Selecting the Number of States in Hidden Markov Models — Pitfalls, Practical Challenges and Pragmatic Solutions

Presentation by Simon Bo Jensen & Johannes Thomsen

8 March 2018

Selecting the Number of States in Hidden Markov Models — Pitfalls, Practical Challenges and Pragmatic Solutions

Jennifer Pohle[1]*, Roland Langrock[1], Floris M. van Beest[2*], Niels Martin Schmidt[2*]

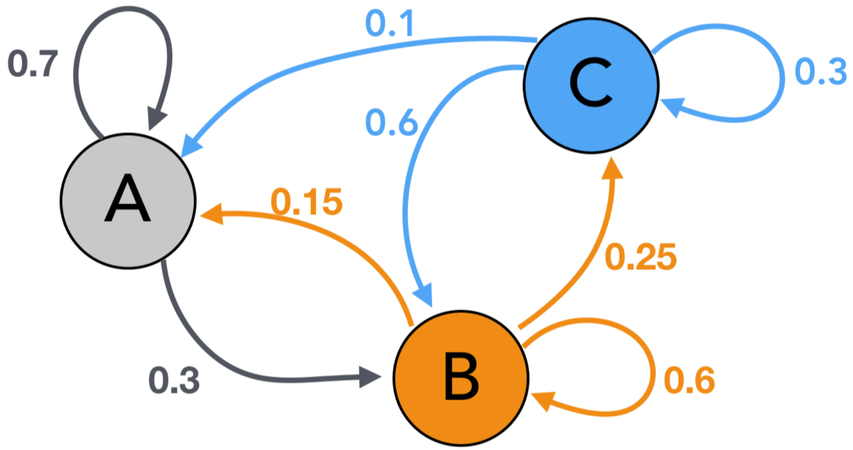
[1]Bielefeld University, Germany
[2]Aarhus University, Denmark

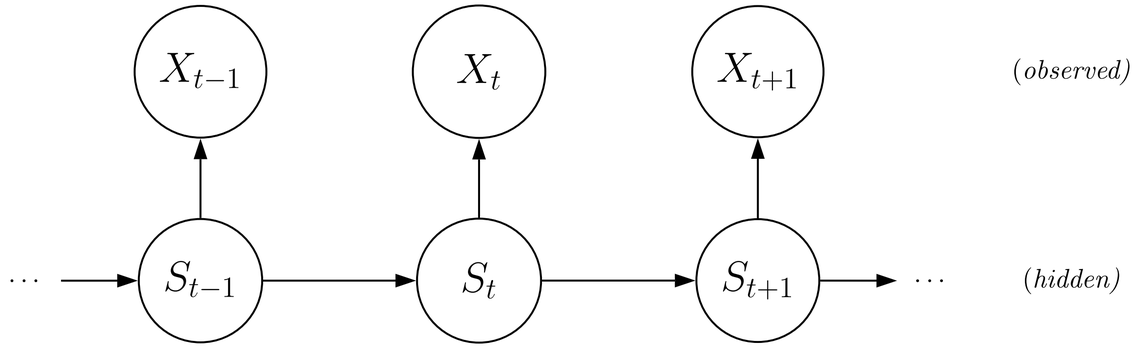
Abstract

We discuss the notorious problem of order selection in hidden Markov models, i.e. of selecting an adequate number of states, highlighting typical pitfalls and practical challenges arising when analyzing real data. Extensive simulations are used to demonstrate the reasons that render order selection particularly challenging in practice despite the conceptual simplicity of the task. In particular, we demonstrate why well-established formal procedures for model selection, such as those based on standard information criteria, tend to favor models with numbers of states that are undesirably large in situations where states shall be meaningful entities. We also offer a pragmatic step-by-step approach together with comprehensive advice for how practitioners can implement order selection. Our proposed strategy is illustrated with a real-data case study on muskox movement.

Keywords: animal movement, information criteria, selection bias, unsupervised learning

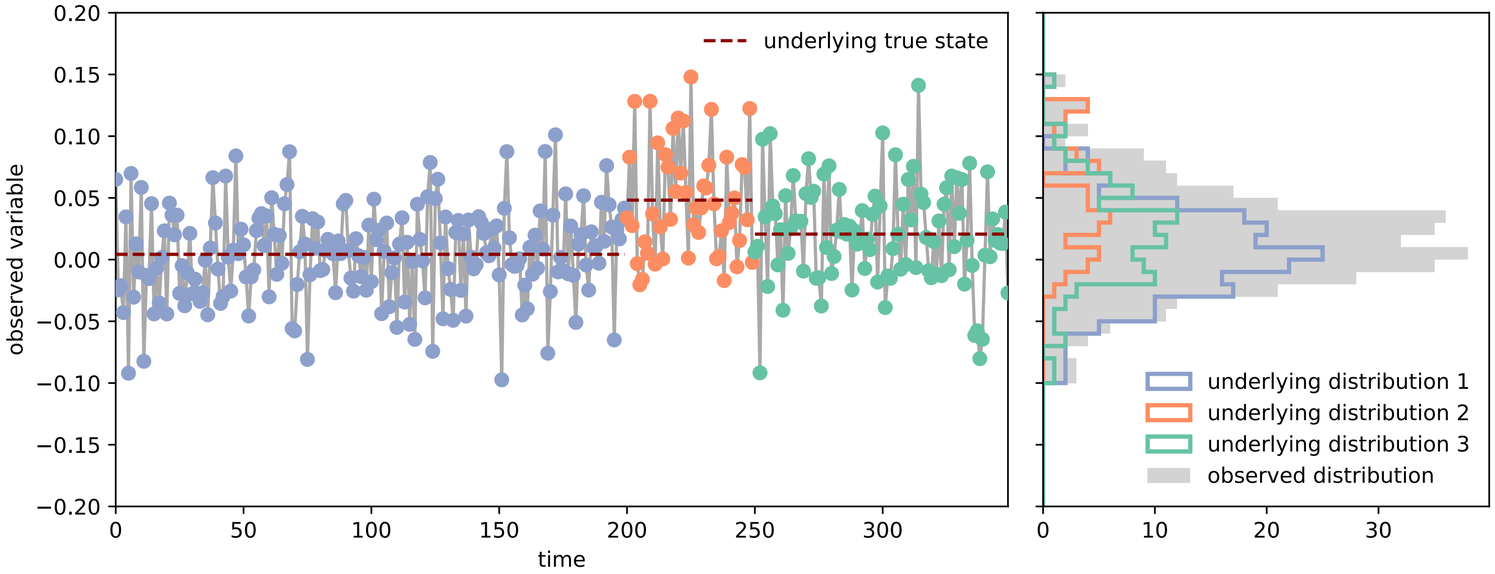
Hidden Markov Model





Real world problem

What is the number of underlying (hidden) states?



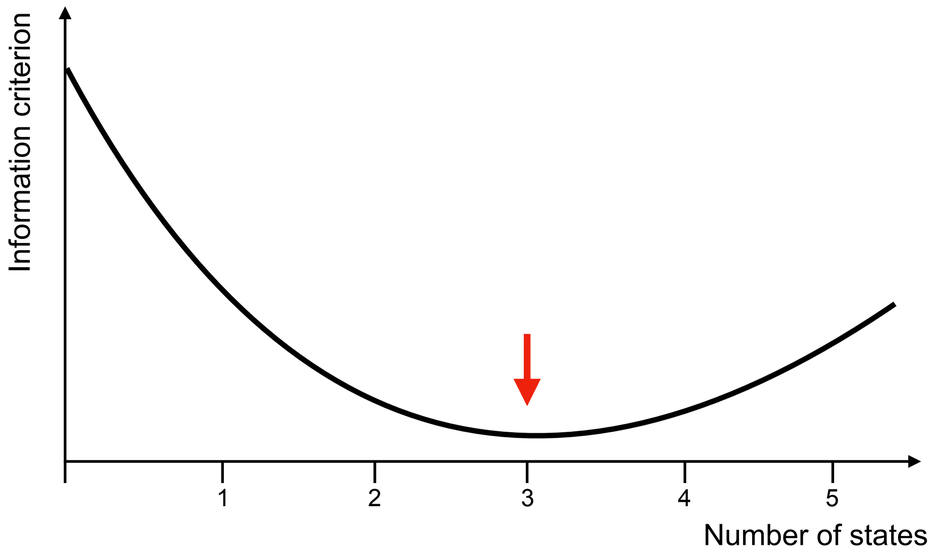
Solution

- Use information criteria to select the number of underlying states:
  - Estimate the goodness-of-fit for *n* underlying states
  - Choose the model which best fits the data

$$\mathrm{AIC} = -2\log\mathcal{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) + 2p,$$

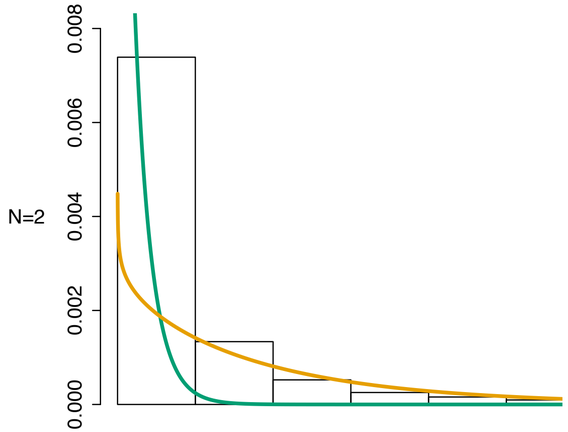
$$\mathrm{BIC} = -2\log\mathcal{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) + p\log(T),$$

$$\mathrm{ICL} = -2\log\mathcal{L}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{x}, \hat{\boldsymbol{s}}) + p\log(T).$$

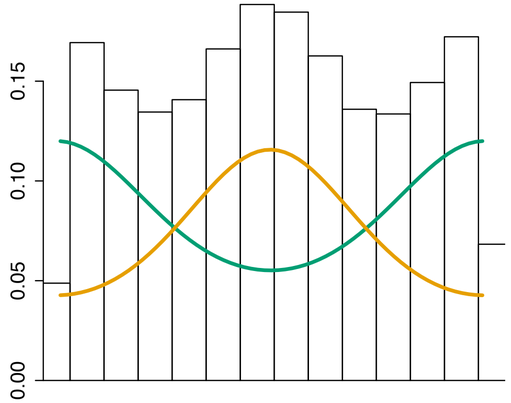


In practice: Studying animal movement

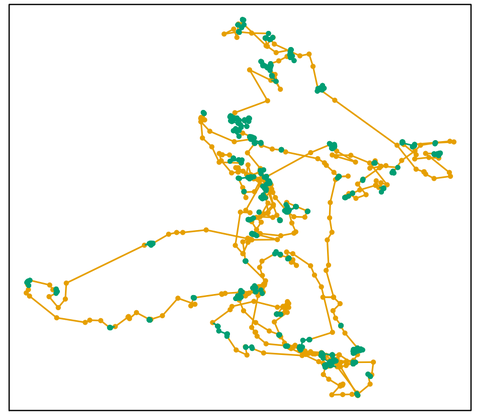
- GPS tracking of a muskox on Greenland
- Hourly GPS coordinates were collected over 3 years
- Ecologists wanted to study behaviour such as “resting”, “feeding” and “moving”



Step length



Turning angle

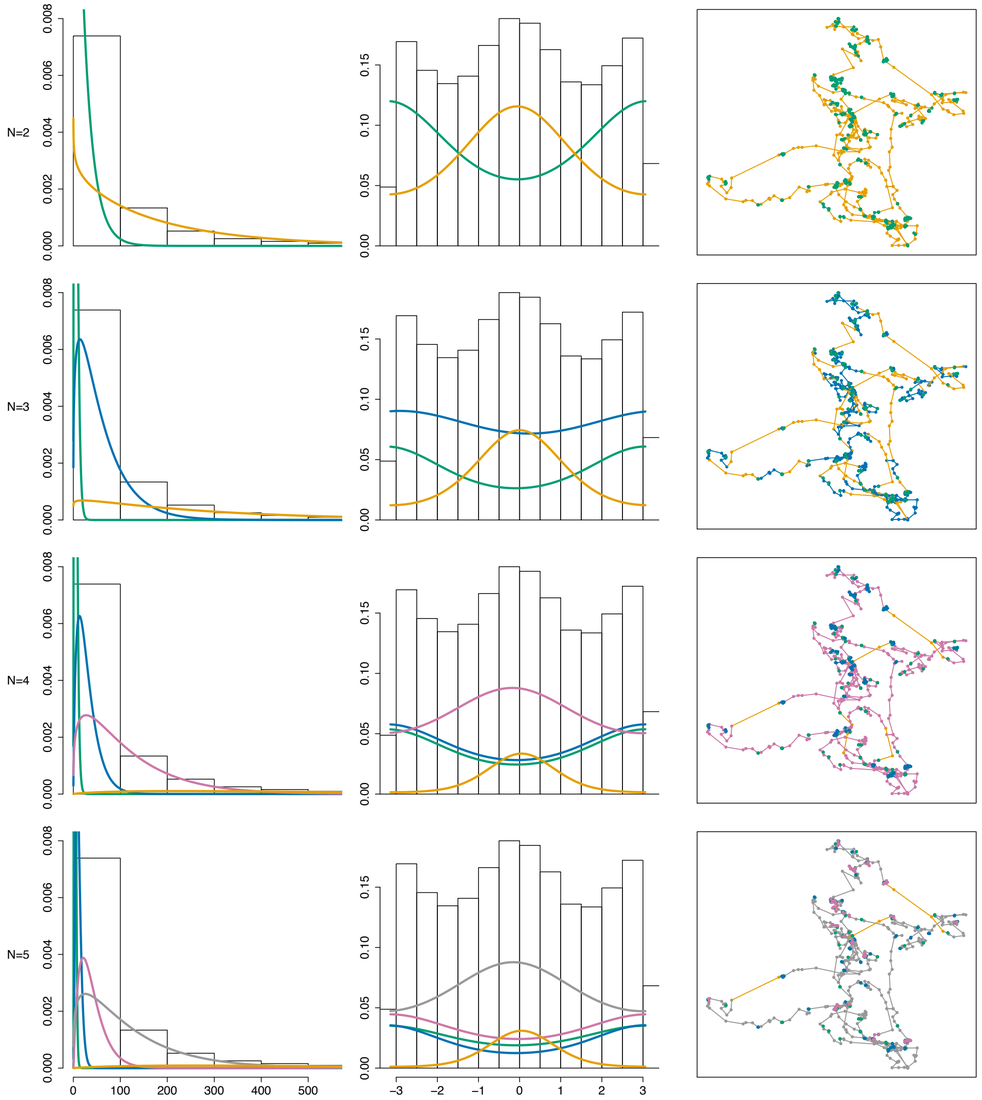


GPS coordinates

In practice: Studying animal movement

- 5 states fit the data best
- 2-3 states are more feasible for the study
- **Which model should be chosen?**

no. states	no. parameters	AIC	BIC	ICL
2	12	350199.3	350296.7	354829.3
3	21	345285.4	345455.8	351544.5
4	32	343404.9	343664.6	**350159.9**
5	45	**342782.0**	**343147.2**	351247.7



A pragmatic solution

- Adding more complexity to the model can make the number of states decrease
- Compromise between best-fit and simplicity of the model
- How many underlying states are sensible? Prior knowledge makes a difference
- Make a conscious decision and be transparent. If two models are equally good, present both.

no. states	no. parameters	AIC	BIC	ICL
2	12	350199.3	350296.7	354829.3
3	21	345285.4	345455.8	351544.5
4	32	343404.9	343664.6	**350159.9**
5	45	**342782.0**	**343147.2**	351247.7

Conclusions

- HMM are used to detect hidden states given a time series of observables
- Selecting the “true” number of hidden states is challenging!
- Information criteria can not always be trusted
- Make an informed decision based on prior knowledge. Be transparent.