Selecting the Number of States in Hidden Markov Models

Johannes Thomsen & Simon Bo Jensen 7/3/2018

Based on the original article *"Selecting the Number of States in Hidden Markov Models — Pitfalls, Practical Challenges and Pragmatic Solutions" - arXiv:1701.08673v2 [stat.ME]* by Pohle et al.

Introduction Hidden Markov models (HMMs) are very flexible time series models based on the idea that an obsevation X_t varying with time is caused by (and only dependent on) an underlying, hidden, state. The probability of being in a certain state is thought to be a Markov process, where the current state S_t is dependent on the previous state S_{t-1}.

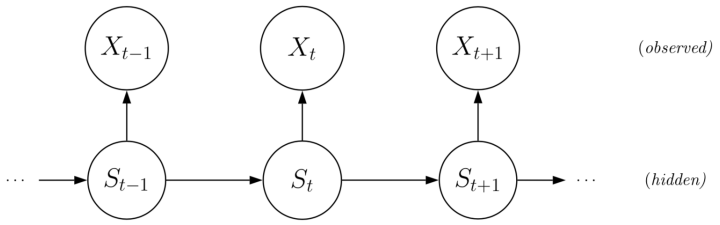

Figure 1: The simplest formulation of a HMM, where a hidden state S gives rise to an observation X.

and a transition probability matrix, which describes the probability of transitting from one state to another, between a given time t and $t+1$. For example, given a two-state model, the transitions between the two states could be described by the matrix

	S_1	S_2
S_1	p_a	p_b
S_2	p_b	p_a

where the probabilities along any horizontal or vertical axis of course should sum up to 1. In practice, a state S corresponds to an underlying distribution, emitting the observations, i.e. an emission distribution. With this information at hand, a HMM assumes independency between states, observations and emission distributions.

Model selection In a supervised learning context, a HMM is a powerful tool, as it can be trained to recognize specific states from noisy distributions, e.g. in speech recognition, where we have the possibility to generate true, underlying training data ourselves. In here, however, we seek to use a HMM to estimate the underlying states, not being able ourselves to generate the true underlying states. As illustrated in the example below (in this case with Gaussian, although *any* distribution is possible in principle) we observe a sequence of observations. By going through all possible observed sequences given a set of transition probabilities, we might conclude that 3 states fit the model best, and that this could be caused by 3 distinct weather patterns, types of animal behavior, enzyme conformations, etc.

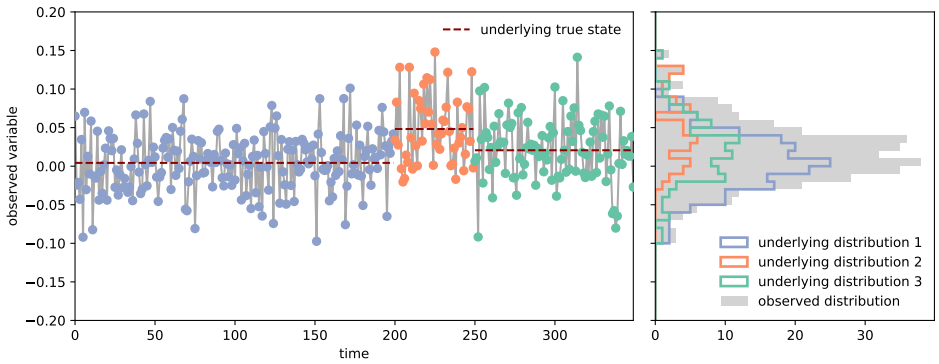

Figure 2: Simulated data showing the application of HMM on observations to deconvolute the underlying states that gives rise to a set of observations

Naturally, more states will make the model a better fit, and to avoid estimating the wrong number of states, several model quality estimators have been developed (herein the authors mention AIC, BIC, and ICL), each based around assumptions. On complex real-world data, some estimators tend to overfit, while others underfit, as some assumptions will always be violated, only that we don't know which ones (e.g. AIC and BIC assume that the true distributions can in fact be represented entirely by the model considered).

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

Figure 3: Table showing how information criteria are applied in practice. The lowest score means "best fit" in terms of model. Note that the scores quite often disagree for real-world data, given (hidden) violations of underlying assumptions.

The authors demonstrate this fact by providing simulation results with various types and numbers of emission distributions, showing how the model estimators will tend to vary wildly at estimating the most likely number of states, and often not agree, depending on the data, making it impossible to simply run the analysis and take the number at face value.

We must therefore strike a balance between the model that best describes the features of the data, while remaining interpretable to humans (which should be the overarching goal of any scientific study). The authors of the original papers outline seven different scenarios, such as data with outliers (so that it's unclear which distribution the datapoint belongs to) and temporal dependence of the emission distributions over the whole observation time.

A Pragmatic solution Through all the different example scenarios outlined in the original paper, the authors make it clear that a HMM out-of-the-box is by no means the most all-encompasing tool there is. For many of the scenarios it should, in principle, be possible to add extra parameters to the model to take the extra assumptions and heterogeneity of the data into account, to more specifically account for the observations. However, it's cautioned that heavily parameterized models may actually distract from the

actual aim of the study, which is not to develop an entirely new, often technically and computationally challenging non-standard model for a specific purpose, but to use the standard HMM in a larger framework of tools, while seeking *"more pragmatic, goal-oriented ways to overcome caveats of information criteria in the context of order selection"*.

The authors also view it as best scientific practice to acknowledge the limits of the model, and publish the different models and their respective outcomes, which is hardly ever done in the field of ecology (as this paper is based on). Based on personal experience in biophysics research, the same can also be said, where BIC is often the figure of merit taken at face value, where people may also seek to validate their models with comparisons to other parameterized models.