6-1-2015

Nonparametric inference in hidden Markov models using P-splines

Roland Langrock
Universität Bielefeld

Thomas Kneib
Georg-August-Universität Göttingen

Alexander Sohn
Georg-August-Universität Göttingen

Stacy L. DeRuiter
Calvin University

Biometrics 71, 520–528
June 2015

DOI: 10.1111/biom.12282

Nonparametric Inference in Hidden Markov Models Using P-Splines

Roland Langrock,[1,*] Thomas Kneib,[2] Alexander Sohn,[2] and Stacy L. DeRuiter[1,3]

[1]University of St Andrews, St Andrews, U.K.
[2]University of Göttingen, Göttingen, Germany
[3]Calvin College, Grand Rapids, Michigan, U.S.A.
[*]*email:* roland.langrock@st-andrews.ac.uk

Summary. Hidden Markov models (HMMs) are flexible time series models in which the distribution of the observations depends on unobserved serially correlated states. The state-dependent distributions in HMMs are usually taken from some class of parametrically specified distributions. The choice of this class can be difficult, and an unfortunate choice can have serious consequences for example on state estimates, and more generally on the resulting model complexity and interpretation. We demonstrate these practical issues in a real data application concerned with vertical speeds of a diving beaked whale, where we demonstrate that parametric approaches can easily lead to overly complex state processes, impeding meaningful biological inference. In contrast, for the dive data, HMMs with nonparametrically estimated state-dependent distributions are much more parsimonious in terms of the number of states and easier to interpret, while fitting the data equally well. Our nonparametric estimation approach is based on the idea of representing the densities of the state-dependent distributions as linear combinations of a large number of standardized B-spline basis functions, imposing a penalty term on non-smoothness in order to maintain a good balance between goodness-of-fit and smoothness.

Key words: Animal movement; B-splines; Forward algorithm; Maximum likelihood; Penalized smoothing.

1. Introduction

1.1. *Hidden Markov Models and Nonparametric Inference*

Due to their versatility and mathematical tractability, hidden Markov models (HMMs) have become immensely popular tools for modeling time series. They have been applied in a range of fields, and in particular in various biological scenarios, including DNA sequence analysis (Durbin et al., 1998), scoring of sleep stages (Langrock et al., 2013), mark-recapture studies (Pradel, 2005), animal abundance estimation (Borchers et al., 2013) and animal movement (Langrock et al., 2012). A basic N-state HMM involves two components, (1) an observed *state-dependent process* and (2) an unobserved *N-state Markov chain*, with the observations of the former assumed to be generated by one of N component distributions as selected by the latter. A key property of HMMs is that dynamic programming algorithms can be used to evaluate the likelihood and to decode the state sequence underlying the observations.

It is usually assumed that each of the N state-dependent distributions is from a parametric family. However, choosing an adequate family can be difficult, for example if the *unknown* true state-dependent distributions are heavy-tailed, skewed or multi-modal. An unfortunate choice of the parametric family can lead, *inter alia*, to a poor fit, to biased estimates of the state transition probabilities, to poor predictive capacity and to wrong conclusions regarding the underlying system to be modeled. More specifically, parametric HMM formulations can lead to higher than adequate numbers of states being selected, for example by information criteria, simply because of the lack of flexibility of the considered state-dependent distributions in capturing the marginal distribution of the observations.

In a recent article, Yau et al. (2011) suggested a nonparametric specification of the state-dependent distributions of an HMM for continuous-valued observations. Their technically challenging approach relies on Dirichlet process mixture priors that allow to specify a hyperprior on the space of potential probability distributions for the state-dependent distribution. Dannemann (2012) developed an alternative frequentist approach based on the expectation–maximization (EM) algorithm, using log-concave densities or smoothing splines in the M-step in order to flexibly estimate the state-dependent distributions. He focused on the special case of two states, with one of the two state-dependent distributions modeled parametrically, arguing that this type of model is most relevant for applications, and that computational and identifiability issues may arise in more difficult scenarios. However, it has recently been shown by Gassiat, Cleynen, and Robin (2013) and Alexandrovich and Holzmann (2014) that identifiability in nonparametric HMMs holds under fairly weak conditions, which in practice will usually be satisfied, namely that the transition probability matrix of the unobserved Markov chain has full rank and that the state-dependent distributions are distinct.

The focus of this work lies in the investigation of practical issues involved when modeling the state-dependent distributions nonparametrically. By means of both simulations and a real data case study, we illustrate that flexible nonparametric

modeling of the state-dependent distributions of an HMM can have substantial advantages, in particular in terms of the resulting complexity of the state process. Notably, we develop a novel nonparametric estimation approach involving an easy-to-implement and computationally feasible estimation algorithm. The main idea is to represent the densities of the state-dependent distributions as linear combinations of a large number of standardized B-spline basis functions, imposing a penalty term in order to arrive at an appropriate balance between goodness-of-fit and smoothness for the fitted densities. The nonparametric approach avoids assumptions regarding the form of the state-dependent distributions and hence is expected to be most useful, if only as an exploratory tool, in scenarios where it is hard to identify a suitable parametric family.

1.2. *Motivating Example: Modeling Vertical Speeds of a Diving Whale*

Blainville's beaked whales have been the focus of a considerable amount of research, motivated by mass strandings that were associated with naval sonar operations (Cox et al., 2006). They seem to be sensitive to acoustic disturbance, altering their diving and foraging behavior in response to military sonar (Tyack et al., 2011). Accurate quantitative description and comparison of undisturbed and disturbed behavior are crucial to measuring the impact of anthropogenic noise, but are challenging given the diverse, sometimes subjective methods commonly used to describe dive behavior (Hooker and Baird, 2001).

We consider a 48-hour time series of depth displacements by a single adult female beaked whale in Hawaii, tagged with a Mk9 time-depth recorder (Wildlife Computers, Redmond, WA, USA) and previously described by Baird et al. (2008). This species performs deep foraging dives and shallow non-foraging dives, with higher vertical speeds during deep dives, especially during descents (Baird et al., 2008). We consider absolute values of depth displacements per minute, hence focusing on speed and ignoring the direction. For modeling purposes, we take the logarithms of those values. Every observation thus gives the logarithm of the absolute vertical displacement of the whale over the previous minute, which is an indicator for the whale's vertical speed in that time period. The resulting time series to be modeled, comprising 2880 observations, is illustrated in the top panel in Figure 1, alongside a histogram of the observations and the sample autocorrelation function (ACF) (bottom left and bottom right panel, respectively). The multimodality depicted in the histogram is a consequence of the whale occupying different behavioral states at different times, and this pattern, together with the strong autocorrelation, motivates the use of HMMs for these data.

However, based on a visual inspection of the histogram, it is anything but clear what family of parametric distributions will be adequate for the state-dependent process. In the literature on animal movement modeling, little attention is given to this issue, with usually only one, often arbitrarily chosen family of distributions being considered. We will demonstrate that this can be problematic, as insufficient flexibility of the state-dependent distributions can lead to unnecessarily high numbers of states being selected. This can have serious consequences on the interpretation, and also makes the corresponding models more difficult to work with in practice, for example if covariates are to be incorporated in the state process. In this regard, a nonparametric approach can have substantial advantages, as the essentially unlimited flexibility obtained for the state-dependent distributions means that inference on the states is driven solely by the correlation structure.

Before we provide details on the models we fitted to the beaked whale data, we introduce our nonparametric estimation approach and demonstrate its feasibility and potential practical advantages in a simulation study.

2. Nonparametric Hidden Markov Models

2.1. *Model Formulation and Penalized Likelihood*

Let the observed state-dependent stochastic process be denoted by $\{X_t\}_{t=1}^T$, and the underlying N-state Markov chain by $\{S_t\}_{t=1}^T$. We assume a basic dependence structure where given the current state of S_t, X_t is conditionally independent from previous and future observations and states, and where the Markov chain is of first order and homogeneous. We summarize the probabilities of transitions between the different states in the transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{ij})$, where $\gamma_{ij} = \Pr\big(S_t = j | S_{t-1} = i\big)$, $i, j = 1, \ldots, N$. The initial state probabilities are summarized in the row vector $\boldsymbol{\delta}$, where $\delta_i = \Pr(S_1 = i)$, $i = 1, \ldots, N$. For such an HMM, with parameter vector $\boldsymbol{\theta}$, the likelihood is given by

$$\mathcal{L}^{\mathrm{HMM}}(\boldsymbol{\theta}) = \sum_{s_1=1}^{N} \cdots \sum_{s_T=1}^{N} \delta_{s_1} \prod_{t=1}^{T} f(x_t|s_t) \prod_{t=2}^{T} \gamma_{s_{t-1},s_t} \,.$$

In this form, the likelihood involves N^T summands, rendering its evaluation infeasible even for a small number of states, N, and a moderate number of observations, T. However, the application of the recursive scheme called the *forward algorithm* leads to a much more efficient way of calculating the likelihood, via the matrix product expression

$$\mathcal{L}^{\mathrm{HMM}}(\boldsymbol{\theta}) = \boldsymbol{\delta}\mathbf{Q}(x_1)\boldsymbol{\Gamma}\mathbf{Q}(x_2), \ldots, \boldsymbol{\Gamma}\mathbf{Q}(x_T)\mathbf{1} \,, \tag{1}$$

where $\mathbf{Q}(x_t) = \mathrm{diag}\big(f_1(x_t), \ldots, f_N(x_t)\big)$, with $f_i(x_t) = f(x_t|S_t = i)$ denoting the density of the ith state-dependent distribution, and where $\mathbf{1} \in \mathbb{R}^N$ is a column vector of ones. The computational cost of evaluating (1) is linear in the number of observations, T, such that a numerical maximization of the likelihood is usually feasible. The similarly popular alternative route to maximum likelihood estimates, via the use of the EM algorithm, is not considered in this work, since we agree with MacDonald (2014) in there being no apparent reasons to prefer it over direct likelihood maximization. In our view, the direct maximization approach is more convenient to work with and more attractive to users, in particular since it has the crucial practical advantage that modifications in the model formulation usually require only very minor alterations in the code used to fit an HMM.

Here, we are concerned with the nonparametric estimation of the densities $f_1, \ldots, f_N$, which we conduct following ideas from Schellhase and Kauermann (2012). More specifically, we suggest to represent each of these densities as a finite linear combination of basis functions $\phi_{-K}, \ldots, \phi_K$, which are (fixed)

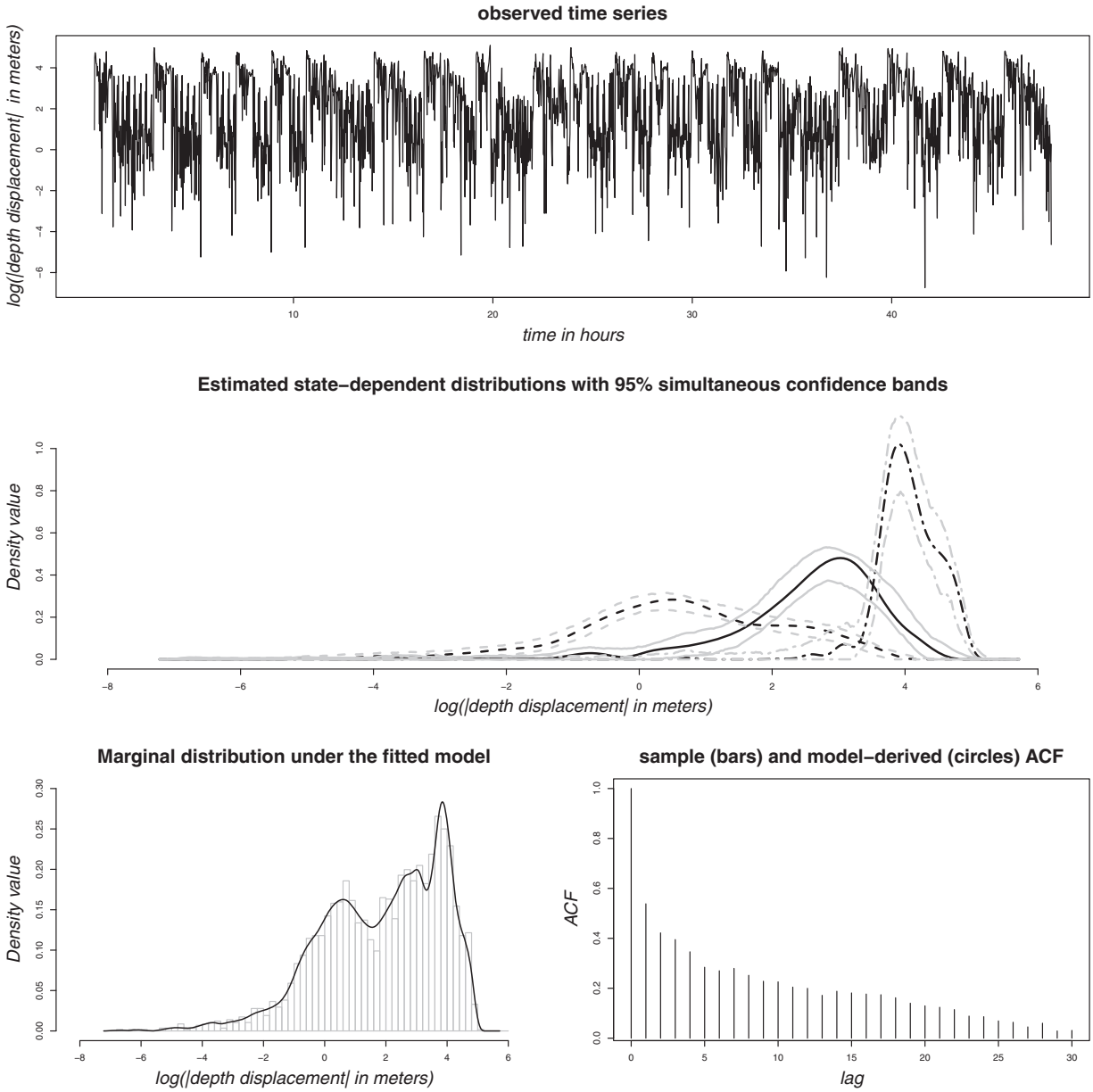

Figure 1. The top plot shows the time series that is modeled. The plot in the middle shows the nonparametrically estimated state-dependent distributions (weighted with their proportion in the mixture according to the stationary distribution of the Markov chain, and together with 95% simultaneous confidence bands). The bottom left plot shows the corresponding marginal distribution (solid line), together with a histogram of the observations (gray bars). The bottom right plot gives the sample autocorrelation function (ACF, vertical bars) and the model-derived ACF (black circles) based on the 3-state nonparametric HMM.

probability density functions, as follows:

$$f_i(x) = \sum_{k=-K}^{K} a_{i,k}\phi_k(x), \quad i = 1, \ldots, N. \tag{2}$$

Throughout this work, we use the same set of basis functions for each state-dependent distribution. Clearly, $f_i(x)$ is a probability density function if $\sum_{k=-K}^{K} a_{i,k} = 1$ and $a_{i,k} \geq 0$ for all $k = -K, \ldots, K$. To enforce these constraints, the coefficients to be estimated, $a_{i,-K}, \ldots, a_{i,K}$, are transformed using the multinomial logit link $a_{i,k} = \exp(\beta_{i,k})/\{\sum_{j=-K}^{K} \exp(\beta_{i,j})\}$, where we set $\beta_{i,0} = 0$ for identifiability. In principle, any set of densities $\phi_{-K}, \ldots, \phi_K$ can be used to approximate $f_i(x)$ as in (2). We follow Schellhase and Kauermann (2012) and use (cubic) B-splines, in ascending order in the basis used in (2), with equally spaced knots and standardized such that they integrate to one. B-splines form a numerically stable, convenient basis for the space of polynomial splines, that is, piecewise polynomials that are fused together smoothly at the

interval boundaries; see de Boor (1978) and Eilers and Marx (1996) for more details. In most cases, cubic B-splines are a suitable default since they are twice continuously differentiable and therefore yield visually smooth density estimates. Each B-spline basis function is still associated with a separate parameter, leading to a model with finite-dimensional parameter space. However, the dimensionality is high and the separate parameters are not of interest. We therefore call our approach nonparametric, which is in line with the standard terminology in the statistical literature on (penalized) spline approaches (see, e.g., Ruppert, Wand, and Carroll, 2003).

To overcome the problem of selecting an optimal number of basis elements, we follow the penalized spline approach by Eilers and Marx (1996) and modify the log-likelihood by including a penalty on the sums of squared (mth order) differences between coefficients associated with adjacent B-splines. Crucially, the characteristic HMM likelihood structure given in (1) remains valid, with the expression given in (2) applying to $f_1(x_t), \ldots, f_N(x_t)$ in the diagonal matrices $\mathbf{Q}(x_t)$, $t = 1, \ldots, T$. For independent realizations $x_1, \ldots, x_T$, the corresponding penalized log-likelihood is given by

$$l_p^{\mathrm{HMM}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \log\bigl(\mathcal{L}^{\mathrm{HMM}}(\boldsymbol{\theta})\bigr) - \left[\sum_{i=1}^{N} \frac{\lambda_i}{2} \sum_{k=-K+m}^{K} \bigl(\Delta^m a_{i,k}\bigr)^2\right], \tag{3}$$

where $\Delta a_k = a_k - a_{k-1}$ and $\Delta^m a_k = \Delta(\Delta^{m-1} a_k)$ are difference operators, the parameter vector $\boldsymbol{\theta}$ comprises the state transition probabilities and the parameters $\beta_{i,k}$ ($i = 1, \ldots, N$, $-K \leq k \leq K$, $k \neq 0$), and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ is a vector of smoothing parameters. The penalty term penalizes roughness of the estimator, and the choice of $\boldsymbol{\lambda}$ determines how much emphasis is put on goodness-of-fit and on smoothness, respectively. In particular, choosing $\boldsymbol{\lambda} = (0, \ldots, 0)$ leads to an unpenalized estimation, whereas for $\lambda_i \to \infty$, $i = 1, \ldots, N$, the penalty will dominate the likelihood and for each i we will obtain a sequence of weights $a_{i,k}$ that follow a polynomial of order $m - 1$ in k. We will use $m = 2$ in the remainder, since this provides an approximation to the integrated squared second derivative penalty that is popular in the context of smoothing splines.

The penalty term allows us to circumvent the problem of selecting an optimal number of basis elements, since it effectively reduces the number of free parameters and yields an adaptive fit to the data, provided that the smoothing parameters are chosen in a data-driven way. We only have to ensure that the number of basis elements is large enough to provide enough flexibility for reflecting the structure of the state-dependent distributions. Once this threshold is passed, a further increase in the number of basis elements does no longer change the fit to the data much due to the impact of the penalty. Allowing for different smoothing parameters across states will be important in some circumstances, for example if the (true) densities for some state-dependent distributions are much more wiggly than for others, or if some states of the Markov chain are visited much less frequently than others, potentially requiring higher penalties on roughness due to less information being available.

2.2. *Model Fitting and Inference*

2.2.1. Parameter estimation. The penalized log-likelihood (3) can be maximized numerically, corresponding to a simultaneous estimation of the Markov chain parameters and the coefficients that determine the state-dependent distributions according to (2). Of the technical issues arising in the numerical maximization, discussed in detail in Zucchini and MacDonald (2009), the most important one is that of local maxima. Particularly for complicated models, for example, such with a relatively high number of states, it will sometimes happen that the numerical search fails to find the maximum penalized likelihood estimate, and returns a local maximum instead. The best way to address this issue seems to be to use a number of different sets of initial values, in order to maximize the chances of finding the global maximum.

2.2.2. Choice of the smoothing parameter vector. Cross-validation techniques can be used for choosing the smoothing parameters. For a given time series, we suggest to generate C random partitions such that in each partition a high percentage of the observations, for example, 90%, form the calibration sample, while the remaining observations constitute the validation sample. For each of the partitions and any given $\boldsymbol{\lambda}$, the model is then fitted (i.e., calibrated) using only the observations from the calibration sample (in a times series of exactly the same length as the original one, treating the data points from the validation sample as missing data). Subsequently, scoring rules can be used on the validation sample to assess the model for the given $\boldsymbol{\lambda}$ and the corresponding calibrated model. We consider the log-likelihood of the validation sample, under the model fitted in the calibration stage, as the score of interest (now treating the data points from the calibration sample as missing data in the time series). From some pre-specified grid $\boldsymbol{\Lambda} \subset \mathbb{R}^N_{\geq 0}$, we then select the $\boldsymbol{\lambda}$ that yields the highest mean score over the C cross-validation samples. The number of samples C needs to be high enough to give meaningful scores (i.e., such that the scores give a clear pattern rather than noise only), but must not be too high to allow for the approach to be computationally feasible. Furthermore, in scenarios where a full grid search over $\boldsymbol{\Lambda}$ is computationally infeasible, we suggest the following pragmatic algorithm for finding an appropriate $\boldsymbol{\lambda}$:

1. choose an initial point $\boldsymbol{\lambda}_0^*$ from the grid $\boldsymbol{\Lambda}$ (and set $k = 0$);
2. calculate the mean score for $\boldsymbol{\lambda}_k^*$ and each direct neighbor of $\boldsymbol{\lambda}_k^*$ on the grid;
3. from these values choose $\boldsymbol{\lambda}_{k+1}^*$ as the one that yielded the highest mean score;
4. repeat 2. and 3. Until $\boldsymbol{\lambda}_{k+1}^* = \boldsymbol{\lambda}_k^*$.

2.2.3. Uncertainty quantification. Uncertainty quantification, on both the estimates of the transition probabilities and on the estimates of the densities of the state-dependent distributions, can be performed using a parametric bootstrap. In particular, from the bootstrap replications one can obtain pointwise confidence intervals for the estimated densities as the corresponding quantiles at a specific point in the support. These pointwise confidence intervals can also be used to obtain simultaneous confidence bands for the complete den-

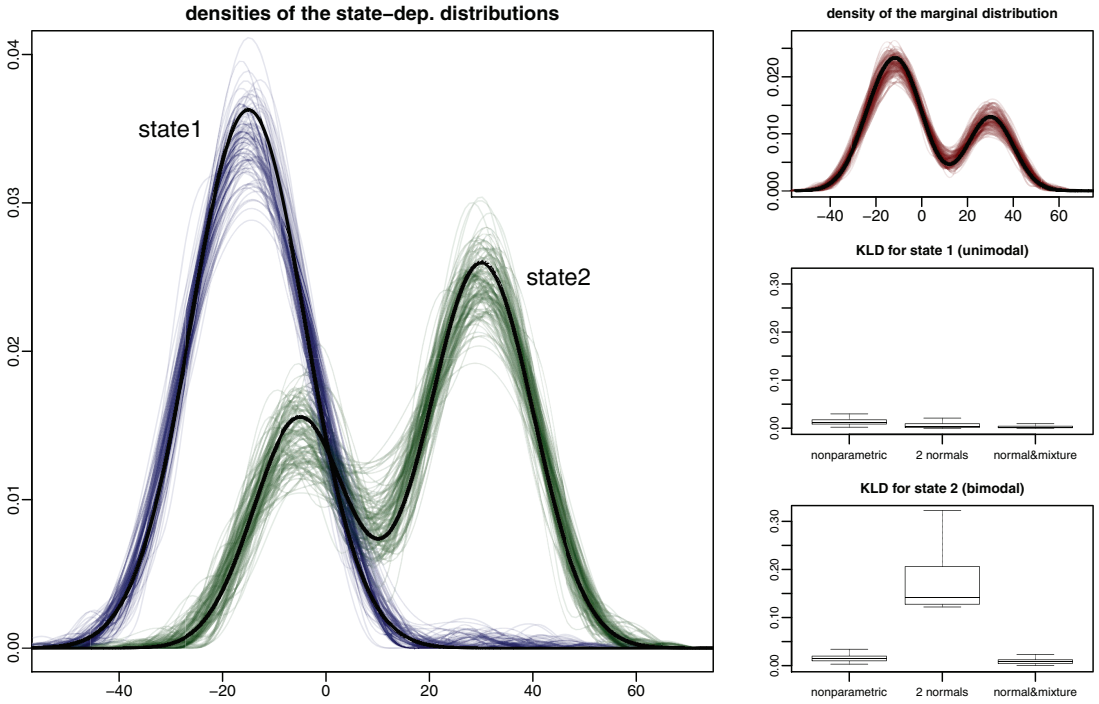

Figure 2. Simulation study: true and estimated densities of the state-dependent distributions estimated in the first 100 simulation runs (left plot, with true densities indicated by thick black lines), corresponding true and estimated densities of the marginal distribution (top right plot, with true density indicated by thick black line), and box plots of Kullback–Leibler divergences (KLDs) for unimodal state 1 (middle right plot) and for bimodal state 2 (bottom right plot).

sity following Krivobokova, Kneib, and Claeskens (2010). The idea is to rescale the pointwise confidence bands with a constant factor until a certain fraction of complete densities from the set of bootstrap replications is contained in the confidence band. By construction, these simultaneous bands use the pointwise intervals to assess local uncertainty about the estimated density and inflate this local uncertainty such that simultaneous coverage statements are possible.

3. Simulation Study

To demonstrate the practicality of the suggested approach, we first present a simulation experiment. We consider a two-state HMM where the state-dependent distributions substantially overlap, with a unimodal conditional distribution in state 1 and a bimodal conditional distribution in state 2; see Figure 2 for an illustration. The states of the Markov chain were generated using the 2×2–t.p.m. with both diagonal entries equal to 0.9. In practice, the chosen configuration would make it difficult to specify an adequate parametric HMM, since the marginal distribution gives no indication for the bimodality of state 2 (cf. the top right panel of Figure 2). It is also not clear a priori if a nonparametric approach can exploit the correlation over time in order to identify the smaller peak of the conditional bimodal distribution in state 2 (at $x = -5$), or if it fails to do so and wrongly allocates the corresponding observations to state 1.

For this model, we conducted 500 simulation runs, with $T = 800$ observations being generated in each run. In each run, the nonparametric approach was applied, maximizing (3) using the optimizer `nlm` in R; corresponding code is given in Web Appendix B. For the cross-validation, we selected the grid of potential smoothing parameter vectors, $\boldsymbol{\Lambda}$, such that as possible values for each of the state-specific smoothing parameters the values 256, 512, 1024, 2048, 4096, 8192, and 16,384 were considered. We used $C = 10$ cross-validation partitions in each simulation run to select the smoothing parameter vector from $\boldsymbol{\Lambda}$. We further set $K = 15$, hence using 31 B-spline basis densities in the estimation.

The sample mean estimates of the transition probabilities γ_{11} and γ_{22} were 0.907 (Monte Carlo standard deviation of estimates: 0.018; average of standard errors obtained in each run via parametric bootstrap: 0.018; coverage of bootstrap 95% confidence intervals obtained in each run: 92.6%) and 0.907 (0.018, 0.019, 95.0%), respectively. The estimated state-dependent distributions from the first 100 simulation runs are visualized in the left panel of Figure 2. All fits are fairly reasonable, although as expected the peaks are slightly underestimated, on average, while the troughs are slightly overestimated. The same pattern can be seen for the marginal distribution, displayed in the top right panel in Figure 2.

We further calculated the Kullback–Leibler divergence (KLD) of the estimated densities from those of the true model, for both state-dependent distributions and in each simulation run. To have a benchmark, we also calculated the corresponding KLDs of densities estimated using either of two parametric HMMs: (1) an incorrect parametric model, assuming normal state-dependent distributions (which one may visually deduce from a histogram of the observations to be appropriate), and (2) the correct parametric model, involving

a normal distribution for state 1 and a mixture of two normal distributions in state 2. Unsurprisingly, the correct parametric model performed best in terms of the KLD. The nonparametric approach yielded the highest average KLD for the conditional distribution in state 1, due to the oversmoothing close to the peak. For the state-dependent distribution in state 2, our nonparametric approach yielded an average KLD of 0.016, which is slightly higher than the corresponding average KLD obtained using the correct parametric specification (0.010), whereas for the incorrectly specified parametric model the corresponding average KLD was obtained as 0.290, indicating the expected much poorer fit. For the correctly specified parametric model, the sample mean estimates of the transition probabilities γ_{11} and γ_{22} were obtained as 0.898 (Monte Carlo standard deviation of estimates: 0.018; average of standard errors obtained in each run via parametric bootstrap: 0.019; coverage of bootstrap 95% confidence intervals obtained in each run: 95.6%) and 0.899 (0.020; 0.021; 94.2%), respectively, while for the incorrectly specified parametric model the corresponding mean estimates were obtained as 0.866 (0.019; 0.040; 63.6%) and 0.821 (0.024; 0.093; 45.6%), respectively. The incorrectly specified model thus led to erroneous inference on the state process, in particular substantially underestimating the persistence in state 2. Furthermore, with the incorrectly specified parametric model, inference on the state process was erroneous also regarding the choice of the number of states: the Akaike information criterion (AIC) selected 3- or 4-state models in all 500 simulation runs, while the Bayesian information criterion (BIC) selected a 3-state model in all runs. Overall, in this simulation study the nonparametric approach performed only slightly worse than the correct parametric specification—unlike the latter leading to a small bias in the estimates of the transition probabilities, and resulting in slightly less accurately estimated state-dependent distributions—and much better than the incorrect parametric specification which one may naively choose based on a histogram of the data or another form of visual inspection.

In each simulation run, we additionally fitted one- and three-state HMMs with nonparametrically modeled state-dependent distributions, in order to illustrate that the cross-validation technique can also be used to select the number of states. In each run, the model selection was based on a comparison of the out-of-sample log-likelihood scores on 10 random validation samples, obtained for the different models fitted to the corresponding 10 calibration samples. This is essentially the multifold cross-validation procedure considered in Celeux and Durand (2008), only that here we obtain estimates via direct maximization of the likelihood rather than using the EM algorithm. In our simulations, this model selection exercise led to a correct identification of the two-state model in 459 out of 500 cases (91.8%), with the three-state model being selected in the other 41 cases. Comparing this to the results presented in Celeux and Durand (2008) for the parametric case, we consider this to be a good performance.

To investigate the estimation performance under different conditions, we experimented with several further model formulations. In particular, we considered (a) different levels of correlation as induced by the 2-state Markov chain (by varying the diagonal entries in the t.p.m.) and (b) different levels of overlap of the two state-dependent distributions (by shifting one of the two). In all those scenarios where there was a reasonable level of correlation induced by the Markov chain (roughly, for both diagonal entries either > 0.75 or < 0.25), the estimation worked well. The estimation performance improved with diagonal entries in the t.p.m. approaching either 1 or 0 (which leads to an increased correlation). This intuitively makes sense, since the stronger the correlation, the clearer becomes the pattern and hence the easier it is for the model to allocate observations to states. Similarly, the estimation performance improved as the overlap of the state-dependent distributions was reduced.

Table 1
Results of fitting HMMs with normal state-dependent distributions to the beaked whale data

N	p	$\log \mathcal{L}$	AIC	BIC	JB p-value
3	12	−4880.00	9784.00	9855.59	0.000
4	20	−4729.08	9498.16	9617.47	0.000
5	30	−4670.15	9400.30	9579.27	0.002
6	42	−4605.44	9294.88	9545.43	0.016
7	56	−4548.02	9208.04	**9542.11**	0.261
8	72	−4492.57	9129.15	9558.67	0.310
9	90	−4455.48	9090.98	9627.87	0.475
10	110	−4422.26	**9064.53**	9720.74	0.429

N, number of states; p, number of model parameters; $\log \mathcal{L}$, maximum of the log-likelihood; AIC, Akaike information criterion; BIC, Bayesian information criterion; JB p-value, p-value of the Jarque–Bera test for normality applied to the pseudo-residuals.

4. Results of Fitting HMMs to the Beaked Whale Dive Data

4.1. *Analysis Using Conventional Parametric HMMs*

We begin our analysis of the beaked whale dive data, described in Section 1.2, by fitting conventional parametric HMMs. For the given time series, HMMs with normal state-dependent distributions constitute an obvious and plausible choice of a parametric family of models. We present the results of fitting such models to the data, acknowledging that this is only one of dozens of plausible parametric HMM formulations that could be considered—a flexibility which is both a blessing and a curse when dealing with HMMs, as the model formulation is in general by no means straightforward.

Table 1 summarizes the results of fitting HMMs with normal state-dependent distributions and 3–10 states to the dive data, including the log-likelihood values, the AIC values, the BIC values and the p-values of Jarque–Bera (JB) tests for normality applied to the models' pseudo-residuals. The latter are distributed standard normal if the model is correct (cf. Section 1 of the Web Appendix, and Zucchini and MacDonald, 2009). Some graphic illustrations of the fitted models are provided in Section 2 of the Web Appendix. The AIC and the BIC select models with 10 and 7 states, respectively. The need for these high numbers of states is confirmed by goodness-of-fit analyses of the fitted models, where at the 5% level normality of the pseudo-residuals is rejected by a JB test for all models with less than 7 states.

Figures A1 and A3 in the Web Appendix illustrate that a crucial problem with the considered parametric formulation is the lack of flexibility of the state-dependent distributions, with the consequence that the marginal distribution cannot be captured adequately with small numbers of states, whereas the correlation structure in the time series is captured well already by a 3-state model (cf. Figure A3 in the Web Appendix). For the 7-state model—which based on the information criteria and the goodness-of-fit test seems to be a plausible model—it is difficult to come up with biologically meaningful interpretations of the states, with the results indicating that some of the HMM states may in fact be lumped together to form a single behavioral state (cf. Figure A5, states 3 and 5).

4.2. *Analysis Using Nonparametric HMMs*

We now turn to the results of applying our nonparametric approach to the dive data. The characteristics of the behavioral states of the whale, most notably "close to the surface," "on the ascent/descent," and "at the bottom of a dive," motivate the use of three states in an HMM for this time series, and here we restrict ourselves to the consideration of such a model. The results of fitting models with 2 and 4 states, respectively, are provided and discussed in Section 3 of the Web Appendix, where we demonstrate that the 2-state model fails to capture important structure, while the 4-state model provides little extra biologically meaningful information.

The (stationary) 3-state nonparametric HMM was fitted via a numerical maximum of the penalized likelihood given in (3). As smoothing parameter vector we selected $\boldsymbol{\lambda} = (65536, 8192, 32)$ via cross-validation as described in Section 2.2.2 (see the Web Appendix for more details). We used 51 standardized B-splines in the estimation, that is, $K = 25$ in (2). On an i7 CPU, at 2.7 GHz and with 4 GB RAM, the parameter estimation took about 20 minutes using R, which we were able to reduce to about 2 minutes by writing the forward algorithm in C++. To minimize the risk of missing the global maximum of the likelihood, 500 randomly chosen sets of initial values were tried. The t.p.m. was estimated as

$$\widehat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.975\ (0.965, 0.983) & 0.007\ (0.001, 0.015) & 0.018\ (0.010, 0.027) \\ 0.017\ (0.005, 0.033) & 0.893\ (0.871, 0.926) & 0.090\ (0.059, 0.113) \\ 0.038\ (0.020, 0.056) & 0.111\ (0.074, 0.138) & 0.851\ (0.821, 0.890) \end{pmatrix}$$

with the 95% confidence intervals (in brackets) obtained using a parametric bootstrap (500 samples). Figure 1 displays (1) the time series that was modeled, (2) the state-dependent distributions of the fitted model, together with 95% simultaneous confidence bands (cf. Section 2.2.3), (3) the marginal distribution of X_t according to the fitted model, together with a histogram of the observations, and (4) the sample ACF alongside the model-derived ACF. An illustration of how the fitted state-dependent distributions are built from the B-spline basis densities is given in the Web Appendix (Figure A7).

A quantile–quantile plot of the pseudo-residuals against the standard normal, and the sample ACF of the series of residuals, are given in the Web Appendix (Figure A9). The plots indicate a very good fit and only a minor correlation in the residuals over time, with the goodness of fit being similar to that of the 7-state parametric HMM. Applying a JB test to the pseudo-residuals yields a p-value of 0.3, such that the null hypothesis of normality cannot be rejected at the 5% level. Thus, the model fits the data well.

To facilitate interpretation of the fitted model, we compared the Viterbi-decoded states to the actual positions of the whale in the water column (which were recorded, but not modeled here); an illustration is given in the Web Appendix (Figure A8). We find that state 1 of the fitted HMM, which is associated with the smallest absolute depth displacements, captures the whale's vertical speeds close to the surface and on very shallow dives (to depths ∼100 m), with the shallow dives causing the second mode in the fitted density (at values slightly higher than 2). This state is occupied about 52% of the time according to the stationary distribution of the fitted Markov chain. State 2, which involves moderate absolute depth displacements, is occupied about 26% of the time and is associated with likely foraging periods at the bottoms of deep dives. State 3 implies the highest absolute depth displacements, is occupied about 22% of the time and only on deep dives, and is occasionally interspersed with state 2 at the bottoms of those dives due to slower movement related to foraging activity.

We note that, just as for parametric HMMs, any biological interpretations of this model have to be made cautiously. In particular, the HMM states are not to be confounded with behavioral states of the animal, as they merely summarize vertical diving speeds into three categories, and the speeds can be similar for distinct behaviors. However, it can nevertheless be stated that the features implied by the fitted nonparametric HMM are consistent with previous research on the species (Baird et al., 2008). Moreover, this exploratory analysis demonstrates the potential of these models as tools for example for objective identification and characterization of foraging periods, which is notoriously challenging with time-depth recorder data (Hooker and Baird, 2001).

5. Discussion

In our case study, applying nonparametric HMMs led to a coherent, succinct summary of the dive data using a small number of states, sensibly partitioning the data into few velocity regimes that are easy to relate to broader behavior categories. The ability to summarize the data accurately without using a large number of states facilitates interpretation, with the nonparametric approach resulting in greater persistence within states, so that intuitive association of the states with broader behavior is more straightforward. In the context of measuring the effects of acoustic disturbance, a parsimonious model that accurately summarizes the data is ideally suited for the purpose of quantifying potential behavioral changes,

as high numbers of parameters in the t.p.m. render it impractical to incorporate covariates, or random effects or both, in the state process.

In general, a nonparametric approach essentially offers unlimited flexibility to capture the marginal distribution, irrespective of the number of states considered, which means that inference on the number of states is based solely on the correlation structure of a given time series. In contrast, both in our simulation experiments and in our case study we have seen that conventional parametric HMMs can lead to high numbers of states being selected due to limited flexibility of the state-dependent distributions considered, leading to state processes that are unnecessarily complex relative to the actual correlation structure, complicating the interpretation. This reveals a potentially substantial benefit of the nonparametric approach, as the disentanglement of the two main reasons for a poor fit of an HMM—failure to capture the marginal distribution and failure to capture the correlation structure–can be quite challenging using conventional methods.

Our likelihood-based estimation approach, which exploits the strengths of the HMM machinery and of penalized B-splines, is relatively simple yet powerful, allowing for comprehensive inferential analyses, including uncertainty quantification, model checking, and state decoding. The choice of the smoothing parameter, local maxima of the likelihood and uncertainty quantification constitute the most challenging issues with the approach. For smoothing parameter selection, as an alternative to cross-validation, one could consider a leverage-based, approximate leave-one-out cross-validation which yields an AIC-type criterion. Regarding local maxima, estimation via the EM algorithm could potentially be more robust (Bulla and Berzel, 2008), but is technically more challenging and likely to be slower. Uncertainty quantification is routinely achieved in the Bayesian approach of Yau et al. (2011) by studying the variability of the posterior samples, whereas we employed computationally expensive bootstrap techniques. The approach is inferior to parametric models in cases where those are adequate, and, in view of the computational challenges, especially local maxima, is practical only for models with small numbers of states ($N \leq 4$ in our case study). If larger number of states and high flexibility are required, then we anticipate that flexible parametric approaches such as the one considered in Holzmann and Schwaiger (in press), where the state-dependent distributions are mixtures of normal distributions, will be preferable. Furthermore, it is clear that the nonparametric approach will work best if the serial correlation in the data is relatively high, since otherwise it is difficult for a flexible nonparametric model to pick up a meaningful pattern. Finally, we note that insufficient flexibility of parametric models can be unproblematic if parsimony and interpretability of the state process are not of major importance.

In general, it will often be the case that multiple time series, for example associated with multiple individuals, are collected. In the given type of application, this will in fact often be necessary in order to adequately address biological questions of interest, for example, regarding the effect of sonar exposure. Models for such longitudinal data need to account for potential variability between the different component series. Zucchini, Raubenheimer, and MacDonald (2008) give a useful overview of the different strategies for modeling heterogeneity in HMM component series. Regarding the state process of an HMM, the techniques provided by Altman (2007) in her framework of mixed HMMs—which allows for random effects, but also covariates that are specific to the component series—and the discrete random effects approach suggested by Maruotti and Rydén (2009) are directly applicable in our nonparametric estimation framework. However, accounting for heterogeneity, as well as the possible incorporation of covariates, in the state-dependent process is anything but straightforward when using the nonparametric approach, and requires further research.

6. Supplementary Material

Web Appendices and Figures referenced in Section 4, and R code to generate data as in the simulation study and to fit, to the generated data, a 2-state nonparametric HMM, are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors would like to thank Walter Zucchini, Antonello Maruotti, two anonymous referees, the Associate Editor and the Editor for very helpful comments on earlier versions of this paper, Robin Baird for providing access to the beaked whale data, and Théo Michelot for writing the HMM forward algorithm in C++. The last author acknowledges financial support from the U.S. Office of Naval Research.

References

Altman, R. M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association* **24**, 201–210.

Alexandrovich, G. and Holzmann, H. (2014). Nonparametric identification of hidden Markov models. *arXiv*:1404.4210.

Baird, R. W., Webster, D. L., Schorr, G. S., Mcsweeney, D. J., and Barlow, J. (2008). Diel variation in beaked whale diving behavior. *Marine Mammal Science* **24**, 630–642.

Bulla, J. and Berzel, A. (2008). Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics* **13**, 1–18.

Borchers, D. L., Zucchini, W., Heide-Jørgensen, M. P., Cañadas, A., and Langrock, R. (2013). Using hidden Markov models to deal with availability bias on line transect surveys. *Biometrics* **69**, 703–713.

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

Cox, T. M., Ragen, T. J., Read, A. J., Vos, E., Baird, R. W., Balcomb, K., Barlow, J., Caldwell, J., Cranford, T., Crum, L., D'Amico, A., D'Spain, G., Fernandez, A., Finneran, J., Gentry, R., Gerth, W., Gulland, F., Hildebrand, J., Houser, D., Hullar, T., Jepson, P. D., Ketten, D., MacLeod, C. D., Miller, P., Moore, S., Mountain, D. C., Palka, D., Ponganis, P., Rommel, S., Rowles, T., Taylor, B., Tyack, P., Wartzok, D., Gisiner, R., Mead, J., and Benner, L. (2006). Understanding the impacts of anthropogenic sounds on beaked whales. *Journal of Cetacean Research and Management* **7**, 177–187.

Dannemann, J. (2012). Semiparametric hidden Markov models. *Journal of Computational and Graphical Statistics* **21**, 677–692.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin, Germany: Springer.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. J. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science* **11**, 89–121.

Gassiat, E., Cleynen, A., and Robin, S. (2013). Finite state space non parametric Hidden Markov Models are in general identifiable. *arXiv*:1306.4657v1.

Holzmann, H. and Schwaiger, F. (in press). Hidden Markov models with state-dependent mixtures: Minimal representation, model testing and applications to clustering. *Statistics and Computing*, DOI:10.1007/s11222-014-9481-1.

Hooker, S. K. and Baird, R. W. (2001). Diving and ranging behaviour of odontocetes: A methodological review and critique. *Mammal Review* **31**, 81–105.

Krivobokova, T., Kneib, T., and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association* **105**, 852–863.

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**, 2336–2342.

Langrock, R., Swihart, B., Caffo, B., Crainiceanu, C., and Punjabi, N. (2013). Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in Medicine* **32**, 3342–3356.

MacDonald, I. L. (2014). Numerical maximisation of likelihood: A neglected alternative to EM? *International Statistical Review* **82**, 296–308.

Maruotti, A. and Rydén, T. (2009). A semiparametric approach to hidden Markov models under longitudinal observations. *Statistics and Computing* **19**, 381–393.

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

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Schellhase, C. and Kauermann, G. (2012). Density estimation and comparison with a penalized mixture approach. *Computational Statistics* **27**, 757–777.

Tyack, P. L. et al. (2011). Beaked whales respond to simulated and actual navy sonar. *PLoS ONE* **6**, e17009.

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society, Series B* **73**, 37–57.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton, FL: Chapman & Hall.

Zucchini, W., Raubenheimer, D., and MacDonald, I. L. (2008). Modeling time series of animal behavior by means of a latent-state model with feedback. *Biometrics* **64**, 807–815.

Received February 2014. Revised November 2014.
Accepted November 2014.