Supplementary materials for this article are available at 10.1007/s13253-017-0283-8.

Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement

Jennifer POHLE, Roland LANGROCK, Floris M. van BEEST, and Niels Martin SCHMIDT

We discuss the notorious problem of order selection in hidden Markov models, that is 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. Supplementary materials accompanying this paper appear online.

Key Words: Animal movement; Information criteria; Selection bias; Unsupervised learning.

1. INTRODUCTION

Hidden Markov models (HMMs) are flexible models for time series of observations driven by underlying, serially correlated states. Originating from speech recognition, they have found applications in various areas such as robotics, finance and social science (Zucchini et al. 2016). Over the last couple of years, they have also emerged as an increasingly popular statistical tool for the analysis of ecological time series data, where they have proven to be natural statistical models for animal movement data (Patterson et al. 2016), general animal behavior data (DeRuiter et al. 2016) and capture–recapture data (Pradel 2005), to name but a few.

In this paper, we discuss order selection in (finite-state) HMMs, that is how to select the number of states. While conceptually order selection appears to be a simple model

Jennifer Pohle (✉) and Roland Langrock, Bielefeld University, Bielefeld, Germany (E-mail: *jennifer.pohle@uni-bielefeld.de*). Floris M. van Beest and Niels Martin Schmidt, Aarhus University, Roskilde, Denmark.

Journal of Agricultural, Biological, and Environmental Statistics
DOI: 10.1007/s13253-017-0283-8

Published online: 05 June 2017

selection task, in practice it remains a notoriously difficult challenge faced by practitioners. In principle, when a maximum likelihood approach is taken, it is conceptually straightforward to use information criteria like the Akaike information criterion (AIC), the Bayesian information criterion (BIC) or variations thereof, to select between models with different numbers of states. However, especially in the ecological literature it has been claimed—and sometimes demonstrated—that traditional model selection criteria, and especially the AIC, often lead to the selection of much larger numbers of states than expected a priori (Langrock et al. 2015; DeRuiter et al. 2016; Li and Bolker 2017). Whether or not order selection involves difficulties depends on the purpose of an HMM-based analysis. We distinguish three main types of applications of HMMs: forecasting, classification (in a supervised learning context) and general inference on the data-generating process (unsupervised learning). Order selection is most challenging in the latter case, and we therefore focus on this application specifically in this paper.

The preference to include many states, particularly in ecological settings, can to some extent be explained by the complexity of such data sets. In addition to the features that actually motivate the use of state-switching models, such as multimodality and autocorrelation, real data often exhibit further structure, such as outliers, seasonality or individual heterogeneity. When neglecting these features in the HMM formulation, then additional states may be able to capture this ignored data structure and therefore provide a better model fit than models with a lower, but (biologically) more realistic number of hidden states (Langrock et al. 2015; Li and Bolker 2017). For analyses where the interest lies on the interpretation of the states, or on the general dynamics of the state process, this behavior of model selection criteria is highly undesirable. For example, in the context of animal movement modeling, the states can intuitively be interpreted as proxies for the behavioral states of an animal (e.g., resting, foraging or traveling), and primary interest usually focuses on identifying the drivers of behavioral processes. Therefore, an HMM applied to an animal's movement data can yield a deeper understanding of the behavior of said animal. However, as outlined above, traditional model selection criteria often point to models with large numbers of states which, crucially, may not be biologically interpretable anymore. In this paper, we suggest a pragmatic approach to choose the number of states in practical applications, which takes into account formal criteria for guidance, but also stresses the importance of the study aim, expert knowledge and model checking procedures.

The paper is organized as follows. In Sect. 2, we lay out the HMM basics, including a brief review of the types of ways in which HMMs are applied, and provide an overview of theoretical and practical aspects related to order selection. In Sect. 3, we use simulation studies to demonstrate how additional states in an HMM can capture neglected structure in the data, leading standard information criteria to often overestimate the true number of hidden states. In Sect. 4, we discuss how to pragmatically choose an adequate number of states and provide practical advice and guidance. Section 5 gives a real-data case study on muskox movement.

2. HIDDEN MARKOV MODEL BASICS

2.1. BASIC FORMULATION OF HMMS

An HMM is a doubly stochastic process in discrete time, the dependence structure of which, in its most basic form, is displayed in Fig. 1. The model assumes the observations, $X_1, \ldots, X_T$, to be driven by an underlying (unobserved) state sequence, $S_1, \ldots, S_T$. The state process is usually assumed to be an N-state Markov chain, such that $\Pr(S_t|S_{t-1}, S_{t-2}, \ldots, S_1) = \Pr(S_t|S_{t-1})$, with $S_t \in \{1, \ldots, N\}$. Throughout the paper, we assume the Markov chain to be stationary, unless explicitly stated otherwise. The state-switching probabilities are summarized in the transition probability matrix (TPM) $\Gamma = \big(\gamma_{ij}\big)$, where $\gamma_{ij} = \Pr(S_t = j \mid S_{t-1} = i)$, $i, j = 1, \ldots, N$. In addition to the Markov property, it is usually assumed that the observations are conditionally independent of past observations and states, given the current state: $p(X_t|X_{t-1}, \ldots, X_1, S_t, \ldots, S_1) = p(X_t|S_t)$. Here p is used as a general symbol to denote either a probability function (if X_t is discrete-valued) or a density function (if X_t is continuous-valued). Thus, the distribution of each observed variable X_t, $t = 1, \ldots, T$, is completely determined by the current state S_t. Together with the Markov property, this completes the basic model structure as directed graph depicted in Fig. 1.

With animal behavior data, the observed process will often be multivariate, $\mathbf{X}_t = (X_{1t}, \ldots, X_{mt})$. In that case, a commonly made additional assumption is that the m variables are also conditionally independent, given the current state, hence $p(\mathbf{X}_t \mid S_t) = \prod_{i=1}^{m} p(X_{it} \mid S_t)$. This is the so-called contemporaneous conditional independence assumption.

In this basic form, the HMM formulation is completed by choosing the number of hidden states, N, and the class(es) of state-dependent distributions (also called emission distributions). In practice, the model parameters will then need to be estimated, which is usually accomplished using either numerical maximum likelihood, the expectation–maximization (EM) algorithm or Markov chain Monte Carlo (Zucchini et al. 2016).

2.2. TYPES OF HMM-BASED ANALYSES

As indicated in Sect. 1, there are three main types of ways in which HMMs can be applied. First, HMMs can be used for forecasting future values of the observed time series,

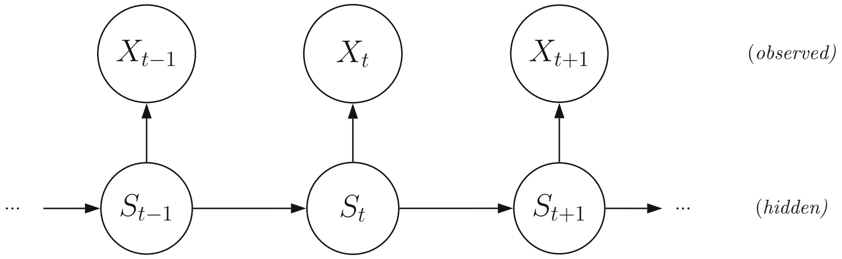

Figure 1. Dependence structure of an HMM in its most basic form.

as typically done in econometric time series analysis. In these instances, HMMs are not usually regarded as good representations of the true data-generating process and instead are used merely as tools to accommodate features of the observed time series that may be difficult to capture otherwise, using standard time series models (e.g., ARIMA or GARCH models). The main purpose of this application of HMMs is usually not to learn something about the process, but instead to accurately predict future observations. A specific example is given by the application of HMMs to financial share returns: here it is not actually reasonable to assume that there is a finite number of (market) states, yet the models can produce good forecast distributions.

Second, in the machine learning literature, HMMs are often used for classification in a supervised learning context (e.g., for speech recognition, gesture recognition). In these settings, an HMM is trained using data where the underlying states are known and subsequently applied to new, unlabeled data with the aim to recognize the underlying states. In ecology, HMMs are sometimes applied in this way to detect animal behavior states (Broekhuis et al. 2014; see also the discussion in Leos-Barajas et al. 2016). However, these applications are relatively rare in ecology, since training data, where the states are directly observed, need to be available, which is usually difficult to realize in the field. In this type of application of HMMs, the choice of the number of states is not an issue, as the states and their meaning are predefined.

Third, in an unsupervised context, HMMs are used to learn something about the data-generating process, without defining the role of the states a priori. Especially in movement ecology, this is the standard way in which HMMs are applied, with the aim of inferring novel aspects related to the behavioral process (Morales et al. 2004; Patterson et al. 2009). While practitioners may have some expectations regarding the number of states also in the unsupervised context, the identification of a suitable number of states in general still remains a primary aim of empirical studies. Thus, the unsupervised learning context is where order selection in HMMs constitutes the biggest challenge, so it is this case that we focus on.

2.3. Model Selection for HMMs

In practical applications of HMMs, users need to (i) specify the dependence assumptions made within the model, (ii) decide on the class of distributions used for the state-dependent process (e.g., normal distributions) and (iii) select the number of states, N. It may also be necessary to (iv) decide which covariates to include in the model. It is our experience that in most practical applications of HMMs, model selection focuses on (iii) and, if applicable, (iv), with (i) and (ii) specified with little or no investigation into the corresponding goodness of fit of the resulting models. For the model selection involved in both (iii) and (iv), when a maximum likelihood approach is taken, then information criteria such as the AIC or the BIC are typically used.

When using the AIC, the focus lies on out-of-sample predictive accuracy. Given a model fitted using maximum likelihood, with corresponding estimate $\hat{\boldsymbol{\theta}}$ for the parameter vector $\boldsymbol{\theta}$, the AIC is defined as $\text{AIC} = -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|x) + 2p$, where $\mathcal{L}(\cdot|x)$ is the likelihood function given the observed time series $x = (x_1, \dots, x_T)$ and p is the number of model parameters (see Zucchini et al. 2016, for details on how to evaluate the likelihood of an HMM). The term

$\log \mathcal{L}(\hat{\boldsymbol{\theta}}|x)$ can be regarded as a simple plug-in estimate of the expected log predictive density (using only the available data to forecast the log likelihood of future data). The log predictive density is one of many examples of a proper scoring rule for assessing predictive accuracy (Gneiting and Raftery 2007). Crucially, the plug-in estimator is biased due to overfitting: on average, the model fits the given sample better than an average sample. Under regularity conditions, it can be shown that in the limit (i.e., asymptotically, as $T \to \infty$), this bias converges to p. For large sample sizes, $\log \mathcal{L}(\hat{\boldsymbol{\theta}}|x) - p$ hence is an approximately unbiased estimator of the expected log predictive density (with the AIC obtained by multiplying this term by -2).

The BIC is defined as $\text{BIC} = -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|x) + p \log(T)$ and differs from the AIC in its form only through the increased penalty term (for $T \geq 8$). However, it is derived from a Bayesian viewpoint and aims at identifying the model that is most likely to be true, instead of maximizing prediction accuracy as does the AIC. Under regularity conditions and for large samples, minimizing the BIC is approximately equivalent to maximizing the posterior model probability (Schwarz 1978). However, for HMMs, consistency of the BIC is not fully established (Celeux and Durand 2008). More comprehensive accounts on the theoretical background of both AIC and BIC, and also their relation to other model selection concepts, are given in Zucchini (2000) and Burnham and Anderson (2002).

Similarly as the BIC, the integrated completed likelihood (ICL) criterion proposed by Biernacki et al. (2013) takes into account model evidence, but additionally considers the relevance of partitions of the data into distinct states, as obtained under the model. The ICL criterion approximates the integrated complete-data likelihood, which is the joint likelihood of the observed values $x = (x_1, \ldots, x_T)$ *and* its associated underlying state sequence $s = (s_1, \ldots, s_T)$ using a BIC-like approximation. As the true state sequence is unknown, it is replaced by the Viterbi-decoded state sequence $\hat{s}$, that is the most probable state sequence under the model considered. With $\mathcal{L}_c(\cdot|x, \hat{s})$ denoting the (approximate) complete-data likelihood, the ICL criterion is defined as $\text{ICL} = -2 \log \mathcal{L}_c(\hat{\boldsymbol{\theta}}|x, \hat{s}) + p \log(T)$. As in case of the AIC and the BIC, the model is chosen that leads to the smallest value of the criterion. In the context of HMMs, the simulation studies provided by Celeux and Durand (2008) indicate that ICL may actually underestimate the number of states of the HMM in certain scenarios. This can be explained by the preference of the ICL criterion for models where the emission distributions do not strongly overlap.

Cross-validated likelihood using a proper scoring rule, as suggested in Celeux and Durand (2008), constitutes another approach to model selection that focuses mostly on predictive performance. Cross-validation methods can become very computationally intensive, which becomes particularly problematic with the increasingly large telemetry data sets collected in movement ecology, where the estimation of a single model can easily take several hours.

From the theoretical perspective, the behavior of any of these criteria, and of ICL in particular, is poorly understood. A key assumption of both AIC and BIC is that the actual process is indeed represented by one of the candidate models. If the candidate models are at least good approximations of the true data-generating process, then this procedure can still be expected to work well (Zucchini 2000). However, when working with complex ecological data, neither an HMM nor in fact any other computationally feasible statistical model can be expected to be a complete representation of the true process, which will typically involve various complex patterns (even if the actual state space is low-dimensional). Thus, intuitively, it is clear that undesirable behavior of say AIC or BIC may occur in applications of HMMs to complex patterns.

2.4. Running Example: Animal Movement

To demonstrate the practical challenges involved when selecting the number of states of an HMM, we use animal movement modeling as a running example. Animal movement is one of the most natural and intuitive applications of HMMs and also constitutes a scenario where overestimation of the number of states is particularly prevalent. In HMM applications to animal movement data, the observed process is usually a bivariate time series comprising step lengths and turning angles between subsequent locations, typically though not necessarily collected using GPS technology. The states of the Markov chain underlying a fitted HMM can then be interpreted as providing rough classifications of the behavioral modes of the animal observed (e.g., exploratory vs. encamped). While we use this particular area of application of HMMs to fix the ideas, the issues and methods discussed in this paper are not restricted to such data.

3. SIMULATION STUDIES

In this section, we present simulation studies to investigate the performance primarily of the AIC and the BIC, but also of the ICL criterion, when it comes to selecting a suitable number of hidden states. Given the asymptotic equivalence of AIC and cross-validated likelihood (Stone 1977)—at least for ordinary linear regression—we did not implement the latter because of the associated substantial computational effort. ICL in contrast offers a conceptually different, classification-based approach to assessing a model's relative suitability, with a stronger focus on separation of classes.

We showcase seven scenarios where there is additional structure in the data that is not accommodated within basic HMM formulations. Each type of additional structure considered may be found in real data, and especially within animal behavior data, where the assumptions made with the basic HMM formulation typically are overly simplistic. What will be shown is that in such cases, the misspecification of the model formulation will often be compensated by additional states which, to some extent, absorb (or "mop up," figuratively speaking) the neglected structure.

3.1. Simulation Scenarios Considered

As a baseline model, we consider a two-state HMM having a gamma-distributed output variable, with two distinct sets of parameters, one for each state of the underlying stationary Markov chain. This type of model is common for analyzing animal movement data, with the two HMM states corresponding roughly to "foraging/resting" and "traveling" behavior, respectively. The state-dependent gamma distributions in the baseline model are shown in Fig. 2. If not explicitly stated otherwise in the description of the scenarios below, then the

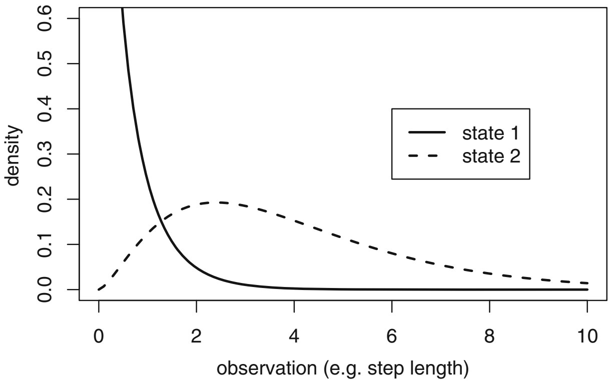

Figure 2. Gamma distributions of the baseline HMM used in the simulation experiments.

gamma distribution within the first state has mean $= 0.5$ and shape $= 0.7$ (resulting in a strictly monotonically decreasing density function), and the gamma distribution within the second state has mean $= 4$ and shape $= 2.5$ (resulting in a density function with mode distinct from zero). In addition, if not explicitly stated otherwise, then the probability of leaving a state in any given time interval is specified to be 0.1 (hence that of remaining in a given state is 0.9).

In each of the seven scenarios considered below, some component of the baseline model formulation will be slightly modified when simulating data. However, crucially, all the scenarios still involve only two genuine states. We will then demonstrate that when not taking the modification into account, hence fitting the slightly misspecified basic models to the data, then additional states will be included as per recommendation of model selection criteria, in order to compensate for the inflexibility of the model to otherwise capture the additional structure.

Scenario 1 (outliers) This first, very simple scenario represents a situation in which some of the data are outliers. In movement ecology, these outliers could, for example, be due to unusually large measurement errors (e.g., as a result of poor satellite coverage when using GPS tags). The corresponding simulated data are generated using the baseline setup described above, but subsequently adding uniformly distributed random errors from the interval $[10, 20]$ to only 0.5% of the data points (25 data points per sample generated). Intuitively it is clear that these few outlying values may cause the two-state baseline model to have a rather poor goodness of fit, since the two gamma state-dependent distributions may not be able to cover the extreme values without losing accuracy for the non-outlying observations. This can potentially be compensated for by including additional states merely to capture the outlying values.

Scenario 2 (inadequate emission distribution) While parametric distributions will often provide good approximations of the actual empirical distribution within a state, in practice it will practically never be the case that the true within-state distributions are actual members of some parametric family. In other words, the parametric family being used (e.g., gamma or normal distributions) will in general only provide an approximation of the actual data-generating process within a state. In this scenario, we illustrate possible consequences of this for order selection.

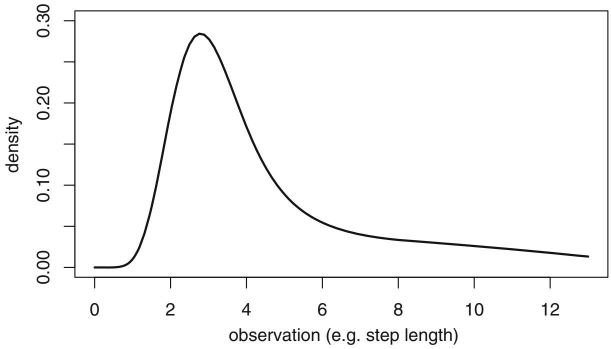

Figure 3. Heavy-tailed emission distribution within state 2, as implemented in Scenario 2.

To do so, the observations within the second state were not generated by a gamma distribution with mean 4 and shape 2.5 anymore, but instead by a similarly shaped but heavy-tailed distribution which we constructed nonparametrically using B-spline basis functions (Langrock et al. 2015). The specific shape of the distribution within state 2 is shown in Fig. 3.

Clearly, the shape of this distribution cannot be fully captured by a single gamma distribution. However, the distribution appears to be such that a two-component mixture of gamma distributions within the second state may be sufficiently flexible to provide a good approximation to the nonparametric distribution. Notably though, a corresponding two-state model, with TPM $\Gamma = \left(\gamma_{ij}\right)$, $i, j = 1, 2$, and mixture weights α and $1 - \alpha$ for the two gamma distributions within state 2, is equivalent to a three-state gamma–HMM with suitably structured TPM,

$$\Gamma = \begin{pmatrix} \gamma_{11} & \alpha(1-\gamma_{11}) & (1-\alpha)(1-\gamma_{11}) \\ (1-\gamma_{22}) & \alpha\gamma_{22} & (1-\alpha)\gamma_{22} \\ (1-\gamma_{22}) & \alpha\gamma_{22} & (1-\alpha)\gamma_{22} \end{pmatrix}.$$

Therefore, a two-state HMM with a gamma mixture in one of the states can be represented by a three-state simple gamma–HMM. It is thus possible that model selection criteria favor models with more than two states, not because there are more than two genuine (behavioral) states, but because with the additional states it is possible to represent more flexible emission distributions.

Scenario 3 (temporal variation) In many ecological time series, there are clear temporal patterns in the data. We use diel (24-h period) variation as an example, but the issue applies to any temporal resolution with variation in the data (e.g., seasonal). Diel patterns could, for example, be present in the transition probabilities, with the corresponding parameters then being cyclic functions of time. The scenario we consider here is a hypothetical setting with a nocturnal animal that is more likely to be active at nighttime than during the day. The state-switching probabilities will then depend on the time of day. We constructed the transition probabilities using trigonometric functions, with a possible state switch occurring every 15 min. The resulting transition probabilities, as a function of time of day, are shown in Fig. 4.

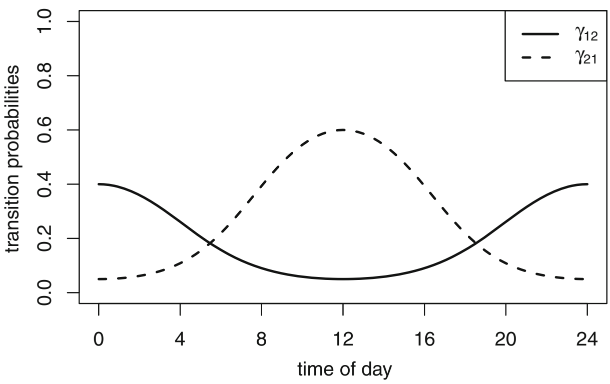

Figure 4. Transition probabilities as a function of time, as implemented in Scenario 3.

In the movement ecology literature, analyses including such temporal components were presented, for example, in Towner et al. (2016) and in Li and Bolker (2017). Such a temporal structure in the data, if neglected, may to some extent be captured by additional hidden states.

Scenario 4 (individual heterogeneity) When observing more than only one individual, it is natural to assume the individuals to differ, *inter alia*, in their personality and fitness. In the context of animal movement, one could imagine younger individuals to move faster when in an active state, or to occupy certain behavioral states more often, than older individuals. In the former case, the natural modeling approach for such a setting would be to consider individual-specific parameters of the gamma distribution within the active state. For parsimony in terms of the number of model parameters, random effects would typically be used.

Not accounting for such individual heterogeneity within the model formulation, hence assuming identical within-state gamma distributions across all individuals, could again lead to information criteria favoring models with more than two states: for example, one of the resulting >2 nominal HMM states may be capturing the active movement of fitter individuals, while another may be associated with the active movement of less fit individuals, with at least a third state capturing the inactive movement behavior. In our simulations, we simulated 10 animal tracks of length 500 each, using a log-normal distribution with the parameters $\mu = \ln 4$ and $\sigma = 0.15$ to generate individual means for the (track-specific) gamma emission distribution within the second state.

Scenario 5 (semi-Markov state process) A first-order Markov chain implies that the times spent within a state are geometrically distributed. For example, in the baseline model, the probability mass function (PMF) for the time k spent in a state (either 1 or 2) is

$$p(k) = 0.1 \cdot 0.9^{k-1}, \quad k = 1, 2, 3, \ldots . \tag{1}$$

In particular, the mode of the PMF of the dwell time is at $k = 1$. This implicit consequence of the Markov property will clearly be inappropriate in some applications. To give just one example, in Langrock et al. (2014) it was shown that the distribution of the time beaked whales spend foraging at the bottom of a deep dive is substantially different from a geometric distribution.

Hidden semi-Markov models are designed to overcome this limitation of HMMs, by explicitly specifying a state dwell-time distribution (e.g., a Poisson or a negative binomial), at the cost of a considerable increase in computational effort. Notably, any given semi-Markov state process can be arbitrarily accurately represented using a (first-order) Markov state process with expanded state space (Zucchini et al. 2016). This implies that when a semi-Markov structure is ignored in the model formulation, then model selection criteria can be expected to point to models with larger number of states than there are genuine (biological) states, with the model states structured such that the semi-Markov structure is reflected.

The synthetic data in this simulation experiment are generated by the baseline model described above, but replacing the geometric dwell-time distribution within state 2, as given in (1), by a Poisson distribution with mean $\lambda = 3$, shifted by one.

Scenario 6 (second-order state process) A similar, yet conceptually different modification of the Markov assumption is to consider higher-order Markov chains for the state process (Zucchini et al. 2016), thus allowing, at any point in time, the future state to depend not only on the present but also on one or more past states. An example application to eruption times of the Old Faithful Geyser is given in Langrock (2012).

Similarly as in case of hidden semi-Markov models, HMMs with underlying higher-order Markov state processes can equivalently be represented as HMMs with first-order Markov state processes with extended state space (Zucchini et al. 2016). If higher-order memory is neglected in the model formulation, then with the identical reasoning as in the previous scenario, we would expect model selection criteria to favor models with overly complex state architectures.

To demonstrate this issue, the data in this scenario were generated from a second-order Markov chain, determined by the following (time-homogeneous) state-switching probabilities:

$$P(S_t = 2 \mid S_{t-1} = 1, S_{t-2} = 1) = P(S_t = 1 \mid S_{t-1} = 2, S_{t-2} = 2) = 0.25;$$
$$P(S_t = 2 \mid S_{t-1} = 1, S_{t-2} = 2) = P(S_t = 1 \mid S_{t-1} = 2, S_{t-2} = 1) = 0.05.$$

This means that switching the state after just having entered it is less likely than when having already stayed in the state for $k > 1$ time units.

Scenario 7 (violation of conditional independence assumption) Conditional independence of the observations, given the states, is one of the key assumptions made in the basic HMM formulation presented in Sect. 2.1. This assumption is violated if there is additional correlation in the observed time series *within a state*. To demonstrate the consequences of not accounting for corresponding structure in the model formulation, in this simulation scenario we consider time-varying mean parameters of the state-dependent gamma distributions, generated using autoregressive processes of order 1, each of them with fairly strong persistence.

Figure 5 depicts an example sequence of mean parameters generated in this way, showing only the first 200 time points for clarity. Within state 1, the mean of the gamma distribution

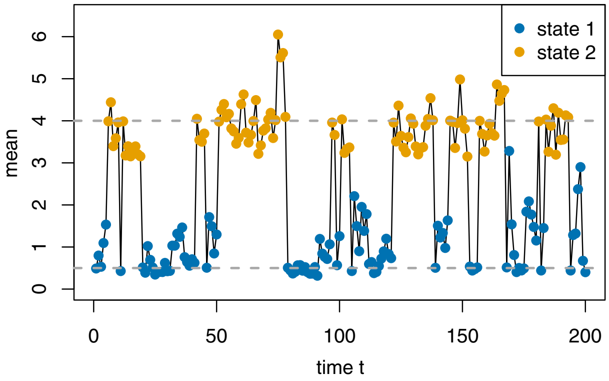

Figure 5. One example (sub-)sequence of gamma mean values generated in Scenario 7.

fluctuates around 0.5, while within state 2 it fluctuates around 4. Thus, the state now determines the average level of the mean, but the exact value of the mean depends not only on the state but also on its previous value. For example, in the context of animal movement, the step length of an animal may depend not only on the current behavioral state, but also on the previous speed/step length. Especially at fine temporal resolutions, this would in fact be expected as there will be a certain momentum in the movement. While such fine resolutions are rarely seen with GPS data, they are nowadays commonly seen in analyses of accelerometer data, which can also be tackled using HMMs (Leos-Barajas et al. 2016).

With the same rationale as outlined in Scenario 4—just replacing heterogeneity across individuals by temporal heterogeneity within individuals—it is intuitively clear that, when the additional correlation is not incorporated in the model, model selection criteria may favor models with more states than adequate.

Scenario 8 (benchmark, correct model specification) As a benchmark, we also consider a scenario where the data were generated from exactly the baseline model.

3.2. Simulation Results

For each scenario, 100 data sets were generated with $T = 5000$ observations each. Within a given scenario, to each of these 100 data sets, simple gamma–HMMs were fitted, which do not take into account the modification implemented. These slightly misspecified models, with 2–5 states, were fitted to the simulated data sets using numerical maximum likelihood.

For each scenario, Table 1 displays the proportions of the 100 runs in which models with 2, 3, 4 or 5 states were favored by AIC/BIC/ICL, respectively. Both AIC and BIC mostly overestimated the number of states that were used to generate the artificial data, in all scenarios that involve model misspecifications. The AIC performs worse than the BIC due to the smaller penalty on model complexity (consistent with the results of Celeux and Durand 2008). The performance of order selection based on AIC and BIC clearly depends on the severity of the assumption violation. For example, if there were less extreme outliers in Scenario 1, then of course the performance of AIC and BIC with respect to order selection would not be as bad as reported here, and similarly in the other scenarios.

Table 1. Percentages of runs in which the models with 2–5 states are chosen by AIC, BIC and ICL, for all simulation scenarios.

Simul. scenario	Criterion	Number of hidden states selected			
		2 (%)	3 (%)	4 (%)	5 (%)
1 (outliers)	AIC	–	47	49	4
	BIC	30	70	–	–
	ICL	58	42	–	–
2 (inadequate emission distribution)	AIC	–	27	60	13
	BIC	–	100	–	–
	ICL	26	71	3	–
3 (temporal variation)	AIC	–	–	57	43
	BIC	14	84	2	–
	ICL	100	–	–	–
4 (individual heterogeneity)	AIC	3	13	77	7
	BIC	79	15	6	–
	ICL	96	1	3	–
5 (semi-Markov state process)	AIC	–	14	74	12
	BIC	–	100	–	–
	ICL	100	–	–	–
6 (second-order state process)	AIC	–	1	90	9
	BIC	2	98	–	–
	ICL	92	2	3	3
7 (violation of cond. indep. assumption)	AIC	–	–	28	72
	BIC	5	95	–	–
	ICL	100	–	–	–
8 (benchmark)	AIC	37	43	20	–
	BIC	100	–	–	–
	ICL	100	–	–	–

In the scenarios considered, the ICL performed very well. This can be explained by the tendency of the ICL to favor non-overlapping solutions, that is HMMs where the state-dependent distributions are clearly distinct. By virtue of the way the state-dependent distributions were defined, this behavior is appropriate in most of the scenarios considered. The obvious exceptions are Scenarios 1 and 2—in the former case due to a third, distinct state-dependent distribution to account for the fairly extreme outliers, and similarly in the latter case, with a third state-dependent distribution accounting for the right tail of the distribution within state 2.

3.3. Discussion of the Simulation Results

In this paper, we focus on analyses where the interest lies in identifying the number of genuine states of the data-generating process, for example due to a focus on drawing ecological inference related to the state process. When faced with the choice between a model with the correct number of states, but a lack of fit in the emission distributions, and a model with too many states that does fit the data well—cf. Scenarios 1 and 2 above, comparing models with two vs. such with three states—we argue that in practice it may then be preferable to choose the former. Of course, the ideal solution would be to re-formulate the model with the correct number of states such that it does fit the data well (see Sect. 4.3). There are other situations, for instance those where the aim is to forecast future observations, where it would be preferable to have an essentially incorrect number of states yet a better model fit. In the Supplementary Material, we further investigate the trade-off between identifying the correct number of states and obtaining unbiased estimators of the emission distributions.

In practice, it may be the case that individual assumption violations are much less dramatic than those presented in our simulation setups. On the other hand, with complex empirical data, we will usually be facing more than just one violation of the assumptions involved in the basic HMM formulation. For example, in the application of HMMs to blue whale dive data reported in DeRuiter et al. (2016), there were indications of a minor violation of the assumption of contemporaneous conditional independence, strong individual heterogeneity and a minor lack of fit of the emission distributions. These deviations of a basic HMM being fitted from the true data-generating process may effectively accumulate, such that order selection may be at least as problematic, if not more, than with just a single, yet stronger assumption violation.

Conceptually, it is of course possible to modify the basic HMM, incorporating the additional structure in the model formulation, before tackling the problem of order estimation. For example, in Scenario 2 the reason for the inclusion of additional states, as per recommendation of the model selection criteria considered, is simply the insufficient flexibility of the state-dependent gamma distribution to adequately capture the observations generated within state 2. Such a mismatch between the distributional family employed and the empirical distribution can be detected using residual analyses, as discussed in Sect. 4 below. A natural and easy-to-implement remedy would then be to use a finite mixture distribution as emission distribution within state 2. Such a model was implemented, for example, in Leos-Barajas et al. (2016), in that case effectively merging two states associated with low activity of eagles. Furthermore, for example GPS measurement error can be accounted for within the model formulation (Patterson et al. 2016), mixed HMMs can be used to accommodate individual heterogeneity (Zucchini et al. 2016), and semi-Markov or second-order state processes can be implemented to better capture the dependence structure (Langrock 2012). When feasible, then improving the model formulation to overcome any substantial lack of fit should be the gold standard. However, each of these modifications is technically challenging and thus difficult to realize for practitioners, and corresponding models are also much more demanding to fit computationally. When viewed in isolation, then each of these extensions will usually still be computationally feasible. However, simultaneously addressing several such patterns will in general be infeasible. Thus, while conceptually it would seem to be most natural to simply overcome the limitations of HMM formulations that cause criteria-based order selection to fail, this is not always a useful strategy in practice. Rather than spending considerable time and effort on technically challenging non-standard HMM formulations, practitioners will probably seek more pragmatic, goal-oriented ways to overcome the caveats of information criteria in the context of order selection.

Finally, while overall positive in the particular scenarios considered, the focus of the ICL on clear separation of the state-dependent distributions is not desirable in general. For example, routine movements of most mammals, that is, movements occurring during an animal's daily activities, can be dissected into three primary behavioral modes: resting, foraging and traveling. While a traveling mode will typically imply movement patterns clearly distinct from those in the other two modes, it is intuitively clear that the movement metrics (e.g., step length) observed in resting and foraging modes, respectively, can actually be very similar, which would typically translate to associated state-dependent distributions that do clearly overlap. In corresponding analyses, both within ecology and in other settings, the sensitivity of the ICL with regard to overlapping state-dependent distributions may hinder inference on genuinely distinct modes. In Supplementary Material, we provide two additional simulation experiments (Scenarios 9 and 10) showing biologically realistic situations where the ICL performs much worse than the BIC due to overlapping emission distributions. Since the general behavior of the ICL criterion is still poorly understood, we do not recommend its uncritical use in practice.

4. PRAGMATIC ORDER SELECTION

Given the difficulties outlined above, we suggest the following pragmatic step-by-step approach to selecting the number of states of an HMM:

Step 1 decide a priori on the candidate models, in particular the minimum and the maximum number of states that seem plausible, and fit the corresponding range of models;

Step 2 closely inspect each of the fitted models, in particular by plotting their estimated state-dependent distributions and by considering their Viterbi-decoded state sequences;

Step 3 use model checking methods, in particular residual analyses, to obtain a more detailed picture of the fitted models, and to validate or invalidate any given candidate model;

Step 4 consider model selection criteria for guidance as to how much of an improvement, if any, is obtained for each increment in the number of states;

Step 5 make a pragmatic choice of the number of states taking into account findings from Steps 2–4, but also the study aim, expert knowledge and computational considerations;

Step 6 in cases where there is no strong reason to prefer one particular model over another (or several other) candidate model(s), results for each of these models should be reported.

The proposed strategy applies only to the unsupervised learning case (cf. Sect. 2.2). The exact procedure we suggest within each step, and the underlying rationale, is detailed below.

4.1. Step 1: Deciding on Candidate Models

Regarding *Step 1*, it is good practice, not only for HMMs, to restrict model selection only to those candidate models that are plausible a priori. Considering additional, implausible models increases the likelihood of an undesired selection bias, where, roughly speaking, a model is selected not by merit but because it got lucky, with the data at hand giving a more favorable picture of the model than would be obtained if more data were available (Zucchini 2000). When applying HMMs in movement ecology in particular, we have experienced that it is seldom useful to consider models with more than four states, (a) because the biologists, who know their study species extremely well, typically expect 2–4 behavioral states to be present, and (b) because models with more than four states typically turn out to be difficult to interpret.

4.2. Step 2: Inspecting the Fitted Models

The aim of *Step 2* is to develop an understanding of the key patterns picked up in the data by fitted candidate models, and how these relate to biological expectations and the study aim. This often goes a long way in helping to make an informed choice on the number of states. For example, consider the standard HMM formulation that is nowadays routinely applied in movement ecology, where each state is associated with a distinct correlated random walk movement behavior (Morales et al. 2004). In these settings, the two-state models almost always exhibit the same key pattern, with one state associated with large step lengths and small turning angles (sometimes labeled the “exploring” or “traveling” state) and the other state associated with much shorter step lengths and many more reversals (the “encamped” or “foraging” state). It is then usually interesting to see what happens when a third state is included. In many cases, this will lead to either the “encamped” state or the “exploratory” state splitting up into two states. In the former case, the two states resulting from splitting the “encamped” state could, for example, correspond to “foraging” and “resting” states, respectively. (We reiterate at this point that these interpretations are not to be taken too literally, as the HMM states are in general not going to correspond exactly to behavioral states.) When further increasing the number of states, it could for example happen that a state is split, but there is no biological reason to distinguish the resulting two states (cf. Scenario 2 in Sect. 3), or that the key structure of the model is unchanged, with the additional state explaining only a handful of observations (cf. Scenario 1 in Sect. 3). In general, it is our experience that the more states are included, the more difficult it becomes to assign biological meaning to the states. Overall, the purpose of *Step 2* thus is to get an overview of the suitability of the models, in relation to biological expectations and intuition, to the study aim, but also to each other (i.e., what additional feature of the data is explained by the model with $N + 1$ states that cannot be explained by the model with N states).

4.3. Step 3: Model Checking

The main purpose of *Step 3* is model validation, that is the assessment if any given candidate model adequately represents the data-generating process. Validation of HMMs via model checking is covered in detail in Chapter 6 in Zucchini et al. (2016), such that here we focus on the investigation of the role of the number of states. In particular, in order to make an informed choice of the number of states, it is important to understand what causes the potential preference for models with many states.

In the context of HMMs, a model check based on pseudo-residuals (Patterson et al. 2009) could, for example, reveal that the right tail of the empirical step length distribution can be better captured by a three-state model than a two-state model. This could be due to inflexibility of the state-dependent distribution assumed, rather than a genuine, that is biologically meaningful, third state responsible for the most extreme step lengths. As a second example, the empirical distribution of the data may be captured accurately by a two-state model, yet the residuals obtained for the two-state model exhibit much stronger autocorrelation than those for the three-state model. This could be an indication of a violation of the dependence assumptions.

In case of an identification of weaknesses in the model within *Step 3*, it needs to be decided whether or not any assumption violation ought to be addressed as part of the model formulation. For example, more flexible distributional families can be employed (e.g., a mixture). When there is only a single such problem to overcome, then it will often be feasible to formulate and fit corresponding more complex models. However, when dealing with complex time series, there is usually a bit of everything: a minor lack of fit caused by inflexible state-dependent distributions, a correlation structure not fully captured by a first-order Markov chain, heterogeneity which is not fully accounted for, etc. Simultaneously, addressing all these features in the data within an HMM will very quickly lead to heavily parameterized models, the estimation of which might be very unstable, if feasible at all. In those cases, it will often be preferable to use a simpler, more stable model which ignores features that *are not pertinent to the aim of the study*. For example, if the focus lies on the effects of environmental covariates on the state-switching dynamics, then a minor lack of fit in the marginal distribution of the observations may not make any difference.

4.4. Step 4: Considering Model Selection Criteria

As demonstrated in Sect. 3, model selection criteria can be misleading when it comes to order selection. Nevertheless, *Step 4* should be implemented to get an overall assessment of the candidate models. If the improvements in say the BIC are large when increasing the number of states, then this could be an indication that the additional states are indeed required. However, it may just as well be an indication that the additional states merely absorb some structure missing in the model, but have no clear (biological) meaning. A large difference in AIC, BIC, ICL or any other criterion alone does not prove that the model with the higher number of states is most suitable. However, if the inclusion of additional states substantially improves the AIC or BIC, then this does usually indicate problems of the simpler model, which should be investigated.

4.5. Step 5: Pragmatic Order Selection

At this point, a lot of information has been gathered which should facilitate the selection of an appropriate number of states. In our view, order selection should take into account:

- realism of the fitted candidate models, assessed using expert knowledge (as per *Step 2*);

- the results from model checking, in particular in relation to the study aim (*Step 3*);
- model selection criteria for guidance (*Step 4*);
- potentially computational considerations, if relevant.

As a consequence, the selection of the number of states will necessarily be somewhat subjective. However, we have experienced that a thorough implementation as detailed in *Steps 1–4* will usually make it fairly easy to pick a suitable N. First of all, the inspection of the fitted models within *Step 2* will often leave only two or perhaps three candidate models as reasonable contestants (cf. Sect. 5). Together with thorough model checks as implemented within *Step 3*, the advantages and disadvantages of the remaining contestants often become apparent, such that a final pragmatic choice of N can be made. In movement ecology, pragmatic order selection was effectively implemented for example in Morales et al. (2004) and DeRuiter et al. (2016).

As in any other statistical modeling exercise, order selection in HMMs can of course be an iterative process. That is, after going through *Steps 1–4*, it may become clear that alternative model formulations need to be considered, for instance modifying the dependence assumptions, such that one needs to return to *Step 1*.

4.6. Step 6: Reporting of Several Models

While it will sometimes be straightforward to make a pragmatic choice as described in *Step 5*, there will certainly also be cases where two or more candidate models may seem more or less equally suitable after following the steps detailed above. In those cases, it is our view that best scientific practice is to report the results of all suitable models (as recommended also by Burnham and Anderson 2002). In the context of order selection for HMMs, this translates to acknowledging uncertainty and presenting results accordingly. However, at least within movement ecology, this is hardly ever done (but see Morales et al. 2004; DeRuiter et al. 2016).

5. CASE STUDY: MUSKOX MOVEMENT

We demonstrate the workflow of our suggested pragmatic approach to order selection using movement data collected for a single adult female muskox in east Greenland, which was observed for a period of nearly 3 years. The raw data set comprises $T = 25103$ hourly location observations obtained using GPS collars (including about 1% missing locations). These were used to calculate the hourly step lengths and turning angles, to which we fitted fairly basic HMMs, using gamma distributions for the step lengths and von Mises distributions for the turning angles.

Following Sect. 4, we first ought to decide a priori on the minimum and maximum number of states that seem plausible (*Step 1*). The most dominant behavioral states are expected to be “resting/ruminating,” “feeding” and “moving.” Thus, from a biological perspective, three or four behavioral states for muskoxen seem most reasonable (Schmidt et al. 2016). Since it is not clear if exactly these behavioral modes will manifest themselves for the given data set,

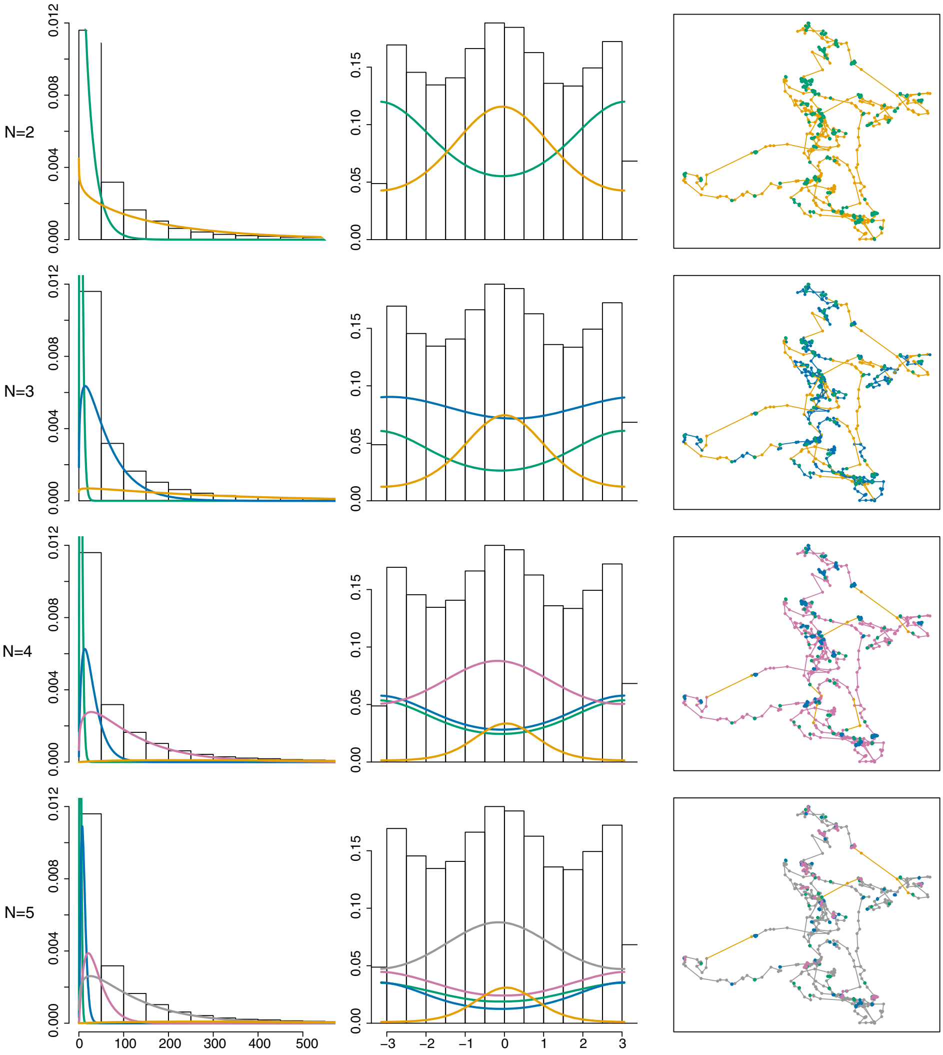

Figure 6. Estimated state-dependent distributions for models with 2–5 states (*one row* for each model, gamma step length distributions in the *left column*, von Mises turning angle distributions in the *middle column*), and associated decoded state sequences (*right column*). The state-dependent distributions are weighted with the proportion of time the corresponding state is occupied according to the stationary distribution of the Markov chain.

we included models with 2–5 states in our candidate set. Parameter estimation was carried out using maximum likelihood as implemented in the R package moveHMM (Michelot et al. 2016).

For *Step 2* (inspecting the fitted models), Fig. 6 displays the emission distributions estimated for the step lengths and the turning angles, respectively, as well as a Viterbi-decoded sequence of states for a subset of observations. For the two-state model, the fitted distributions exhibit the standard pattern typically found with these models, with one state involving large steps and directed movement and the other state much smaller steps and many reversals.

When adding a third state, the state associated with the smallest steps now involves hardly any movement (Fig. 6). This corresponds well to the (biological) expectation of a "resting"/"ruminating" state. With the low-activity state focusing on this virtually stationary behavior, the other two states now provide a more nuanced differentiation of active movement behavior. More precisely, one of the more active states involves relatively long steps (mean $\approx$ 286 m) on average and few turnings (directed movement), while the other active state involves moderate steps (mean $\approx$ 59 m) and many turnings. The former pattern is strongly indicative of a movement mode without any clear foraging activity, while the latter suggests foraging behavior with small localized search movements. These results make biological sense and agree with previous empirical findings (Schmidt et al. 2016).

The four-state model can most easily be compared to the two-state model described above. Roughly speaking, when going from two to four states, both the more active and the less active modes are split into two separate states. While the split of the inactive state seems biologically sensible (see above discussion), it is much less clear if the distinction between "moving" and "long-distance traveling" is really necessary and useful. As discussed in Sect. 3.3, in such a situation it would be natural to try to effectively merge these states via the use of a more flexible emission distribution, for instance a mixture distribution.

Adding a fifth state does not add any clear value and muddles interpretation (Fig. 6). In the corresponding model, there are now two states which involve hardly any movement (with hourly step length means of 3.4 and 10.8 m, respectively), which could simply be artifacts of temporal variation in GPS measurement error (here expected to be ca. 10 m).

We restrict model checking (*Step 3*) to the pseudo-residuals of the step lengths, since turning angles are not as easily amenable to a residual analysis due to their circular nature (Langrock et al. 2012). Figure 7 shows the quantile–quantile plots of the pseudo-residuals and their autocorrelation functions (ACFs), for each estimated model. The pseudo-residuals of the two-state model indicate a fairly substantial lack of fit in both the upper and the lower tail, while the models with three states or more appear to provide a reasonable fit of the marginal distribution of the step lengths. However, for all models considered, the respective ACFs clearly indicate another source of lack of fit, namely that there is some diel variation in the data, which is not taken into account within our models. As we have seen in the simulation experiments (Sect. 3), this may already cause criteria to point to models with more states than biologically sensible. And indeed, the sample ACF of the residuals obtained in case of the five-state model shows a less strongly marked diel pattern than the ACF of the residuals for the two-state model.

For *Step 4* (considering criteria), Table 2 displays the AIC, BIC and ICL values for each model fitted. Both AIC and BIC favor the five-state model. In fact, the two are further improved when considering even more states (results not shown). This could indicate that there is much more structure in the data than assumed by a basic HMM as the ones fitted here. Indeed, muskox move in a highly seasonal and dynamic environment (the Arctic) where environmental conditions can change rapidly over time (e.g., weather patterns) and space (e.g., heterogeneity in availability of vegetative cover). The movement patterns of muskoxen are therefore too complex to be fully captured with a simple three-state HMM with strong dependence assumptions. Notably, the ICL does not point to the most complex model being fitted but to the four-state model.

At this point, there are two options, and it depends on the aim of the study which of the two should be pursued. First, it may be relevant to explicitly account for the diel patterns

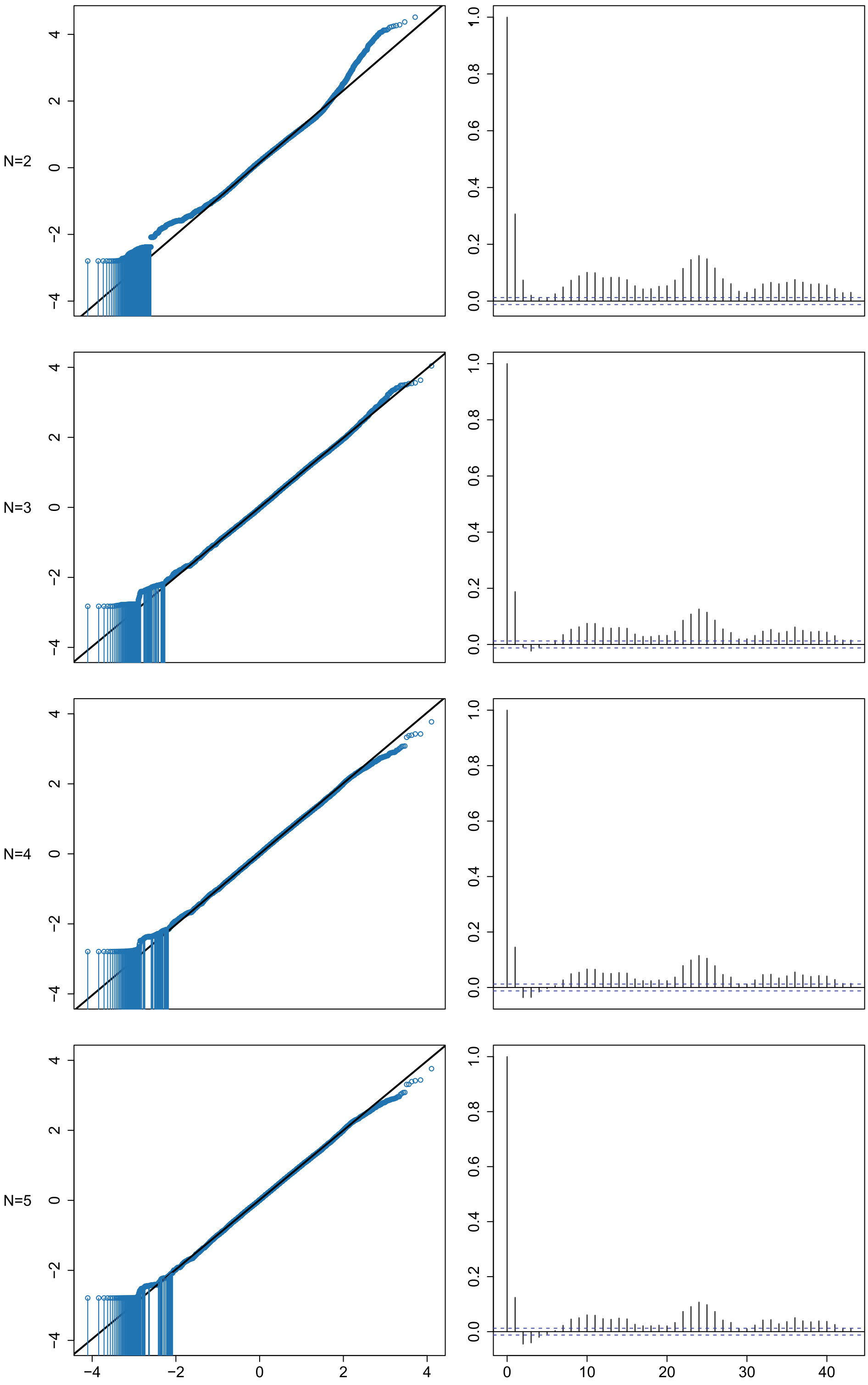

Figure 7. Quantile-quantile plots and autocorrelation functions of the pseudo-residuals obtained for the four different models fitted (with 2–5 states, from *top* to *bottom*) to the muskox movement data.

Table 2. AIC, BIC and ICL values obtained for the different models fitted to the muskox movement data.

No. states	No. parameters	AIC	BIC	ICL
2	12	350,199.3	350,296.7	354,829.3
3	21	345,285.4	345,455.8	351,544.5
4	32	343,404.9	343,664.6	**350,159.9**
5	45	**342,782.0**	**343,147.2**	351,247.7

The models selected by the different criteria are highlighted in bold face font.

exhibited by the muskox, say when investigating the state-switching dynamics in relation to internal and external drivers (Li and Bolker 2017). In that case, one needs to return to *Step 1* and formulate corresponding candidate models and then proceeding with *Steps 2–6*. It could also be worthwhile to investigate whether more flexible emission distributions, for instance mixtures, would substantially improve the fit of the models with only two or three states. Second, it may be the case that the diel variation and any minor lack of fit of the emission distributions can be neglected because it does not interfere with the study aim. For example, the primary interest may lie in identifying the spatial regions in which an animal is most likely to forage during a specific time window in which diurnal or environmental variation is low (i.e., high Arctic summer with 24 h of daylight and abundant vegetation). In such a case, whether or not the exact correlation structure of the state process is captured will likely have very little influence on the state decoding.

As this case study is supposed to merely illustrate the workflow of the pragmatic order selection approach suggested, we do not pursue the former of the two options. If the latter route is taken, then taking into account the findings from above, it is clear that either the three-state or the four-state model constitutes a pragmatic yet justified choice (*Steps 5, 6*).

6. DISCUSSION

In this paper, we have demonstrated why model selection criteria are problematic with respect to choosing the number of states of an HMM. More specifically, AIC and BIC tend to favor models with too many states, since any structure in the data that is neglected in the model formulation will, to some extent, be absorbed ("mopped up") by additional model states that do not have a clear interpretation anymore. Since the performance of model selection criteria strongly depends on the severity of any deviations of the HMM formulation from the *unknown* true process, we believe that no one-size-fits-all objective and universally applicable criterion can be developed for order selection in HMMs. The ICL criterion appears to overcome several of the problems associated with the more established AIC and BIC, yet it does not come without its own limitations, namely a sensitivity to overlapping state-dependent distributions. Therefore, we proposed a pragmatic step-by-step approach to order selection which, while lacking objectivity, we believe is the best possible practical solution.

We focused exclusively on inference based on maximum likelihood and did not investigate order selection under a Bayesian paradigm. In principle, the possibility to use priors to effectively exclude models with undesirably large numbers of states seems appealing.

Robert et al. (2000) proposed reversible jump Markov chain Monte Carlo algorithms in particular for choosing the number of states. Notably, those authors also find that "adding structure [...] to a model pushes the posterior distribution of k towards smaller values" (where k is the number of states), which indicates that the conceptual problem arises in both inferential frameworks. Overall, we do believe that the Bayesian framework may in principle offer opportunities for formalizing the concept of pragmatic order selection, but also that a certain level of subjectivity is unavoidable—to use priors to enforce a small number of states in our view effectively just shifts the problem.

Many of the key principles brought forward in this paper have previously been presented in the area of cluster analysis (Hennig 2015). In the context of finding the "true" number of clusters for a given data set, Hennig argues that there is a "misguided desire for uniqueness and context-independent objectivity." Similarly, there may also be different concepts of what constitutes a "true" state of an HMM. Corresponding definitions may be based on the data alone, on external a priori information, or on HMM fitting results, and are to be seen in relation to the study aim. As also pointed out in Hennig (2015), the researcher's modeling decisions need to be made transparent. While the pragmatic approach to order selection presented in this paper clearly depends on subjective decisions made by the researcher, a corresponding analysis nevertheless will be as scientific, if not more scientific, than any allegedly objective choice.

The somewhat problematic notion of a "true" state is in fact exacerbated within movement ecology, where the meaning of a state may strongly depend on the time interval the Markov chain operates on, and hence on the sampling protocol. Nevertheless, there is a strong desire within the community to use statistical models to infer actual behavioral states from tracking data. In those cases, pragmatic order selection will often lead to the choice of a model with a small number of interpretable states, at the expense of a lack of fit of the corresponding model, and potentially biased estimation in particular of the state-dependent distributions. The muskox case study illustrates this trade-off: here a model with five or more states, while superior in the goodness of fit, would not be useful in say determining drivers of resting and foraging behavior, simply because there is no clear correspondence between model states and behavioral states. In other situations, for instance when interest lies in forecasting, it can of course be preferable to use a larger model, essentially with too many states, which fits the data well.

Overall, the selection of the number of states clearly is an important yet challenging issue, which requires statistical expertise and modeling experience, but also a good understanding and intuition of the data and research question at hand. Within statistical ecology, this underlines the need for statisticians and ecologists to closely collaborate in all stages of an analysis.

[Received January 2017. Accepted May 2017.]

REFERENCES

Biernacki, C., Celeux, G. & Govaert, G. (2013), Assessing a mixture model for clustering with the integrated completed likelihood *IEEE Transactions on pattern analysis and machine intelligence*, **22**, 719–725.

Broekhuis, F., Grünewälder, S., McNutt, J.W. & Macdonald, D.W. (2014), Optimal hunting conditions drive circalunar behavior of a diurnal carnivore. *Behavioral Ecology*, **25**, 1285–1275.

Burnham, K.P. & Anderson, D.R. (2002), *Model Selection and Multimodel Inference*, Second Edition, Springer, New York.

Celeux, G. & Durand, J.-B. (2008), Selecting hidden Markov model state number with cross-validated likelihood. Computational Statistics, **23**, 541–564.

DeRuiter, S.L., Langrock, R., Skirbutas, T., Goldbogen, J.A., Calambokidis, J., Friedlaender, A.S. & Southall, B.L. (in press), A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *Annals of Applied Statistics*, **11**, 362–392.

Gneiting, T. & Raftery, A.E. (2007), Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Hennig, C. (2015), What are the true clusters? *Pattern Recognition Letters*, **64**, 53–62.

Langrock, R. (2012), Flexible latent-state modelling of Old Faithful's eruption inter-arrival times in 2009. *Australian and New Zealand Journal of Statistics*, **54**, 261–279.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D. & Morales, J.M. (2012), Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology*, **93**, 2336–2342.

Langrock, R., Kneib, T., Sohn, A. & DeRuiter, S.L. (2015), Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, **71**, 520–528.

Langrock, R., Marques, T.A., Baird, R.W. & Thomas, L. (2014), Modeling the diving behavior of whales: a latent-variable approach with feedback and semi-Markovian components. *Journal of Agricultural, Biological and Environmental Statistics*, **19**, 82–100.

Leos-Barajas, V., Photopoulou, T., Langrock, R., Patterson, T.A., Watanabe, Y.Y., Murgatroyd, M. & Papastamatiou, Y.P. (in press), Analysis of animal accelerometer data using hidden Markov models. *Methods in Ecology and Evolution*, **8**, 161–173.

Li, M. & Bolker, B.M. (2017), Incorporating periodic variability in hidden Markov models for animal movement *Movement Ecology*, 5, DOI:10.1186/s40462-016-0093-6.

Michelot, T., Langrock, R. & Patterson, T.A. (2016), moveHMM: An R package for analysing animal movement data using hidden Markov models. *Methods in Ecology and Evolution*, **7**, 1308–1315.

Morales, J.M., Haydon, D.T., Frair, J., Holsinger, K.E. & Fryxell, J.M. (2004), Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, **85**, 2436–2445.

Patterson, T.A., Basson, M., Bravington, M.V. & Gunn, J.S. (2009), Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, **78**, 1113–1123.

Patterson, T.A., Parton, A., Langrock, R., Blackwell, P.G., Thomas, L. & King, R. (2016), Statistical modelling of animal movement: a myopic review and a discussion of good practice. arXiv:1603.07511.

Pradel, R. (2005), Multievent: an extension of multistate capture–recapture models to uncertain states, *Biometrics*, **61**, 442–447.

Robert, C.P., Rydén, T. & Titterington, D.M. (2000), Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society Series B*, **62**, 57–75.

Schmidt, N.M., van Beest, F.M., Mosbacher, J.B., Stelvig, M., Hansen, L.H. & C. Grøndahl. (2016), Ungulate movement in an extreme seasonal environment: Year-round movement patterns of high-arctic muskoxen. *Wildlife Biology*, **22**, 253–267.

Schwarz, G. (1978), Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Stone, M. (1977), An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society Series B*, **39**, 44–47.

Towner, A., Leos-Barajas, V., Langrock, R., Schick, R.S., Smale, M.J., Jewell, O., Kaschke, T. & Papastamatiou, Y.P. (2016), Sex-specific and individual preferences for hunting strategies in white sharks. *Functional Ecology*, **30**, 1397–1407.

Zucchini, W. (2000), An introduction to model selection. *Journal of Mathematical Psychology*, **44**, 41–61.

Zucchini, W., MacDonald, I.L. & Langrock, R. (2016), *Hidden Markov Models for Time Series: An Introduction using R*, Second Edition, Chapman & Hall/CRC, Boca Raton.