Expectation-Maximization Binary Clustering for Behavioural Annotation

Joan Garriga[1], John R. B. Palmer[2], Aitana Oltra[1], Frederic Bartumeus[1,2] *

1 ICREA Movement Ecology Laboratory (CEAB-CSIC), Cala Sant Francesc, 14, 17300, Blanes, Spain,
2 Centre for Ecological Research and Forestry Applications (CREAF), Cerdanyola del Vallès, 08193, Barcelona, Spain

☯ These authors contributed equally to this work.
* fbartu@ceab.csic.es

Abstract

The growing capacity to process and store animal tracks has spurred the development of new methods to segment animal trajectories into elementary units of movement. Key challenges for movement trajectory segmentation are to (i) minimize the need of supervision, (ii) reduce computational costs, (iii) minimize the need of prior assumptions (*e.g.* simple parametrizations), and (iv) capture biologically meaningful semantics, useful across a broad range of species. We introduce the Expectation-Maximization binary Clustering (EMbC), a general purpose, unsupervised approach to multivariate data clustering. The EMbC is a variant of the Expectation-Maximization Clustering (EMC), a clustering algorithm based on the maximum likelihood estimation of a Gaussian mixture model. This is an iterative algorithm with a closed form step solution and hence a reasonable computational cost. The method looks for a good compromise between statistical soundness and ease and generality of use (by minimizing prior assumptions and favouring the semantic interpretation of the final clustering). Here we focus on the suitability of the EMbC algorithm for behavioural annotation of movement data. We show and discuss the EMbC outputs in both simulated trajectories and empirical movement trajectories including different species and different tracking methodologies. We use synthetic trajectories to assess the performance of EMbC compared to classic EMC and Hidden Markov Models. Empirical trajectories allow us to explore the robustness of the EMbC to data loss and data inaccuracies, and assess the relationship between EMbC output and expert label assignments. Additionally, we suggest a smoothing procedure to account for temporal correlations among labels, and a proper visualization of the output for movement trajectories. Our algorithm is available as an R-package with a set of complementary functions to ease the analysis.

Introduction

Current movement research is undergoing a revolution. The growing capacity to collect high-resolution spatio-temporal movement data and radical improvements in data management and processing are unprecedented in this field and reminiscent of the bioinformatics revolution

of genomics and proteomics two decades ago [1]. A key challenge now is the analysis of massive movement datasets with largely different sampling rates and accuracies (e.g. high resolution GPS, standard GPS, Argos satellite, geolocation). In particular, it is critical to identify, in an unsupervised manner, movement trajectories' functional units, known as behavioural modes [1–3]. The behavioural mode is to the movement trajectory what gene is to the DNA sequence [4].

Animal movement can be analysed as a set of measurable behavioural responses to a combination of internal states, environmental factors, and evolutionary/biological constraints. Such behavioural responses or modes are manifestations of the organism's decision mechanism, providing information about the cognitive process driving the movement [5]. Identifying behaviourally significant movement modes is crucial to bringing research beyond mere statistical descriptions of movement patterns and unravelling the underlying biological processes that determine animals movement and behaviour. Establishing robust connections between patterns and processes is important for the biological interpretation of the data but also for nurturing models of movement with larger predictive capacity.

Classic approaches to movement trajectory segmentation focus on the trajectory's structure by using local measures based on tortuosity [6], first-passage time [7], residence time [8], and positional entropy [9, 10]. Other relatively simple procedures include the use of cumulative sums methods [5, 11]. On the other side, more sophisticated procedures involve Bayesian estimates of the space-time probability of being in a given behavioural mode or state [12, 13]. These can include location errors as well as environmental information [14–17]. Recent examples clearly show the suitability of state-space models for estimation and inference of behavioural modes. Especially promising have been hidden Markov models (HMMs) and some HMM variants considering autocorrelations among variables or context-dependent transition probabilities [18–23] as well as some multi-state models (MSM) [24, 25]. State-space approaches provide mechanistic and statistically sound insights about movement patterns but rely on strong *a priori* assumptions about the underlying distributions governing movement states and state transitions in time, usually requiring species-specific and time-consuming parameter estimations. Indeed, there are several frameworks for state-space modelling and the criteria to identify the best framework for a given problem still remain unclear [26]. Behavioural Change Point Analysis [27] or t-Stochastic Neighbouring Embedding (tSNE) [28, 29] do not require as many prior assumptions as state-space modelling but may also be limited by the fact that behaviours are described in a continuous parameter space which is not easy to interpret or discretize into behavioural units or modes. Many of the current behavioural annotation procedures require intense computational resources or heavy data-specific supervision (*e.g.* [19, 20, 29]) limiting its use with massive amounts of data or in comparative ecology studies (*i.e.* patterns across different populations, species or tracking methodologies).

Among the future challenges for behavioural annotation of movement trajectories is to devise scalable and minimally supervised methods capable of keeping results understandable on the basis of a sufficiently generalized and robust statistical methodology. With this aim we here develop a generalized, computationally efficient method to identify behavioural modes in movement trajectories. The method is based on a minimally-supervised multi-variate clustering algorithm that takes into account both the correlations and the uncertainty of the variables (input features), making sense of multiple time-scale behavioural events. The underlying assumption is that behavioural modes can be described by a mixture of Gaussian distributions over a binary partition of the input space. Other assumptions are just aimed at minimizing biases and sensitivity to initial conditions. The method stands out in accomplishing a good compromise between the statistical significance and the biological interpretability (semantics) of the output.

In the following we introduce the basic EMC framework and the EMbC variant for behavioural annotation, both its main conceptual features and implementation. Next we compare the EMbC with the EMC framework and basic HMM (using synthetic datasets) and with supervised expert labelling (using empirical datasets). The aim of these comparisons is not to rank the methods but simply to illustrate their relative strengths and weaknesses. Based on our results, we finally discuss the main features of the EMbC and we clarify when and why the EMbC might be most useful in the context of behavioural annotation in comparison to similar approaches.

Models

Gaussian mixture models are not new in animal movement modelling. As an example, the assumption of Gaussian mixtures is key in composite Brownian models currently very much used in movement ecology [30, 31]. Also, modelling approaches that assume Gaussian mixtures for movement variables such as speed have already proved useful for classifying animal tracking data into discrete movement modes [32, 33]. The novelty here is to use Gaussian mixtures into an EMC framework that is specifically modified to ease the interpretation of the output classification. This in turn should facilitate biologically meaningful annotation of movement trajectories.

Expectation-Maximization Clustering

The Expectation-Maximization (*EM*) algorithm [34, 35] is a well sounded, general, and iterative procedure for the maximum likelihood estimate of a parametric distribution underlying some given data, the latter eventually incomplete or showing missing values. A particular case of this algorithm is the parameter estimation of a Gaussian Mixture Model (GMM) when the generating Gaussian of each observation is unknown, commonly known as *Expectation-Maximization Clustering* (EMC), a well known methodology for the identification of clusters (*i.e.* different classes or patterns) in a multivariate data set.

The EMC formal statement is the following:

- given a dataset $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where each data point $\boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(m)})$, $(1 \leq i \leq n)$, is a vector of values corresponding to m variables, the EMC fits a k multivariate-Gaussian Mixture Model defined by the parametric set $\Theta = \{\boldsymbol{\mu}_1, \Sigma_1, \pi_1, \ldots, \boldsymbol{\mu}_k, \Sigma_k, \pi_k\}$, where $\boldsymbol{\mu}_j, \Sigma_j, \pi_j$, $(1 \leq j \leq k)$, are respectively the vector of means, the covariance matrix and the mixing coefficient of multivariate Gaussian j.

EMC is a two step iterative optimization method to estimate Θ^*, alternating between estimates of the probability of a particular observation belonging to each cluster, and estimates about the parameters Θ that maximize the likelihood of these probabilities. For a GMM, the maximization equations have a forward analytical solution [36, 37] that greatly simplifies the optimization procedure. In a few words, the algorithm proceeds as follows:

1. Initialization: take a guess Θ^g over the set of parameters;

2. Iteration loop: given the current guess Θ^g,

    a. Expectation step: for each data point i and each cluster j, compute the likelihood weight w_{ij}, (a *posterior* probability), of $\boldsymbol{x}_i$ being generated by Gaussian j, given by,

$$w_{ij} \equiv p(y_i = j \mid \boldsymbol{x}_i, \Theta^g) = \frac{\mathcal{N}(\boldsymbol{x}_i \mid \mu_j, \Sigma_j) \, \pi_j}{\sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}_i \mid \mu_j, \Sigma_j)} \qquad (1)$$

where $\mathcal{N}(\boldsymbol{x_i} \mid \mu_j, \Sigma_j)$ is the multivariate Gaussian density function:

$$\mathcal{N}(\boldsymbol{x}_i \mid \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{m/2} |\Sigma_j|^{1/2}} \, e^{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)}$$

b. Maximization step: compute a new set of parameters Θ^{new} that maximizes the likelihood [36] of these weights, given by the expressions,

$$\pi_j^{new} = \frac{1}{n} \sum_{i=1}^{n} w_{ij} \tag{2}$$

$$\mu_j^{(l),new} = \frac{\sum_{i=1}^{n} w_{ij} x_i^{(l)}}{\sum_{i=1}^{n} w_{ij}} \tag{3}$$

$$\sigma_j^{(r,s),new} = \frac{\sum_{i=1}^{n} w_{ij} \left(x_i^{(r)} - \mu_j^{(r),new}\right) \left(x_i^{(s)} - \mu_j^{(s),new}\right)}{\sum_{i=1}^{n} w_{ij}} \tag{4}$$

where $\mu_j^{(l),new}$, $1 \le l \le m$, are the components of the mean vector μ_j^{new}, and $\sigma_j^{(r,s),new}$, $1 \le r \le m$, $1 \le s \le m$ are the variances and covariances of the covariance matrix Σ_j^{new}.

c. Use Θ^{new} as the parametric guess for the next iteration, that is, take $\Theta^g \equiv \Theta^{new}$.

3. Output classification: at the end of the process, each data point is assigned to its most probable cluster.

This iterative procedure is theoretically guaranteed to increase the likelihood at each step and, although the algorithm does not promise to reach a global maximum of the likelihood function, it is indeed guaranteed to converge to a local maximum, dependent on the initial conditions [36, 38, 39]. In practice, it is common to start it from multiple random initial guesses and select the one with the largest likelihood. Usually, the process is stopped after a prefixed number of iterations or when the increments of likelihood are less than a prefixed δ.

In a typical EMC implementation, the number of desired output clusters must be specified, and the algorithm will return that number of clusters. A value $\sigma_{min}^{(r,r)} > 0$, $(1 \le r \le m)$ must be specified to limit the minimum variance of each variable. This parameter avoids errors derived from indefinite covariance matrices along the optimization process. In practical terms, σ_{min} can be directly related to measurement errors (or maximum resolution) of the variables and will limit the minimum range of variability (*i.e.* minimum standard deviation) within the clusters obtained.

EMbC Algorithm

The *generalized EM algorithm* is a family of variants of the EM algorithm aimed at overcoming particular problems (e.g. difficult E-step/M-step computations, slow convergence [38]). The general behaviour of these variants is not always clear and they may not yield monotonic increases in the log likelihood over iterations [38]. The *Expectation-Maximization binary Clustering* (EMbC) algorithm is a variant of the EMC algorithm [34, 35] aimed to address: (i) clustering interpretability and, (ii) the variability in data reliability, two key issues in behavioural

annotation of movement. The novelty is that the clustering is driven towards a statistically meaningful classification that should be easier to interpret by experts and that, similarly to other methods [40, 41] it can take into account uncertainties associated to the data points.

Clustering semantics: the delimiters. In any unsupervised clustering procedure, one should distinguish cluster identification from cluster *semantics*, the intuitive interpretation of the obtained clusters. Classical implementations of the EMC can generate statistically sound clustering configurations that are difficult to interpret in behavioural terms, that is, at the cost of clear semantics.

In the EMbC algorithm semantically meaningful clustering is achieved by introducing a set of parameters, denoted as *delimiters*. A delimiter is a value that splits the range of a variable into a binary discretization. The whole set of delimiters defines a partition of the variable space into regions where each variable takes either low (L) or high (H) values. The binary nature of this partition is what favours the link between elementary and semantically meaningful labelling. As an example, classical behavioural annotation is commonly based on velocity and a turning behaviour estimate (*e.g.* turning angle, angular correlation, tortuosity). In this case, a binary labelling has a direct intuitive interpretation: low velocities and low turns (LL) can be interpreted as *resting*, low velocities and high turns (LH) as *intensive search*, high velocities and low turns (HL) as *travelling* or *relocation*, and high velocities and high turns (HH) as *extensive search*. The semantic annotation is however variable-dependent and species-specific.

In a general multivariate case, each delimiter is associated to two adjacent clusters having the same combination of low and high values for all variables, except for the splitting variable, which takes low values in one and high values in the other. In other words, we have one delimiter for each variable and each combination of highs and lows of the rest of the variables. For m variables this makes a total of $m2^{m-1}$ delimiters. We use a multivariate notation denoting delimiters by r_Z where Z is a length m sub-index, based on an ordered sequence of the variables. Each element of the sub-index is either L or H except for the splitting variable for which we use a dot, according to the combination of values at both adjacent clusters. As an example, in a 3-variate case, $r_{L.H}$ denotes the delimiter for the second variable, separating the two clusters LLH and LHH, in which the first and the third variables take low and high values respectively.

Conceptually, the delimiters are related to the frontier of equiprobability between two adjacent clusters, and are used to bound the computation of the Gaussian means within the regions that they delimit. In such a way, the mean of each cluster can not drift away from its associated binary region. To illustrate this issue, we show a comparison of two bivariate (velocity and turning angle) clusterings of the same trajectory (Fig 1): one resulting from a classical EMC implementation (left panel), and the other resulting from the EMbC variant, the dashed lines depicting the final value of the delimiters computed by the EMbC algorithm (right panel). Starting with exact initial conditions, these two algorithms yield output clusterings corresponding to different local optima. In the left panel it is difficult to obtain a clear semantics based on velocity/turn. In the right panel the clustering shows a meaningful partition of the variable space into LL/LH/HL/HH regions, accounting for a clear cluster semantics.

At each EM iteration the delimiters depict the current state of the binary regions, and the subset of data points in each region is used to compute the mean of the corresponding Gaussian component. As the Gaussians spread beyond the limits of the binary regions, at any step (including the final output) data points are assigned to the most probable cluster, regardless of the values of the delimiters. This is the reason for the few mismatches that can be observed between the clusters and the binary regions in the right panel. Only in case of equal probabilities, the delimiters constitute a further criterion to assign labels to data-points. Importantly, there is no guarantee that either of the algorithms (EMC and EMbC) will always be better in terms of likelihood. Both EMC and EMbC will just reach the best optimum attainable from
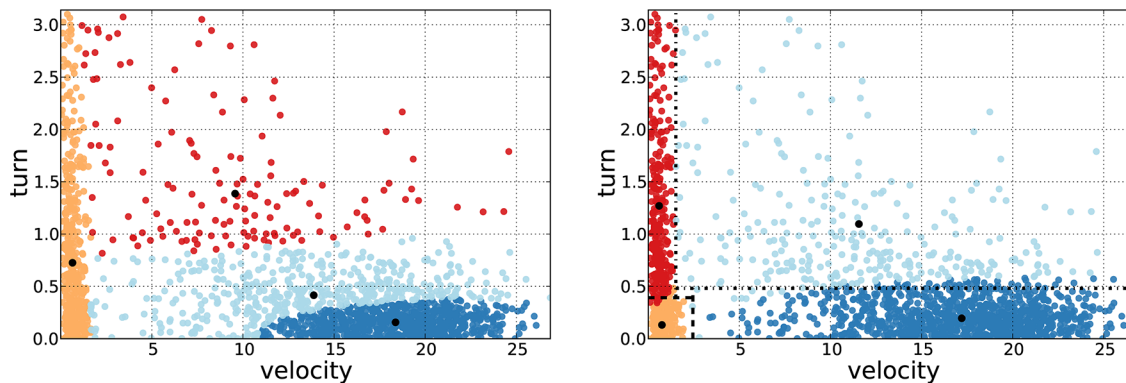
Fig 1. Cluster semantics. Comparison of the EMC (left) and EMbC (right) algorithms. Bivariate (velocity/turn) scatter-plots showing the clustering reached by each algorithm, corresponding to the same trajectory and exact initial conditions. Clusters are shown in different colours. In the right panel, the EMbC delimiters determining the final L/H binary regions are depicted as dashed lines (r_L, $r_{.L.}$) and dot-dashed lines ($r_{.H}$, r_H.). The centroids of each cluster are shown as black dots. Left: the EMC yields an output clustering that is difficult to link to a clear semantics. Right: the EMbC is driven by the delimiters, forcing the centroids to lay within the associated binary regions, yielding a final clustering that can be clearly interpreted in terms of L/H values of the variables (orange:LL, red:LH, cyan:HL and blue:HH). The matching among binary regions and clusters is not perfect because data-points are assigned to clusters depending on their weights, not on the delimiter values. In this case, the EMbC performs better (the clustering log likelihoods are -3.3368 for the EMC and -3.2180 for the EMbC), but this result can not be generalized.

doi:10.1371/journal.pone.0151984.g001

any given starting point. In our example, the local optimum based on EMbC is better ([Fig 1](#) right panel), but this must not be generalized. The concern here is not to reach higher likelihood partitions but rather to reach meaningful partitions even at some cost in likelihood.

Data accuracy and reliability. Movement trajectory data is associated to different sources of uncertainty: (i) spatial errors due to technical limitations of the systems used (*e.g.* GPS, Argos) or interferences in signal transmissions of the geopositioning system; and (ii) temporal errors due to difficulties in inferring a location at a given time, which generates irregular time steps. Therefore, estimated variables such as velocity or turn, which depend on the sampling rate and on the locations themselves (Section A in [S1 Text](#)), present different degrees of reliability or accuracy.

Similarly to [[40](#), [41](#)], the reliability of the data is implemented as an additional weighting coefficient in Eqs ([3](#)) and ([4](#)), giving less weight to the less accurate values in the estimation of the Gaussian parameters, and favouring the more accurate ones. These coefficients should be given by a *reliability* function that can not be generalized, as it will be variable-specific and dependent on the source of error considered. For the general case we denote them as,

$$u_i^{(l)} = \mathcal{E}(x_i^{(l)}) \tag{5}$$

where $\mathcal{E}(x_i^{(l)})$ is a function that returns a normalized reliability coefficient for the data value $x_i^{(l)}$ based on the source (or multiple sources) of error that might be operating upon the input variable *l*. In [S1 Text](#) Section C we suggest an example of a reliability function that can be used to take into account the reliability of the values of velocity and turn computed from a real trajectory with heterogeneous time intervals.

Implementation. Implementing the modifications described above to account for clustering semantics and data reliability imply relevant changes in the maximization step (M-step) of the algorithm:
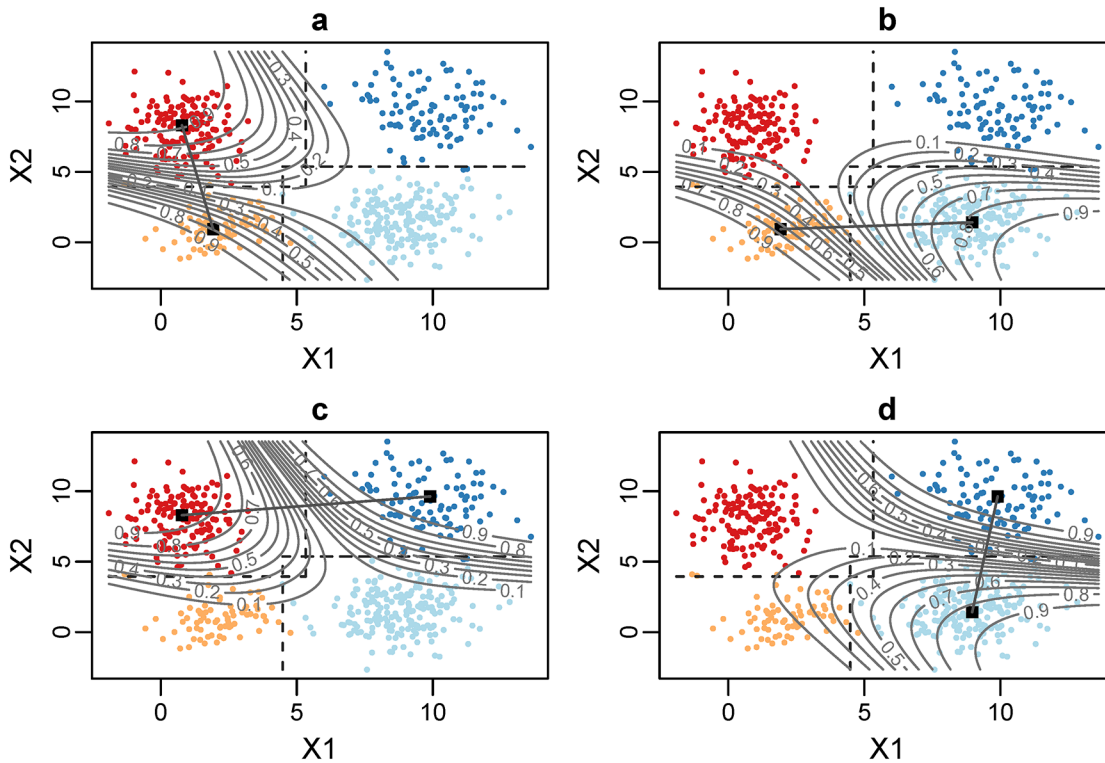
Fig 2. Computation of the delimiters. This is a synthetic example with data drawn from a bivariate ($X1$, $X2$) GMM with four components, showing the state of the clustering at the third iteration. Current delimiters are shown as dashed lines. To depict the idea, we defined a grid $\mathcal{G}$ over the range of the scatter-plot and we computed the likelihood weights $w_{LL}(\mathcal{G})$, $w_{LH}(\mathcal{G})$, $w_{HL}(\mathcal{G})$, $w_{HH}(\mathcal{G})$. We show the contour lines corresponding to the differences in likelihood weight and the lines connecting the means of the two adjacent clusters: panel a) shows the line (μ_{LL}, μ_{LH}) and the contour $abs(w_{LL}(\mathcal{G}) - w_{LH}(\mathcal{G}))$; panel b) shows the line (μ_{LL}, μ_{HL}) and the contour $abs(w_{LL}(\mathcal{G}) - w_{HL}(\mathcal{G}))$; panel c) shows the line (μ_{LH}, μ_{HH}) and the contour $abs(w_{LH}(\mathcal{G}) - w_{HH}(\mathcal{G}))$; and panel d) shows the line (μ_{HL}, μ_{HH}) and the contour $abs(w_{HL}(\mathcal{G}) - w_{HH}(\mathcal{G}))$. In each case, we can observe that the delimiter crosses the corresponding line between means at the point with minimum likelihood difference; panel a) $r_{L.}$ (turn splitting value for low values of speed); panel b) $r_{.L}$ (speed splitting value for low values of turn); panel c) $r_{.H}$ (speed splitting value for high values of turn); panel d) $r_{H.}$ (turn splitting value for high values of speed).

doi:10.1371/journal.pone.0151984.g002

1. Foremost, the delimiters have to be computed. The values of the delimiters correspond to the point of minimum difference in likelihood weight in between two adjacent clusters. At each iteration the delimiters are computed by projecting the data points onto the straight line connecting the current means of the adjacent clusters. The likelihood weights of the projected data points are computed for both clusters with the current parameters, and the delimiter is set to the value of the data point for which the difference in those likelihoods is minimum ([Fig 2](#)).

2. For each cluster j, we must determine the set $\mathcal{R}_j$ of points lying in the region determined by the corresponding delimiters. We note that in the general case, the delimiters will not constitute a perfectly definite partition of the variables space, and some points may belong to different regions at the same time, as shown in [Fig 3](#), contributing to the computation of the Gaussian means of all related clusters.
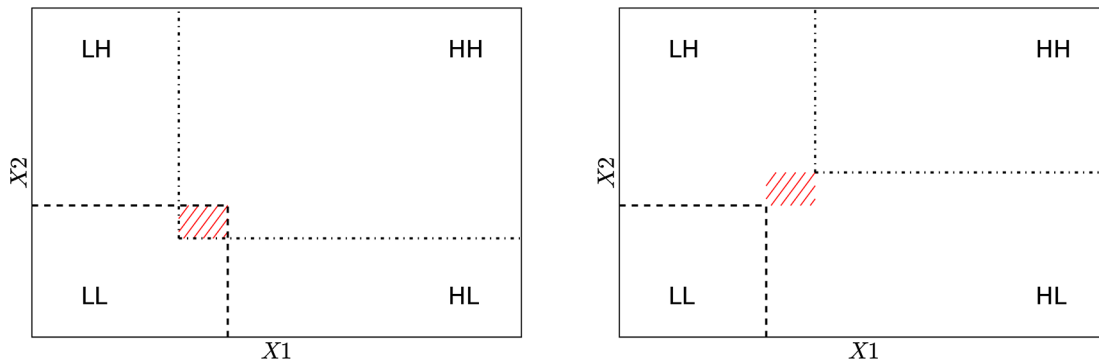
Fig 3. Definition of the binary regions. At each iteration step, the most common situation is that the delimiters do not determine a perfect partition of the variable space. We show two typical cases for the bivariate ($X1$, $X2$) case with delimiters overlapping (left panel) and non-overlapping (right panel). Delimiters are shown as dashed lines (r_L, $r_{L.}$) and as dot-dashed lines (r_H, $r_{H.}$). Left: the data points in the middle red area may belong either to $\mathcal{R}_{LL}$ or $\mathcal{R}_{HH}$, hence they are considered in the computation of both μ_{LL} and μ_{HH}. Right: with non-overlapping delimiters, we can still figure out an overlapping area between regions $\mathcal{R}_{LH}$ and $\mathcal{R}_{HL}$ by extending the delimiters (middle red area), hence data points in this area are considered in the computation of both μ_{LH} and μ_{HL}.

3. Finally, we recompute the Gaussian parameters, bounding Eq (3) to the sets $\mathcal{R}_j$ and including the reliability function in Eqs (3) and (4), that is,

$$\mu_j^{(l),new} = \frac{\sum_{i \in \mathcal{R}_j} u_i^{(l)} w_{ij} x_i^{(l)}}{\sum_{i \in \mathcal{R}_j} u_i^{(l)} w_{ij}} \tag{6}$$

$$\sigma_j^{(r,s),new} = \frac{\sum_{i=1}^{n} u_i^{(r,s)} w_{ij} \left(x_i^{(r)} - \mu_j^{(r),new} \right) \left(x_i^{(s)} - \mu_j^{(s),new} \right)}{\sum_{i=1}^{n} u_i^{(r,s)} w_{ij}} \tag{7}$$

where $u_i^{(r,s)}$ weights the combined effect of uncertainty on variables r and s, and is computed as the normalized length,

$$u_i^{(r,s)} = \frac{1}{\sqrt{2}} \sqrt{\left(u_i^{(r)} \right)^2 + \left(u_i^{(s)} \right)^2}$$

The delimiters become the essential part of the parametric set Θ, and therefore, the model is no longer a standard GMM but a constrained variant. The optimization through the likelihood space is driven by the new conditions imposed, which force each Gaussian to have its mean inside meaningful regions, restricting the potential positions of the cluster centroids and the type of clusterings allowed.

A major consequence is that Eqs (6) and (7) do not correspond to maximization expressions. This change however, does not jeopardize the convergence of the EMbC algorithm. The effect of our modifications in the EMC algorithm is an increasing likelihood optimization process, interspersed with likelihood drops at sporadic iterations. Every drop in likelihood can be considered a restart in the likelihood landscape from a new (but not so blind) guess, with the likelihood being lower compared to the previous step but higher compared to the likelihood value from which we started. A steady likelihood decrease at some stage of the optimization process is an indication of some discrepancy between the binary and the optimal likelihood

partitions, and that the input data might not be suited for a binary partition. In such cases, the algorithm may get stuck in a cycle balancing between both possible solutions. This situation (more likely to occur on the last iterations) is automatically detected and the algorithm stops returning a corresponding warning message. The likelihood dynamics are further discussed and illustrated with some examples in S1 Text Section E.

Unlike the EMC algorithm, the number of output clusters is given here by the number of variables used, $k = 2^m$. However, during the process of likelihood optimization, some clusters can vanish while being absorbed by adjacent clusters. Thus, $k = 2^m$ is only an upper bound to the final number of clusters. This limitation in the number of clusters is not a drawback but rather a consistency with our main motivation of favouring the semantic interpretation of the final clustering. Although there is no restriction on the number of variables (we are not considering computational limitations) the EMbC is intended to be used with not more than 5 or 6 variables, yielding a maximum of 32 or 64 clusters, what is usually far beyond the number of potential behaviours of interest in any biological application, and far beyond the number of behaviours that an expert might easily handle. The key point here is to determine a few variables conveying the right information to decode the set of behaviours of interest. In case of using feature selection or dimensionality reduction methodologies (*e.g. principal component analysis*) it is important that the selected input features can be well understood in order not to compromise the interpretability of the output clustering, which is the ultimate purpose of the EMbC algorithm. Input features should be selected based on their physical or biological meaning.

The parameter σ_{min} is variable-specific and determines the minimal resolution of the clusters. It can be set by default to orders of magnitude lower than the expected variances (*e.g* σ_{min} = 2.22e − 16 or whatever it is the double-precision of the computer) for each of the variables or else be used to limit the minimum range of variability (*i.e*. minimum standard deviation) within the clusters. Rather than a subjective question, this is usually related to the physical concept expressed or measured by the variable under consideration. For instance, regarding to movement variables like velocity and turn, σ_{min} can be directly related to physical/biological constraints as well as to measurement errors (or maximum resolution) of geolocation devices. Thus, values of σ_{min} in the order of 0.01 m/s for velocity and 0.087 rad (5 degrees) for turns, would work for a wide range of species. For this reason, σ_{min} should be regarded as a variable-specific factor that sets the analysis resolution rather than a user free parameter.

Also key in the algorithm implementation is to minimize prior assumptions, biases and sensitivity to initial conditions. With this aim, the EMbC starting point is automatically set as the most uninformative condition, that is: (i) each data point is assigned a uniform probability of belonging to each cluster, (ii) the prior marginal distribution of the clusters is also uniform (each cluster starts with the same number of data points), (iii) the starting partition, *i.e*. the initial delimiters position, is selected based on a global maximum entropy criterion, thus conveying the minimum information possible. The latter condition is computed by sequentially selecting the variable such that its median value splits the related set of data into high and low subsets with maximum entropy. This is a simple algorithm for the 2D case but slightly more complex for the general case of d dimensions.

Analysis

We use simulated and empirical trajectories to asses the EMbC algorithm and illustrate its main features. Our examples are mostly based on local measures of velocity and turn but we also show an example with other movement variables (Sections A and B in S1 Text).

Synthetically generated and annotated movement trajectories allow us to measure the performance of the algorithm and compare it with closely related methods such as EMC [34, 35]

Table 1. Empirical datasets.

Common name	Scientific name	Type	Tracks	Fixes	$\hat{\tau}$ secs.	Context
Osprey	*Pandion haliateus*	Argos	1	594	3600	migration
Straw-coloured fruit bat	*Eidolon helvum*	GPS	1	434	299	roosting/foraging
Cory's shearwater	*Calonectris diomedea*	HR-GPS	1	2543	155	foraging
Nematode	*Caenorhabditis elegans*	HR-video	6	10203	3	search

Empirical datasets. Fixes indicate the total number of locations (data points) in each data set. $\hat{\tau}$ refers to the mean time interval (in seconds) between fixes.

doi:10.1371/journal.pone.0151984.t001

and HMM, commonly used for modelling of animal movement data [19, 21, 42]. We use the implementations of EMC and HMM included in the R-packages EMCluster [43, 44] and DepmixS4 [45], respectively.

Synthetic trajectories are generated by assuming four clusters (behavioural modes) with different degrees of mixture or overlap, γ = {0.01, 0.05, 0.1}, where the lower the value of γ the more blurred are the clusters. The trajectories are of different lengths n = {50, 100, 200, 400, 800, 1600} and the sequence of behavioural modes or states is constructed either by sampling states from a 4×4 transition matrix (Markov-chain sampling) or else by sampling states using the parameters of the prior distributions π_j, $1 \leq j \leq 4$ (prior-mixture sampling), (see Section F in S1 Text for more details).

Our empirical tracks (see Table 1) cover different ecological contexts and a variety of tracking technologies including high-resolution GPS (shearwater), standard GPS (bat), Argos (osprey) and video recorded (nematode) datasets. The data sets are further described in S2 Text and are included in S1 Data.

EMbC outputs are shown in different ways (*i.e.* scatter-plots, labelling profiles) including a bursted visualization of annotated trajectories based on the conversion into segments of all consecutive locations sharing the same label (S1 Text Section D). In the case of supervised datasets (*i.e.* synthetic datasets and also the bat dataset, which included an expert annotation), we use confusion matrices to yield a numerical assessment of the performance of the algorithm with respect to the supervised labelling. Commonly used performance measures based on the confusion matrix are *recall*, *precision* and *F-measure* (S1 Text Section G). Beyond a pure numerical assessment of performance, a confusion matrix offers a clear depiction of how EMbC annotation is redistributed into expert label assignments, reflecting any departure of the information conveyed by the input features from the information used by experts (*e.g.* velocity and turn *versus* visual information). Thus, from the analysis of the confusion matrix one can gain much knowledge about the suitability of the selected features and on the behaviour of the algorithm itself.

Other aspects analysed are: (i) the coarse-graining of EMbC behavioural annotation, and (ii) the robustness of the EMbC with respect to potential sources of error like data loss and data inaccuracy.

The results that we show are mostly direct outputs of the EMbC R-package. In the S2 Text file we spell out the code used to generate them and we work through them further to give a brief overview of the use of the package. The empirical data sets used in the examples are included in S1 Data.

Smoothing of annotated trajectories

The EMbC algorithm generates *local* behavioural annotations without considering their temporal context. Based on EMbC, labels are given for each observed location and reveal any small change in behaviour irrespective of how this change is framed in a broader temporal context (*e. g.* a long-term predominant behavioural mode). If coarse-grained patterns are desired, the EMbC provides two means for smoothing the output:

1. Pre-processing of the trajectory using running windows to compute averaged local measures, with the length of the running window representing a behavioural scale of interest.

2. Post-processing of the output based on the temporal behavioural correlations, a feature explicitly implemented in state-space segmentation algorithms [12, 13, 19, 20].

In the latter case, the EMbC smoothing function makes use of the likelihood weights w_{ij} of location i belonging to cluster j, information provided by the EMbC algorithm. In its most basic implementation, the function looks for *singles*, that is, locations with labels that differ from equally labelled neighbouring locations, and checks the condition $(w_{ic} - w_{in}) \leq \delta_w$, where i is the *single* location index, w_{ic} and w_{in} are the likelihood weights with respect to its current and its neighbouring assignments (clusters c and n, respectively), and δ_w is a threshold parameter expressing the user's will to accept the change. The subjectiveness of this parameter is our reason for keeping this smoothing function apart from the overall clustering procedure.

Robustness to data loss and data inaccuracy

We studied the robustness of EMbC annotation to data loss by removing data points from the set of velocity/turn pairs (Fig 4a). Note that we cannot study data loss by eliminating locations from the trajectory because this would simply change the values of velocity and turn (which depend on actual sampling intervals), leading to a totally *new* dataset. In the sub-sampling process it is important to preserve the underlying behavioural distribution. Thus, sub-sampled datasets were generated by assigning a uniform random value $0 \leq p_i \leq 1$ to each data point and removing all those points with $p_i < k_{dl}$, with $0 < k_{dl} \leq 1$ being k_{dl} the data loss factor. For
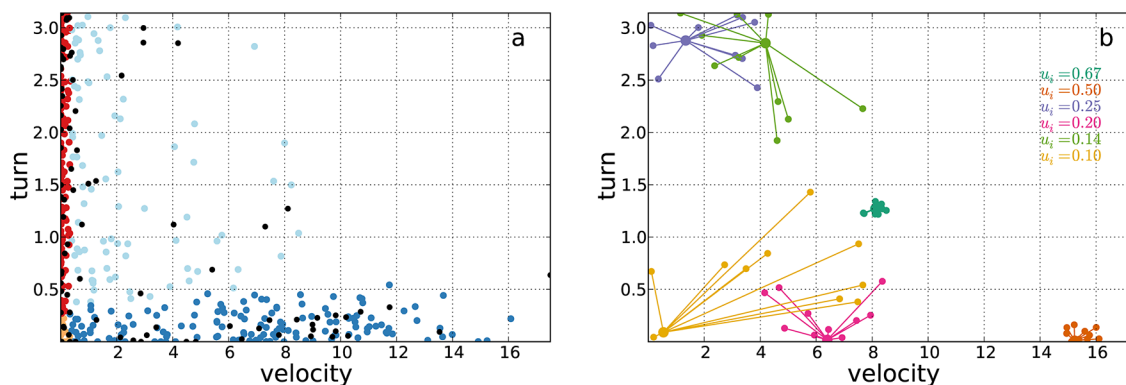

Fig 4. Procedures for robustness tests. Procedures used in the robustness tests. a) Data-loss: sub-sampled datasets are generated by assigning a uniform random value $0 \leq p_i \leq 1$ to each data-point and removing all those points with $p_i < k_{dl}$, with $0 \leq k_{dl} \leq 1$ being the data loss factor, (in this example $k_{dl} = 0.2$, removed points are marked as black dots). b) Data-inaccuracy: jittered datasets are generated by jittering the data-points using a noise function based on the associated uncertainties Eqs (8) and (9); we show some example data points connected with several jittered versions of themselves with $k_{di} = 0.05$, using different colours to identify the correspondences, and also to relate each one with its associated reliability u_i indicated in the legend; note that the more unreliable is a data point the more different could have been its observed value.

doi:10.1371/journal.pone.0151984.g004

each empirical trajectory, we generated datasets with different values of k_{dl}. After running the EMbC on the subsampled datasets we compared the output labels with the corresponding labels in the original (full) dataset, the latter considered the ground truth for comparative purposes.

We also explored the effect of including the reliability of the data Eq (5) in Eqs (6) and (7). In particular, we considered the effect of sampling rate heterogeneity (*i.e.* the larger the time gap between two successive locations, the larger the probability of inaccurate velocity and turn estimates), and to what extent our approach decreased the sensitivity of the final clustering to this source of inaccuracy (see Eq 10 in S1 Text Section C as an instance of Eq (5) devised for this particular case). We did so by jittering the data points in the scatter plot (Fig 4b) using a noise function based on a uniform distribution over an area around the data point proportional to the associated time gap,

$$max(min(X), \boldsymbol{x}_i - \boldsymbol{\Delta}_i) < \hat{\boldsymbol{x}}_i < min(max(X), \boldsymbol{x}_i + \boldsymbol{\Delta}_i) \quad , \tag{8}$$

where X is the (multivariate) dataset, $\boldsymbol{x}_i$ and $\hat{\boldsymbol{x}}_i$ (vectors) are the original and jittered data points, and $\boldsymbol{\Delta}_i$ (vector) is computed as,

$$\boldsymbol{\Delta}_i = k_{di} \, max(X) \, (\tau_i - \tilde{\tau})/\tilde{\tau} \quad , \tag{9}$$

where $0 < k_{di} \ll 1$ is a data inaccuracy factor determining a jittering range $k_{di} \, max(X)$. Thus, $\boldsymbol{\Delta}_i$ is a fraction of the jittering range proportional to the relative length of the time interval $\tau_i = t_{i+1} - t_i$ with respect to the most frequent time interval $\tilde{\tau}$ (the mode of the τ distribution). For each empirical trajectory, and for different k_{di} values, we compared the labellings obtained in jittered datasets with the corresponding non-jittered labellings, with and without implementing a reliability function in Eqs (6) and (7). This is only a particular example focused on inaccuracies derived from sampling heterogeneity but the same could apply to other sources of uncertainty, such as geopositioning errors. The effects would be similar, since the higher the uncertainty of the values, the less their influence in determining the final clustering.

Results

We used synthetic trajectories and empirical datasets to evaluate the performance and illustrate the outputs of the EMbC algorithm. The sequences of movement states generated in the synthetic trajectories come from two sampling schemes: Markov-chain and mixture-prior. The empirical datasets covered different tracking technologies (e.g. GPS, Argos) and a wide range of sampling heterogeneity.

Simulated trajectories

The EMbC algorithm recovered the modelled clusters but with some expected sensitivity to both the level of mixture of the clusters γ, and the size of the data set n. In general (Fig 5), the performance was above 90% for $n \geq 200$ decreasing around 80% for the shortest trajectories, *i. e.* $n \leq 100$.

With synthetic trajectories derived from the Markov-chain sampling scheme, where the sequence of states comes from a transition probability matrix, (Fig 5 upper panels), the three algorithms (EMC, EMbC and HMM) showed a similar behaviour. For $\gamma = \{0.01, 0.05\}$ (well-mixed clusters) the performance of the EMbC was in between the one of the HMM (the best) and the EMC. However, for $\gamma = 0.1$ (well-defined clusters) and $n \geq 200$ the EMbC outperformed the HMM. Compared to EMC and EMbC, HMM works best when the binary partitions are blurred but the temporal sequence of states is well-defined, according to a transition matrix, and it can adequately exploit this information to improve state assignment.
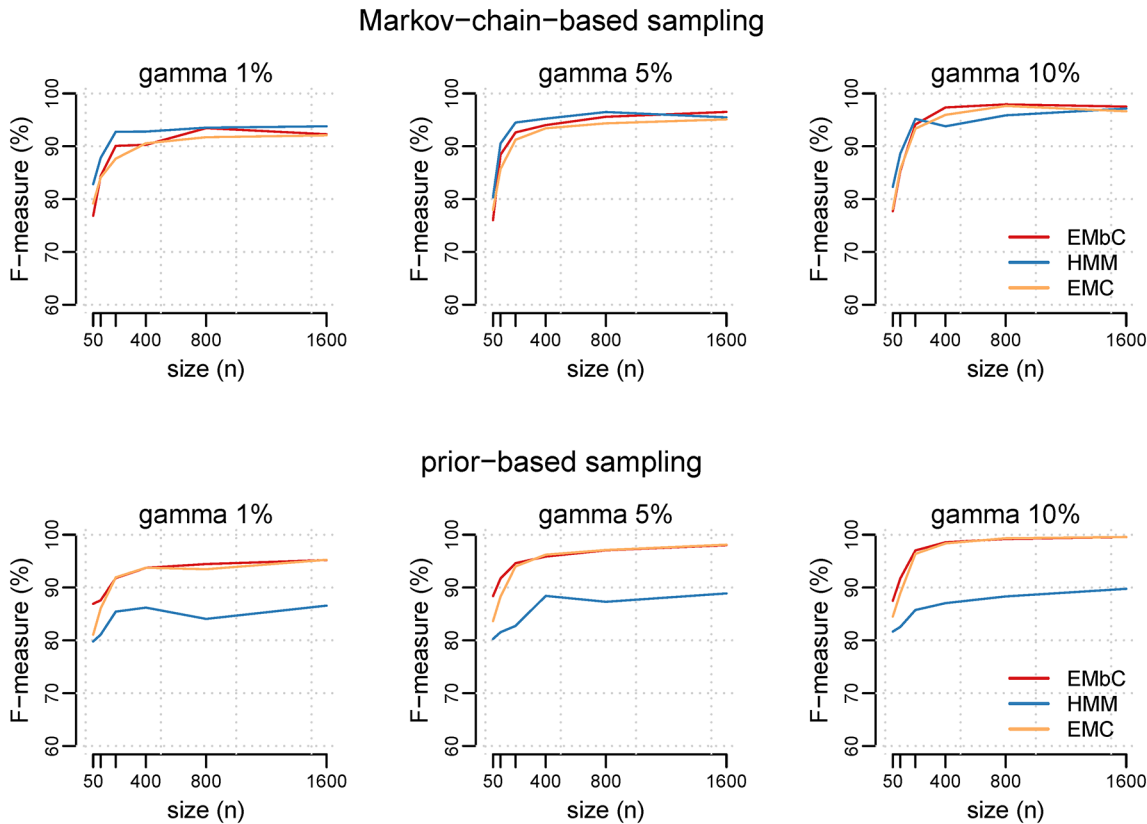
Fig 5. Performance comparison among EMbC, EMC, and HMM. Average performance of EMbC, EMC and HMM on 100 synthetic trajectories drawn from a GMM (4 components) using a Markov-chain-based sampling scheme (top panel) and a prior-based sampling scheme (bottom panel), for different trajectory sizes ($n = \{50,100,200,400,800,1600\}$) and definition of the binary regions ($\gamma = \{0.01,0.05,0.10\}$). Values of performance are given in terms of F-measure.

With synthetic trajectories derived from the mixture-prior sampling scheme, where the sequence of states comes from the set of prior cluster distributions (Fig 5 lower panels), the EMbC and the EMC presented similar results. Expectation-Maximization clustering procedures do not take into account the temporal correlation of states but take the best out of the synthetically generated binary partitions, even if the clusters are well-mixed or blurred. The EMbC performed slightly better than EMC for low values of γ and n. In contrast, the performance of the HMM was clearly much lower, both compared to EMC and EMbC, and also compared to the results obtained when HMMs are applied to trajectories based on Markov-chain sampling schemes.

The larger the size of the data set, the more evidence about the partition of the input space, and the better the performance of the three algorithms for both sets of trajectories (Markov-chain and mixture-prior sampling schemes). Tables B and C in S1 Text reinforce the idea that EMbC performs better when the information is compromised, either because the clusters are not well-defined (small γs) or because the amount of information is small (low ns). In addition, the EMbC leads straightforwardly to the binary partition and keeps a high stability in the results with the lowest values of *root mean square error* (RMSE), (see Tables B and C in S1 Text). In contrast, the EMC and the HMM algorithms must be fed with a starting seed and can be extremely sensitive to it, specially when data sets are sparse. A negative effect of this
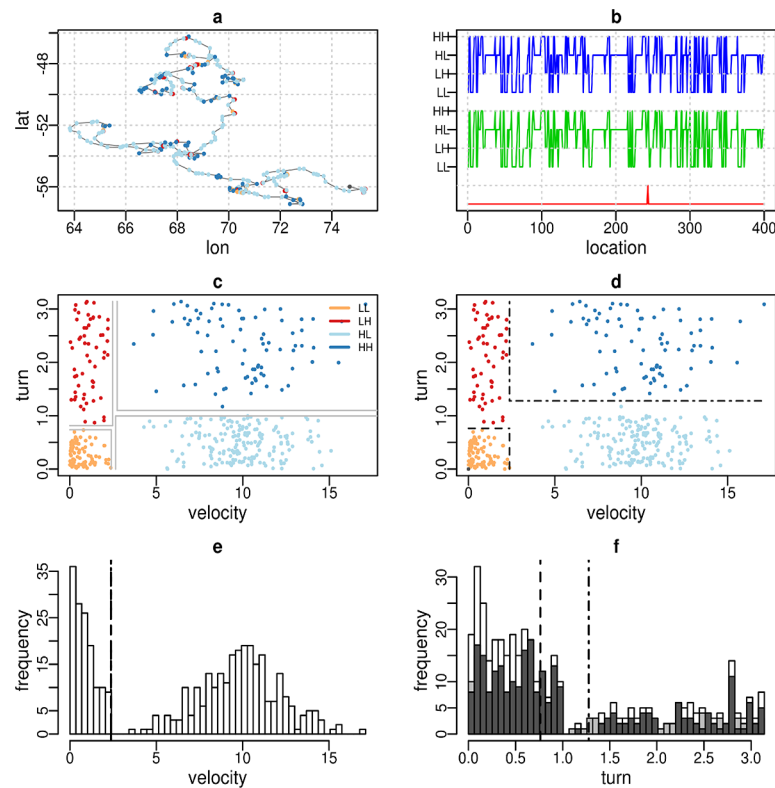
Fig 6. Synthetic trajectory. Simulated trajectory with $N = 400$ and $\gamma = 0.05$. Panels: (a) trajectory grid plot, (b) temporal behaviour profile with output labelling (top), reference labelling (centre) and labelling differences between them (bottom), (c) reference velocity/turn scatter plot showing the limits of the binary regions (grey lines), (d) output velocity/turn scatter-plot showing the resulting delimiters $r_{.L}, r_{L.}$ (dashed lines) and $r_{.H}, r_{H.}$ (dot-dashed lines), (e) velocity, and (f) turning angle frequency distributions (white colour). The turn distribution for low/high values of velocity is shown separately in light/dark grey colours, respectively. Bottom: Confusion matrix and performance measures.

sensitivity is that some seeds can lead to a final clustering that does not correspond to a binary partition despite of the underlying binary clustered distribution forced in the input data. In other cases, the EMC and HMM algorithms (as implemented in the packages we used) were not able to reach a stable solution because they incurred in decreasing likelihoods and they stopped.

As an example, Fig 6 shows the EMbC output for a Markov-chain sampled trajectory ($n = 400$, $\gamma = 0.05$), where the clusters were perfectly recovered.
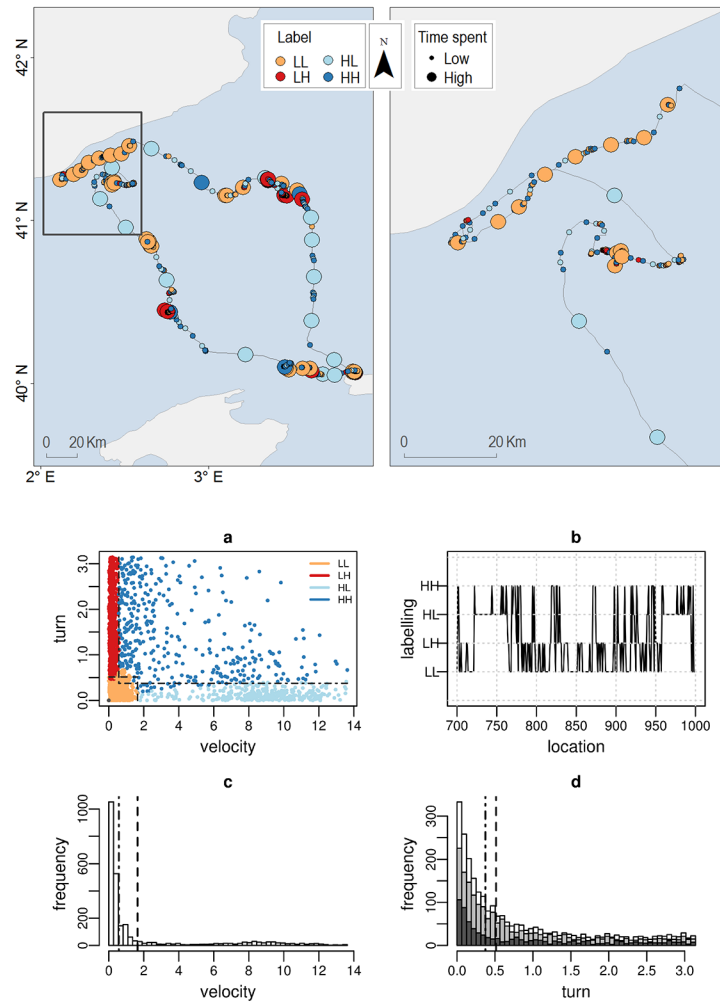
Fig 7. Cory's shearwater (*Calonectris diomedea*) foraging trajectory. Upper panel: Burst-wise labelled foraging trajectory. Symbolization by label (LL:orange, LH:red, HL:cyan, HH:blue) and time spent (symbol size: 2 natural jenks). Scale bar is only approximate. Due to copyright restrictions the figure is for representative purposes only. Source: Made with Natural Earth; Free vector and raster map data @ naturalearthdata.com. Bottom panels: (a) velocity/turn scatter plot (clustering colour code: LL:orange, LH:red, HL:cyan, HH:blue), (b) temporal behavioural profile (from location 700 to 1000) (c) velocity (*m/s*) and (d) turning angle (*rad*) frequency distributions (white colour). The turn distribution for low/high values of velocity is shown separately in light/dark grey colours, respectively. The black lines in panels a, c and d depict the delimiters $r_{.L}$, $r_{L.}$ (dashed lines) and $r_{.H}$, $r_{H.}$ (dot-dashed lines).

doi:10.1371/journal.pone.0151984.g007

Empirical trajectories

The Cory's shearwater (*Calonectris diomedea*, Fig 7) and the Osprey (*Pandion haliateus*, Fig 8) datasets show a clustering layout with similar velocity/turn partition and similar semantic labelling, regardless of the ecological context (*i.e.* migration, foraging) or the sampling resolution (*i.e.* Argos, high-resolution-GPS), although with different proportions in the LL (resting) and HL (relocation) modes according to the ecological context (*i.e. LL* = 37%, *HL* = 13% for the Cory's shearwater versus *LL* = 18%, *HL* = 30% for the Osprey, see further details in S2 Text). In both, the velocity distribution (Figs 7, 8 panel c) shows bi-modality to some extent
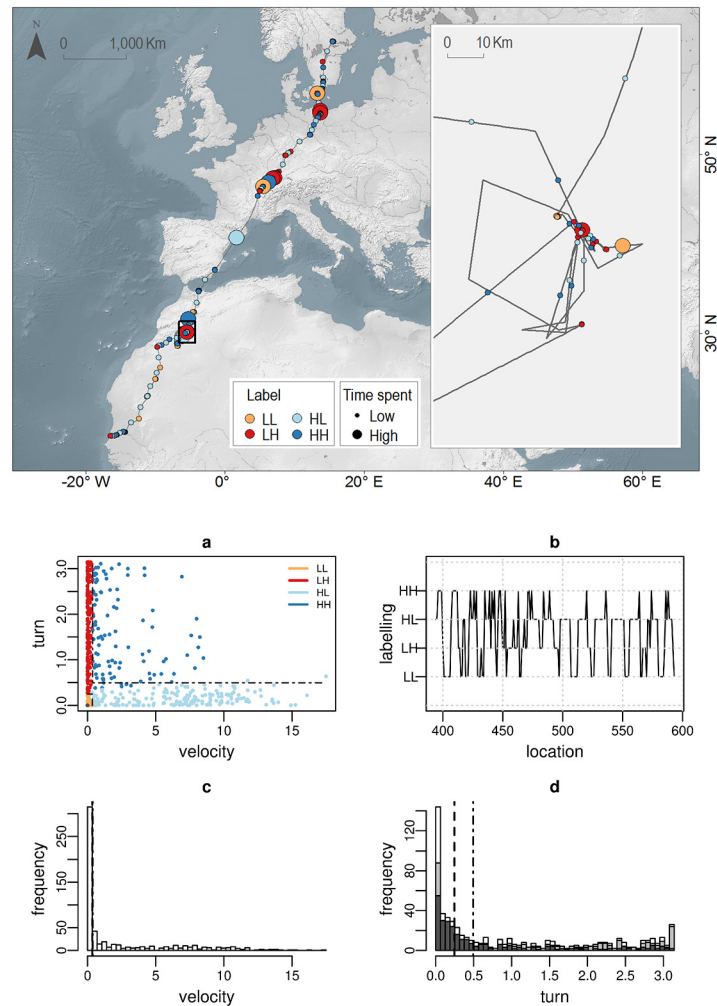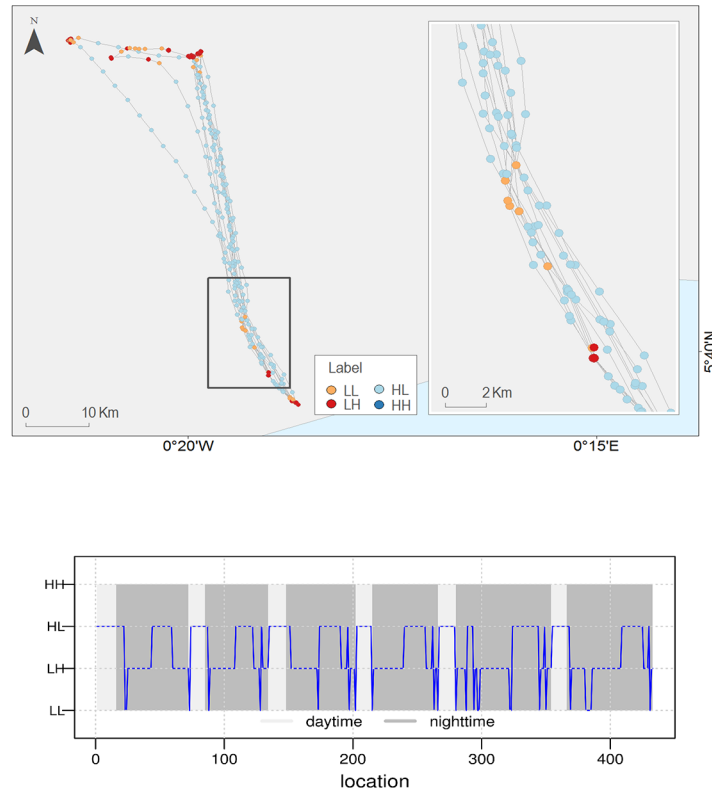
Fig 8. Osprey (*Pandion haliateus*) migratory trajectory. Upper panel: Burst-wise labelled migration trajectory. Symbolization by label (LL:orange, LH:red, HL:cyan, HH:blue) and time spent (symbol size: 2 natural jenks). Scale bar is only approximate. Due to copyright restrictions the figure is for representative purposes only. Source: Made with Natural Earth; Free vector and raster map data @ naturalearthdata.com. Bottom Panels: (a) velocity/turn scatter plot (clustering colour code: LL:orange, LH:red, HL:cyan, HH:blue), (b) temporal behavioural profile (from location 400 to 593) (c) velocity (*m/s*) and (d) turning angle (*rad*) frequency distributions (white colour). The turn distribution for low/high values of velocity is shown separately in light/dark grey colours, respectively. The black lines in panels a, c and d depict the delimiters $r_{.L}, r_{L.}$ (dashed lines) and $r_{.H}, r_{H.}$ (dot-dashed lines).

(although hardly apparent in Fig 8 because of the relative high frequency of low values), thus being the binary partition assumption particularly suitable. Within these standard layouts, the HH (dark blue) labelling is usually subject to more subtle semantic interpretation. In Fig 7a, the distribution of HH in the scatter plot suggests the existence of two possible sub-modes, one more closely related to foraging (low velocity and wide turn range) and the other more closely related to relocation (high velocity and low turns). A partition with only three clusters (LL,LH, and HL), with the HH cluster absorbed partly by the LH and partly by the HL clusters, would probably represent a better behavioural classification. However, the likelihood pay-off of this solution prevents the algorithm to reach it. Conversely, Fig 8a shows an homogeneous HH

Fig 9. Bat (*Eidolon helvum*) foraging trajectory. Upper: Point-wise labelled foraging trajectory. Symbolization by label (LL:orange, LH:red, HL:cyan, HH:blue) and time spent (symbol size: 2 natural jenks). Scale bar is only approximate. Due to copyright restrictions the figure is for representative purposes only. Source: Made with Natural Earth; Free vector and raster map data @ naturalearthdata.com. Centre: smoothed temporal behavioural profile with daytime/night-time (light/dark grey) background indication. Bottom: Confusion pie showing expert *vs*. EMbC labelling. Column titles *mrg.*, *rcl.*, *prc.* and *Fms* stand for *marginals*, *recall*, *precision* and *F-measure* respectively.

doi:10.1371/journal.pone.0151984.g009

cluster. A visual check of these data points on the landscape map reveals that they correspond to long relocations within stopover areas, thus justifying their assignment to a different behaviour.

The Straw-coloured fruit bat roosts in the colony during the day and moves for foraging in a very directed manner to individual fruiting trees during the night (Fig 9). The GPS was turned off during the day and fixes occurred when the animal moved during the night. In this example we used the post-processing smoothing procedure. The behavioural labelling profile (Fig 9, central panel) shows a quite regular behavioural pattern. It is worth noting that after the

smoothing procedure, some LL labels still remain suggesting the existence of a real but short transient state (LL) occurring between HL and LH state. A few more LL labels appear also in between relocation periods (as shown in Fig 9 upper panel inset). These locations seem to correspond to specific landmarks in the daily relocations of the animal and might have some biological relevance. Indeed, it is characteristic of the EMbC algorithm to capture behaviours showing strong correlations among movement variables despite being short in time and happening only intermittently. The decision on whether to consider or else to smooth out these type of behaviours relies on the expert decision.

For this trajectory we compared our results with an expert's labelling identifying behavioural modes (*i.e. roosting, forage, commuting*) stemming from GPS and acceleration data. From the visualization of the annotated trajectory (Fig 9, top panel) we can easily assimilate the *forage* mode with the LH cluster and the *commuting* mode with the HL cluster and we can therefore build the confusion matrix shown in Fig 9 bottom table. The expert classification embeds reasonably well into the EMbC classification. However, it is clear that *roosting* behaviour is not well defined in terms of velocity and turn.

As an example of a trivariate clustering with different input features, Fig 10 shows a subset of results obtained when applying the EMbC algorithm at the population-level on solitary nematode crawling in an agar plate. Tracks are highly resolved (32Hz) and last for about 90 minutes. We computed three movement variables combining information about the shape of the trajectory and the speed of the individual (i.e. average straightness, average velocity, and net displacement) over 5 minute windows (Section B in S1 Text). With 3 variables, the number of potential clusters is $2^3 = 8$. Because the number of clusters is limited, the larger the pool of individual trajectories, the more the clustering will tend to favour generic behaviours to the detriment of individual specific behaviours. In addition, only a subset of the population-level clusters are recovered in each individual, unveiling the presence of individual-level behavioural variability (Fig 10). Accordingly to our input features, the semantic labelling of the output clusters must be considered in terms of looping behaviour or intensity of local search and the will of the individual to move to a different location. The statistics of the clustering (Table A in S1 Text Section B) reveal that on average, the C.elegans parsimoniously deploys all of its motor repertoire resulting in a gradual increase of the straightness, the mean velocity and the net displacement of its movement. The result is a complex movement behaviour evolving through phases of decreasing degrees of local search, going from a strong intensive search phase to a more straight-lined motion, resembling relocation behaviour (Fig 10).

Robustness to data loss and data inaccuracy

We assessed the robustness to data loss of the EMbC algorithm. In general, the larger the dataset, the more robust is the EMbC labelling to data loss (Fig 11a). However, the absence or presence of strong heterogeneities in the marginal distribution of clusters also plays a role. For example, although the Osprey and the Straw-coloured fruit bat datasets are both small ($n = 594$ and $n = 434$, respectively) the former is more robust to data loss. Interestingly, Osprey data shows more uniform posterior marginal distribution of clusters ($LL = 17.51\%$, $LH = 35.02\%$, $HL = 29.97\%$, $HH = 17.34\%$) than the bat data ($LL = 10.60\%$, $LH = 43.09\%$, $HL = 46.08\%$, $HH = 0.00\%$). As it is a nocturnal bat, the daily resting in the roost was intentionally skipped by a pre-fixed sampling scheme, and therefore, the LL cluster commonly associated to resting behaviour is misrepresented ($LL = 10.60\%$). Accordingly, neither the position of the LL cluster nor its mean velocity of 2.26 m/s (see the scatter plot and statistics in S2 Text) suggest such type of semantics. In general, pre-assigned GPS fixes scheduling will bias the sampling
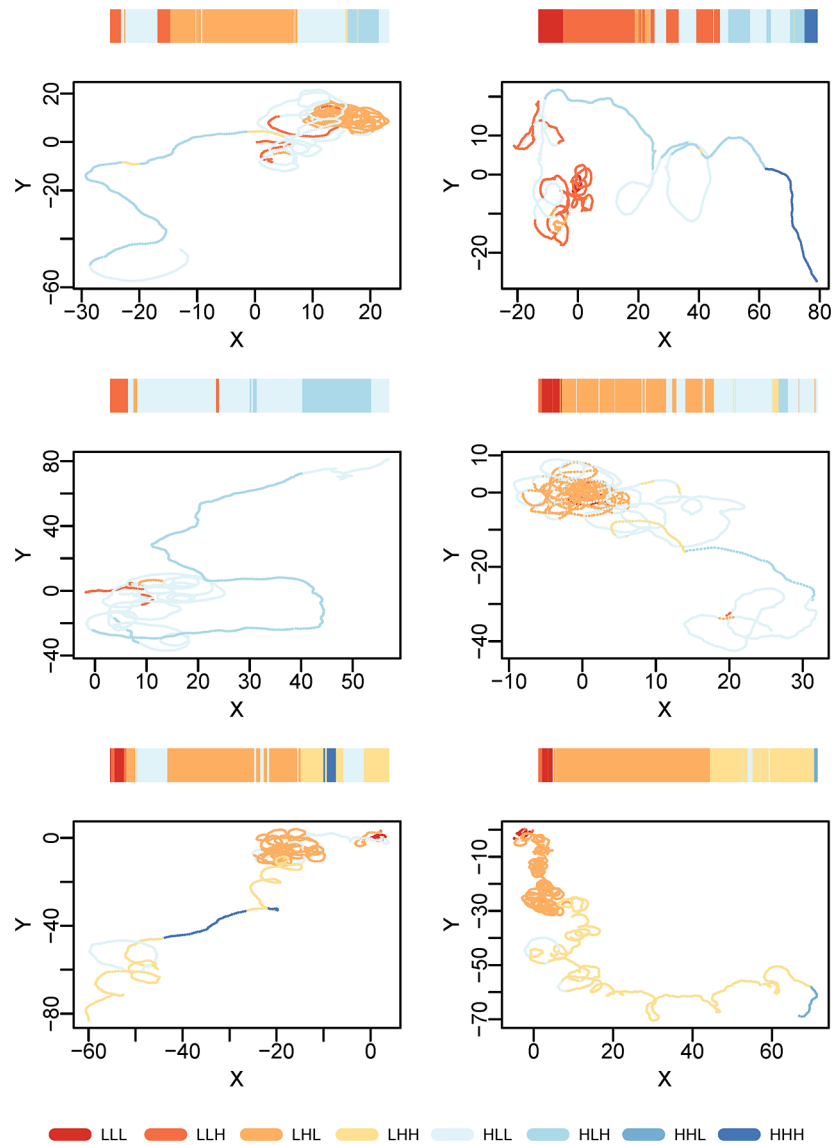
Fig 10. C.elegans (*Caenorhabditis elegans*) search trajectory. Trivariate clustering of 6 solitary nematode trajectories crawling in an agar plate (with a sampling frequency of 32 Hz and 90 minutes of trajectory time). The clustering is performed at population level (all data points at the same time) and is afterwards visualized on each individual trajectory. The algorithm captures different behaviours in terms of intensity of local search, looping behaviour and relocation.

distribution of behaviours, thus conditioning both the labelling outcome and the robustness of the results to data loss.

Fig 11b shows the robustness of the EMbC algorithm to data inaccuracy when weighting or not the contribution to the clustering of each data point on the basis of a reliability function Eq (5). In high resolution GPS datasets (*e.g.* Cory's shearwater), the algorithm shows a strong robustness because inaccuracies due to sampling heterogeneity are expected to be low, so that the effect of accounting for data reliability is almost non-significant. However, accounting for data reliability in datasets with large sampling heterogeneities (i.e. ARGOS Osprey dataset) or
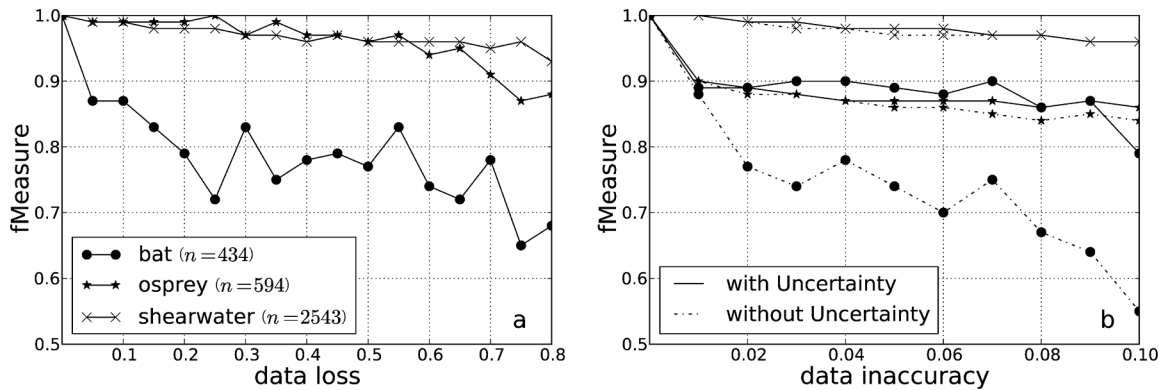
Fig 11. Data loss and data accuracy. EMbC robustness results for a data loss range of $0 \leq k_{dl} \leq 0.8$ and for a jittering range of $0 \leq k_{di} \leq 0.1$, see Eq (9). For each trajectory the values shown correspond to the average F-measure after 10 different runs.

doi:10.1371/journal.pone.0151984.g011

prefixed sampling schedules (e.g. GPS Straw-coloured fruit bat dataset) favours the robustness of the labelling to data inaccuracy.

Discussion

Splitting a trajectory into its most basic components is essential for studying and understanding mechanisms of movement [1, 46, 47]. Currently, trajectory segmentation algorithms constitute an essential component of ecological spatio-temporal analyses that seek to mechanistically understand organisms' interactions with the environment.

Current behavioural annotation methods show gradients of complexity, supervision requirements, and sensitivity to initial conditions, as well as to sampling rate and data accuracy [18–24, 48]. A possible classification of methods could result from the kind of underlying assumptions: (i) assumptions about the input feature distributions (EMC family), (ii) assumptions about time-dependency and context-dependency among behavioural states (HMM family) [18, 21, 22], and (iii) assumptions about the autocorrelation structure of the movement variable [23, 48]. We completely subscribe to the idea that for each particular problem there will be an optimal approach to follow, or in other words, there is not an overall best methodology. From this point of view, the EMBC do have its own room within the unsupervised clustering domain, and the researcher's self criteria must be the only valid judgment upon its suitability for the problem at hand. Making a thorough comparison across different methodologies would not be consequent with the former idea and can easily lead to wrong or interested conclusions. Such comparative analysis makes no sense unless accompanied by an extensive discussion with respect to prior assumptions, final objectives, parameterizations used and also about how the results obtained through the different methodologies have been matched, what constitutes a daunting task out of the scope of this paper. In this context, we are essentially concerned about utility and simplicity of the models in the line of reasoning that *all models are wrong, but some are useful* [49, 50]. In our case, the idea of usefulness stems from the fact that the most elementary partition of the input space (i.e. a binary partition into High/Low values of the variables) can be very informative, in many circumstances sufficient, to characterize behavioural patterns. The EMbC algorithm finds a solution for this binary partition based on a likelihood criterion under the assumption of a multivariate Gaussian mixture space. Such a key

and simple concept helps reaching a compromise between interpretable behavioural annotation and adequate statistical performance.

In terms of implementation, the EMbC algorithm implies iteratively computing the centroids of the clusters within regions determined by explicit *delimiters* Eq (6) in order to provide easy and interpretable semantics (i.e. LL, HL, LH, HH). Nonetheless, at each iteration, we keep the computation of cluster covariances unbounded to incorporate information about the correlation landscape provided by the whole variable space Eq (7). The choice of a binary partition of variables into H/L clusters, restricts the maximal number of clusters to 2^m, where m is the number of input variables used. This restriction over the number of clusters indeed avoids *aprioristic* decisions on the number of clusters in the n-dimensional space, and facilitates their interpretation.

The initial assumptions implemented in the EMbC algorithm aim at minimizing biases and sensitivity to initial conditions: (i) each data point is assigned a uniform probability of belonging to each cluster, (ii) the prior mixture distribution is uniform (each cluster starts with the same number of data points), (iii) the starting partition, (*i.e.* the starting delimiters position), is selected based on a global maximum entropy criterion, thus conveying the minimum information possible. A single parameter σ_{min} controls the minimal resolution at which clusters will be accepted. In practical terms, σ_{min} can be directly related to measurement errors (or maximum resolution) of the variables and will limit the minimum range of variability (*i.e.* minimum standard deviation) within the clusters obtained. Based on intuitive physical and biological considerations, we set $\sigma_{min} = 0.01$ m/s and $\sigma_{min} = 0.087$ rad (5 degrees), respectively for velocity and turn.

The algorithm deals very intuitively with data reliability: the larger the uncertainty associated with the values, the smaller the leverage of those values in the clustering. We considered two elementary sources of uncertainty: sampling heterogeneity (for those variables whose reliability depends on the sampling interval), and the measurement error of the devices. In the present work we computed velocity and turn based on the sampling intervals. Although it is worth incorporating the measurement error of the devices, here, for the sake of simplicity we considered that the major source of error comes from sampling interval lengths. However, the algorithm is multivariate and can deal with any type of movement/behavioural variables and error sources. For example, one could also use instantaneous speed [51] or tri-axial accelerometer data [52, 53] and take into consideration the errors associated to them.

Without intending to be exhaustive, we have presented a comparison of the EMbC with similar state-of-the-art algorithms like the EMC and basic HMMs, in terms of their performance in clustering synthetically generated datasets based on GMMs. Being simpler, the EMbC yields a similar degree of performance without the need of multiple restarts or initial parametrization. Of note, HMMs are more complex in that behavioural states are defined based on both the states' definitions (the distribution of input features) and the transitions among them (the Markov-chains). Compared to HMM and EMC, we have shown that the EMbC proves particularly useful as long as: (i) we can expect bi-modality, to some extent, in the distribution of the input features, (ii) we can expect that forcing a binary partition of the input space can provide useful information, and (iii) we cannot guarantee that the temporal state dynamics can be associated to a Markov-chain process. A basic HMM is equivalent to assuming that, for an individual in a particular state, the probability of changing to a different state remains constant as it keeps moving. In terms of movement data this is almost equivalent to assuming a memoryless individual with stationary internal states. Additionally, in movement trajectories the first-order dependence condition of a Markov-chain is easily violated because of the heterogeneity in empirical time series due to large gaps, or prefixed sampling scheduling. To overcome this problem, either we use more complex HMM approaches taking

into account sampling heterogeneities [21] and/or introducing explicit time or spatial dependencies among states [18, 23], or else, we disregard any assumption about state dependence on time (EMbC). The main message, however, is that regardless of their complexity, HMM approaches are always forced to make estimates on state transitions. When the assumptions related to these transitions are not fulfilled this may impair substantially their ability to correctly identify the states. This is the reason for the low performance of HMMs in fitting trajectory states generated from a mixture-prior sampling scheme rather than from well-defined transition matrices (see Fig 5).

The results obtained for different empirical datasets suggest that the EMbC algorithm behaves reasonably well for a wide range of tracking technologies, species, and ecological contexts (*e.g.* migration, foraging). Different layouts in the scatter plots may emerge depending on the probability distribution of the input data, the underlying mixture prior distribution and the specific set of behaviours performed by the animals along the trajectories. We also show the possibility of running the algorithm at the population level (by applying the algorithm on a pooled set of trajectories from a given population) to define *average modes*, that can be visualized in single trajectories.

Importantly, the degree of match between the EMbC output and a particular set of behaviours will be highly dependent on the amount of information conveyed by the input variables regarding these behaviours, but not so much on their relative proportions/durations. Indeed, it is worth mentioning that the EMbC is good at identifying behaviours that are hardly represented in the dataset. Here we focus on velocity/turn as key movement variables for trajectory behavioural annotation *e.g.* [11, 27] but it is certainly possible to choose other behavioural or physiological data (e.g. accelerometry, stereotyped turns, body postures, metabolic rates) that correlates adequately with the behavioural classification the analyst is looking for. Even though the determination of the final semantics is variable-dependent and species-specific, the method is general and robust enough to to be used across species and tracking methodologies.

Improvements on the current EMbC implementation could come from exploring other *reliability* functions and their relative contribution to the final clustering. Also, one of the drawbacks of not explicitly considering the temporal correlations in the segmentation procedure (as in HMM) is that, at coarse-scales, behavioural labels may not appear as continuous as desired, and some local labels may be considered neglectable or meaningless by the expert. We suggest considering the EMbC algorithm as a fine-scale behavioural segmentation method, with optional pre/post smoothing alternatives, the latter explicitly taking into account the weights of the labels relative to each cluster and their temporal dependencies in order to generate an aggregated coarse-grained behavioural annotation of the trajectory. We also suggest the possibility of running the algorithm at the population level (by applying the algorithm on a pooled set of trajectories from a given population). This approach smooths potential individual variability, which might be of non-interest for certain analyses. Certainly, all these alternatives (i.e. coarse-graining, population-level analysis) may have implications in the estimations of the number and durations of the behavioural states. But this kind of problem is intrinsic to any segmentation method and only the expert's criteria and the specific goals of the study can help to justify the choices taken. All in all, behaviour should be interpreted in relative and not absolute terms as it is a multidimensional and multi-scale phenomena, and the description of behaviour will inevitably be biased by the scientific goals, the observational constraints, and the methodological choices.

To ease the complex analysis of movement behaviour and segmentation to interested users we released a ready-to-use tool (the EMbC R-package) including not only the EMbC algorithm itself but also a set of side functions for straightforward analysis (clustering statistics, clustering scatter-plot, temporal behavioural profile, smoothing function) and visualization (burst and

point-wise kml doc generation) of its output. The package is compatible with the Move R-package developed by the Movebank team [54].

Supporting Information

S1 Text. Supporting Information.
(PDF)

S2 Text. Generating the Results with the EMbC R-package.
(PDF)

S1 Data. Data sets zip file.
(ZIP)

Acknowledgments

Author Contributions

Conceived and designed the experiments: JG FB. Performed the experiments: JG. Analyzed the data: JG JRP AO. Contributed reagents/materials/analysis tools: JG JRP AO. Wrote the paper: JG JRP AO FB.

References

1. Nathan R. An emerging movement ecology paradigm [Editorial Material]. Proceedings of the National Academy of Sciences of the United States of America. 2008 DEC 9; 105(49):19050–19051. doi: 10.1073/pnas.0808918105 PMID: 19060197

2. Giuggioli La, Bartumeus Fb. Animal movement, search strategies and behavioural ecology: A cross-disciplinary way forward. Journal of Animal Ecology. 2010; 79(4):906–909. Cited By (since 1996)13. PMID: 20337757

3. Morales JMa, Moorcroft PRb, Matthiopoulos Jc, Frair JLd, Kie JGe, Powell RAf, et al. Building the bridge between animal movement and population dynamics. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010; 365(1550):2289–2301. Cited By (since 1996)44. doi: 10.1098/rstb.2010.0082

4. Nathan R, Getz WM, Revilla E, Holyoak M, Kadmon R, Saltz D, et al. A movement ecology paradigm for unifying organismal movement research [Article]. Proceedings of the National Academy of Sciences of the United States of America. 2008 DEC 9; 105(49):19052–19059. doi: 10.1073/pnas.0800375105 PMID: 19060196

5. Thiebault A, Tremblay Y. Splitting animal trajectories into fine-scale behaviorally consistent movement units: breaking points relate to external stimuli in a foraging seabird [Article]. Behavioral Ecology and Sociobiology. 2013 JUN; 67(6):1013–1026. doi: 10.1007/s00265-013-1546-1

6. Benhamou S. How to reliably estimate the tortuosity of an animal's path: Straightness, sinuosity, or fractal dimension? Journal of Theoretical Biology. 2004; 229(2):209–220. Cited By (since 1996)104. doi: 10.1016/j.jtbi.2004.03.016 PMID: 15207476

7. Fauchald P, Tveraa T. Using first-passage time in the analysis of area-restricted search and habitat selection [Article]. Ecology. 2003 FEB; 84(2):282–288. doi: 10.1890/0012-9658(2003)084%5B0282:UFPTIT%5D2.0.CO;2

8. Barraquand F, Benhamou S. Animal movements in heterogeneous landscapes: identifying profitable places and homogeneous movement bouts [Article]. Ecology. 2008 DEC; 89(12):3336–3348. doi: 10.1890/08-0162.1 PMID: 19137941

9. Roberts S, Guilford T, Rezek I, Biro D. Positional entropy during pigeon homing I: application of Bayesian latent state modelling [Article]. Journal of Theoretical Biology. 2004 MAR 7; 227(1):39–50. doi: 10.1016/j.jtbi.2003.07.002 PMID: 14969706

10. Guilford T, Roberts S, Biro D, Rezek L. Positional entropy during pigeon homing II: navigational interpretation of Bayesian latent state models [Article]. Journal of Theoretical Biology. 2004 MAR 7; 227(1):25–38. doi: 10.1016/j.jtbi.2003.07.003 PMID: 14969705

11. Knell AS, Codling EA. Classifying area-restricted search (ARS) using a partial sum approach [Article]. Theoretical Ecology. 2012 AUG; 5(3):325–339. doi: 10.1007/s12080-011-0130-4

12. Jonsen ID, Myers RA, James MC. Identifying leatherback turtle foraging behaviour from satellite telemetry using a switching state-space model. Marine Ecology Progress Series. 2007; 337:255–264. Cited By (since 1996)55. doi: 10.3354/meps337255

13. Bailey H, Shillinger G, Palacios D, Bograd S, Spotila J, Paladino F, et al. Identifying and comparing phases of movement by leatherback turtles using state-space models. Journal of Experimental Marine Biology and Ecology. 2008; 356(1–2):128–135. <ce:title>Sea turtles: physiological, molecular and behavioural ecology and conservation biology</ce:title> doi: 10.1016/j.jembe.2007.12.020

14. Morales JM, Haydon DT, Frair J, Holsinger KE, Fryxell JM. Extracting more out of relocation data: building movement models as mixtures of random walks. Ecology. 2004; 85(9):2436–2445. doi: 10.1890/03-0269

15. Forester JDae, Ives ARa, Turner MGa, Anderson DPa, Fortin Db, Beyer HLc, et al. State-space models link elk movement patterns to landscape characteristics in Yellowstone National Park. Ecological Monographs. 2007; 77(2):285–299. Cited By (since 1996)54. doi: 10.1890/06-0534

16. Ovaskainen O. Analytical and numerical tools for diffusion-based movement models. Theoretical Population Biology. 2008; 73(2):198–211. Cited By (since 1996)11. doi: 10.1016/j.tpb.2007.11.002 PMID: 18199463

17. Dalziel BDa, Morales JMb, Fryxell JMa. Fitting probability distributions to animal movement trajectories: Using artificial neural networks to link distance, resources, and memory. American Naturalist. 2008; 172(2):248–258. Cited By (since 1996)43. doi: 10.1086/589448 PMID: 18598199

18. Bestley S, Patterson TA, Hindell MA, Gunn JS. Predicting feeding success in a migratory predator: integrating telemetry, environment, and modeling techniques [Article]. Ecology. 2010 AUG; 91(8):2373–2384. doi: 10.1890/08-2019.1 PMID: 20836459

19. Dean B, Freeman R, Kirk H, Leonard K, Phillips RA, Perrins CM, et al. Behavioural mapping of a pelagic seabird: combining multiple sensors and a hidden Markov model reveals the distribution of at-sea behaviour [Article]. Journal of the Royal Society Interface. 2013 JAN 6; 10(78).

20. Freeman R, Dean B, Kirk H, Leonard K, Phillips RA, Perrins CM, et al. Predictive ethoinformatics reveals the complex migratory behaviour of a pelagic seabird, the Manx Shearwater [Article]. Journal of the Royal Society Interface. 2013 JUL 6; 10(84). doi: 10.1098/rsif.2013.0279

21. Joo R, Bertrand S, Tam J, Fablet R. Hidden Markov Models: The Best Models for Forager Movements? PLoS ONE. 2013 08; 8(8):e71246. doi: 10.1371/journal.pone.0071246 PMID: 24058400

22. Charles C, Gillis D, Wade E. Using hidden Markov models to infer vessel activities in the snow crab (Chionoecetes opilio) fixed gear fishery and their application to catch standardization. Canadian Journal of Fisheries and Aquatic Sciences. 2014; 71(12):1817–1829. doi: 10.1139/cjfas-2013-0572

23. Gloaguen P, Mahévas S, Rivot E, Woillez M, Guitton J, Vermard Y, et al. An autoregressive model to describe fishing vessel movement and activity. Environmetrics. 2015; 26(1):17–28. doi: 10.1002/env.2319

24. Jackson CH. Multi-State Models for Panel Data: The msm Package for R. Journal of Statistical Software. 2011; 38(8):1–29. doi: 10.18637/jss.v038.i08

25. van Gils JA, van der Geest M, De Meulenaer B, Gillis H, Piersma T, Folmer EO. Moving on with foraging theory: incorporating movement decisions into the functional response of a gregarious shorebird. Journal of Animal Ecology. 2015; 84(2):554–564. doi: 10.1111/1365-2656.12301 PMID: 25283546

26. Tremblay Y, Robinson PW, Costa DP. A Parsimonious Approach to Modeling Animal Movement Data [Article]. PLoS ONE. 2009 MAR 5; 4(3). doi: 10.1371/journal.pone.0004711

27. Gurarie E, Andrews RD, Laidre KL. A novel method for identifying behavioural changes in animal movement data [Letter]. Ecology Letters. 2009 MAY; 12(5):395–408. doi: 10.1111/j.1461-0248.2009.01293.x PMID: 19379134

28. Maaten L. Learning a Parametric Embedding by Preserving Local Structure. In: Dyk DV, Welling M, editors. Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS-09). vol. 5. Journal of Machine Learning Research—Proceedings Track; 2009. p. 384–391.

29. Berman GJ, Choi DM, Bialek W, Shaevitz JW. Mapping the stereotyped behaviour of freely moving fruit flies. Journal of The Royal Society Interface. 2014; 11(99). doi: 10.1098/rsif.2014.0672

30. Jansen VAA, Mashanova A, Petrovskii S. Comment on Lévy walks evolve through interaction between movement and environmental complexity. Science. 2012; 335(6071):918. doi: 10.1126/science.1215903 PMID: 22362991

31. Reynolds A. Distinguishing between Lévy walks and strong alternative models. Ecology. 2012; 93(5):1228–1233. doi: 10.1890/11-1815.1 PMID: 22764508

32. Guilford TCa, Meade Ja, Freeman Rab, Biro Da, Evans Ta, Bonadonna Fc, et al. GPS tracking of the foraging movements of Manx Shearwaters Puffinus puffinus breeding on Skomer Island, Wales. Ibis. 2008; 150(3):462–473. Cited By (since 1996)27. doi: 10.1111/j.1474-919X.2008.00805.x

33. Freeman Rab, Dennis Tc, Landers Tcd, Thompson De, Bell Ef, Walker Mc, et al. Black Petrels (Procellaria parkinsoni) patrol the ocean shelf-break: GPS tracking of a vulnerable procellariiform seabird. PLoS ONE. 2010; 5(2). Cited By (since 1996)4. doi: 10.1371/journal.pone.0009236

34. Hartley H. Maximum likelihood estimation for incomplete data. Biometrics. 1958; 14:174–194. doi: 10.2307/2527783

35. Dempster A, Laird N, Rubin D. Maximum likelihood for incomplete data via EM algorithm. Journal of the Royal Statistical Society B. 1977; 39(1):1–38.

36. McLachlan G, Krishnan T. The EM algorithm and extensions. 2nd ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley; 2008.

37. Gupta M, Chen Y. Theory and use of the EM Algorithm. Foundations and Trends in Signal Processing. 2010; 4(3):223–296. doi: 10.1561/2000000034

38. McLachlan GJ, Krishnan T, Ng SK. The EM algorithm. Papers/Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE); 2004.

39. Bilmes JA. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Computer Science Division, Department of Electrical Engineering and Computer Science, U.C. Berkeley; 1998. TR-97-021

40. Kim J, Min S, Na S, HS C, SH C. Modified GMM training for inexact observation and its application to speaker identification. Speech Sciences. 2007; 14:163–175.

41. Tariquzzaman MD, Kim JY, Na SY, Kim HG. Reliability-Weighted HMM Considering Inexact Observations and its Validation in Speaker Identification. International Journal of Innovative Computing, Information and Control. 2012 July; 8(7(B)).

42. Langrock R, King R, Matthiopoulos J, Thomas L, Fortin D, Morales JM. Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. Ecology. 2012; 93(11):2336–2342. doi: 10.1890/11-2241.1 PMID: 23236905

43. Chen WC, Maitra R, Melnykov V. EMCluster: EM Algorithm for Model-Based Clustering of Finite Mixture Gaussian Distribution; 2012. R Package, http://cran.r-project.org/package=EMCluster.

44. Chen WC, Maitra R, Melnykov V. A Quick Guid for the EMCluster Package; 2012. R Vignette, http://cran.r-project.org/package=EMCluster.

45. Visser I, Speekenbrink M. depmixS4: An R Package for Hidden Markov Models. Journal of Statistical Software. 2010; 36(7):1–21. doi: 10.18637/jss.v036.i07

46. Fronhofer EA, Hovestadt T, Poethke HJ. From random walks to informed movement. Oikos. 2012; 122:857–866. doi: 10.1111/j.1600-0706.2012.21021.x

47. Shamoun-Baranes Ja, Van Loon EEa, Purves RSb, Speckmann Bc, Weiskopf Dd, Camphuysen CJe. Analysis and visualization of animal movement. Biology Letters. 2012; 8(1):6–9. Cited By (since 1996) 3. doi: 10.1098/rsbl.2011.0764 PMID: 21865243

48. Metzner C, Christoph M, Julian S, Lena L, Franz S, Ben F. Superstatistical analysis and modelling of heterogeneous random walks. Nat Commun. 2015; 6:7516. doi: 10.1038/ncomms8516 PMID: 26108639

49. Box GEP. Science and Statistics. Journal of the American Statistical Association. 1976; 71(356):791–799. doi: 10.1080/01621459.1976.10480949

50. Box GEP, Hunter WG, Hunter JS. Statistics for experimenters: an introduction to design, data analysis, and model building. Wiley series in probability and mathematical statistics. New York, Chichester, Brisbane: J. Wiley & Sons; 1978.

51. Kamran S, Kranstauber B, Weinzierl R, Griffin L, Rees E, Cabot D, et al. Flying with the wind: scale dependency of speed and direction measurements in modelling wind support in avian flight. Movement Ecology. 2013; 1:4. doi: 10.1186/2051-3933-1-4

52. Resheff Y, Rotics S, Harel R, Spiegel O, Nathan R. AcceleRater: a web application for supervised learning of behavioral modes from acceleration measurements. Movement Ecology. 2014; 2(1):27. doi: 10.1186/s40462-014-0027-0 PMID: 25709835

53. Bom R, Bouten W, Piersma T, Oosterbeek K, van Gils J. Optimizing acceleration-based ethograms: the use of variable-time versus fixed-time segmentation. Movement Ecology. 2014; 2(1):6. doi: 10.1186/2051-3933-2-6 PMID: 25520816

54. Kranstauber B, Smolla M. move: Visualizing and analyzing animal track data; 2013. R package version 1.1.441. Available from: http://CRAN.R-project.org/package = move