Noname manuscript No.
(will be inserted by the editor)

Topic modeling of behavioral modes using sensor data

Yehezkel S. Resheff · Shay Rotics · Ran Nathan · Daphna Weinshall

Received: date / Accepted: date

Abstract The field of Movement Ecology, like so many other fields, is experiencing a period of rapid growth in availability of data. As the volume rises, traditional methods are giving way to machine learning and data science, which are playing an increasingly large part it turning this data into science-driving insights. One rich and interesting source is the bio-logger. These small electronic wearable devices are attached to animals free to roam in their natural habitats, and report back readings from multiple sensors, including GPS and accelerometer bursts. A common use of accelerometer data is for supervised learning of behavioral modes. However, we need unsupervised analysis tools as well, in order to overcome the inherent difficulties of obtaining a labeled dataset, which in some cases is either infeasible or does not successfully encompass the full repertoire of behavioral modes of interest. Here we present a matrix factorization based topic-model method for accelerometer bursts, derived using a linear mixture property of patch features. Our method is validated via comparison to a labeled dataset, and is further compared to standard clustering algorithms.

Invited Extended version of a paper [21] presented at the international conference *Data Science and Advanced Analytics*, Paris, France, 19-21 OCtober 2015

Yehezkel S. Resheff
Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, 91914, Israel
E-mail: yehezkel.resheff@mail.huji.ac.il

Shay Rotics
Movement Ecology Lab, Department of Ecology, Evolution and Behavior, The Hebrew University of Jerusalem

Ran Nathan
Movement Ecology Lab, Department of Ecology, Evolution and Behavior, The Hebrew University of Jerusalem

Daphna Weinshall
School of Computer Science and Engineering, The Hebrew University of Jerusalem

Keywords Behavioral Modes · Topic Model · Movement Ecology · MS-BoP

1 Introduction

Wearable devices with various sensors are becoming increasingly popular, with ongoing research into applications to health monitoring [18] and context detection [12]. Many fields of animal behavior and conservation have also began to utilize similar devices in order to remotely monitor the whereabouts and behavior of their research subjects [20], and this has especially been the case in the field of Movement Ecology.

The aim of Movement ecology is to unify research of movement of organisms and aid in the development of a general theory of whole-organism movement [16]. Recent technological advances in tracking tools and especially the appearance of cheap and small GPS devices [9], have driven the field into a period of rapid growth in knowledge and insight [11], and have led to the emergence of various methods of analyzing movement patterns [24].

Nevertheless, movement data, however accurate, is unlikely to suffice for inference on the links between behavioral, ecological, physiological, and evolutionary processes driving the movement of individuals, and link these subjects which have traditionally been researched separately in their respective fields. Thus, promoting movement ecology research and the desirable unification across species and movement phenomena requires the development of additional data sources: sensors and tools providing simultaneous information about the movement, energy expenditure and behavior of the focal organisms, together with the environmental conditions they encounter en route [17].

One such tool, which has been introduced into the field of movement ecology, is the accelerometer-biologger (ACC). These sensors allow the determination of the acceleration of the tagged animal's body, and are used as a means of identifying moment-to-moment behavioral modes [30], and estimating energy expenditure [29].

ACC loggers typically record in 1-3 dimensions, either continuously or in short bouts in a constant window [20]. Their output is used to infer behavior, most commonly through supervised machine learning techniques, and energy expenditure using the Overall Dynamic Body Acceleration (ODBA) or related metrics [8,29]. When combined with GPS recordings, acceleration sensors add fine scale information on the variation in animal behavior, and energy expenditure in space and time (see [2] for a recent review).

ACC-based analysis has been used to compute many measures of interest in the field of Movement Ecology, including behavior-specific body posture, movement and activity budgets, measures of foraging effort, attempted food capture events, mortality detection, classifying behavioral modes and more [2]. These measures have facilitated movement-related research for a wide range of topics in ecology and animal behavior [24,2,26,25] as well as other fields of research such as animal conservation and welfare [26,3] and biomechanics [10, 23].

In recent years there has been considerable interest in the analysis of behavioral modes using ACC data and supervised learning techniques. The protocol for using ACC data for supervised learning of behavioral modes consists of several steps. First, a sensor calibration procedure is preformed in a controlled environment: before deployment, the response of each tag to $\pm 1G$ acceleration on each axis is recorded, in order to fit the tag-specific linear transformation from the recorded values (mV) to the desired units of acceleration. Next, the calibrated tags are given a recording schedule and mounted on the focal animals, after these are captured. Finally, the data is retrieved using RF (radio) methods, Cellular transmission, or physically reacquiring the device.

Once the data is retrieved, before supervised machine learning models can be used, a labeled dataset is collected through field observations. This time and labor intensive stage requires the researcher to observe the animal, either in its natural habitat or in captivity, and relate the actual behavioral modes to the time-stamp of the ACC recordings. Since some behavioral modes tend to be less common, or are performed predominantly at specific times, recording a sufficient number of such behavior-measurement samples may be tricky. Furthermore, for aquatic and nocturnal species, observations may not be feasible. In the final stage, models are trained using the labeled data, and the entire dataset is then labeled.

Supervised machine learning methods have been applied to ACC data from many species, and for a diverse range of behavioral modes. However, there are several drawbacks to the supervised approach. Observations, even if perfectly accurate, may not be adequately representative of the behavioral pattern throughout the period of the research (which is desirably the lifetime of the animal), for several reasons: field work is inherently confined to a specific time and place; moreover, only some of the animals are observed, and the presence of the observer may in some cases have an impact on the behavior of the observed animals. Furthermore, the need for observations limits the scope of such research projects to observable species and to research labs with the necessary resources (in money, manpower, and knowledge) to carry out all the steps listed above.

In this paper we present a framework for unsupervised analysis of behavioral modes from ACC data. First we suggest a patch-codebook descriptor (MS-BoP) of ACC signals reminiscent of "bag of visual words" descriptors in Computer Vision (see [4,31]). Next, we present a simple topic model for behavioral modes incorporating a linear mixture property of the MS-BoP features, and demonstrate how it can be used for unsupervised analysis of behavioral modes.

The rest of the paper is organized as follows: The next section describes related work both in Movement Ecology and in matrix factorization for clustering and topic modeling. In section 3 we introduce the features and model. Finally, in section 4 we present the results of an analysis on a large real-world dataset and the comparison to other methods.

2 Previous Work

Previous work on behavioral mode analysis using ACC data focused predominantly on supervised learning, with an emphasis on constructing useful features and finding the right classifiers for a specific use, such as monitoring dairy cows [6], or determining the flight type of soaring birds [28].

While this line of work proved very successful, both in terms of classifier performance and of scientific discovery that it was able to drive, it still suffers from the inherent limitations of supervised learning, compounded by the very high cost of obtaining labeled data for behavioral observations of wild animals. It remains the case that for some animals (nocturnal or sea species for instance), obtaining a labeled dataset is currently infeasible. Thus, in order to use all available ACC data for behavioral mode analysis in the field of Movement Ecology, an unsupervised framework is essential.

To the best of our knowledge, there have been two attempts at such an approach. In [22], K-means was applied to a representation of the ACC data, to achieve behavior-mode clusters. In [7,15] a Gaussian Mixture Model (GMM) variant was used to cluster a low-dimensional representation of ACC signals into a small number of useful behavioral modes. Our method goes one step further by allowing samples to be a mixture (more precisely, a convex combination) of behavioral modes, accounting for the observation that ACC samples do indeed tend to be mixed this way (Figure 1).

Non-Negative Matrix Factorization (NNMF) has been studied extensively in the context of clustering [27,13] and topic modeling [1]. Connections have been shown to various popular clustering algorithms such as K-means and spectral clustering [5]. Our proposed method is essentially topic modeling with NNMF, based on theoretical justification that incorporates the nature of our signals and the features under consideration.

3 Methodology

3.1 Feature generation

In the field on Natural Language Processing (NLP), textual documents are commonly described as word-count histograms. These descriptors are generally known as *bag-of-word* representations (BoW), since during their creation all the words in a document are (figuratively speaking) thrown into a bag, loosing all proximity information, then each word in a pre-defined dictionary is assigned the number of times it repeats in the bag. The final representation of the document is a vector of these counts.

The BoW representation was adopted in recent years into Computer Vision for the representation of images. Since images are not naturally divided into discrete elements (like words in a document), the first step is to transform the image into a series of word-analogues which can then be thrown into a bag. This discretization process is often achieved by clustering *patches* of images,

then assigning each patch the index of its cluster. The resulting feature vector for a given image is the histogram of the cluster associations of its patches. The cluster centroid are often referred to as the *codebook*, and the method as Bag of Visual Words (BoVW).

Here, we adapt the BoVW method to be used with the ACC signal. We start by defining the notion of a patch of an ACC signal.

3.1.1 definition: patch in an ACC signal

Let:

$$s = [s_1, ..., s_N]$$

be an ACC signal of length N. The patch of length l starting at index i of s is the sub-vector:

$$[s_i, ..., s_{i+l-1}]$$

thus, there are $N - l + 1$ distinct patches in s.

3.1.2 Codebook Generation

As in the BoVW case, ACC signals and patches do not consist of discrete elements. In order to count and histogram types of patches, we must first construct a patch-codebook. We suggest the following construction: given a codebook size k and a patch length l, for each ACC signal in the dataset, extract and pool all of the l-length-ed patches. Next, using K-means cluster the patches into k clusters. The resulting k centroids will be called the codebook. The intuition behind using patches to describe an ACC signal, is that behavioral modes should be definable by the distribution of short-time-scale movements that they are comprised of. Since different behavioral modes occur at various characteristic timescales, we would like to repeat the process for more than one patch length, in order to efficiently capture all ACC patterns of relevance. Thus, we generate a separate codebook for several time-scales in the appropriate range, depending on the behavioral modes we are interested in (Alg. 1).

3.1.3 Feature Transformation

Once we have constructed the codebook for all of the scales, we are ready to transform our ACC signals into the final Multi-Scale Bag of Patches (MS-BoP) descriptor. For each ACC record in the dataset, and for each scale, we extract all patches of the signal at that scale, and assign each one the index of the nearest centroid in the appropriate codebook. For each scale we then histogram the index values to produce a (typically sparse) vector of the length of the codebook. The final representation is the concatenation of histograms for the various scales (Alg. 2).

Algorithm 1 Multi Scale Codebook Generation

input:

$\{s_i\}_{i=1}^{p}$	the set of raw acceleration measurements
$l_1, .., l_m$	list of scales to use
$k_1, ..., k_m$	list of corresponding sizes per codebook

output:

$CB_1, ..., CB_l$	the generated codebooks. $CB_i[j]$ is the $j-th$ word in the $i-th$ codebook $(i = 1, ..., l; j = 1, ..., k_i)$

for scale := 1,...,l **do**
    patches := list of all patches of scale l_{scale} in $\{s_i\}_{i=1}^{p}$
    CB_i := K_means(patches, k_{scale}).centroids
end for
return $CB_1, ..., CB_l$

Algorithm 2 MS-BoP feature transformation

input:

$CB_1, ..., CB_l$	The l codebooks, output of Alg. 1.
$l_1, .., l_m$	list of the patch scales that were used in Alg. 1.
s	an ACC signal to transform

output:

f	The MS-BoP representation of signal s

for scale := 1,...,l **do**
    f_{scale} := a zeros vector of the same length as CB_{scale}
    patches := list of all patches of scale l_{scale} in s
    for each p in patches **do**
        idx := index of the closest word to p in the codebook CB_{scale}
        increment $f_{scale}[idx]$ by 1
    end for
end for
f := stack_vectors($f_1, ..., f_l$)
return: f

3.2 Mixture property of patch features

In order to motivate the proposed model (next section), we present the mixture property of patch features. We assume that our signals have the property that a large enough part of a sample from a certain behavioral mode will have distribution of patches that is the same as the distribution in the entire sample. The meaning of this assumption is that each behavioral mode has a distribution of patches that characterizes it at each scale.

Intuitively, if a signal s is constructed by taking the first half of a signal s_a and the second half of an equal length signal s_b, then the distribution of patches in s will be approximately an equal parts mixture of those in s_a and in s_b. The reason for this is that a patch in s is either (a) completely contained in s_a and will then be distributed like patches in s_a or, (b) completely in s_b, and will then be distributed like patches in s_b or, (c) starts in s_a and continues into s_b, in which case we know little about the patch distribution and consider

it as noise. The key point is that the number of patches of type (c) is at most twice the length of the patch, and thus can be made small in relation to the total number of patches which is in the order of the length of the signal. More formally:

Let s be an ACC signal composed of a concatenation of t_1 consecutive samples during behavioral mode a and t_2 consecutive samples during behavioral mode b (see Figure 1). Denote $p_{mode}(v)$ the probability of a patch v of length l in behavioral $mode \in \{a, b\}$. Let v be a patch drawn uniformly from s, then:

$$\begin{aligned} p(v) &= Pr(A)p(v|A) + Pr(B)p(v|B) + Pr(C)p(v|C) \\ &\geq Pr(A)p_a(v) + Pr(B)p_b(v) \\ &= \frac{t_1 - l}{t_1 + t_2}p_a(v) + \frac{t_2 - l}{t_1 + t_2}p_b(v) \\ &= \frac{t_1}{t_1 + t_2}p_a(v) + \frac{t_2}{t_1 + t_2}p_b(v) - \epsilon \end{aligned}$$

where events A, B, C denote the patch being all in s_1, all in s_2 and starting in $s1$ and ending in s_2 respectively, and:

$$\epsilon = \frac{l}{t_1 + t_2}[p_a(v) + p_b(v)]$$

ϵ can be made arbitrarily small by making $t_1 + t_2$ large and keeping l constant, meaning that for patches small enough in relation to the length of the entire signal, the distribution of patches of the concatenated signal is a mixture (convex combination) of the distributions of the parts, with mixing coefficients proportional to the part lengths. We note that this result can easily be extended to a concatenation of any finite number of signals, as long as each one is sufficiently long in comparison to the patch width.

Since behaviors of real-world animals may start and stop abruptly, and a recorded ACC signal is likely to be a concatenation of signals representing different behavioral modes (typically 1-3), the above property inspires a model that is able to capture such mixtures in an explicit fashion. Furthermore, the resulting mixture coefficients may provide some insight into the nature of the underlying behaviors and the relationships between them – for example, which often appear alongside each other, and which are more temporally separated.

3.3 The proposed model

Let k denote the number of behavioral modes under consideration, and p the dimension of the representation of ACC observations. Following the mixture property presented in the previous section, we assume that every sample is a convex combination of the representation of a "pure" signal of the various behavioral modes. Further, we assume the existence of a matrix $F \in R^{pk}$, the *factor matrix*, such that the $i-th$ column of F is the representation of a pure

signal of the $i-th$ behavioral mode, which we will call the factor associated with the $i-th$ behavioral mode. Let s be an ACC sample, then:

$$s = F\alpha + \epsilon \tag{1}$$

where $\epsilon \in R^p$ is some random vector. In other words, we say that the sample s is a linear combination of the factors associated with each of the behavioral modes with some remainder term. For the full dataset, we then have:

$$S = FA + \epsilon \tag{2}$$

where F is the same matrix, $A's$ columns are the factor loadings for each of the samples denoted α in (1), and $\epsilon \in R^{pN}$ is a random matrix. Since our features are non-negative histograms, and we would like the factor loadings to be non-negative, we constrain the matrices F, A to have non-negative values. We solve for F, A using a least squares criterion:

$$\begin{aligned} &\underset{F,A}{\text{argmin}} && \|FA - S\|_F^2 \\ &\text{subject to} && F_{i,j}, A_{i,j} \geq 0 \ \forall i,j \end{aligned} \tag{3}$$

This is by now a standard problem, which can be solved, for instance, using alternating non-negative least squares [27]. The idea behind the algorithm (Algorithm 3) is that while the complete problem is not convex, and not easily solved, for a set A it becomes a simple convex problem in F, and vice versa. This inspires the simple block-coordinate-descent algorithm which minimizes alternately w.r.t each of the matrices. Since this procedure generates a (weakly) monotonically decreasing series of values of the objective (3), it is guaranteed to converge to a local minimum[1].

Algorithm 3 Alternating Non-Negative Least Squares

input:

S the complete matrix $S \in R^{pN}$
k factorization rank

output:

F, A matrices $F \in R^{pk}$, $A \in R^{kN}$

1: $F :=$ random initialization
2: $A :=$ random initialization
3: **while** not converged **do**
4: $\quad F := \underset{F}{\text{argmin}} \|FA - S\|_F^2 \ \ s.t.\ F_{i,j} \geq 0 \ \ \forall i,j$
5: $\quad A := \underset{A}{\text{argmin}} \|FA - S\|_F^2 \ \ s.t.\ A_{i,j} \geq 0 \ \ \forall i,j$
6: **end while**
7: **return** F, A

[1] The objective is bounded from below by 0

3.4 Speed-up via sampling

Since this method may potentially be applied to large datasets (containing at least hundreds of millions of records and many billions of patches), it is worth mentioning that all parameter-learning steps of the algorithm can be processed (identically to the original method) on a sample of the dataset. During codebook generation, records in the dataset and/or patches in each used record could be sampled to reduce the number of resulting patches we have to cluster. Next, fitting F and A on a sample of the records gives us the factor matrix, but not the factor loadings per record of the dataset. However, once we have F the optimization problem (3) turns into:

$$\begin{aligned} \underset{A}{\text{argmin}} \quad & \|FA - S\|_F^2 \\ \text{subject to} \quad & A_{i,j} \geq 0 \ \forall i, j \end{aligned} \tag{4}$$

a simple convex problem in which the factor loadings per record (columns of A) can be minimized independently for each record s in the dataset, as follows:

$$\begin{aligned} \underset{\alpha}{\text{argmin}} \quad & \|F\alpha - s\|^2 \\ \text{subject to} \quad & \alpha_i \geq 0 \ \forall i \end{aligned} \tag{5}$$

3.5 Extension to the multi-sensor case

Thus far we have constructed a topic model applicable for data derived from a single (albeit possibly multi-dimensional) sensor. The multi-sensor (or sensor-integration) case is of particular interest in this case because many devices containing accelerometers also include other sensors such as gyroscopes and magnometers. Since each of these is recording at different frequencies, we can't simply consider them to be extra dimensions in the same time-series produced by the 3D accelerometer. The integrative framework we suggest assumes that the same behavioral modes are manifested in distinct patterns for each of the sensors. Thus, we will have separate factor matrices:

$$F^1, ..., F^l$$

for the l sensor types, and a single shared factor loading matrix A. Denoting the features matrices of the MS-BoP features for each of the l sensor types:

$$S^1, ..., S^l$$

we now look for matrices:

$$A, F^1, ..., F^l$$

such that:

$$\forall i : S^i \approx F^i A$$

which we encode in the following optimization problem:

$$\begin{aligned} \underset{F^1,\ldots,F^l,A}{\text{argmin}} \quad & \frac{1}{l}\sum_{i=1}^{l}\|F^iA - S^i\|_F^2 \\ \text{subject to} \quad & F^k_{i,j}, A_{i,j} \geq 0 \ \forall i,j,k \end{aligned} \tag{6}$$

This problem is solvable using the same type of method. Specifically, we will now show that this new problem can be re-written in the same form as (3), with both the sample and factor matrices stacked. Denote:

$$F = \begin{bmatrix} [F^1] \\ \vdots \\ [F^l] \end{bmatrix}$$

and:

$$S = \begin{bmatrix} [S^1] \\ \vdots \\ [S^l] \end{bmatrix}$$

then (6) becomes::

$$\begin{aligned} \underset{F,A}{\text{argmin}} \quad & \|FA - S\|_F^2 \\ \text{subject to} \quad & F_{i,j}, A_{i,j} \geq 0 \ \forall i,j \end{aligned}$$

since the $\frac{1}{l}$ scaling factor makes no difference to the *argmin*. In summary, the multi-sensor case where a separate factor matrix is allocated to each sensor, with a joint factor-loading matrix, is identical to the single-sensor case when the MS-BoP features for each sensor are stacked vertically.

3.6 Extension to the supervised and semi-supervised cases

Supposing observation (or any other mechanism) allowed us to obtain "pure" ACC signals for some (or all) of the behavioral modes. Using the mean MS-BoP representation of the signals in each of these modes for the corresponding column of F, we are left with a convex problem similar to (3), where the optimization is over the remaining elements of F only.

In the extreme case, when we have labeled samples for a pure ACC signal for all the behavioral modes under consideration, and thus all of F is predetermined, the resulting problem is equivalent to (4). Namely, we are left with the task of obtaining the factor loadings for the remaining (unlabeled) data.

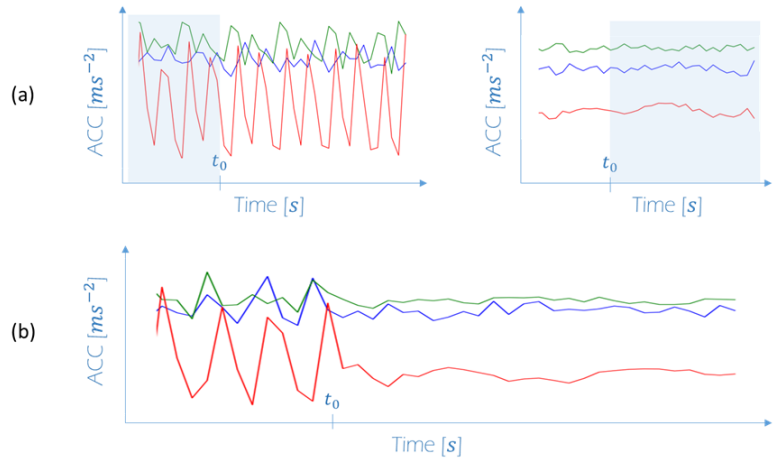

Fig. 1 Pure and mixed triaxial ACC signals. Pure ACC signals (panel A) are measured during a single behavioral mode. However, in most cases a single measurement contains a mixture of more than one behavioral mode (Panel b), and may be viewed as a concatination of the beginning/end of two pure signals. The colors represent each of the three acceleration dimensions.

3.7 Limitations

Consider a solution, matrices F, A that minimize objective (3), so that:

$$S \approx FA$$

Clearly, for any Orthogonal matrix Q (of the appropriate dimensions):

$$FA = FQQ^TA = (FQ)(A^TQ)^T$$

thus, the solution:

$$F' = FQ$$
$$A' = (A^TQ)^T$$

is also a minimizer of objective (3), iff the matrices F', A' obey the constraints:

$$F'_{i,j}, A'_{i,j} \geq 0 \ \forall i, j \tag{7}$$

While this clearly holds if Q is a permutation matrix, there are (always) orthogonal matrices Q which contain negative elements for which the constraints in (7) hold. From the construction of F' and A', we can interpret them as an entanglement of our factors and loadings (technically, what we find is the span of the correct factors, but not the factors themselves). We note that while this property limits the ability to recover factors that generate the data, in practice the factors themselves are useful for analysis of behavioral topics, as demonstrated in the section below.

We leave to future research the issue of the disentanglement, which should be achieved via regularization with respect to A in the original optimization problem (3).

4 Results

In this section we present experiments designed to compare our method to alternatives, and derive insights about the data. Results are then discussed in the next section.

Data for these experiments consists of $3D$ acceleration measurements from bio-loggers which were recorded during 2012. Each measurement consists of 4 seconds at $10Hz$ per axis, giving a total of 120 values.

A ground truth partitioning of the data was obtained using standard machine learning techniques (see [20,17] for more details regarding the methodology), based on 3815 field observations each of which was assigned one of 5 distinct behavioral modes (Walking, Standing, Sitting, Flapping, Gliding). Experiments were conducted using stratified sampling of $100,000$ measurements ($20,000$ per behavioral mode).

Matrix factorization was preformed using the scikit-learn [19] python software library (see [14] for method details). In all experiments the results were stable across repetitions, leading to essentially zero standard deviation, and therefore the reported results correspond to single repetitions.

The purpose of these experiments is to assess to what extent the soft-partitioning via our topic model method relates to the hard, ground truth partitions. Our method is compared to the following:

Random partitioning: each sample is assigned a value drawn uniformly from the set of possible partitions $\{1, 2, .., k\}$

Uniform partition: each sample is assigned the same distribution of $\frac{1}{k}$ per partition, over the k partitions.

Kmeans: the sample are partitioned using Kmeans.

Gaussian Mixture Models (GMM): GMM is used to assign samples k partition coefficients.

where (a) and (b) are used as controls, (c) and (d) are used as representative hard and soft clustering methods, respectively.

The data is then divided randomly into two equal parts designated train and test. Using the training-set we learn the partitioning of the data for each of the methods (random, uniform, Kmeans, GMM, and NNMF). Next, for each method separately, we assign each of the partitions one of the semantic labels (Flapping, Gliding, Walking, Standing, Sitting). In order to do this we group the data in the training-set according to the semantic label it received (the supervised annotation), and compute the average loading for each label in the partition. The final assignment for the partition is the label with the highest mean loading in it (see schematic in Fig. 2).

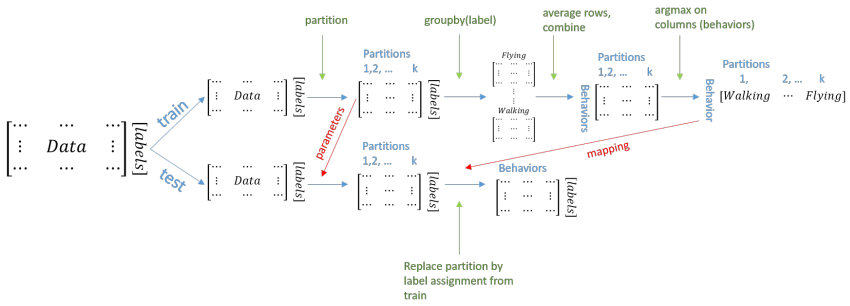

Fig. 2 Schematic flow of partition evaluation

Table 1 Mean label association per ground-truth behavioral mode. NNMF with 30 factors. Normalized rows.

Ground truth / Assignment	Flapping	Gliding	Walking	Standing	Sitting
Flapping	**51.25%**	13.66%	13.37%	4.33%	17.39%
Gliding	0.75%	**49.98%**	8.49%	3.95%	36.84%
Walking	2.41%	19.71%	**43.92%**	20.56%	13.41%
Standing	0.86%	13.30%	1.04%	**74.93%**	9.88%
Sitting	0.01%	30.88%	0.15%	10.46%	**58.50%**

The evaluation stage is preformed on the test-set only. Resemblance to the ground-truth is measured using log-loss (Figure 3) and $0-1$ loss (Figure 4), after partition values are converted to soft label assignments using the mapping derived from the training set (see schematic in Fig. 2). For an assignment $l_1, ..., l_5$ for the 5 behavior labels, where the ground-truth label is i, we use the $0-1$ loss:

$$l_{0-1} = \begin{cases} 0 & i = argmax\{l_1, ..., l_5\} \\ 1 & otherwise \end{cases} \tag{8}$$

and the log-loss:

$$l_{log} = -log(l_i) \tag{9}$$

Table 1 shows the average distribution of supervised (ground-truth) behavioral modes for partitions assigned each of the labels, in the form of a confusion matrix. Partitions were obtained using non-negative matrix-factorization (NNMF) with $k = 30$, and associations between partitions and labels as described above. Data is presented after row normalization to facilitate between-row comparison.

5 Discussion

As expected, both $0-1$ and log-loss error plots are monotonically decreasing in the number of clusters (we use the term clusters here for cluster/partition/topic

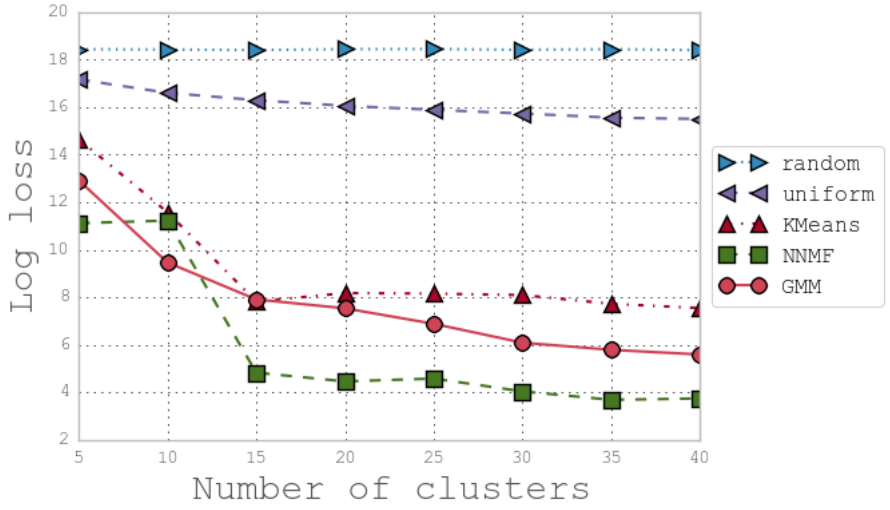

Fig. 3 Log loss of soft-assignment to each of the ground-truth classes using each of the methods under consideration. (NNMF: non-negative matrix factorization, GMM: Gaussian mixture model)

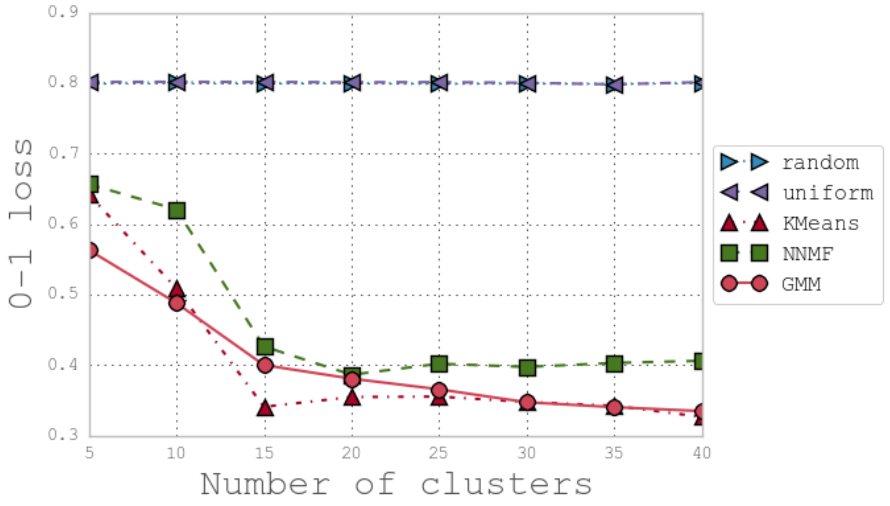

Fig. 4 0-1 loss of hard-assignment to each of the ground-truth classes using each of the methods under consideration. For the soft-assignment partitioning methods, hard-assignment is achieved using argmax. (NNMF: non-negative matrix factorization, GMM: Gaussian mixture model)

depending on the method under consideration). The most striking result is that while the matrix-factorization topic model method preforms well compared to the other methods with respect to the log-loss metric (Figure 3), it is not quite as good with respect to the 0-1 loss (Figure 4).

In order to better understand this phenomena, we take a closer look at the data. Consider an observation where the animal takes a single step during the 4-second acceleration measurement window, and stands still for the rest of it. In order not to dramatically underestimate the amount of walking, an observer will label this sample as Walking (In fact, most samples are probably mixtures).

From the mixture property of the MS-BoP features (see Methodology section), when using the matrix factorization topic model approach we would expect to get a Walking factor proportional to the time spent doing so in the measurement windows. Thus, for a sample with some walking (say, less than 50%) we get a miss in the 0-1 loss metric, but a better score in the log-loss which is more sensitive to assignment of low probabilities to the correct class.

Table 1 sheds more light on the aforementioned result by showing average assignment of factors for each of the ground-truth classes, in the form a confusion matrix. Flapping samples indeed received the highest weight, on average, on Flapping factors (51.25%), but the Gliding and Walking factors get over 13% each. This may be due to the fact that Storks indeed glide between wing flaps, and may have walked prior to taking off during the observations which are inherently biased to behavior close to the ground (where the observer is). Conversely, none of the other behavioral modes include a significant amount of Flapping factors.

This result may also point to the tendency (or strategy) of field observers to assign the more active behavior to mixed samples (In which case a sample where the bird flaps for a part of the duration of the measurement would be assigned to Flapping, in the same sense that a step or two would qualify an otherwise stationary sample as Walking).

We note that the Sitting factors received factor weights higher than expected in all other behavioral modes. It might be interesting to try and overcome this sort of systematic error using a column normalization. We defer this to future research.

6 Conclusions

In this paper we describe a matrix factorization based topic model approach to behavioral mode analysis from accelerometer data and demonstrate its qualities using a large Movement Ecology dataset. While clustering and topic modeling with matrix factorization is by no means a new idea, the novelty here is in the integration with patch features (MS-BoP) that theoretically motivate the method in the context of time-series sensor readings for behavioral mode analysis.

The main contribution of this paper is in presenting a framework that will allow for a widespread use of behavioral mode analysis in Movement Ecology, and related fields where determining movement patterns from remote sensor readings is necessary. Further, we introduce the MS-BoP features, which may be applicable for many continuous sensor readings, and show that a linear mixture model is justified when using such features.

Acknowledgment

This work was supported in part by a grant from the Israel Science Foundation (ISF) to Prof. Daphna Weinshall.

References

1. S. Arora, R. Ge, and A. Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012.
2. D. D. Brown, R. Kays, M. Wikelski, R. Wilson, and a. Klimley. Observing the unwatchable through acceleration logging of animal behavior. *Animal Biotelemetry*, 1(1):20, 2013.
3. S. Cooke. Biotelemetry and biologging in endangered species research and animal conservation: relevance to regional, national, and IUCN Red List threat assessments. *Endangered Species Research*, 4(January):165–185, Jan. 2008.
4. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *Workshop on statistical learning in computer vision, ECCV*, 1(1-22):1–2, 2004.
5. C. H. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, volume 5, pages 606–610. SIAM, 2005.
6. J. A. V. Diosdado, Z. E. Barker, H. R. Hodges, J. R. Amory, D. P. Croft, N. J. Bell, and E. A. Codling. Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system. *Animal Biotelemetry*, 3(1):15, 2015.
7. J. Garriga, J. R. Palmer, A. Oltra, and F. Bartumeus. Expectation-maximization binary clustering for behavioural annotation. *arXiv preprint arXiv:1503.04059*, 2015.
8. A. C. Gleiss, R. P. Wilson, and E. L. C. Shepard. Making overall dynamic body acceleration work: on the theory of acceleration as a proxy for energy expenditure. *Methods in Ecology and Evolution*, 2(1):23–33, 2011.
9. M. Hebblewhite and D. T. Haydon. Distinguishing technology from biology: a critical review of the use of GPS telemetry data in ecology. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1550):2303–12, July 2010.
10. A. Hindle, D. Rosen, and A. Trites. Swimming depth and ocean currents affect transit costs in Steller sea lions Eumetopias jubatus. *Aquatic Biology*, 10(2):139–148, Aug. 2010.
11. M. Holyoak, R. Casagrandi, R. Nathan, E. Revilla, and O. Spiegel. Trends and missing parts in the study of movement ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49):19060–5, Dec. 2008.
12. N. Kern, B. Schiele, and A. Schmidt. Multi-sensor activity context detection for wearable computing. In *Ambient Intelligence*, pages 220–232. Springer, 2003.
13. T. Li and C. Ding. The relationships among various nonnegative matrix factorization methods for clustering. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 362–371. IEEE, 2006.
14. C.-b. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

15. M. Louzao, T. Weigand, F. Bartumeus, and H. Weimerskirch. Coupling instantaneous energy-budget models and behavioural mode analysis to estimate optimal foraging strategy: an example with wandering albatrosses. *Mov Ecol*, 2(8), 2014.
16. R. Nathan and W. Getz. A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49):19052–19059, 2008.
17. R. Nathan, O. Spiegel, S. Fortmann-Roe, R. Harel, M. Wikelski, and W. M. Getz. Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures. *The Journal of experimental biology*, 215(6):986–996, 2012.
18. A. Pantelopoulos and N. G. Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(1):1–12, 2010.
19. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
20. Y. S. Resheff, S. Rotics, R. Harel, O. Spiegel, and R. Nathan. AcceleRater: a web application for supervised learning of behavioral modes from acceleration measurements. *Movement Ecology*, 2(1):25, 2014.
21. Y. S. Resheff, S. Rotics, R. Nathan, and D. Weinshall. Matrix factorization approach to behavioral mode analysis from acceleration data. In *Data Science and Advanced Analytics (DSAA), 2015 International Conference on*. IEEE, 2015.
22. K. Q. Sakamoto, K. Sato, M. Ishizuka, Y. Watanuki, A. Takahashi, F. Daunt, and S. Wanless. Can ethograms be automatically generated using body acceleration data from free-ranging birds? *PloS one*, 4(4):e5379, Jan. 2009.
23. W. I. Sellers and R. H. Crompton. Automatic monitoring of primate locomotor behaviour using accelerometers. *Folia primatologica; international journal of primatology*, 75(4):279–93, 2004.
24. P. E. Smouse, S. Focardi, P. R. Moorcroft, J. G. Kie, J. D. Forester, and J. M. Morales. Stochastic modelling of animal movement. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1550):2201–11, July 2010.
25. O. Spiegel, R. Harel, W. M. Getz, and R. Nathan. Mixed strategies of griffon vultures (Gyps fulvus) response to food deprivation lead to a hump-shaped movement pattern. *Movement Ecology*, 1(1):5, 2013.
26. M. Takahashi, J. R. Tobey, C. B. Pisacane, and C. H. Andrus. Evaluating the utility of an accelerometer and urinary hormone analysis as indicators of estrus in a Zoo-housed koala (Phascolarctos cinereus). *Zoo biology*, 28(1):59–68, 2009.
27. Y.-X. Wang and Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *Knowledge and Data Engineering, IEEE Transactions on*, 25(6):1336–1353, 2013.
28. H. Williams, E. Shepard, O. Duriez, and S. Lambertucci. Can accelerometry be used to distinguish between flight types in soaring birds? *Animal Biotelemetry*, 3(1):1–11, 2015.
29. R. P. Wilson, C. R. White, F. Quintana, L. G. Halsey, N. Liebsch, G. R. Martin, and P. J. Butler. Moving towards acceleration for estimates of activity-specific metabolic rate in free-living animals: the case of the cormorant. *Journal of Animal Ecology*, 75(5):1081–1090, 2006.
30. K. Yoda, K. Sato, Y. Niizuma, M. Kurita, C. Bost, Y. Le Maho, and Y. Naito. Precise monitoring of porpoising behaviour of Adélie penguins determined using acceleration data loggers. *Journal of Experimental Biology*, 202(22):3121–3126, 1999.
31. K. Zagoris, I. Pratikakis, A. Antonacopoulos, B. Gatos, and N. Papamarkos. Distinction between handwritten and machine-printed text based on the bag of visual words model. *Pattern Recognition*, 47(3):1051–1062, 2014.