Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 227 (2004) 39–50
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

Positional entropy during pigeon homing I: application of Bayesian latent state modelling

Stephen Roberts[a,*], Tim Guilford[b], Iead Rezek[a], Dora Biro[b]

[a] *Machine Learning Research Group, University of Oxford, UK*
[b] *Animal Behaviour Research Group, University of Oxford, UK*

Received 7 April 2003; received in revised form 15 July 2003; accepted 18 July 2003

Abstract

In these two companion papers, we introduce a new approach to the analysis of bird navigation which brings together several novel mathematical and technical applications. Miniaturized GPS logging devices provide track data of sufficiently high spatial and temporal resolution that considerable variation in flight behaviour can be observed remotely from the form of the track alone. We analyse a fundamental measure of bird flight track complexity, spatio-temporal entropy, and explore its state-like structure using a probabilistic hidden Markov model. The emergence of a robust three-state structure proves that the technique has analytical power, since this structure was not obvious in the tracks alone. We propose the hypothesis that positional entropy is indicative of underlying navigational uncertainty, and that familiar area navigation may break down into three states of navigational confidence. By interpreting the relationship between these putative states and features on the map, we are able to propose a number of hypothetical navigational strategies feeding into these states. The first of these two papers details the novel technical developments associated with this work and the second paper contains a navigational interpretation of the results particularly with respect to visual features of the landscape.

Keywords: Bird navigation; Hidden Markov model; Bayesian analysis; Entropy

1. Introduction

The navigational abilities of birds are impressive, and much has been learnt about the mechanisms involved from more traditional experimental approaches involving the manipulation of potential cue systems, or their sensory availability, and subsequent release from unfamiliar sites. Indeed, it is the homing pigeon that has provided the bulk of these discoveries. Nevertheless, a detailed understanding of how birds navigate within their own familiar area, where they are able to access a familiar area map representation of some kind, has eluded Biologists. This is for two reasons. Visual landmarks probably play the dominant role (Guilford, 1993), yet they are effectively impossible to manipulate meaningfully on the scale required outside the laboratory, especially in the kind of visually rich environments in which most birds live. Second, it is starting to become clear that birds can use multiple strategies in homing, not single mechanisms as was once thought. The dominance of any one mechanism may vary with the availability of the cue system during early experience (Braithwaite and Guilford, 1995), or, more crucially, with far more immediate factors causing strategy switching during the homing event itself. The structure of a bird's homing trajectory, however, must contain exquisite information about the complexity of each bird's representations and the dynamics of its navigational decision making, if only we have the methods to observe and analyse this structure in sufficient detail.

The removal of selective availability on GPS signals and the recent development of miniaturized (35 g) GPS logging devices small enough to be carried by homing pigeons, has allowed for the first time their detailed tracking (Biro et al., 2002; Steiner et al., 2000). As a model species, the homing pigeon is of special value because it will attempt to return directly home after release from a site to which its has been displaced

*Corresponding author. Department of Engineering Science, Parks Road, Oxford OX1 3PJ, UK. Fax: +44-1865-283301.
E-mail address: sjrob@robots.ox.ac.uk (S. Roberts).

0022-5193/$ - see front matter © 2003 Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2003.07.002

artificially, and is very tolerant of handling, allowing the tracking of defined flight attempts from a known release point to a known target goal, and under experimentally controlled manipulations. Our prototype devices record geographical position coordinates every second, storing up to 100,000 positions, to within 4 m accuracy. This extraordinary degree of resolution provides a rich information source on the relationship between a navigating animal's moment-by-moment behaviour, and its position with respect to both the external landscape and internal factors such as the time since release. Whilst these data are clearly potentially deeply insightful of the bird's navigational decision making processes, they are far beyond what has been handled by traditional biological approaches to studying animal navigation systems.

Our working hypothesis is that the complex navigational behaviour of birds may be modelled around a small set of prototype behavioural states which define different navigation strategies. We further consider the navigation of the birds to switch between these strategy states due to internal or external factors. Given this working hypothesis, we may formulate a mathematical generative model of the observed navigation patterns (i.e. the birds' positional trajectories) based on a set of hidden (latent) variables which define the unknown navigation strategies. We believe that the separation of these strategy states allows for a better understanding of processes of navigational decision making and the underlying map-like representations involved.

As our working hypothesis is that navigation requires the formation of a set of latent (or hidden) 'strategy' states, extensions of *hidden Markov models* (HMMs) form the basis for our investigations. The HMM is attractive given that it is a dynamic switching state model and hence naturally encodes the notion of time-dependent state changes. The HMM architecture has the added advantage that it may be drawn as an acyclic graphical model and thus may be modelled in a fully probabilistic manner. The advantages of probabilistic models are well known, in particular, the principled handling of uncertainty in data and model (Jordan, 1999). Inference in standard HMMs requires the estimation of state transition probabilities and generative observation models for each state. The number of states required to explain the patterns of bird navigation is unknown. Although some clues could come from zoological studies, this information is necessarily vague and here we explore the observation that a fully probabilistic approach allows the structure in the data to direct the number of states. We, therefore, do not run the risk of forcing a switiching-state model to the data if the data do not support it. We return to this important consideration later in this paper. This fully probabilistic approach may be achieved via a sampling paradigm (Rezek and Roberts, 2000) or, more recently, via the computationally less demanding approach of variational learning and it is this latter approach which we introduce here.

The cross-disciplinary range of the current study is considerable with novel technical and analytical methods or their applications in both mathematics and animal bahaviour. We, therefore, divide the study into two companion papers, the first dealing primarily with the mathematical approach, the second with the biological issues and their interpretation.

The rest of this paper is organized as follows. We first introduce the basic concepts of embedding a multi-dimensional time series and estimating its stochastic complexity. Various desirable properties of the stochastic complexity measure are introduced and examples given on synthetic data. An overview of the use of variational Bayesian learning of latent state models is given in Sections 3 and 4.2. Section 6 gives a presentation of the results. We conclude with a summary of this work and a discussion.

Our companion paper gives considerably more results and details of their biological interpretation; this paper offers 'proof of concept' results only. We, therefore, suggest the following reading strategy for the *biologically* motivated reader: our companion paper followed by the non-mathematical parts of Section 2, the introduction to Section 3, Section 4 and finally Section 5.2.3. For the *technically* oriented reader, we suggest reading Sections 2–4 and 6 in this paper, followed by our companion paper.

2. Stochastic complexity

2.1. Why complexity?

There are a wide range of potential measures which may be inferred from flight-track data sequences. It has been observed empirically that measures based on the *tortuosity* of the tracks bear some relation to hypothesized navigational behaviour (Biro et al., 2002). We argue that such measures, whilst important, are actually attempting to measure a more fundamental quality of the data. We argue that this is the data complexity, i.e. the intrinsic positional variability associated with short sections of the flight data, and that this is better measured using alternative techniques which we introduce here.

2.2. There is no absolute 'complexity'

Many different definitions, methods and measures of signal, or system, 'complexity' have been proposed. Some are not well suited to analysis of small-sample signals, as they are notoriously variant in the presence of noise and non-stationarities (the classic examples being

the correlation dimension (Grassberger and Procaccia, 1983) and the Lyapunov exponent (Wolf et al., 1985). Several methods of assessing the 'complexity' of data sequences were developed, presented and compared in (Rezek and Roberts, 1998). The methods investigated were: optimal model order and prediction error under autoregressive modelling, Spectral entropy (SE) (Porta et al., 1998), Approximate entropy (ApEn) (Pincus, 1991) and stochastic complexity based on embedding space decomposition (ESD). On a variety of biological and non-biological data sets, it was observed empirically that autoregressive modelling (a general linear dynamical model) failed to give a reliable estimate of system complexity, ApEn and SE gave similar results (ApEn is computationally very expensive—so SE was preferred for this reason) and ESD methods gave good results over a variety of data sequences (Rezek and Roberts, 1998). We concentrate in this paper therefore on the ESD method, which is detailed next. We note, however, that we do not imply this to be the canonical representation of system, or data, complexity; merely a robust measure.

2.3. Temporal information

It is clear that the data we investigate constitute a time-evolving dynamical system. We would like, therefore, to base our understanding of its properties on these intrinsic dynamics. Although we could formulate a parametric model for this, we choose, in this paper, to consider a generic approach first developed for the analysis of chaotic systems (Broomhead and King, 1986). This approach, referred to as *embedding space decomposition* is detailed in this subsection.

We first define the construction of the embedded data matrix, $\mathbf{X}$, which we obtain in the standard way using the method of delays (Takens, 1981). We consider some window on the data sequence as containing p samples taken at intervals of J samples from the observed q-dimensional time series $\mathbf{x}_t \in \mathbb{R}^q$. The elements in the window represent components of an *embedding space* in $\mathbb{R}^{p\times q}$. As the data sequence (time series) is repeatedly windowed, the series of vectors obtained constitutes the trajectory, or embedding, matrix. This is depicted schematically in Fig. 1.

Defining

$$\vec{\mathbf{x}}_i = (\mathbf{x}_i, \mathbf{x}_{i+J}, \ldots, \mathbf{x}_{i+(p-1)J})^{\mathsf{T}}, \tag{1}$$

we obtain an embedding matrix, over an N-sample window, which may be written as

$$\mathbf{X} \in \mathbb{R}^{Np\times q} = (\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \ldots, \vec{\mathbf{x}}_{N-(i-1)})^{\mathsf{T}}. \tag{2}$$

The value p is referred to as the embedding dimension and must satisfy

$$p \geqslant 2d + 1, \tag{3}$$

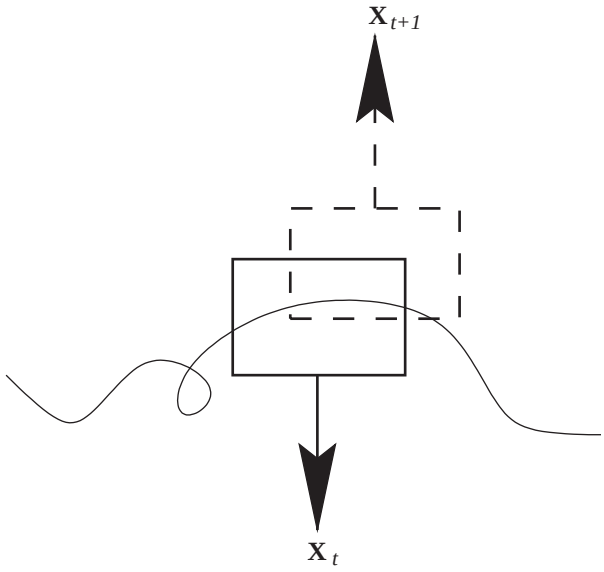

Fig. 1. Schematic example of forming the embedding matrix. A simple two-dimensional trajectory (thin line) consists of a set of data vectors, $\mathbf{x} = (x_1, x_2)^{\mathsf{T}}$ which evolves with time. The embedding matrix, $\mathbf{X}_t$, is formed from concatenation of all $\mathbf{x}$ which lie within a window centred at time t (solid rectangle). As the window slides along the data, to $t+1$, say (dashed rectangle) so another embedding matrix, $\mathbf{X}_{t+1}$, is formed.

which sets the lower bound for p given a d-dimensional manifold in the phase space. Since d is not known a priori means that in practice the embedding dimension p is chosen large enough such that redundancy in the embedding results. This redundancy manifests itself as a rank deficiency in the embedding matrix $\mathbf{X}$. We may investigate this redundancy by means of a singular value decomposition, whereby we decompose $\mathbf{X}$ via Eq. (A.1) in Appendix A.

For a noiseless system with redundancy (i.e. p is larger than the intrinsic dimensionality of the embedded data sequence) some *singular values*, σ_i, will zero. In real-world situations, however, the observed time series will be corrupted by noise (including quantization noise). In the case of white noise, this results in a shifting of the singular values such that

$$\sigma_i^2 \leftarrow \sigma_i^2 + \langle \xi^2 \rangle, \tag{4}$$

where $\langle \xi^2 \rangle$ is the expected noise variance. Hence, no singular value will be zero. It is noted that noise, in this context, is taken to be any process whose dynamics is more complex than the upper limit imposed by Eq. (3) and hence has no preferred direction in the embedding space, giving rise to a noise floor in the singular value spectrum. This is in contrast to those singular values associated with the deterministic (lower complexity) system which will be significantly larger (see Broomhead and King, 1986). Unless stated otherwise, all the examples shown in this paper used an embedding dimension of $p = 5$ over windows of $N = 30$ samples

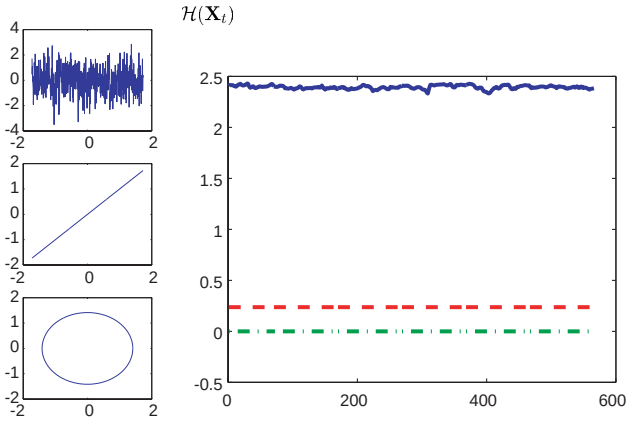

Fig. 2. Left: from top to bottom. Trajectory with Gaussian noise y-component, linear trajectory, circular trajectory. Right: stochastic complexity (in bits) from random (solid), linear (dash–dot) and circular (dashed) trajectories. The x-axis in the right-hand plot is in samples. All three sequences had the same number of samples.

(corresponding to a window of 30 s of flight) and lag $J = 1$. For a more detailed discussion regarding these parameters see Kember and Fowler (1993).

To define a measure of dynamic complexity, we note that, as the number of apparent dynamic components[1] increases, so the number of singular values above the noise floor increases. In the limit, when the system becomes complex enough to be indistinguishable from any additive noise, all singular values lie on the noise floor and have similar magnitude. We may define, therefore, a pragmatic measure of this trend using the entropy of the singular value spectrum (Rezek and Roberts, 1998). We apply a normalization to the embedding matrix, $\mathbf{X}$ such that each column has zero mean and unit variance. As discussed in Section 2.4, this ensures desirable invariancy properties for the complecity measure. Defining

$$\bar{\sigma}_i(\mathbf{X}) = \sigma_i(\mathbf{X}) / \sum_{i'} \sigma_{i'}(\mathbf{X}),$$

so the stochastic complexity is defined as

$$\mathscr{H}(\mathbf{X}) = -\sum_{i=1}^{N} \bar{\sigma}_i(\mathbf{X}) \log \bar{\sigma}_i(\mathbf{X}). \tag{5}$$

We choose this form so that systems of decreasing complexity will give rise to decreasing entropy measures. We will use this entropy-based measure as an indicator of system complexity. Note that if $\log_2$ is taken the resultant entropy is in *bits*. In all the examples in this paper we adopt this convention. Fig. 2 shows the application of this approach to three sections of synthetic trajectory. As expected, the entropy measures show highest values for a random trajectory, intermediate (but low) for a circular orbit and close to zero for a straight line.

[1] This may not, of course, be the true number of components—the apparent number is conditional on the sample size, the embedding dimension, etc.

2.4. Invariancy of stochastic complexity

All tracking data consists of arrays of coordinates. It is clear that many consistent coordinate frames exist and it is desirable that the measure with which we represent the track complexity be as invariant as possible to the precise choice of coordinate frame. In this subsection, we show that the measure obtained for stochastic complexity has these desired invariance properties.

Consider an embedding matrix $\mathbf{X} \in \mathbb{R}^{Np \times q}$. We may decompose $\mathbf{X}$ into its singular value decomposition, $\mathbf{X} = \mathbf{USV}^{\mathsf{T}}$. The squares of the singular values hence correspond to the eigenvalues of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, and it is often algorithmically easier to deal with the eigenvalue properties than directly with the singular values.

Theorem. *The stochastic complexity measure, $\mathscr{H}(\mathbf{X})$, is invariant under translation, scaling and rotation of $\mathbf{X}$.*

Proof. For two embedding matrices, $\mathbf{X}$ and $\mathbf{Z}$, the sets of singular values (and hence the stochastic complexity measure), $\mathbf{S}_X, \mathbf{S}_Z$ are equal if the probability density functions $g_X(\mathbf{X}) = g_Z(\mathbf{Z})$. It is sufficient, therefore, to demonstrate the invariance of the stochastic complexity measure to some operation, f on the data $\mathbf{Y} \stackrel{\text{def}}{=} f(\mathbf{X})$ by showing that, after the normalization operation n, $\mathbf{Z} \stackrel{\text{def}}{=} n(\mathbf{Y})$ the densities over $\mathbf{X}, \mathbf{Z}$ are equal, i.e. $g_Z(\mathbf{Z}) = g_X(\mathbf{X})$.

1. *Translation*: Let the $q \times 1$ translation vector, acting on each dimension of $\mathbf{X}$ be denoted by $\mathbf{c}$. The translation operation is then

   $$\mathbf{Y} = \mathbf{X} + \mathbf{1}_{Np}\mathbf{c}^{\mathsf{T}} = \mathbf{X} + \mathbf{C},$$

   where $\mathbf{1}_{Np}$ is an $Np \times 1$ vector of ones and $\mathbf{C} \stackrel{\text{def}}{=} \mathbf{1}_{Np}\mathbf{c}^{\mathsf{T}}$. The distribution of $\mathbf{Y}$, $g_Y(\mathbf{Y})$ in terms of $g_X(\mathbf{X})$ is then given by

   $$g_Y(\mathbf{Y}) = g_X(\mathbf{X} + \mathbf{C}).$$

   Hence, any detrending operation on the columns of the embedding matrix $\mathbf{Y}$ gives

   $$\mathbf{Z} = \mathbf{Y} - \mathbf{1}_{Np}\hat{\mathbf{c}}^{\mathsf{T}} = \mathbf{Y} - \hat{\mathbf{C}},$$

   where $\hat{\mathbf{c}}$ is the $q \times 1$ vector of estimated means, $\hat{\mathbf{C}} \stackrel{\text{def}}{=} \mathbf{1}_{Np}\hat{\mathbf{c}}^{\mathsf{T}}$. Hence

   $$g_Z(\mathbf{Z}) = g_Y(\mathbf{Y} - \hat{\mathbf{C}})$$

   and the distribution of $\mathbf{Z}$, $g_Z(\mathbf{Z})$, in terms of $g_X(\mathbf{X})$ is then given by

   $$g_Z(\mathbf{Z}) = g_Y(\mathbf{Y} - \hat{\mathbf{C}}) = g_X(\mathbf{X} + \mathbf{C} - \hat{\mathbf{C}}).$$

Since, by definition, $\mathbf{C} = \hat{\mathbf{C}}$, so

$g_Z(\mathbf{Z}) = g_X(\mathbf{X})$.

2. *Scaling*: Let the $q \times 1$ scaling vector, acting on each dimension of $\mathbf{X}$ be denoted by $\mathbf{d}$. The translation operation is then

$\mathbf{Y} = \mathbf{X}(\mathbf{1}_m\mathbf{d}^{\mathsf{T}}) = \mathbf{XD}$,

where $\mathbf{D} \stackrel{\text{def}}{=} \mathbf{1}_{Np}\mathbf{d}^{\mathsf{T}}$. The distribution of $\mathbf{Y}$, $g_Y(\mathbf{Y})$ in terms of $g_X(\mathbf{X})$ is then given by

$g_Y(\mathbf{Y}) = g_X(\mathbf{XD})|\mathbf{D}|^{-m}$.

Hence, any variance rescaling operation on the columns of the embedding matrix $\mathbf{Y}$ gives

$\mathbf{Z} = \mathbf{Y}(\mathbf{1}_m\hat{\mathbf{d}}^{\mathsf{T}})^{-1} = \mathbf{Y}\hat{\mathbf{D}}^{-1}$ and

$g_Z(\mathbf{Z}) = g_Y(\mathbf{Y}\hat{\mathbf{D}}^{-1})|\hat{\mathbf{D}}|^m$,

where $\hat{\mathbf{d}}$ is the $q \times 1$ vector of estimated variances. The distribution of $\mathbf{Z}$, $g_Z(\mathbf{Z})$ in terms of $g_X(\mathbf{X})$ is then given by

$g_Z(\mathbf{Z}) = g_Y(\mathbf{Y}\hat{\mathbf{D}}^{-1})|\hat{\mathbf{D}}|^m = g_X(\mathbf{XD}\hat{\mathbf{D}}^{-1})|\hat{\mathbf{D}}|^m|\mathbf{D}|^{-m}$.

For zero mean densities $\hat{\mathbf{D}} = \mathbf{D}$ giving

$g_Z(\mathbf{Z}) = g_X(\mathbf{X})$.

3. *Rotation:* Let $\mathbf{R}$ be a $q \times q$ rotation matrix. Hence

$\mathbf{Y} = \mathbf{RX}$.

As the normalization operation does not affect rotation, $\mathbf{Z} = \mathbf{Y}$. The distribution of $\mathbf{Z}$, $g_Z(\mathbf{Z})$ in terms of $g_X(\mathbf{X})$ is then given by

$g_Z(\mathbf{Z}) = g_X(\mathbf{R}^{-1}\mathbf{Z}) = |det\,\mathbf{R}|g_X(\mathbf{X})$.

Since, for any rotation matrix, $|det\,\mathbf{R}| = 1$, so

$g_Z(\mathbf{Z}) = g_X(\mathbf{X})$.

We conclude, therefore, that with mean removal and variance normalization, the resultant entropic complexity measure on $\mathbf{X}$ has invariancy to shift, scaling and rotation. These invariancies are vitally important when analysing data which are expressed in terms of arbitrary (but consistent) coordinate frames, such as those obtained from maps or GPS systems.

3. State modelling—HMMs

HMMs (see Rabiner (1989) for an excellent overview) are well-established models with a wide range of applications. The two main components of the HMM are its hidden state sequence, S_t, which encodes abrupt changes in the data, and a set of observation models, which model the within-state dynamics of the data. Each state is associated with an observation model, which generates (from the model's perspective) the observed data, $\mathbf{Y}_t$ (Fig. 3). Note that this observation is of the stochastic complexity measure, i.e. $\mathbf{Y}_t \stackrel{\text{def}}{=} \mathscr{H}(\mathbf{X}_t)$. Traditionally, the HMM parameters are estimated in the maximum-likelihood framework (Rabiner, 1989). Maximum-likelihood approaches, however, suffer from well-known problems of over-fitting (the likelihood always increases with model complexity) and thus the number of states in the HMM must be known a priori or estimated by application of (possibly inconsistent) penalty terms to the likelihood score. A full Bayesian approach to learning avoids these problems. This paper describes such a scheme and its application to the analysis of biological time series.

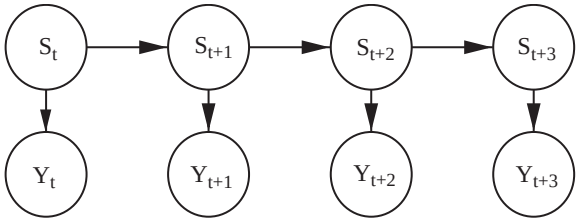

Fig. 3. A graphical model representation for the standard HMM. The set of hidden (latent) states S_t form a sequence which evolves under a first-order Markov process. Each state generates an observation, $\mathbf{Y}_t$, according to its observation model.

4. Implementation—variational Bayes HMMs

4.1. Why Bayesian learning?

From the perspective of fitting a model to a finite data sequence, it is clear that the fitting error decreases with increasing model complexity. When the model is as complex as the data then the error reduces close to zero but we no longer have a model in the strict sense, more a data look-up table. From both an information-theoretic and probabilistic learning perspective, we would like to reduce the joint uncertainty in our representation of the data (i.e. how well we model it) and the parameters of the model itself (i.e. large models have many parameters which we are unable to infer without high uncertainty). Bayesian (or probabilistic) learning consists of a principled approach to inference in the presence of uncertainty (especially important when dealing with real-world, finite and noisy data sequences) and may be broadly seen as finding a tradeoff between maximizing the probability of the data and the model simultaneously. This gives rise to a principled mathematical encoding of *Occam's razor* in which simpler models, which still have high explanatory power on the data, are naturally favoured. In the context of this paper, this means that overcomplex models (i.e. with too many assumed states) are suppressed if not supported by the data itself. The theoretical development by which this

approach is applied to the HMM is discussed in detail in the following section.

4.2. Overview of variational learning

Variational inference is a relatively new method for probabilistic inference. In this framework, integration over computationally or analytically intractable models is solved by minimizing the distance between an approximate but tractable model distribution and the exact but intractable true model distribution (see Jaakkola and Jordan (2000) and Jordan et al. (1999) for excellent tutorials). In this paper, the distribution to be approximated is the full posterior probability distribution over all hidden variables (parameters or hidden states). By approximating the full posterior, one can reap the benefits of Bayesian analysis, such as full Bayesian model estimation and automatic penalties for over-complex models thus avoiding over-fitting.[2]

Variational learning aims to minimize the so-called variational free energy (Jaakkola and Jordan, 2000) between the (intractable) model posterior P and a simpler (analytic) approximating distribution Q. The free energy is given as the Kullback–Leiber (KL) divergence between Q and P, i.e.

$$\begin{aligned}\mathscr{F} &= D(Q(H)\|P(H,V)) \\ &= \int Q(H)\log\frac{Q(H)}{P(H|V)}\,\mathrm{d}H + \log P(V),\end{aligned} \tag{6}$$

where the distribution $Q(H)$ is defined over the hidden variables H, such as parameters or hidden states and V represent the visible variables, such as the data. Since the first term on the right-hand side is always nonnegative, the divergence is an upper bound to the true log-probability of the data, i.e. the evidence. Integral (6) is maximized with respect to the individual distributions.

Given a set of hidden variables $H = \{H_i\}$, the method known as 'mean field' variational approximation assumes that the Q-distributions factorize, i.e.

$$Q(H) = \prod_i Q(H_i) \tag{7}$$

with the additional constraint that $\int Q(H_i)\,\mathrm{d}H_i = 1$. Under the mean-field assumption, the distributions $Q(H_i)$ which maximize the free energy integral (6) can be shown to be (Haft et al., 1999)

$$Q(H_i) = \frac{1}{Z}\exp\int Q(\bar{H}_i)\log P(H_i|\bar{H}_i)\bar{H}_i, \tag{8}$$

where $\bar{H}_i = H\backslash H_i$ (the set of all H excluding H_i) and Z is just a normalization constant.

In this paper, we deviate from the mean field approach in that, while keeping the mean field assumption for all model parameters θ_i, we drop the assumption for the hidden state variables which, importantly, we wish to retain their Markov structure (the conditional dependence on past hidden states). Thus, we assume the Q's to be of the following form:

$$\begin{aligned}Q(H) &\stackrel{\text{def}}{=} Q(S)Q(\boldsymbol{\theta}) \\ &= \left(\prod_{j=1}^{M} Q(\theta_i)\right)\left(Q(S_0)\prod_{t=1}^{T} Q(S_t|S_{t-1})\right),\end{aligned}$$

where S denote the hidden states and $\boldsymbol{\theta}$ the hidden Markov model parameters.

5. Variational learning of HMMs

5.1. Definitions

The HMM free energy integral to be minimized is

$$\mathscr{F} = \int Q(S)Q(\boldsymbol{\theta})\log\frac{Q(S)Q(\boldsymbol{\theta})}{P(S,\mathbf{Y}|\boldsymbol{\theta})P(\boldsymbol{\theta})}\,\mathrm{d}S\,\mathrm{d}\boldsymbol{\theta}, \tag{9}$$

where, as before, S denotes the hidden states, $\mathbf{Y}$ the data, and $\boldsymbol{\theta}$ the Markov model parameters. For an HMM in which the state variables can take on M distinct values, the model parameters consist of the initial state probability $\boldsymbol{\pi}_0$, the state transition probabilities $\pi = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_m, \ldots, \boldsymbol{\pi}_M\}$, and the parameters of the observation model. Here we use K-dimensional Gaussian observation models with mean vectors $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m, \ldots, \boldsymbol{\mu}_M\}$ and precision (inverse covariance) matrices $\boldsymbol{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_m, \ldots, \boldsymbol{C}_M\}$. Thus, the complete data model likelihood is given by

$$\begin{aligned}P(S,\mathbf{Y}|\boldsymbol{\theta}) &= P(S_0)\prod_{t=1}^{T} P(S_t|S_{t-1})P(\mathbf{Y}_t|S_t,\boldsymbol{\theta}) \\ &= P(S_0|\boldsymbol{\pi}_0)\prod_{t=1}^{T} P(S_t|S_{t-1},\boldsymbol{\pi}) \\ &\quad P(\mathbf{Y}_t|S_t,\boldsymbol{\mu},\mathbf{C}).\end{aligned}$$

The model parameter priors are assumed to be conjugate and thus we use (Bernardo and Smith, 1994):

- for an initial state probability $\boldsymbol{\pi}_0$, an M-dimensional Dirichlet density

$$Dir(\boldsymbol{\pi}_0) = \frac{\Gamma(\sum_l \kappa_l)}{\prod_l \Gamma(\kappa_l)}\prod_{m=1}^{M}\boldsymbol{\pi}_{0_m}^{\kappa_m-1},$$

  where $\Gamma(.)$ is the standard Gamma function (Press et al., 1991);
- for the transition probabilities $\boldsymbol{\pi}_m$, $M \times M$-dimensional Dirichlet densities

$$P(\boldsymbol{\pi}) = \prod_{m=1}^{M}\frac{\Gamma(\sum_l \lambda_{m_l})}{\prod_l \Gamma(\lambda_{m_l})}\prod_{n=1}^{M}\boldsymbol{\pi}_{m_n}^{\lambda_{m_n}-1};$$

[2] Being an approximation, the optimal model is chosen from the class of approximated and thus suboptimal models.

- for the observation model means $\boldsymbol{\mu}_m$, K-dimensional Normal densities ($m = 1, \ldots, M$)

  $$P(\boldsymbol{\mu}_m) \propto \mathrm{e}^{-(1/2)(\boldsymbol{\mu}-\boldsymbol{\mu}_{m_0})^{\mathrm{T}}\mathbf{C}_{m_0}(\boldsymbol{\mu}-\boldsymbol{\mu}_{m_0})};$$

- for the observation model precision matrices $\mathbf{C}_m$, K-dimensional Wishart densities ($m = 1, \ldots, M$)

  $$P(\mathbf{C}_m) \propto |\mathbf{C}_m|^{\alpha_m-((K+1)/2)}\mathrm{e}^{-tr(\mathbf{B}_m\mathbf{C}_m)},$$

  where $\mathbf{B}_m$ is the scale hyperparameter matrix for the distribution (Bernardo and Smith, 1994). Note that, as the complexity measure is one dimensional, in this application $K = 1$ and the Wishart collapses to a gamma distribution.

As mentioned in the previous section, we take the Q-distributions to factorize as

$$Q(\boldsymbol{\theta}, S) = Q(\boldsymbol{\theta})Q(S),$$

in which, for Gaussian observation models

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\pi}_0)\prod_{m=1}^{M} Q(\boldsymbol{\pi}_m)Q(\boldsymbol{\mu}_m)Q(\mathbf{C}_m).$$

The distributions for $Q(\boldsymbol{\theta})$ are identical in functional form to the priors and so, to avoid confusion, we denote the parameters of $Q(\boldsymbol{\theta})$ with tildes, e.g. $Q(\boldsymbol{\mu}_m) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_{m0}; \tilde{\mathbf{C}}_{m0})$.

5.2. Estimation

5.2.1. Model parameters

By taking the derivatives of the free energy with respect to the distributions of the unknown parameters, we obtain a set of update formulae for the parameters of the distributions. Full update equations are detailed in Appendix A.

5.2.2. Hidden states

The hidden variables (i.e. the state sequence) can be estimated using standard forward–backward message passing (Rabiner, 1989), conditioned on the data and the expectations of the model parameters under the Q-distributions. The use of the forward–backward recursions is justified by the fact that the message passing equations are fixed point equations of the free energy when the Q-distributions are assumed to be of the form given in Eq. (9) (Yedidia, 2002).

5.2.3. State space dimension

Estimation is performed over several state space dimensions. Given a fixed state space dimension estimation involves iterative application of forward–backward message passing, update of the model parameters, and estimation of the free energies. To illustrate the profound difference in inference between standard (maximum-likelihood) and fully Bayesian HMMs we apply models of varying number of hypothesized states to a synthetic data sequence in which there are two actual states. These data are shown in Fig. 4(a). Plots (b)–(d) show the resultant state labels, for 2, 4 and 10 states, inferred from the (two-state) data in plot (a) of the figure. Note that, even with 10 states at its disposal, the variational Bayes HMM does not overfit to the data. This result is in direct contrast to maximum-likelihood approaches, in which data overfitting is not naturally penalized. This is seen in Fig. 5 in which an $M = 10$ HMM overfits the state transition sequence. The data are as in Fig. 4(a).

6. Results

We present here a set of results which are intended, primarily, as 'proof of concept' rather than as an exhaustive set. Considerably more detail is presented in the companion paper, which concentrates on the biological interpretation rather than the technical details.

6.1. Number of states

The data set we now analyse consists of 48 precision GPS tracks from 12 homing pigeons released from each of four different familiar sites arranged around their home loft at the University Field Laboratory, Wytham, Oxford. Further details of the subjects and the training techniques are presented in our companion paper (Guilford et al., 2003).

Global analysis of all the data (i.e. 48 tracks) was performed using a 10th-order model. This gave rise to the mean state occupancy probabilities of Fig. 6. Note that all states save for three are insignificantly visited. This implies that, at least for the current data set, the birds' behaviour is naturally organized into three states. In all further analysis, we therefore restrict the model to $M = 3$.

6.2. Entropy

We begin by presenting the stochastic entropy measure, which forms the basis of our subsequent analysis, over the local area map. This subsection also details the approach taken in the graphical presentation of subsequent results. Fig. 7 (left) shows the local area map. Shown on this map are the four release sites (filled circles) and the loft (filled square). The coordinates shown on the map are UK grid references (Ordnance Survey). The local area map was divided into a 50×50 grid with linear spacing. Each grid element thus corresponded to a 150 m $\times$ 150 m square. Bird flight trajectories which passed through a grid element

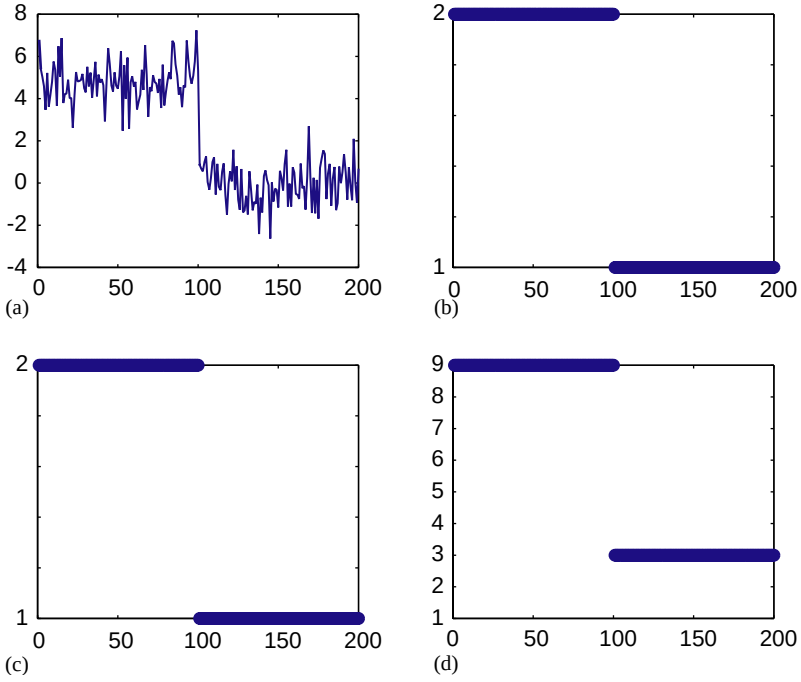

Fig. 4. (a) Synthetic two-state data. (b)–(d) states from M = 2, 4, 10 HMMs. Note that in each case only two states have non-zero occupancy. In all plots the x-axis is in samples. In plots (b)–(d) the y-axis shows the state index.

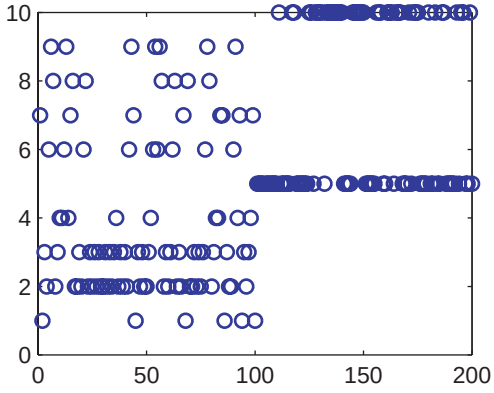

Fig. 5. Standard maximum likelihood HMM with M = 10. Note the severe overfitting to data, i.e. over partitioning of the state sequence.

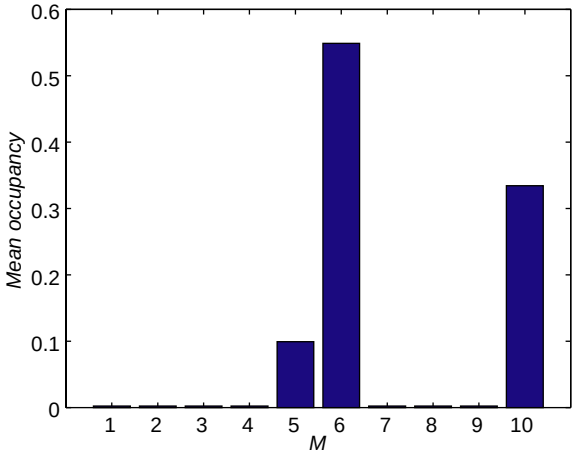

Fig. 6. Mean occupancy of states over track entropy data. Note that only three states show significant occupancy. The occupancies are normalized to sum to unity.

contributed their entropy measure (calculated at that trajectory point) to the grid element. Each grid element has an entropy value associated with it which is the mean entropy of all contributing trajectories.

We subsequently apply a spatial smoothing to the matrix of grid elements. This smoothing is achieved by a weighted average of each 150 m × 150 m grid square along with its eight nearest neighbours. The smoothing is computed using a smoothing kernel, **K**, which is chosen to be a 3 × 3 discretized Gaussian with elements given by

$$\mathbf{K} = \frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}.$$

This choice of this smoothing kernel is, of course, arbitrary but represents a prior belief in the range of visual influence of local features on the birds, namely 300 m in each direction.

If we denote **G** as the matrix of grid elements then the smoothing operation is a convolution given by

$$\mathbf{G} \leftarrow \mathbf{G} * \mathbf{K}. \tag{10}$$

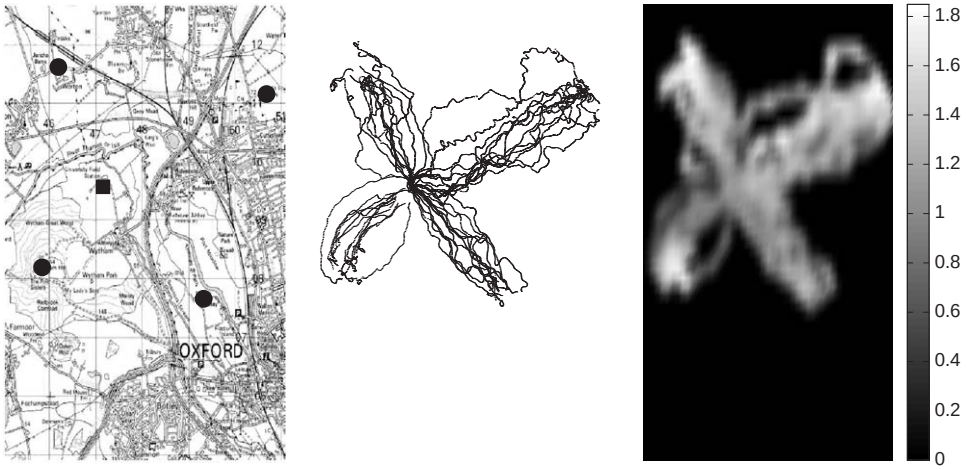

Fig. 7. Left: Local area map. Filled circles indicate the four release sites and the filled square the loft. Middle: Flight tracks of the 48 releases considered in this paper. Right: Entropy (in bits) over the local area map. Bright white indicates highest entropy and dark lowest as indicated by the scale bar to the right of the plot.

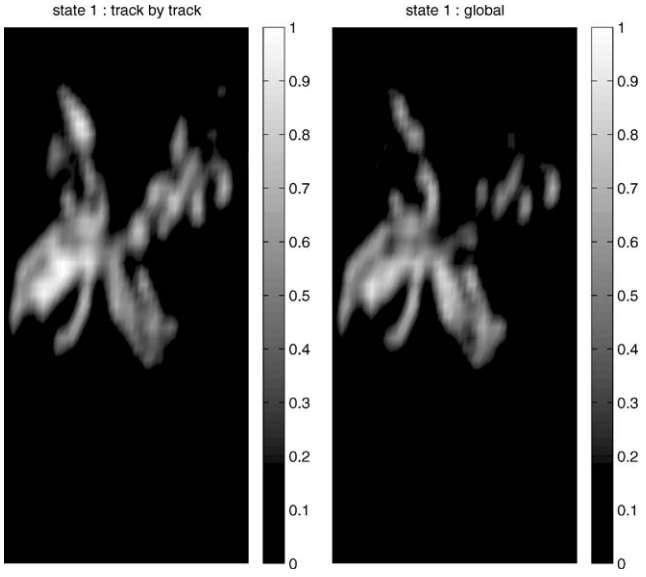

Fig. 8. State 1 probability—low entropy. Left: track-by-track analysis. Right: Global analysis.

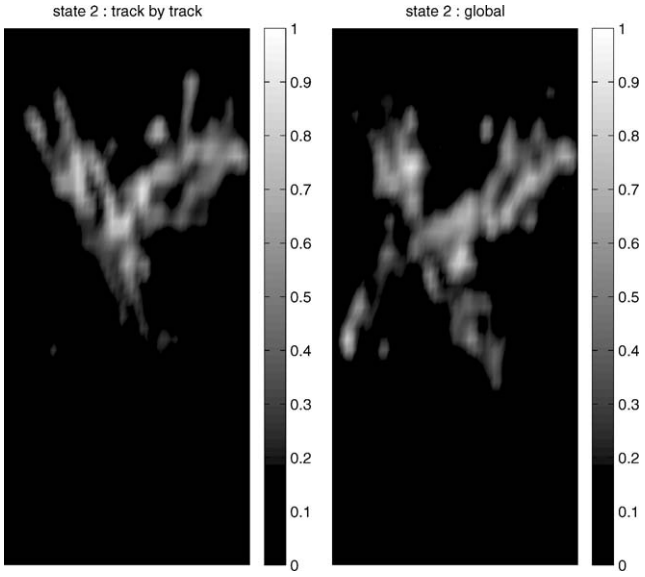

Fig. 9. State 2 probability—intermediate entropy. Left: track-by-track analysis. Right: Global analysis.

6.3. State analysis

The Markov model gives explicit posterior probabilities for each state on a sample-by-sample basis, namely

$$\gamma_t(m) \stackrel{\text{def}}{=} P(S_t = m|\mathbf{Y}).$$

We may calculate the mean state probabilities over each element in the 50×50 spatial map for all flight tracks. Following the same arguments as in the previous subsection, we apply spatial smoothing using the same smoothing kernel as in Eq. (10). We present results for the three state probabilities over the spatial map using two different protocols. Firstly, the state analysis is calculated on a track-by-track basis (i.e. an HMM is inferred for *each* track separately) and secondly the states are inferred from the *entire* 48-track data sequence as a global data set (i.e. a *single* HMM is applied to all data). These results are shown in Figs. 8–10.

We note that similar structure appears in track-by-track and global state probability maps for all three states. Notable is evidence of low-entropy (state 1) flight corridors closer to the loft (in particular, in birds released from the south-western release site) and high entropy (state 3) close to release in all cases. A more detailed analysis and interpretation is given in the companion paper. There are marked differences between the state probabilities, however, for birds released

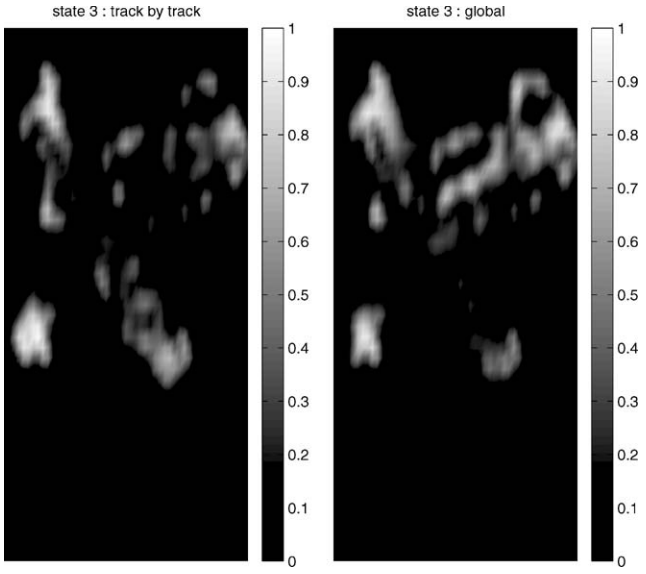

Fig. 10. State 3 probability—high entropy. Left: track-by-track analysis. Right: Global analysis.

from different sites. In particular, we note that intermediate state probabilities are close to zero for birds released from the south-western site.

7. Conclusions

We have presented a model-based analysis of bird flight tracks based on a measure of track complexity and subsequent latent-state inference using a probabilistic hidden Markov model. We observe marked changes in the track complexity from release to home in all birds and these changes appear to support the existence of self-similar states of flight behaviour (as measured using entropic complexity). Subsequent latent-state anlaysis supports the existence of three states, which correspond to high-complexity behaviour close to release, intermediate complexity and low-complexity behaviour on flight corridors leading to the loft. The main point of this paper is to conclude that the approaches introduced provide a powerful method for analysing a complex set of track data. The companion paper looks at the biological significance. The analysis detailed in this paper is by no means restricted to the current data set, nor to data from birds. Indeed, we have tried to make the approach as generic as possible, making minimal use of assumptions regarding the data. We envisage the technique will be applicable to a wide variety of other data in the biological domain.

Acknowledgements

This work is funded by the UK Engineering and Physical Sciences Research Council and the Royal Society for whose support we are most grateful. We thank Christian Brenninkmeyer, Marian Dawkins, Mike Gibbs, Kam Keung Lau, Jess Meade, Evangelos Roussos and Lyndsey Pickup for many thought-provoking discussions and comments on draft manuscripts, and Drs. George and Davis for providing stimulation.

Appendix A

A.1. Singular value decomposition

We consider an arbitrary (real) matrix $\mathbf{X}$ (not normally square). We may decompose any such $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}, \tag{A.1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, i.e. $\mathbf{V}^{\mathsf{T}} = \mathbf{V}^{-1}$. The matrix $\mathbf{U}$ is the matrix of projections of $\mathbf{X}$ onto the eigenvectors of $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and $\mathbf{S}$ is diagonal with elements $S_{ii} = \sigma_i$ where σ_i^2 is the i-th eigenvalue of $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and $\sigma_i \geqslant 0$. The resultant decomposition is usually referred to as singular value decomposition (SVD).

A.2. Update equations

In the following we make use of the notation introduced in Rabiner (1989), specifically $\gamma_t(m) \stackrel{\text{def}}{=} P(S_t = m|\mathbf{Y})$ and $\xi_t(m,n) \stackrel{\text{def}}{=} P(S_t = n, S_{t-1} = m|\mathbf{Y})$.

For Gaussian observation models we have for the posterior means $q(\boldsymbol{\mu}_m) \sim \mathcal{N}ormal(\tilde{\boldsymbol{\mu}}_{m0}, \tilde{\mathbf{C}}_{m0})$,

$$\tilde{\boldsymbol{\mu}}_{m0} = (\bar{\gamma}_m \tilde{\alpha}_m \tilde{\mathbf{B}}_m^{-1} + \mathbf{C}_{m0})^{-1}(\bar{\alpha}_m \tilde{\mathbf{B}}_m^{-1} \bar{\mathbf{Y}}_m + \mathbf{C}_{m0}\boldsymbol{\mu}_{m0}),$$
$$\tilde{\mathbf{C}}_{m0} = (\bar{\gamma}_m \tilde{\alpha}_m \tilde{\mathbf{B}}_m^{-1} + \mathbf{C}_{m0}),$$

in which $\bar{\mathbf{Y}} = \sum_{t=1}^{T} \gamma_t(m)\mathbf{Y}_t$ and $\bar{\gamma}_m = \sum_{t=1}^{T} \gamma_t(m)$. For the posterior precisions $q(\mathbf{C}_m|\tilde{\alpha}_m, \tilde{\mathbf{B}}_m) \sim \mathcal{W}ishart(\tilde{\alpha}_m, \tilde{\mathbf{B}}_m)$,

$$\tilde{\alpha}_m = \tfrac{1}{2}\bar{\gamma}_m + \alpha_m,$$
$$\tilde{\mathbf{B}}_m = \frac{1}{2}\sum_{t}^{T} \gamma_t(m)(\mathbf{Y}_t - \tilde{\boldsymbol{\mu}}_{m0})(\mathbf{Y}_t - \tilde{\boldsymbol{\mu}}_{m0})^{\mathsf{T}} + \frac{1}{2}\bar{\gamma}_m \tilde{\mathbf{C}}_{m0}^{-1} + \mathbf{B}_m.$$

The posterior initial state and transition probabilities are Dirichlet distributed with parameters, respectively,

$$\tilde{\kappa}_m = \gamma_{t=1}(m) + \kappa_m,$$
$$\tilde{\lambda}_{m_n} = \sum\nolimits_t \xi_t(m,n) + \lambda_{m_n}.$$

A.3. Variational free energy

The free energy, F, is given by

$$F = -H(S) - \mathscr{L}_{Avg} + D(Q(\boldsymbol{\theta})\|P(\boldsymbol{\theta})), \tag{A.2}$$

where $H(S)$ is the negative entropy of the hidden (state) variables, i.e.

$$H(S) = H(S_{t=0}) + \sum_{t=1}^{T} H(S_t|S_{t-1})$$

and $\mathscr{L}_{avg}$ is the average log-likelihood under the Q-distribution,

$$\begin{aligned}\mathscr{L}_{avg} = &\sum_m \gamma_{t=0}(m)(\Psi(\tilde{\kappa}_m) - \Psi\left(\sum_{l=1}^{M} \tilde{\kappa}_l\right)\\ &+ \sum_{t,m,n} \xi_t(m,n)\left(\Psi(\tilde{\lambda}_{m_n}) - \Psi\left(\sum_{r_s}^{M} \tilde{\lambda}_{r_s}\right)\right)\\ &+ \frac{1}{2}\sum_{t,n} \gamma_t(m)\left(\sum_{k=1}^{K} \Psi\left(\frac{1}{2}(2\tilde{\alpha}_m + 1 - k)\right) - \log|\tilde{\mathbf{B}}_m|\right.\\ &- \frac{K}{2}\log(2\pi) - tr(\tilde{\alpha}_m\tilde{\mathbf{B}}_m^{-1}\tilde{\mathbf{C}}_{m0}^{-1})\\ &\left.- (\mathbf{Y}_t - \tilde{\boldsymbol{\mu}}_{m0})^{\mathsf{T}}\tilde{\alpha}_m\tilde{\mathbf{B}}_m^{-1}(\mathbf{Y}_t - \tilde{\boldsymbol{\mu}}_{m0}\right)\Bigg),\end{aligned}$$

where $\Psi(.)$ is the standard digamma function (Press et al., 1991). The term $D(Q(\boldsymbol{\theta})\|P(\boldsymbol{\theta}))$ in Eq. (A.2) is the sum of all the KL-divergences between the Q- and the prior distributions, which for Gaussian densities are

$$\begin{aligned}D_{\mathscr{N}}(Q\|P) = &\frac{1}{2}\left(\log\frac{|\mathbf{C}_Q|}{|\mathbf{C}_P|} - K + tr(\mathbf{C}_P\mathbf{C}_Q^{-1})\right.\\ &\left.+ (\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)^{\mathsf{T}}\mathbf{C}_P(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)\right)\end{aligned}$$

for K-dimensional Wishart densities,

$$\begin{aligned}D_{\mathscr{W}i}(Q\|P) = &\sum_{k=1}^{K} \log\frac{\Gamma(\frac{1}{2}(2\alpha_m + 1 - k))}{\Gamma(\frac{1}{2}(2\tilde{\alpha}_m + 1 - k))}\\ &+ (\tilde{\alpha}_m - \alpha_m)\sum_{k=1}^{K} \Psi\left(\frac{1}{2}(2\tilde{\alpha}_m + 1 - k)\right)\\ &+ \tilde{\alpha}_m \log\frac{|\tilde{\mathbf{B}}_m|}{|\mathbf{B}_m|} + \tilde{\alpha}_m(tr(\mathbf{B}_m\tilde{\mathbf{B}}_m^{-1}) - K)\end{aligned}$$

and for M-component Dirichlet densities,

$$\begin{aligned}D_{\mathscr{D}ir}(Q\|P) = &\log\left(\frac{\Gamma(\sum_{m=1}^{M}\tilde{\alpha}_m)}{\Gamma(\sum_{m=1}^{M}\alpha_m)}\right)\\ &+ \sum_{m=1}^{M}(\tilde{\alpha}_m - \alpha_m)\left(\Psi(\tilde{\alpha}_m) - \Psi\left(\sum_{m=1}^{M}\tilde{\alpha}_m\right)\right)\\ &+ \sum_{m=1}^{M}\log\frac{\Gamma(\alpha_m)}{\Gamma(\tilde{\alpha}_m)}.\end{aligned}$$

Algorithm

The algorithm may be written in pseudocode as

```
initialize (using K-means for example);
repeat until convergence (e.g. free energy change <
threshold)
    estimate S using forward–backward recursions;
    estimate C̃_m0 and μ̃_m0 using
        C̃_m0 = (Σ_{t=1}^T γ_t(m) α̃_m B̃_m^{-1} + C_m0)
        μ̃_m0 = C̃_m0^{-1}(α̃_m B̃_m^{-1} Σ_{t=1}^T γ_t(m) Y_t + C_m0 μ_m0)
    estimate α̃_m, B̃_m using
        α̃_m = ½ Σ_{t=1}^T γ_t(m) + α_m
        B̃_m = ½ Σ_{t=1}^T γ_t(m)(Y_t − μ̃_m0)(Y_t − μ̃_m0)^T
              + ½ Σ_{t=1}^T γ_t(m) C̃_m0^{-1} + B_m
    estimate κ̃_m and λ̃_{m_n} using
        κ̃_m = γ_{t=0}(m) + κ_m
        λ̃_{m_n} = Σ_{t=1}^T ξ_t(m,n) + λ_{m_n}
end
```

References

Bernardo, J., Smith, A., 1994. Bayesian Theory. Wiley, New York.

Biro, D., Guilford, T., Dell'Omo, G., Lipp, H.-P., 2002. How the viewing of familiar landscapes prior to release allows pigeons to home faster: evidence from GPS tracking. J. Exp. Biol. 205, 3833–3844.

Braithwaite, V., Guilford, T., 1995. A loft with a view: exposure to the natural land-scape during development may encourage adult pigeons to use visual landmarks during homing. Anim. Behav. 49, 251–253.

Broomhead, D., King, G., 1986. Extracting qualitative dynamics from experimental data. Physica D 20, 217–236.

Grassberger, P., Procaccia, I., 1983. Measuring the strangeness of strange attractors. Physica D 9D, 189–208.

Guilford, T., 1993. Homing mechanisms in sight. Nature 363, 112–113.

Guilford, T., Roberts, S., Biro, D., Rezek, I., 2003. Positional entropy during pigeon homing II. J. Theor. Biol., submitted for publication.

Haft, M., Hofmann, R., Tresp, V., 1999. Model-independent mean field theory as a local method for approximate propagation of information. Comput. Neural Systems 10, 93–105.

Jaakkola, T.S., Jordan, M.I., 2000. Bayesian parameter estimation via variational methods. Stat. Comput. 10, 25–37.

Jordan, M.I. (Ed.), 1999. Learning in Graphical Models. MIT Press, Cambridge, MA.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1999. An introduction to variational methods for graphical models. In: Jordan, M.I. (Ed.), Learning in Graphical Models. MIT Press, Cambridge, MA.

Kember, G., Fowler, A., 1993. A correlation function for choosing time delays in phase portrait reconstructions. Phys. Lett. A 179 (2), 72–80.

Pincus, S., 1991. Approximate entropy as a measure of system complexity. Proc. Natl. Acad. Sci. USA 88, 2297–2301.

Porta, A., Baselli, G., Liberati, D., Montano, N., Cogliati, C., Gnecchi-Ruscone, T., Malliani, A., Ceruti, S., 1998. Measuring regularity by means of a corrected conditional entropy in sympathetic outflow. Biol. Cybernet. 78, 71–78.

Press, W., Flannery, B., Teukolsky, S., Vetterling, W., 1991. Numerical Recipes in C. Cambridge University Press, Cambridge, MA.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–284.

Rezek, I., Roberts, S., 1998. Stochastic complexity measures for physiological signal analysis. IEEE Trans. Biomed. Eng. 44 (9), 1186–1191.

Rezek, I., Roberts, S., 2000. A comparison of Bayesian and maximum likelihood learning of coupled hidden Markov models. IEE Proc. Sci. Technol. Measur. 147 (6), 345–350.

Steiner, I., Brgi, C., Werffeli, S., delli Omo, G., Valenti, P., Truster, G., Wolfer, D., Lipp, H., 2000. A GPS logger and software for analysis of homing in pigeons and small mammals. Physiol. Behav. 71, 589–596.

Takens, F., 1981. In: Rand, D.A., Young, L.S. (Eds.), Detecting Strange Attractors in Turbulence: In Dynamical Systems and Turbulence, Lecture Notes in Mathematics, Vol. 898. Springer, Berlin, pp. 366–381.

Wolf, A., Swift, J., Swiney, H., Vastano, J., 1985. Determining Lyapunov exponents from a time series. Physica D 16D (3), 285–317.

Yedidia, J., 2002. An Idiosyneratie journey beyond mean field theory. In: Opper, M., Saad, D. (Eds.), Advanced Mean Field Methods. MIT Press.