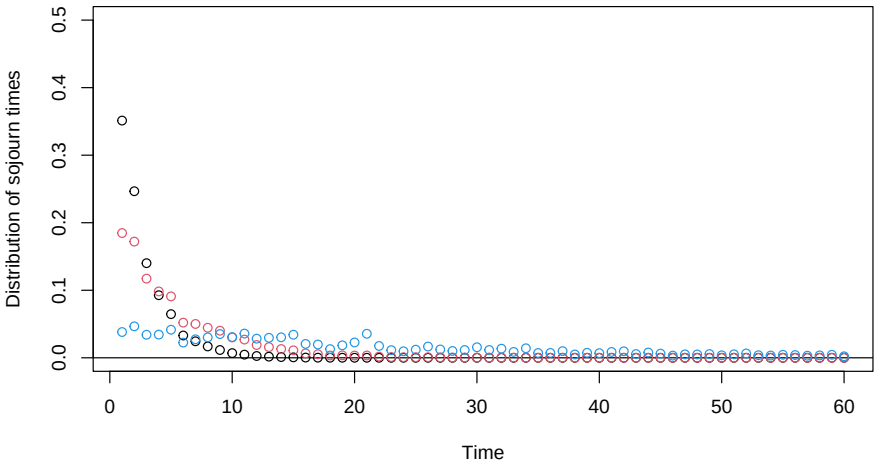
Here the x-axis is the **temperature** ! The tbf is just (1)



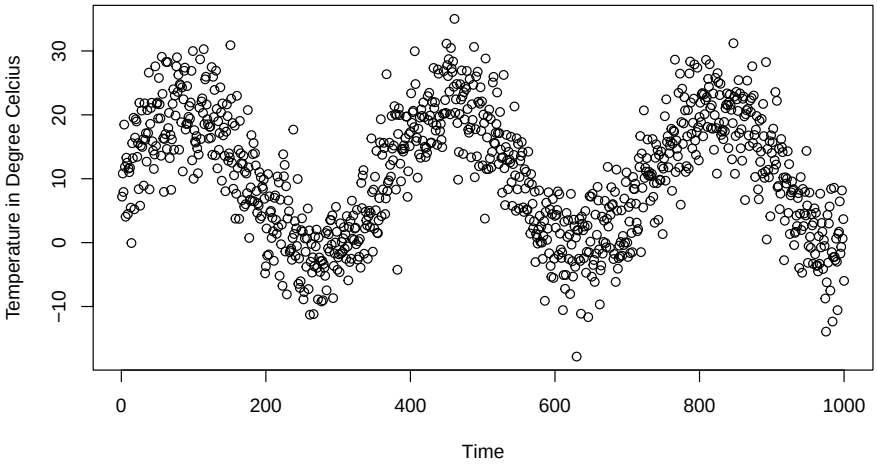
$R \to S$: Winter Temperature Conditions



$R \rightarrow S$: Sojourn Time Distributions, Different Scenarios in Previously Simulated Winter Temperature Conditions

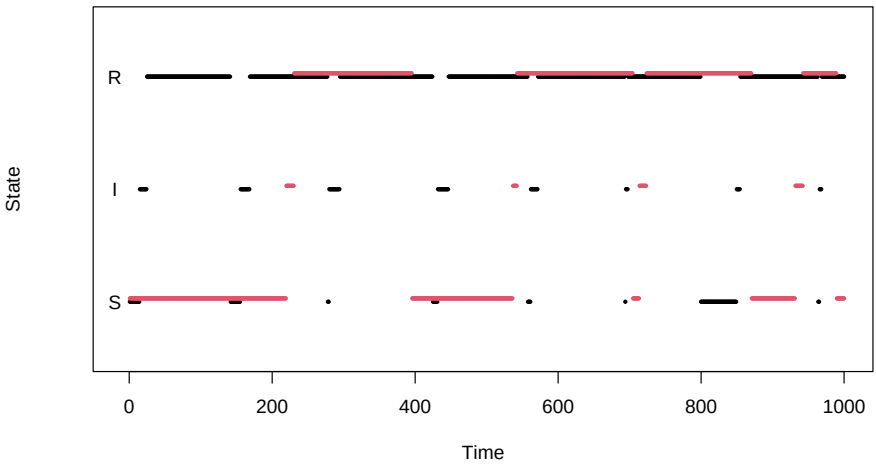


Simulation of a Temperature Series



Simulation of a State Series for 2 Individuals

A **global** behavior emerges from **local** transition probabilities
One individual gets contaminated less easily and keeps immunity longer, guess who's who



1. Introduction + motivating examples

2. Model formulation

3. Inference

Some reasons to choose Bayesian inference for this model (aside of unwholesome interest towards over-complicated MCMC schemes)

Disclaimer

- Frequentist inference has the interest to be generally lighter and faster (understatement)
- If objective is fast prediction / decoding / filtering / smoothing this is great

Frequentist approach (seen by a compulsive Bayesian)

1. Estimate the Maximum Likelihood Estimator $\underset{\theta,\psi,\gamma}{\arg\max}\ P_{Y_1,\dots,Y_T \mid \Theta,\Psi,\Gamma}(y_1,\dots,y_T \mid \theta,\psi,\gamma)$
2. Do stuff like smoothing, Viterbi with the MLE

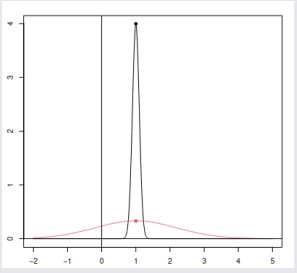
Bayesian approach

1. Get the *A Posteriori* *** distribution *** $P_{\Theta,\Psi,\Gamma \mid Y_1,\dots,Y_T}(\theta,\psi,\gamma \mid y_1,\dots,y_T) \propto \underbrace{P_{Y_1,\dots,Y_T \mid \Theta,\Psi,\Gamma}(y_1,\dots,y_T \mid \theta,\psi,\gamma)}_{\textit{likelihood}} \underbrace{P_{\Theta,\Psi,\Gamma}(\theta,\psi,\gamma)}_{\textit{prior}}$
2. Do stuff like smoothing **not with just one estimate of γ,θ,ψ but with all the posterior distribution**

Some reasons to go to the dark side

What if we do want to keep the incertitude ?

- Interpretation of parameters (see on the right)
- Inject some undue precision in a follow-up modelization



What if we have some extra information to feed the model ?

- We can know that some transitions never happen, e.g. $S \to R$, $R \to I$, etc
- We can have an idea of the emission parameters associated with a latent state, e.g. a S individual will have fever
- This helps with label permutation

Challenging Points for Interpretation

How to Interpret the Model Parameters ?

- How to retrieve simple Softmax parameters from time basis functions ? $\rightarrow$ with the induced low-rank structure
- Another problem : it is not straightforward to infer probability changes directly from Softmax parameters because it also depends on x
  - Try with different plausible configurations of x and calculate $P(W_{t+1}|W_t, U_t, x_{t+1})$ and especially $\partial P(W_{t+1}|W_t, U_t, x_{t+1})/\partial x$
  - Use parsimony
  - Simpler when the probability of staying in the same state is high due to the exponential interpretation of Softmax coefficients

How to Move from Local to Global ?

Once the problems inherent to any categorical model are overcome, we can have an idea of how $P(W_{t+1}|W_t, X_t, U_t, \Theta)$ behaves. But what global behavior emerges from the aggregation of local behaviors ?

- Sojourn time ?
- Which exit state ?

There is no simple solution, but we can improvise

- When the model does not depend on covariates but only on the intercept we can calculate global behaviors such as sojourn times (example : $I \rightarrow R$ and $R \rightarrow S$)
- Otherwise, test with different reasonable configurations of the variables (example : $S \rightarrow I$)

Computational Aspects

One Ingredient

- **Forward** algorithm that directly gives $P(Y|\Theta, \Gamma, \Psi)$
- Recurrence relation with sums and products
- Product rule in differentiation $\rightarrow$ gradient and Hessian calculated by recurrence
- We eliminate latent states $\rightarrow$ we are not encumbered as in EM or Gibbs
- Exponential cost in the number of coupled latent state chains $\rightarrow$ we keep **independent chains**

Several possible delicious Bayesian recipes

- Langevin algorithm on Riemannian manifold ? (Hessian-informed by definition, little assumption concerning the posterior)
- Hybrid Monte-Carlo ? (Possibly informed using the Hessian at the mode, little assumption concerning the posterior)
- Metropolis-within-Gibbs ? (Possibly informed using the Hessian at the mode, little assumption concerning the posterior)
- Deterministic exploration in the style of *INLA* ? (Informed using the Hessian at the mode by definition, assuming lumpoidal posterior)
- Self-normalized Importance Sampling ? (Not realistic without using the Hessian at the mode, assuming lumpoidal posterior)

Balance to find between the striking power of the Hessian and its cost

None

That's all folks !

Questions ?

References

Hanna Bacave.
Extension des modèles de (semi-) Markov cachés et algorithmes pour estimer la dynamique de (méta) populations partiellement observables.
PhD thesis, Université Paul Sabatier (Toulouse), 2024.

Jean Peyhardi, Catherine Trottier, and Yann Guédon.
A new specification of generalized linear models for categorical responses.
Biometrika, 102(4) :889–906, 2015.