moveHMM: An R package for the statistical modelling of animal movement data using hidden Markov models

3 authors, including:

Théo Michelot
University of St Andrews
31 PUBLICATIONS   584 CITATIONS

SEE PROFILE

Toby Alexander Patterson
The Commonwealth Scientific and Industrial Research Organisation
108 PUBLICATIONS   3,861 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Optimality of vertical migrations View project

Arctic cetaceans in a warming ocean: behavioural responses to climate change View project

moveHMM: An R package for the statistical modelling of animal movement data using hidden Markov models

Théo Michelot[1,2], Roland Langrock[3], Toby A. Patterson[4]

[1]*University of Sheffield, UK;* [2]*University of St Andrews, UK;* [3]*Bielefeld University, Germany;* [4]*CSIRO Oceans and Atmosphere, Australia*

Summary

1. Due to the substantial progress in tracking technology, recent years have seen an explosion in the amount of movement data being collected. This has led to a huge demand for statistical tools that allow ecologists to draw meaningful inference from large tracking data sets.

2. The class of hidden Markov models (HMMs) matches the intuitive understanding that animal movement is driven by underlying behavioural modes and has proven to be very useful for analyzing movement data. For data that involve a regular sampling unit and negligible measurement error, these models usually are sufficiently flexible to capture the complex correlation structure found in movement data, yet are computationally inexpensive compared to alternative methods.

3. The R package moveHMM allows ecologists to process GPS tracking data into series of step lengths and turning angles, and to fit an HMM to these data, in particular allowing for the incorporation of environmental covariates. The package includes assessment and visualization tools for the fitted model.

4. We illustrate the use of moveHMM using (simulated) movement of the legendary wild haggis *Haggis scoticus*. Our findings illustrate the role our software, and movement modelling in general, can play in conservation and management by illuminating environmental constraints.

Key-words: behavioural state; maximum likelihood; random walk; state-switching model; telemetry data

1 Introduction

Quality and quantity of animal movement data are rapidly increasing because of substantial improvements of telemetry technology made in recent years. Although increasingly sophisticated statistical methods are being developed in movement ecology (Patterson et al., 2016), the extensive analysis of large data sets is still a computational challenge, especially because the development of user-friendly software is lagging behind.

Among the variety of models that have been used to analyse animal tracking data, hidden Markov models (HMMs) have stood out in recent years because of their appealing combination

of model flexibility, clear interpretability and computational tractability (e.g. Franke et al., 2004; Holzmann et al., 2006; Patterson et al., 2009; Langrock et al., 2012; van de Kerk et al., 2015). The HMM approach to movement modelling is part of the wider family of state-switching models, which focus on the decomposition of the movement process into distinct underlying states. Such an approach matches our intuitive understanding that most animals' movement is driven by switches in underlying behavioural modes.

Several general packages exist for fitting HMMs to data, including depmixS4 (Visser and Speekenbrink, 2010), HiddenMarkov (Harte, 2015) and msm (Jackson, 2011). However, these packages are under-utilized by the ecological community, because they do not offer an easy workflow to process and analyse animal movement data. Other R packages can be used for the analysis of animal tracking data, such as adehabitatLT (Calenge, 2006), crawl (Johnson, 2015), and ctmm (Fleming and Calabrese, 2015), but none of the models that they implement focuses on drawing inference on the animals' behavioural state-switching mechanisms. The package bsam (Jonsen et al., 2013) does include multistate models, but it focuses on the situation where location error is significant. This is a difficult problem requiring detailed and computationally challenging analysis solely for the purpose of inferring switching and most likely movement paths. For high precision data, much more flexible and efficient use of HMMs is possible, yet until now, there has been no dedicated R package for corresponding movement data.

The new package moveHMM for the R environment (R Core Team, 2015) implements HMMs for the analysis of animal movement data. The stable version of the package is available on the Comprehensive R Archive Network (CRAN), at `https://cran.r-project.org/package=moveHMM`, under a license GNU GPL 3, and its latest version – as well as the unit tests used during development – can be found on the Github repository of the project, at `https://github.com/TheoMichelot/moveHMM`. Additional information on the functions included in the package can be found in its vignette. There, we also provide the code for analyzing the elk data set from Morales et al. (2004).

In this paper, we briefly describe HMMs and explain how they can be used to draw inference from animal movement data. Subsequently, we give an overview of the features of the package moveHMM, and demonstrate them on an example using (simulated) wild haggis movement data.

2 Hidden Markov models for animal movement

An HMM is a time series model comprising two processes: the series of (possibly multivariate) observations, $\mathbf{Z}_1, \ldots, \mathbf{Z}_T$, and an underlying non-observable state sequence $S_1, \ldots, S_T$. In the basic model formulation, the process $\{S_t\}$ takes values from $\{1, \ldots, N\}$ and satisfies the Markov property – i.e. it is a finite-state Markov chain of first order. At any given time t, the realization of $\mathbf{Z}_t$ is assumed to have been drawn from one of N (possibly multivariate) component distributions, as determined by the value of the state process at time t. This dependence structure is illustrated in Figure 1. In the context of animal movement, the underlying state process is often interpreted as a proxy for the behavioural state of the animal, e.g. more active ("exploratory") and less active ("encamped"), as in Morales et al. (2004).

Empirical movement data usually consist of time series of coordinates of the animal's locations on the plane, $(x_1, y_1), \ldots, (x_T, y_T)$, from which various movement metrics can be derived. In the standard HMM approach, the observation process is a bivariate time series with $\mathbf{z}_t = (l_t, \phi_t)$ comprising the step length l_t (the Euclidean distance between the locations (x_t, y_t) and (x_{t+1}, y_{t+1})) and the turning angle ϕ_t (the change of direction between the directions of travel during the intervals $[t-1, t]$ and $[t, t+1]$, respectively) of the animal. For the HMM approach to be applicable to animal
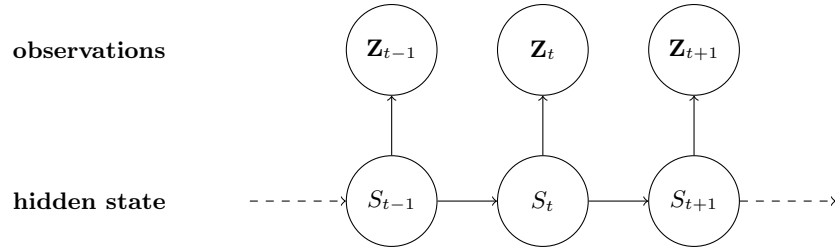
Figure 1: Dependence structure in a hidden Markov model.

movement data, and in particular in order to use the package moveHMM, the spatial measurement error in the data needs to be negligible relative to the magnitude of relocation activity observed. In addition, the data should be collected such that the observations are sampled on a regular basis (usually at regular time intervals, though there can be other regular sampling units, e.g. 'a dive' if positions are observed each time a marine mammal comes to the sea surface; see, e.g., DeRuiter et al., 2016). Data that are missing at random on an otherwise regular sequence can easily be handled.

The state process $\{S_t\}$ is characterized by its state transition probabilities and, less importantly, by its initial state distribution. The transition probability matrix of an N-state HMM, $\mathbf{\Gamma}^{(t)}$, is defined as $\mathbf{\Gamma}^{(t)} = (\gamma_{ij}^{(t)})$, where $i, j = 1, \ldots, N$ and $\gamma_{ij}^{(t)} = \Pr(S_{t+1} = j | S_t = i)$. In some cases, the state transition probabilities will be constant over time, i.e. $\gamma_{ij}^{(t)} = \gamma_{ij}$ for all t, but more often it will be of interest to relate the transition probabilities to time-varying environmental (or other) covariates (see, e.g., Dragon et al., 2012; McKellar et al., 2015). The transition probability at time t, $\gamma_{ij}^{(t)}$, can be linked to the associated p covariates, $\omega_1^{(t)}, \ldots, \omega_p^{(t)}$, via a multinomial logit link:

$$\gamma_{ij}^{(t)} = \frac{\exp(\eta_{ij}^{(t)})}{\sum_{k=1}^{N} \exp(\eta_{ik}^{(t)})},$$

where

$$\eta_{ij}^{(t)} = \begin{cases} \beta_0^{(ij)} + \sum_{l=1}^{p} \beta_l^{(ij)} \omega_l^{(t)} & \text{if } i \neq j; \\ 0 & \text{otherwise.} \end{cases}$$

For a given set of model parameters, the likelihood of the data can be calculated using a recursive algorithm (known as the forward algorithm), which in a very effective way considers all possible state sequences that might have given rise to the observed time series, exploiting the dependence structure of the model. (In fact, the availability of this and other recursion techniques is probably the key reason for the popularity of HMMs.) This makes numerical maximization of the (log-)likelihood, and hence maximum likelihood estimation, feasible in most cases. For a fitted model, one can easily obtain the most likely sequence of states and also examine various variables of interest such as the average rate of movement in each state. For a comprehensive account of the HMM machinery, we recommend Zucchini et al. (2016). Several extensions, with a focus on their usefulness in movement modelling, are discussed in Langrock et al. (2012). In movement ecology, HMMs have successfully been used to analyse the movement of, *inter alia*, caribou (Franke et al.,

2004), fruit flies (Holzmann et al., 2006), tuna (Patterson et al., 2009), panthers (van de Kerk et al., 2015), woodpeckers (McKellar et al., 2015) and white sharks (Towner et al., 2016).

3   Overview of the features of moveHMM

We now describe the main features of the package moveHMM. For further details on the implementation, we refer to the package's documentation and vignette (Michelot et al., 2015). The main functions of the package are listed in Table 1.

Function	Description
AIC.moveHMM	AIC of one or several models
CI	Confidence intervals for a model's parameters
fitHMM	Fit an HMM to data
plot.moveData	Plot summary graphs of movement data
plot.moveHMM	Plot summary graphs of a movement model
plotPR	Plot time series, qq-plots, and sample ACFs of pseudo-residuals
plotStates	Plot the (Viterbi-)decoded states and state probabilities
prepData	Preprocess tracking data
pseudoRes	Pseudo-residuals of a model
simData	Simulate movement data using an HMM
stateProbs	State probabilities
viterbi	Most likely state sequence (using the Viterbi algorithm)

Table 1: Main functions of the package moveHMM.

3.1   Data preprocessing – prepData

Tracking data usually consist of series of coordinates (either Easting-Northing or longitude-latitude). To fit an HMM using moveHMM, it is necessary to derive the series of step lengths and turning angles, as described in Section 2. This can be performed with the function prepData. The user needs to store the tracking data in a data frame, as typically output by the R functions read.csv or read.table, with a column for each coordinate. Both Easting-Northing and longitude-latitude coordinates can be used with the function prepData, as specified by an additional argument, "type". If longitude and latitude values are used, the step lengths are computed using the function spDistsN1 from the package sp (Bivand et al., 2013), and are measured in kilometers. If Easting and Northing coordinates are used, the step lengths are measured in the unit of the data. A column 'ID' must be provided if several animals were tracked; if 'ID' is not specified, all observations are assumed to come from a single track. All additional columns are considered as covariates, and should only contain numerical values. Categorical covariates can however also be considered: for a categorical covariate taking K levels, one simply needs to add $K-1$ dummy variables as columns to the data matrix, the j-th of which indicating whether or not the covariate takes the j-th level at the different observation times (such that level K is the reference category). The function prepData returns a data frame of the movement data: IDs, coordinates, steps, angles, and optionally covariates (if provided).

3.2 Model fitting – `fitHMM`

The function `fitHMM` fits an HMM to the data via numerical optimization of the likelihood using the R function `nlm`. The likelihood computation is based on the forward algorithm (Zucchini et al., 2016, Chapter 3), and is implemented in C++ for speed, using Rcpp (Eddelbuettel and Francois, 2011). To initiate the numerical maximization, starting values for the model parameters need to be specified. This choice is crucial, and the algorithm might not find the global maximum of the likelihood function, i.e. the maximum likelihood estimate (MLE), if the initial parameters are poorly chosen. The standard remedy is to run the maximization many times using a range of different starting values, possibly randomly selected. Such an approach gives an idea of the roughness of the likelihood surface and the reliability of the candidate MLE.

A variety of models can be fitted in moveHMM. Gamma, Weibull, log-normal and exponential distributions can be specified for the step lengths, and von Mises and wrapped Cauchy distributions can be specified for the turning angles. The choice of the family of state-dependent distributions can affect the goodness of fit (see, e.g., the comparison of different distributions in Langrock et al., 2012). Either the gamma or the Weibull distribution is usually used for the step lengths, while the von Mises and the wrapped Cauchy distribution are equally popular for modelling the turning angles (see Codling et al., 2008, for an introduction to these circular-valued distributions). Zero-inflation can be added to the state-dependent distributions of the step lengths to allow for steps of length zero (see, e.g., the analysis of woodpecker movement in McKellar et al., 2015). Corresponding distributions are mixtures of a point mass on zero and a strictly positive distribution (such as the gamma distribution). A model can also be fitted to one-dimensional data of step lengths only, using the option `angleDist="none"`.

It is possible to fit a model to a data set comprising several time series, possibly collected on several different animals. In such a case, all individuals are assumed to share the same movement model, i.e. the same parameters (an approach usually referred to as "complete pooling"; see Langrock et al., 2012). The package does not currently allow for fitting HMMs including random effects.

The state transition probabilities can be specified as functions of one or more covariates, as described in Section 2, using the argument 'formula' of the function `fitHMM`. This argument is an R model formula with default value `formula=~1`, corresponding to the case where no covariate is included. For a more complex example of the use of 'formula', see Section 4.

It is possible to fit a stationary HMM to the data, using the argument 'stationary' in `fitHMM`. If covariates are included in the model, then the transition probabilities are time-dependent such that the HMM cannot be stationary. Thus, this option can only be used with covariate-free models. By default, the initial distribution of the hidden Markov process is estimated; however, when `stationary=TRUE`, the initial distribution is chosen to be the stationary distribution $\boldsymbol{\delta}$ defined as the solution of $\boldsymbol{\delta} = \boldsymbol{\delta}\boldsymbol{\Gamma}$ subject to $\sum_{i=1}^{N} \boldsymbol{\delta}_i = 1$. Using the stationary distribution saves $N - 1$ model parameters, but is only justifiable when at the start of the time series the animal can reasonably be assumed to be in steady state (which, for example, would not be the case if the tagging interferes with the behaviour of the animal, and recordings are made immediately after the tagging). Unless the time series considered are very short, the choice between using the stationary distribution as initial distribution or estimating the latter has little practical relevance (Zucchini et al., 2016).

Model selection techniques can be used to choose an appropriate number of states of the HMM, or to determine if a covariate should be included in the model. The function `AIC` can be applied to one or several fitted movement models in moveHMM, returning the scores of the Akaike Information

Criterion (AIC) for the models considered, in ascending order. Users should not blindly follow such information criteria, especially with regard to the selection of the number of states – it is our experience that such formal model selection approaches tend to favour overly complex models, often to an extent such that selected models become challenging to interpret and very difficult to work with in practice (Langrock et al., 2015). The main reason for this is that, while very natural models for animal movement data, HMMs are still simplistic relative to the actual data-generating process underlying an animal's movement track. As a consequence, additional states are often included in a model to compensate for a simplistic model formulation, or to address artefacts such as outliers. While it is therefore difficult to make general recommendations on how to select the number of states, we advise to use criteria such as the AIC for guidance only, and additionally using both biological intuition and comprehensive model assessments as described in the subsequent section. For example, such investigation could reveal that a two state model may appear inferior to a three-state model due to a single outlier rather than a genuine pattern that requires a third state.

Exact confidence intervals for the MLEs of the parameters of an HMM are not available but, using asymptotic theory, they can be approximated based on the Hessian of the negative log-likelihood (i.e. the observed Fisher information). Given a fitted model, the function CI computes approximate confidence intervals for the model's parameters. Note that these are based on asymptotic normality results for the MLE and are not reliable for small sample sizes (Zucchini et al., 2016, Chapter 3).

3.3   Model checking – `pseudoRes`

Goodness-of-fit of a fitted HMM can be assessed using pseudo-residuals, which are also known as quantile residuals (Zucchini et al., 2016, Chapter 6). If the model is the true data-generating process, then the pseudo-residuals follow a standard normal distribution. For a model that describes the data well, the pseudo-residuals should be approximately normally distributed. Model adequacy can be checked by testing the pseudo-residuals for normality, or by considering qq-plots of the pseudo-residuals against the theoretical quantiles of the standard normal distribution. Patterson et al. (2009), Basson et al. (2012), Langrock et al. (2012) and Langrock et al. (2015) use pseudo-residuals in the context of animal movement modelling.

In the package moveHMM, the function `pseudoRes` returns one series of pseudo-residuals for the step lengths and one for the turning angles, given a fitted model. In addition, the function `plotPR` plots the time series of the pseudo-residuals, as well as qq-plots of the pseudo-residuals against the standard normal quantiles and the sample autocorrelation functions (ACFs) of the pseudo-residuals. As with regression models, in the qq-plots, a deviation from the identity line indicates a deviation from normality, and hence a lack of fit of the model. The sample ACFs can reveal residual autocorrelation, i.e. correlation in the data that was not captured by the model. The latter could be an indication that more states are required, or that a first-order Markov chain is not appropriate for describing the evolution of the states.

3.4   State decoding – `viterbi` and `stateProbs`

To analyse the process of behavioural state-switching, it is useful to decode the sequence of states of the underlying unobserved Markov chain most likely to have generated the observations, given a fitted model. In the context of HMMs, this sequence is computed using the Viterbi algorithm (Zucchini et al., 2016, Chapter 5), implemented in the function `viterbi` of the package moveHMM.

It is also possible to calculate the probabilities of the process having been in the different states, $1, \ldots, N$, at any given time point, given a fitted model. For an N-state HMM and an observed (bivariate) time series of length T, the function `stateProbs` returns a $T \times N$ matrix with the j-th element in row t giving the probability of the Markov chain having been in state j at time t. The most likely states and the state probabilities can be plotted for each animal using the function `plotStates`.

The decoded state sequences can be very useful in providing information on how animal behaviour varies in space. For example, in Towner et al. (2016), the decoded behavioural states of white sharks were superimposed on their movement tracks, which clearly highlighted spatial variation in the animals' behavioural patterns. Similarly, Lidgard et al. (2012) used HMMs to categorize movement data into behavioural states, and link foraging activity to at-sea associations of grey seals.

3.5 Model visualization – `plot`

A fitted model, as output by `fitHMM`, can be plotted with the generic function `plot`. This displays histograms of the step lengths and turning angles, above which are plotted the estimated probability density functions in each state. The densities are weighted by the relative frequency of each state in the most likely state sequence, decoded using the Viterbi algorithm. If covariates are included in the model, plots of the transition probabilities as functions of each covariate are also generated. If several covariates are included, then the transition probabilities are plotted against each covariate separately, fixing the remaining at their respective mean values. Finally, the `plot` function also outputs maps of the animals' tracks, color-coded according to the most likely state sequence.

Figures 3 and 4 illustrate the plots produced by `plot` when given a fitted model as an argument.

3.6 Simulation tool – `simData`

The function `simData` simulates movement data from an HMM, given the parameters provided. Like the preprocessing function `prepData`, it returns a data frame of the movement data. In particular, once a model has been fitted to real data with `fitHMM`, it can be passed as an argument to `simData`, so that the estimates of the model parameters are used in the simulation model. In this way, the patterns in the simulated movement can be compared to those of the real movement of the animal, corresponding to a visual assessment of the fit. While this does not constitute a formal check of a fitted model, in our experience this is a useful informal way to identify potential weaknesses of a given model formulation.

However, it is important to note that, in general, it is not sensible to use such a simulation engine for further inference, e.g. to try to predict the influence that future changes in the covariate values would cause on the animal's behaviour. This is because the simulations, even if conditional on estimated parameters, are essentially "data-free"; whereas in the fitted model, the data are crucial in constraining the state switches even when the strict structure of the model constitutes a limited description of the behavioural switching process. There are exceptions to this rule-of-thumb, but in general we urge caution in making inference from simulation results.

One area where simulations can be validly used, in addition to the informal model checking described above, is in a parametric bootstrap for quantifying estimator uncertainty.

4 Case study: (simulated) wild haggis movement

We illustrate the functionality of the package using wild haggis *Haggis scoticus* movement data. The data set involves movement tracks of 15 individual wild haggises, with 400 locations per track[1]. In the following we illustrate how the package moveHMM can be used to investigate the movement strategies adopted by the wild haggis in response to environmental conditions. Of particular interest is the role played by the animals' leg length. The wild haggis' left ipsilateral pair of legs is considerably longer than its right counterpart (King et al., 2007), an evolutionary adaptation enabling the animal to efficiently circle along topographic contours of the steep hills in the Scottish Highlands (though only clockwise, as seen from above[2]). In this illustrative example, we investigate how wild haggis movement strategies vary with respect to mountain slope. The underlying rationale for our analysis is to characterize an optimal topographic slope, or range of slopes, at which this species' anatomy allows for high movement speeds, while at other slopes the animals might have to resort to primitive crawling strategies. It seems obvious that net-energetic returns will be reduced in sub-optimal environments. Such predictions will be useful for reserve planning or other forms of spatial management for this species.

The 15 tracks, comprised of the locations and associated covariate values, are stored in a csv file 'rawHaggises.csv', and the code used to simulate the data is provided in the file 'simulate_haggis_data.R'. The code used to process and analyse the data with moveHMM is available as 'case_study.R'. All three files are provided in the supplementary material.

Having loaded the package moveHMM, we load and preprocess the data in R (version 3.2.0):

```
rawHaggis <- read.csv("rawHaggises.csv")
processedHaggis <- prepData(rawHaggis,type="UTM")
```

The resulting data frame 'processedHaggis' comprises the columns 'ID', 'step', 'angle', 'x', 'y', 'slope' and 'temp'. The latter two are time-varying environmental covariates giving the mountain slopes (in degrees) and the temperatures (in degree Celsius) associated with the observations. A subset of the data, namely the observations made for the haggis with 'ID=2', is visualized in Figure 2 – these graphics were obtained using the command `plot(processedHaggis)`. The haggis' characteristic (clockwise) circling can clearly be seen in the left plot.

Considering gamma step length distributions and von Mises turning angle distributions, the following models were fitted: $N = 2$ states, no covariates influencing the state transition probabilities ('m1', `formula=~1`); $N = 2$, linear predictor with covariate 'slope' ('m2', `formula=~slope`); $N = 2$, linear predictor with covariates 'slope' and 'temp' ('m3', `formula=~slope+temp`); $N = 2$, quadratic predictor with covariate 'slope' ('m4', `formula=~slope+I(slope^2)`). In each case, a common model was fitted to all animals, i.e. all individuals were assumed to share the same movement parameters.

On an octa-core i7 CPU, at 3.6 GHz and with 8 GB RAM, fitting the simplest model ('m1') took 10 seconds, while fitting the more complex model 'm4' took 53 seconds. From the given set of candidate models, the AIC (obtained with `AIC(m1,m2,m3,m4)`) strongly favours 'm4', and there is no indication that the temperature (covariate 'temp') affects the state-switching dynamics.

For the fitted model 'm4', Figures 3 and 4 display the estimated state-dependent distributions and the estimated effect of the covariate 'slope' on the state transition probabilities. These plots

---

[1]Because we encountered unexpected but insurmountable difficulties when trying to tag individuals in the wild, we resort to simulated data in this work, where we tried to generate movement patterns that we believe are realistic for wild haggises.

[2]The haggises may also move up and down altitudinal gradients, but only sideways to remain upright.
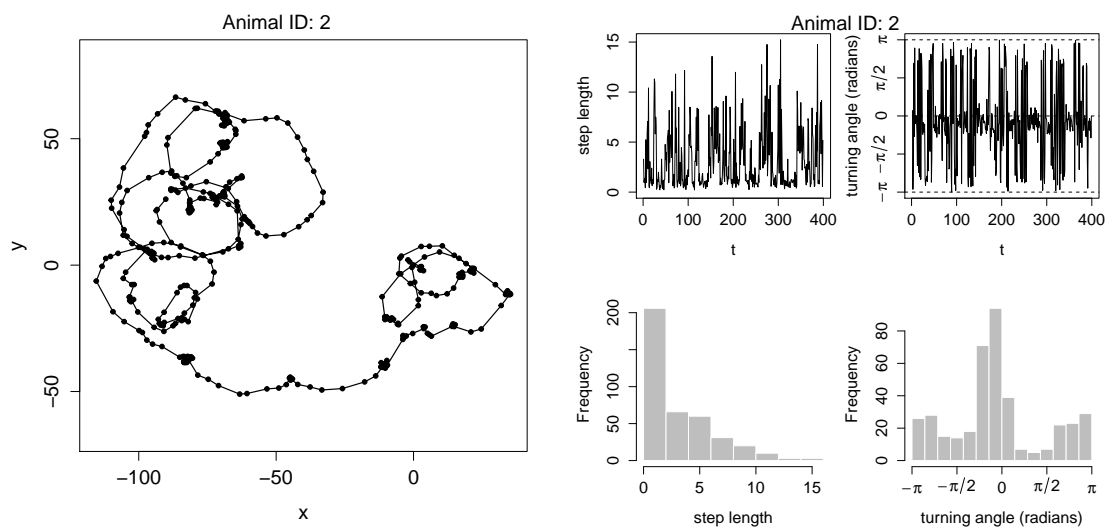
Figure 2: Haggis case study: example plot of the data object 'processedHaggis'. Note the obvious circling motion of the path which is likely due to asymmetric limbs.
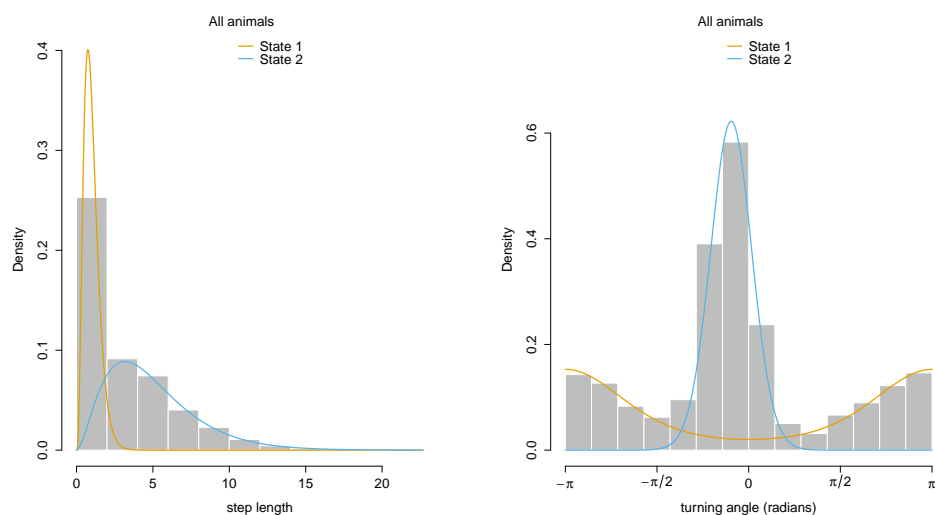

Figure 3: Haggis case study: State-dependent distributions in the fitted model 'm4'.

were obtained using the command `plot(m4)`. From Figure 3, we see that one state involves mostly short steps and many turnings (state 1) and the other state involves longer steps and less frequent

turnings (state 2). Thus, the haggises appear to exhibit the typical movement pattern that has previously been identified for various other species, including, *inter alia*, elk (Morales et al., 2004), fruit flies (Holzmann et al., 2006), woodpeckers (McKellar et al., 2015) and white sharks (Towner et al., 2016). Notably, the mean turning angle in state 2 was estimated as -0.299 – with 95% confidence interval $[-0.312, -0.287]$, obtained using `CI(m4)` – hence confirming the tendency of the haggises to circle clockwise (when active).
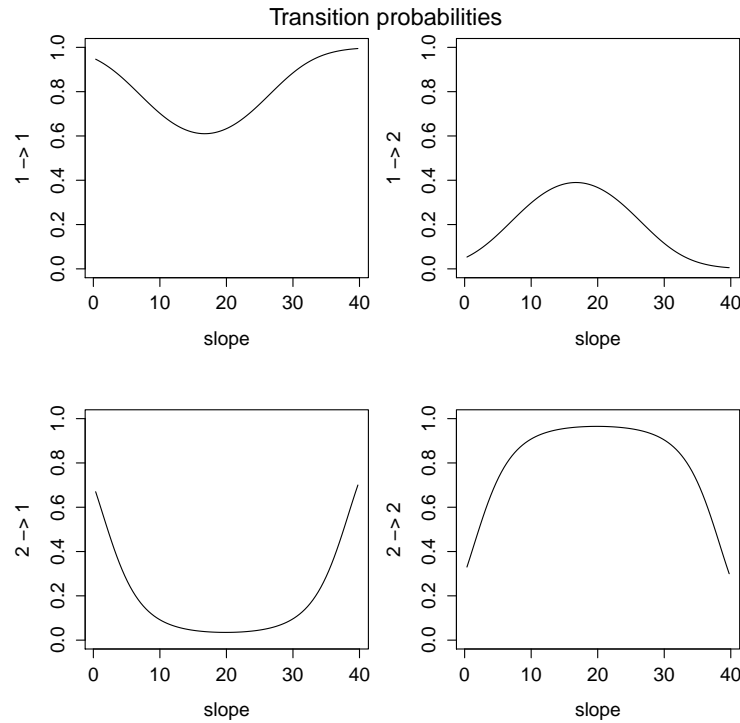

Figure 4: Haggis case study: Effect of the covariate 'slope' on the transition probabilities according to the fitted model 'm4'.

From Figure 4 it can be seen that there is a range of relatively moderate mountain slopes, roughly from $10°$ to $25°$, at which the haggises are highly likely to be in state 2 – due to both a high probability of a switch to state 2 when in state 1 and a high probability of remaining in state 2 when in state 2 – in which higher movement rates and more directed movement is observed. In contrast, when encountering slopes outside this range, i.e. either relatively flat terrain or very steep hillsides ($> 30°$), then the haggises are very likely to leave state 2 and spend prolonged periods in state 1, the state involving very little and undirected movement activity. A possible explanation is that the anatomy of this species only allows for efficient and hence fast movement at moderate slopes. In flat terrain, the haggises might have to resort to primitive crawling strategies, while in very steep terrain, they will likely fall over. State 1 would then encompass the haggises' clumsy attempts to get back to more favourable terrains, while state 2 captures the haggises' running

around the hills. According to the most likely state sequence under the fitted model 'm4', decoded using `viterbi(m4)`, the haggises spent 55.5% of the time running around the hills (i.e. in state 2). The map of the movement track for animal 2, colored by Viterbi-decoded states, is given in Appendix S1, part A.

Jarque-Bera tests for normality of the pseudo-residuals – obtained with `pseudoRes(m4)` – yield p-values of 0.75 and 0.65 for the residuals associated with the step lengths and the turning angles, respectively, such that normality is not rejected in either case. There is also no indication of residual autocorrelation (shown in Appendix S1, part B). The model 'm4' hence seems to fit the data well.

An example of a more extensive simulation study is provided in Appendix S1, part C, to assess the performance of parameter estimation and state decoding in moveHMM.

5   Conclusion and future improvements

We have developed the package moveHMM to facilitate the analysis of animal movement data with HMMs. Collection of such data is nowadays done routinely, and we hope that our package will make the very intuitive HMM framework accessible to applied researchers. Fitting HMMs can be computationally intensive if the likelihood calculation is coded inefficiently, and our implementation of the likelihood function should help increasing the utility of the HMM approach in movement ecology for large data sets. The current version can deal with very large data sets on standard desktop PCs (e.g. fitting a covariate-free HMM to a time series of a million observations in less than 20 minutes).

The package allows for statistical analyses of data indicating multiple phases of movement. More generally, the HMM formulation implemented in the package is suited to bivariate time series with one positive linear and one circular variable (step lengths and turning angles in case of animal movement data), driven by underlying non-observable states, and as such can also be used e.g. to analyze vessel monitoring data to characterize the spatial distribution of fishing effort (Peel et al., 2011; Charles et al., 2014), to model wind speed and direction (Holzmann et al., 2006) and to model wave heights and directions (Bulla et al., 2012).

The package is under further development, and we are considering the option of including additional features. In particular, several non-trivial extensions of the model formulation could potentially be implemented, such as random effects (Altman, 2007; McKellar et al., 2015), semi-Markov state processes (Langrock and Zucchini, 2011) or biased random walk behaviour (Langrock et al., 2012). However, implementing these at a high level of generality in the package moveHMM is anything but straightforward, such that we are counting on users' feedback to help us prioritize the most useful extensions.

We welcome bug reports, feedback, and suggestions for the development of moveHMM.

Acknowledgments

Data accessibility

The wild haggis movement data used in Section 4 is available in the supporting material, as 'rawHaggises.csv'.

References

Altman, R. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210.

Basson, M., Hobday, A., Eveson, J., and Patterson, T. (2012). Spatial interactions among juvenile southern bluefin tuna at the global scale: a large scale archival tag experiment. FRDC Project No: 2003/002.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Bulla, J., Lagona, F., Maruotti, A., and Picone, M. (2012). A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological and Environmental Statistics*, 17(4):544–567.

Calenge, C. (2006). The package adehabitat for the R software: tool for the analysis of space and habitat use by animals. *Ecological Modelling*, 197:1035.

Charles, C., Gillis, D., and Wade, E. (2014). Using hidden Markov models to infer vessel activities in the snow crab (Chionoecetes opilio) fixed gear fishery and their application to catch standardization. *Canadian Journal of Fisheries and Aquatic Sciences*, 71(12):1817–1829.

Codling, E. A., Plank, M. J., and Benhamou, S. (2008). Random walk models in biology. *Journal of the Royal Society Interface*, 5:813–834.

DeRuiter, S., Langrock, R., Skirbutas, T., Goldbogen, J., Calambokidis, J., Friedlaender, A., and Southall, B. (2016). A multivariate mixed hidden markov model to analyze blue whale diving behaviour during controlled sound exposures. *arXiv*, arXiv:1602.06570v1.

Dragon, A.-C., Bar-Hen, A., Monestiez, P., and Guinet, C. (2012). Comparative analysis of methods for inferring successful foraging areas from argos and gps tracking data. *Marine Ecology Progress Series*, 452:253–267.

Eddelbuettel, D. and Francois, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Fleming, C. H. and Calabrese, J. M. (2015). *ctmm: Continuous-Time Movement Modeling*. R package version 0.3.0.

Franke, A., Caelli, T., and Hudson, R. (2004). Analysis of movements and behavior of caribou (Rangifer tarandus) using hidden Markov models. *Ecological Modelling*, 173(2-3):259–270.

Harte, D. (2015). HiddenMarkov: Hidden Markov models. R package version 1.8-4.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2006). Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, 13(3):325–347.

Jackson, C. H. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, 38(8):1–29.

Johnson, D. S. (2015). *crawl: Fit Continuous-Time Correlated Random Walk Models to Animal Movement Data*. R package version 1.5.

Jonsen, I., Basson, M., Bestley, S., Bravington, M., Patterson, T., Pedersen, M. W., Thomson, R., Thygesen, U. H., and Wotherspoon, S. (2013). State-space models for bio-loggers: a methodological road map. *Deep Sea Research Part II: Topical Studies in Oceanography*, 88:34–46.

King, A., Cromarty, L., Paterson, C., and Boyd, J. (2007). Applications of ultrasonography in the reproductive management of *dux magnus gentis venteris saginati*. *Veterinary Record*, 160(3):94–96.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., and Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology*, 93(11):2336–2342.

Langrock, R., Kneib, T., Sohn, A., and DeRuiter, S. (2015). Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, 71(2):520–528.

Langrock, R. and Zucchini, W. (2011). Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics and Data Analysis*, 55(1):715–724.

Lidgard, D. C., Bowen, W. D., Jonsen, I. D., and Iverson, S. J. (2012). Animal-borne acoustic transceivers reveal patterns of at-sea associations in an upper-trophic level predator. *PloS one*, 7(11):e48962.

McKellar, A., Langrock, R., Walters, J., and Kesler, D. (2015). Using mixed hidden Markov models to examine behavioral states in a cooperatively breeding bird. *Behavioral Ecology*, 26(1):148–157.

Michelot, T., Langrock, R., Patterson, T., and Rexstad, E. (2015). *moveHMM: Animal Movement Modelling using Hidden Markov Models*. R package version 1.1.

Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., and Fryxell, J. M. (2004). Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, 85(9):2436–2445.

Patterson, T. A., Basson, M., Bravington, M. V., and Gunn, J. S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, 78(6):1113–1123.

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

Peel, D., Good, N. M., and Quinn II, T. (2011). A hidden Markov model approach for determining vessel activity from vessel monitoring system data. *Canadian Journal of Fisheries and Aquatic Sciences*, 68(7):1252–1264.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

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

van de Kerk, M., Onorato, D., Criffield, M., Bolker, B., Augustine, B., McKinley, S., and Oli, M. (2015). Hidden semi-Markov models reveal multiphasic movement of the endangered Florida panther. *Journal of Animal Ecology*, 84(2):576–585.

Visser, I. and Speekenbrink, M. (2010). depmixS4: An R package for hidden Markov models. *Journal of Statistical Software*, 36(7):1–21.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R, 2nd Edition.* Chapman & Hall/CRC press, Boca Raton, FL.