Impact of some model misspecification in multi-response HSMM - application to fruit tree.

Sarah Assaf† - Jean-Baptiste Durand‡ - Sandra Plancade†

Processus markoviens, semi-markoviens et leurs applications
5-7 juin 2023 - Montpellier

Project supported by ANR HSMM-INCA

† INRAE, Unité MIA, Toulouse, ‡ Cirad, UMR AMAP, Montpellier

Outline of presentation

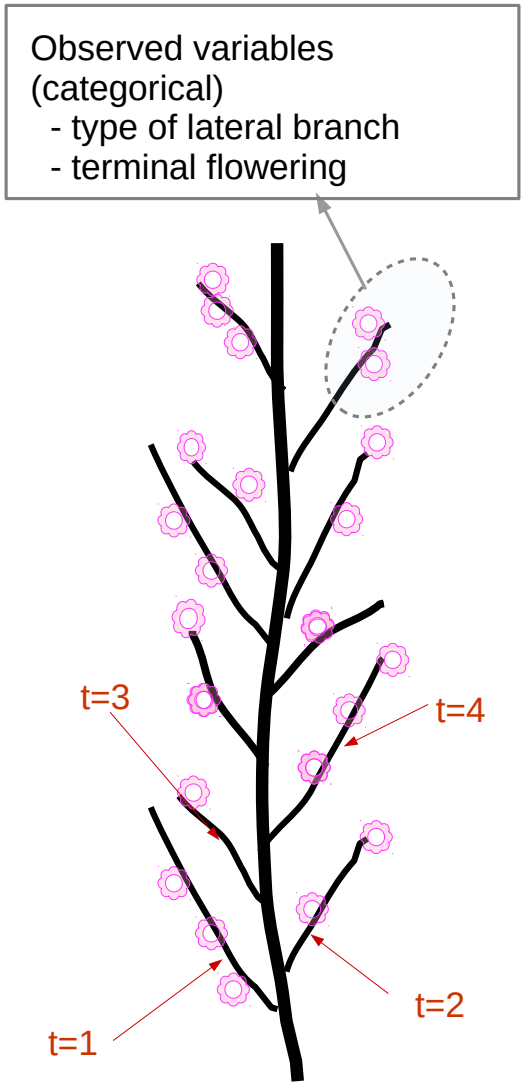
Biological model, motivation and problem statement

Illustration of errors of interpretation in case of model misspecification

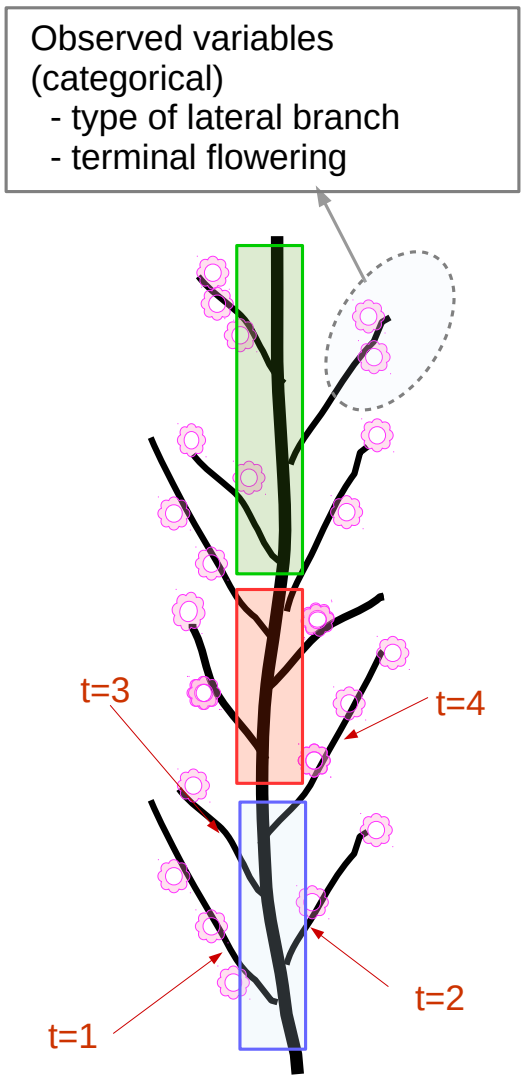
Model selection by BIC

Biological model of apple tree flowering

- Nodes on trunk $\Leftrightarrow$ discrete index process

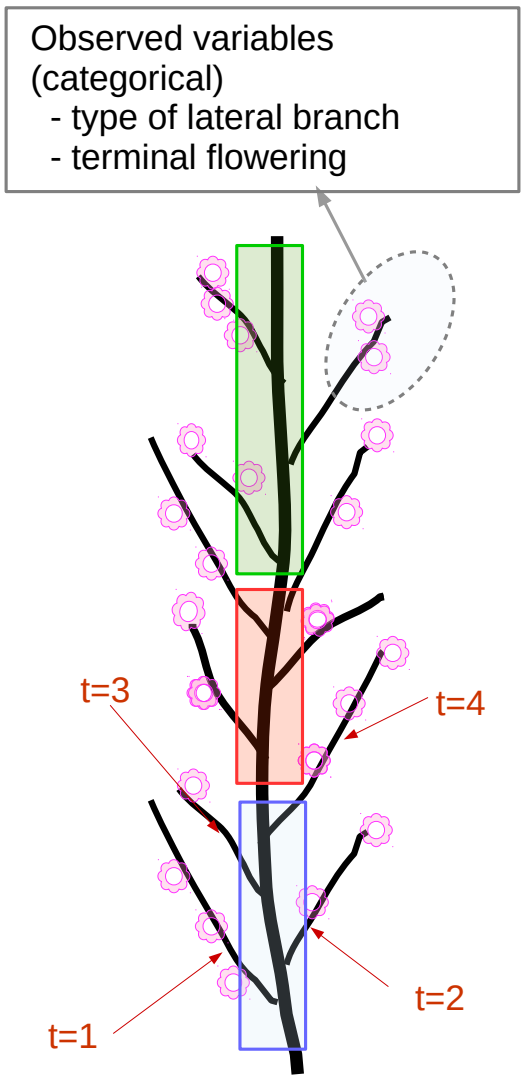


Biological model of apple tree flowering



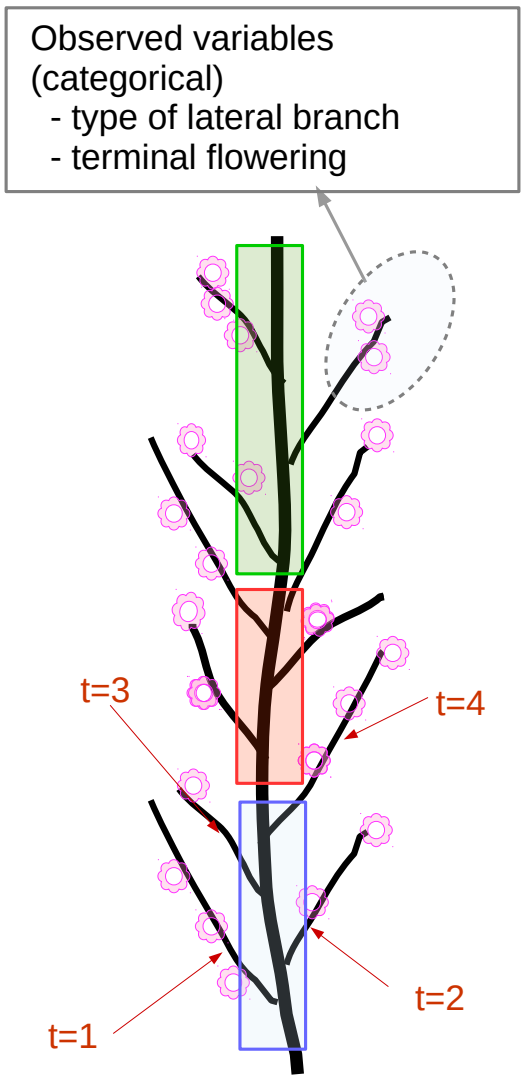
- Nodes on trunk $\Leftrightarrow$ discrete index process
- Biological assumption: phases of growth that impact simultaneously all observed variables

Biological model of apple tree flowering



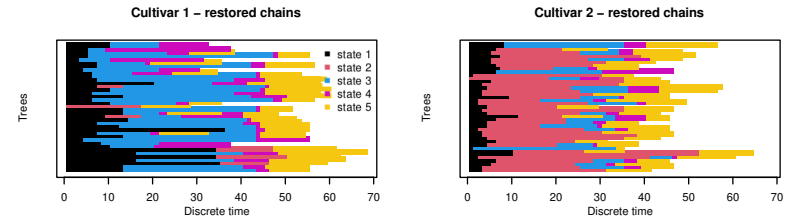
- Nodes on trunk $\Leftrightarrow$ discrete index process
- Biological assumption: phases of growth that impact simultaneously all observed variables
- Mathematical model: multi-response HSMM (mHSMM) with
  - hidden states = growth phases
  - observations = categorical variables
- Question: effect of cultivar (among 2) on growth phase lengths and/or compositions in terms of observations.

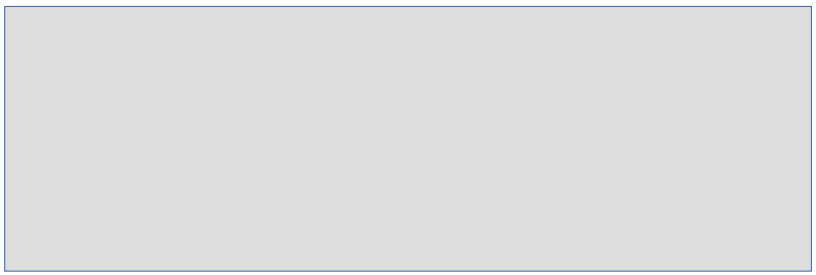
Biological model of apple tree flowering

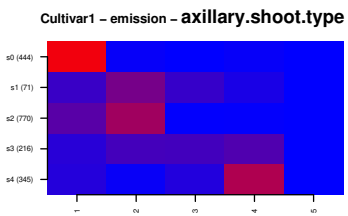


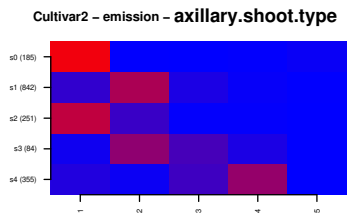
- Nodes on trunk $\Leftrightarrow$ discrete index process
- Biological assumption: phases of growth that impact simultaneously all observed variables
- Mathematical model: multi-response HSMM (mHSMM) with
  - hidden states = growth phases
  - observations = categorical variables
- Question: effect of cultivar (among 2) on growth phase lengths and/or compositions in terms of observations.
- Approximate inference
  - Inference of mHSMM parameters regardless of cultivar (several trajectories)
  - Restoration of hidden chains (Viterbi)
  - Estimation of emission and state duration distributions: empirical distribution given restored hidden states (counts)

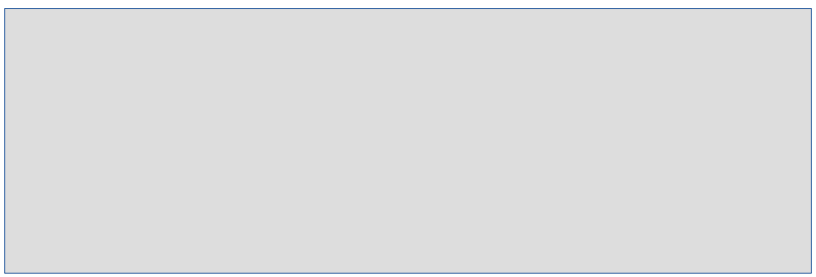
Data analysis based on the mHSMM model

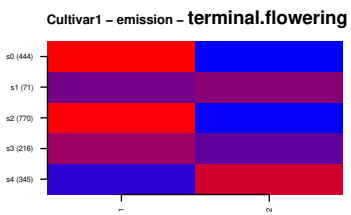


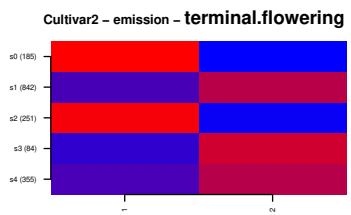










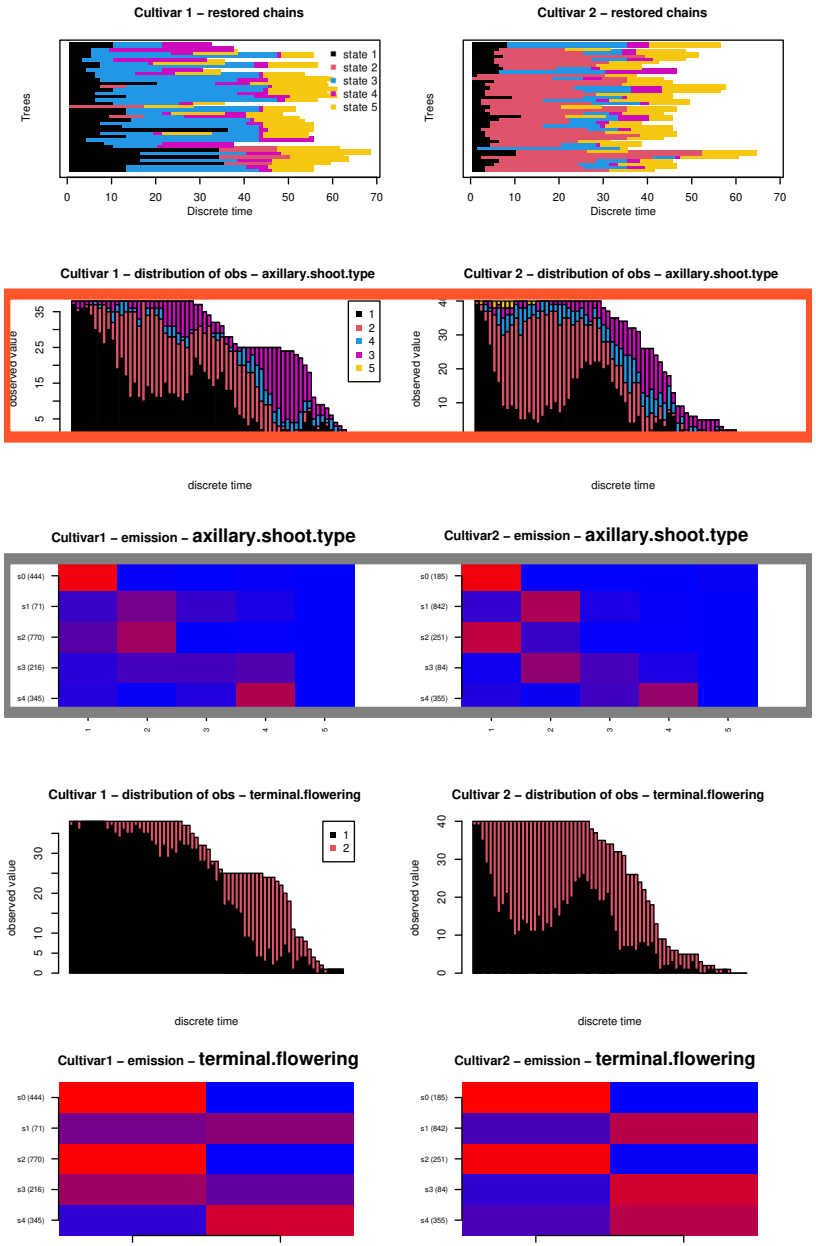


Conclusion

- Cultivar effect on hidden chains (state durations)
- Cultivar effect on "terminal flowering"
- Cultivar effect on "shoot type"

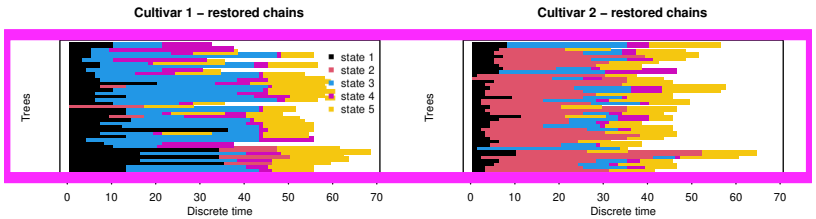
[2]: M. Mészáros, Y. Guédon, B. Krška and E. Costes, 'Modelling the bearing and branching behaviors of1-year-old shoots in apricot genotypes,' PLoS ONE, Public Library of Science, 2020.

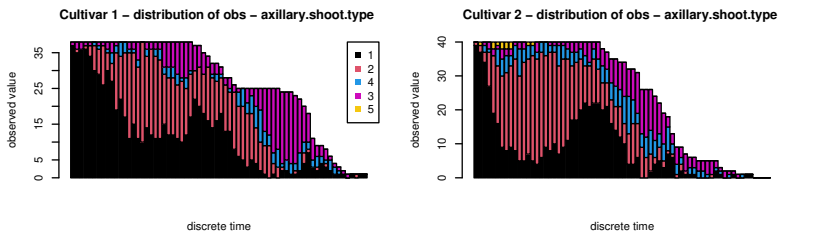
Contradiction with observations on raw data

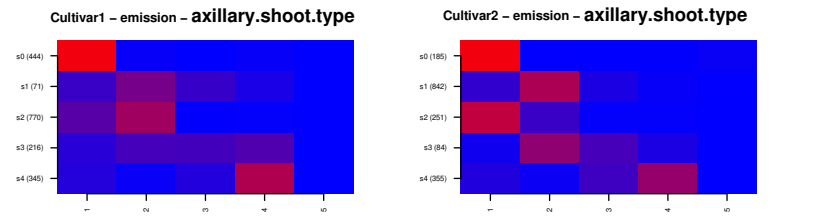


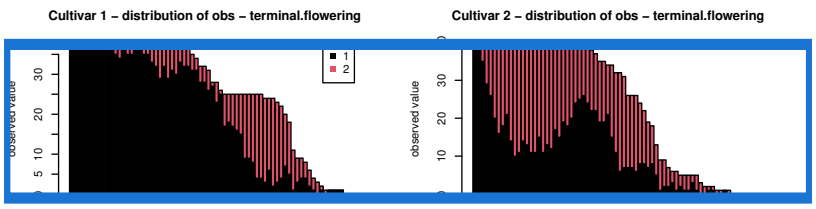
- Consistency between observations and parameters on "terminal flowering"
- Contradiction on "shoot type"
  - No cultivar effect on observed variable.
  - Cultivar effect on emission from parameter estimates.

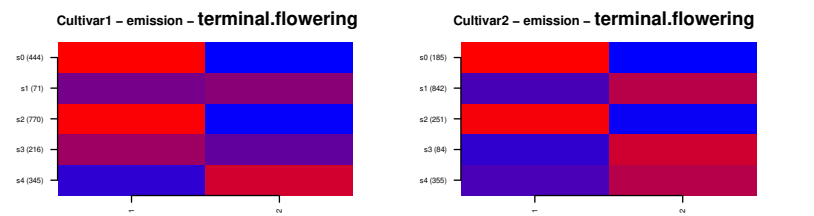
Contradiction with observations on raw data





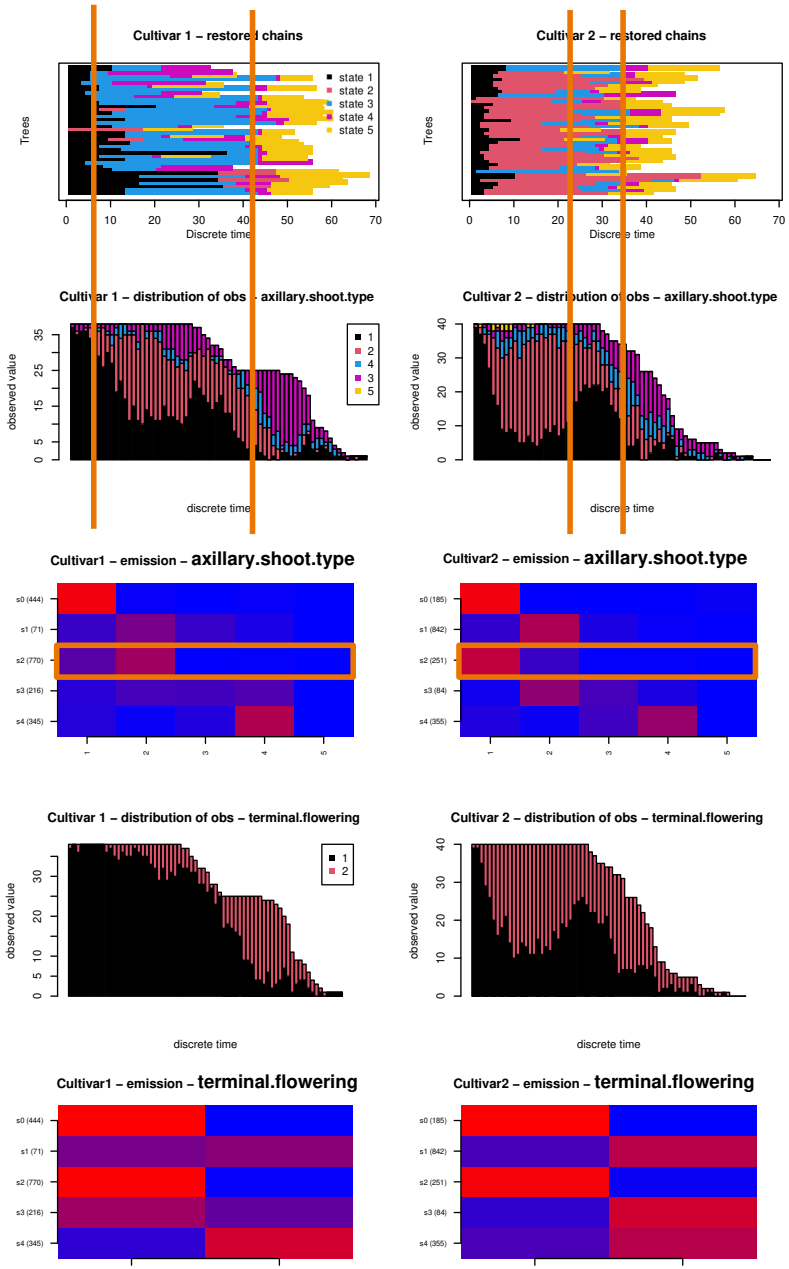






- Consistency between observations and parameters on "terminal flowering"
- Contradiction on "shoot type"
  - No cultivar effect on observed variable.
  - Cultivar effect on emission from parameter estimates.
- Heuristical justification
  - Variable "terminal flowering" different between cultivars $\Rightarrow$ generate different semi-Markov chains.

Contradiction with observations on raw data



- Consistency between observations and parameters on "terminal flowering"
- Contradiction on "shoot type"
  - No cultivar effect on observed variable.
  - Cultivar effect on emission from parameter estimates.
- Heuristical justification
  - Variable "terminal flowering" different between cultivars $\Rightarrow$ generate different semi-Markov chains.
  - Different semi-Markov chains $\Rightarrow$ Different time frame for a given state $\Rightarrow$ Difference in variable distribution given state

Two potential classes of models

State $S_{i,t}$, observation $O_{i,t}^{(c)}$, covariate z_i for sequence (tree) i.

(mHSMM)
- (1) $(S_{i,t})_t$ unidirectional (Left-right) SMM with state duration potentially dependent on z_i : $\mathbb{P}[S_{i,[t+1:t+d]} = s|z_i] := f(d|s, z_i)$
- (2) $(O_{i,t}^{(1)})_t$ independent given $(S_{i,t})_t$ with emission distribution potentially dependent on z_i : $\mathbb{P}[O_{i,t}^{(1)} = x|S_{i,t} = s, z_i] = g_1(x|s, z_i)$
- (3) $(O_{i,t}^{(2)})_t$ independent given $(S_{i,t})_t$ with emission distribution potentially dependent on z_i : $\mathbb{P}[O_{i,t}^{(2)} = x|S_{i,t} = s, z_i] = g_2(x|s, z_i)$

(reg-HSMM)
- (1) $(S_{i,t})_t$ unidirectional SMM with state duration potentially dependent on z_i : $\mathbb{P}[S_{i,[t+1:t+d]} = s|z_i] := f(d|s, z_i)$
- (2) $(O_{i,t}^{(1)})_t$ independent given $(S_{i,t})_t$ with emission distribution potentially dependent on z_i : $\mathbb{P}[O_{i,t}^{(1)} = x|S_{i,t} = s, z_i] = g(x|s, z_i)$
- (3) $(O_{i,t}^{(2)})_t$ follows a regression model on time, does not depend on $(S_{i,t})_t$ and may depends on z_i : $O_{i,t}^{(2)} \sim \mathcal{D}(\theta(t))$

State transitions are necessarily $s \to s + 1$

Goals

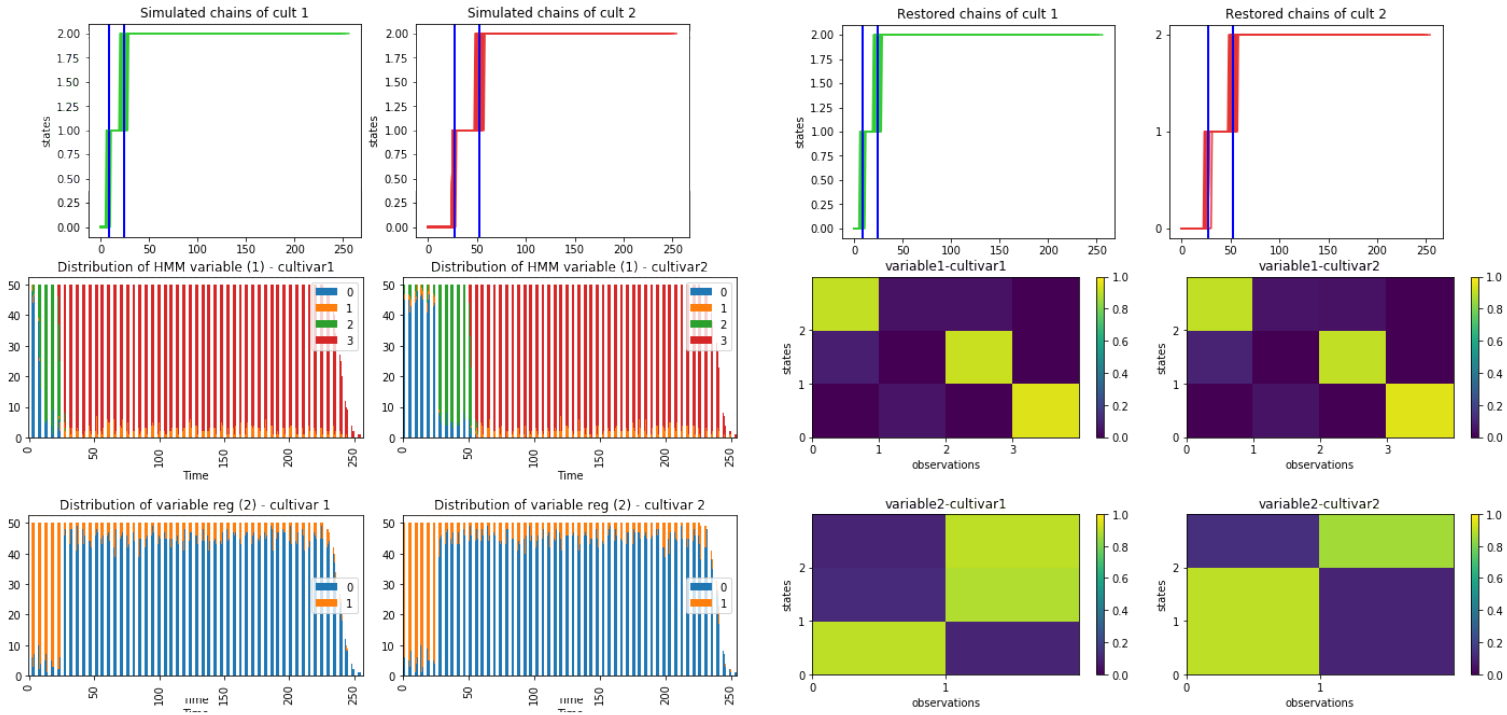
- To emphasize using simulations that considering wrong models may lead to wrong interpretation.
- Illustrate that BIC may select the correct model
- Application to our true data set

Remark

In (reg-HSMM) we will consider a simple parametric regression model: two-piece "histogram" (Bernoulli single change-point model).

Simulations under true (reg-HSMM) and estimation based on (mHSMM)

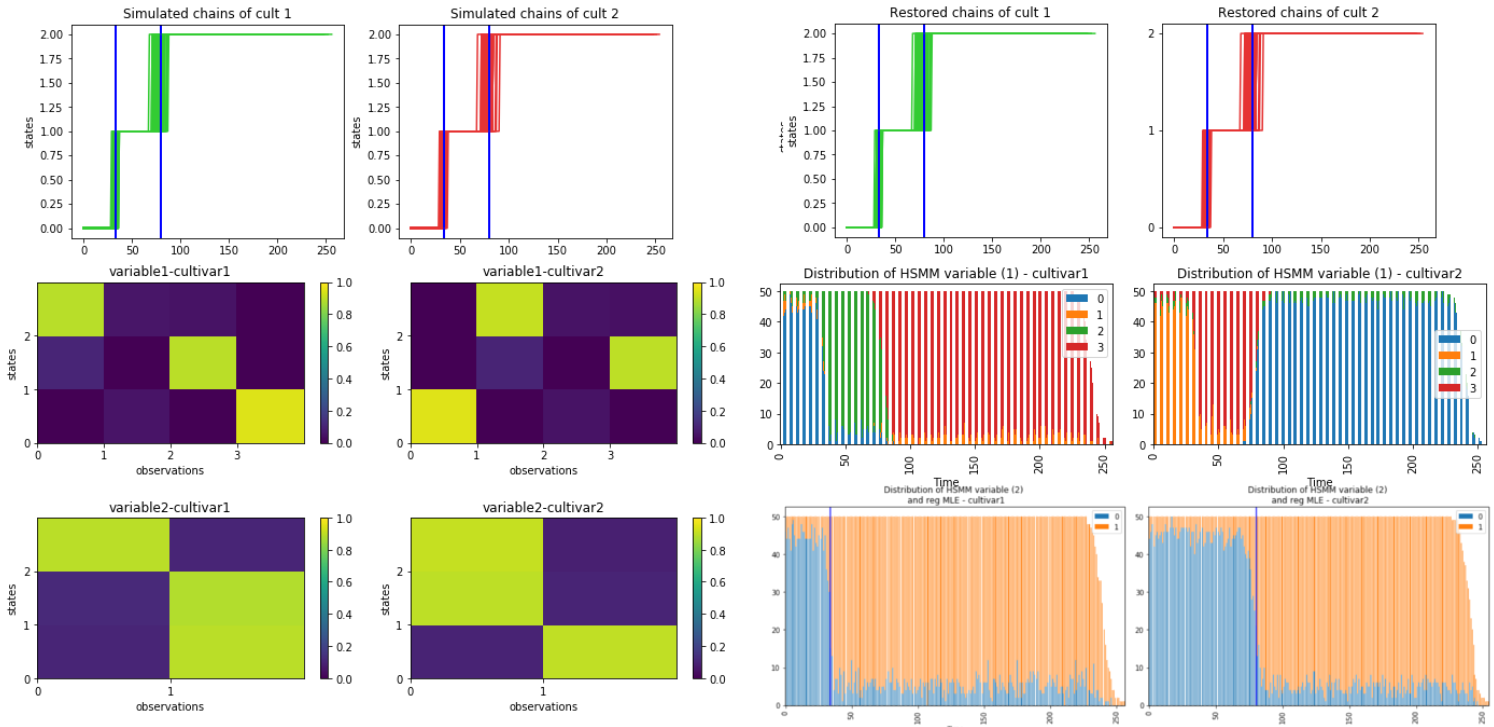
Simulations under (reg-HSMM) with	Estimation with (mHSMM) with
- SMM **depends on cultivar (duration probabilities)**	- SMM depends on cultivar (duration probabilities)
- Observed process 1 $O^{(1)}$ (HSMM variable) **independent on cultivar**	- $O^{(1)}$ (HSMM variable) depends on cultivar
- Observed process 2 $O^{(2)}$ (regression variable) **independent on cultivar**	- $O^{(2)}$ (HSMM variable) depends on cultivar



(Erroneous) conclusion: Emission distribution for $O^{(2)}$ **depends on cultivar $\Rightarrow$ cultivar effect on $O^{(2)}$.**

Simulations under true (mHSMM) and estimation based on (reg-HSMM)

Simulations under (mHSMM) with	Estimation with (reg-HSMM) with
- SMM **independent on cultivar**	- SMM depends on cultivar
- Observed process $O^{(1)}$ (HSMM variable 1) depends on cultivar	- O^1 (HSMM variable) depends on cultivar
- Observed process $O^{(2)}$ (HSMM variable 2) depends on cultivar	- O^2 (reg variable) depends on cultivar



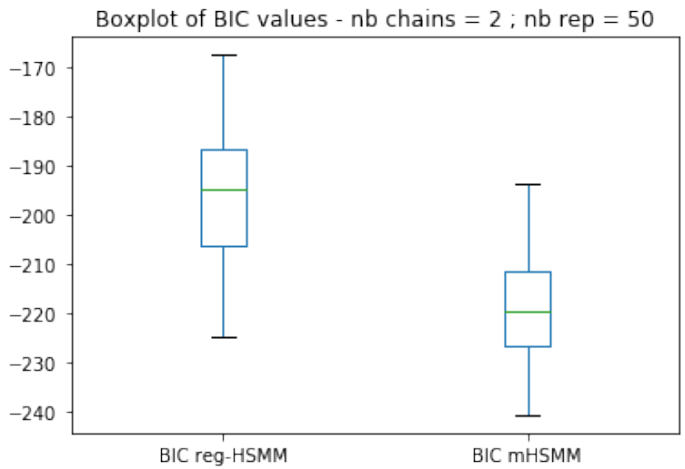
(Erroneous) conclusion: dynamics for $O^{(2)}$ **depends on cultivar** $\Rightarrow$ **cultivar effect on zone length with low flowering probability on** $O^{(2)}$

BIC for model selection

BIC = log(likelihood) - 0.5 ν log(nb obs)
(maximize over models; ν number of free parameters)

- 50 simulations of n trajectories of length T under either (mHSMM) or (reg-HSMM) (parameters as in previous examples).
- Estimate both models and compute BIC.

Simulation under (reg-HSMM);
$n = 50, 10, 8, 6, 4, 3, 2$; $T = 240, 50$



Boxplot of BIC values over 50 simulations ($n = 2$, $T = 50$)

nb trajectories n	4	3	2
correct selection rate (BIC)	100 %	100 %	100 %

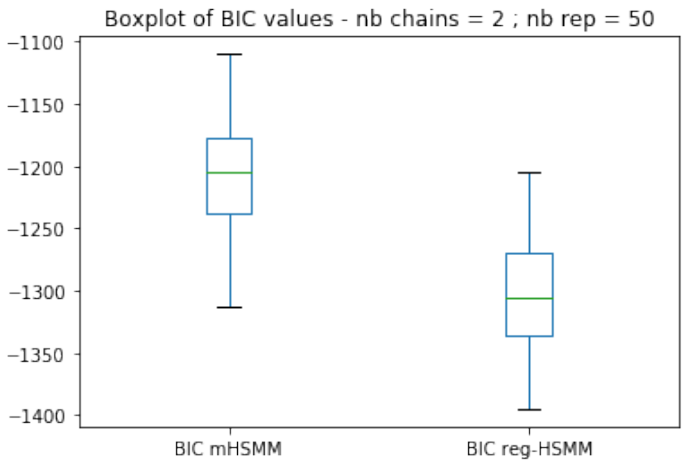
Rate of BIC selecting (reg-HSMM)

BIC for model selection: simulation under mHSMM

BIC = log(likelihood) - 0.5 ν log(nb obs)
(maximize over models; ν number of free parameters)

- 50 simulations of n trajectories of length T under either (mHSMM) or (reg-HSMM) (parameters as in previous examples).
- Estimate both models and compute BIC.

Simulation under (mHSMM); $n = 50, 10, 8, 6, 4, 3, 2$; $T = 240, 50$



Boxplot of BIC values over 50 simulations ($n = 2$, $T = 50$)

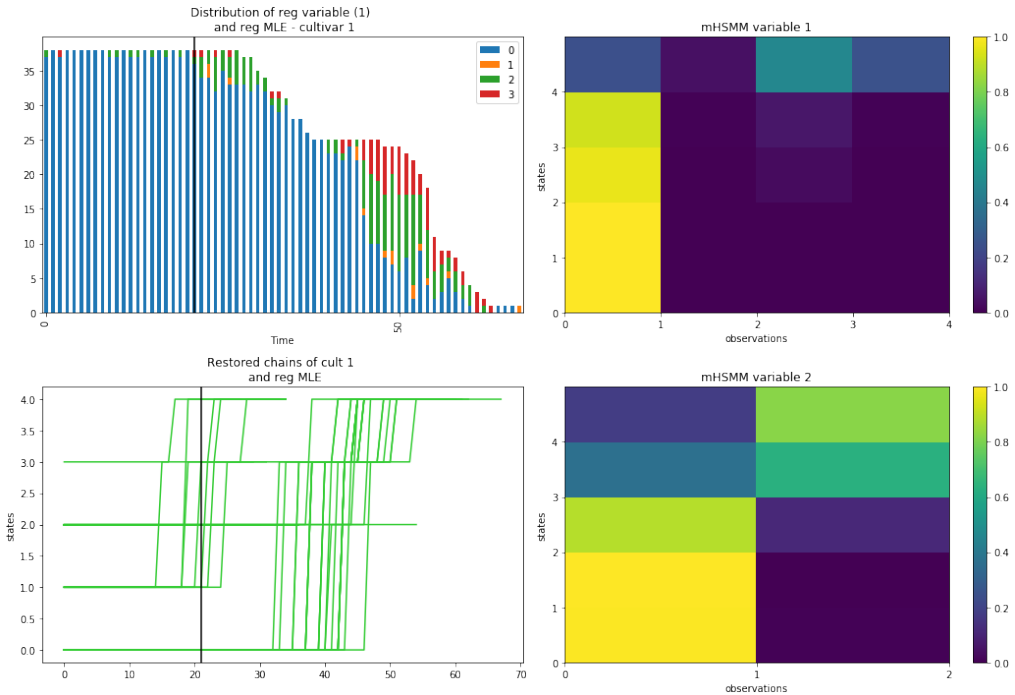
nb trajectories n	4	3	2
correct selection rate (BIC)	100 %	100 %	100 %

Rate of BIC selecting (mHSMM)

Results on true data set

Example: cultivar 1, variable 1 as (reg-HSMM), variables 2 and 3 as (mHSMM)

- Categorical reg variable with 5 categories, one change point.
- BIC mHSMM = -2,755 $<$ **BIC regHSMM = -2,542**
- Suggests (either?) unique, synchronous change for variable 1, independent on HSMM states
- Complicates model selection



Apple tree data set: cultivar 1. Marginal distribution of (reg) variable 1, emission distribution for mHSMM variables 2 and 3 and restored state sequences (3 states: sufficient?)

Conclusion et perspectives

- Summary
  - Classic model for fruit tree flowering (and other plant/organ development) : multi-response HSMM
  - Observations on our data: potential alternative model
  - Model misspecification can lead to erroneous biological conclusions
  - BIC for model selection has promising behaviour
- Perspectives
  - Evaluation of BIC on a realistic model (parameters estimated from data, multiple change-point models with common parameters between trajectories or other regression models)
  - Study of the consistency of BIC in model selection
  - Exploration of the combinatorics of models on true data (which variables are reg-, which ones are mHSMM and then, with how many change-points / states...?)