Segmenting accelerometer data from daily life with unsupervised machine learning

Dafne van Kuppevelt[1], Joe Heywood[2], Mark Hamer[3], Séverine Sabia[4,5], Emla Fitzsimons[2,6], Vincent van Hees[1]

1. Netherlands eScience Center, Amsterdam, The Netherlands
2. Centre for Longitudinal Studies, UCL Institute of Education, London, United Kingdom
3. School Sport, Exercise and Health Sciences, Loughborough University, UK
4. INSERM, U1018, Centre for Research in Epidemiology and Population Health, France
5. Department of Epidemiology and Public Health, University College London, UK
6. Institute for Fiscal Studies, London

Corresponding author: Prof. Dr. Emla Fitzsimons, e-mail: e.fitzsimons@ucl.ac.uk

Abstract

Purpose: The cut-points approach to segment accelerometer data is widely used in physical activity research but requires resource expensive calibration studies and does not make it easy to explore the information that can be gained for a variety of raw data metrics. To address these limitations, we present a data-driven approach for segmenting and clustering the accelerometer data using unsupervised machine learning.

Methods: The data used came from five hundred fourteen-year-old participants from the Millennium cohort study who wore an accelerometer (GENEActiv) on their wrist on one weekday and one weekend day. A Hidden Semi-Markov Model (HSMM), configured to identify a maximum of ten behavioral states from five second averaged acceleration with and without addition of x, y, and z-angles, was used for segmenting and clustering of the data. A cut-points approach was used as comparison.

Results: Time spent in behavioral states with or without angle metrics constituted eight and five principal components to reach 95% explained variance, respectively; in comparison four components were identified with the cut-points approach. In the HSMM with acceleration and angle as input, the distributions for acceleration in the states showed similar groupings as the cut-points categories, while more variety was seen in the distribution of angles.

Conclusion: Our unsupervised classification approach learns a construct of human behavior based on the data it observes, without the need for resource expensive calibration studies, has the ability to combine multiple data metrics, and offers a higher dimensional description of physical behavior. States are interpretable from the distributions of observations and by their duration.

1 Introduction

Accelerometers are increasingly used for studying daily physical activity. A common technique to process accelerometer data is the so called ‘cut-points’ approach. In the cut-points approach a researcher calculates the time spent with acceleration between certain thresholds to define physical activity levels (sedentary, light, moderate, vigorous), and some criteria on the amount of time spent at the different levels [1]. The threshold values are calibrated relative to an indirect calorimetry derived metabolic equivalent (MET) level which, in brief, is a proxy for energy expenditure relative to rest [2,3]. Results from the cut-points approach are easy to report and reproduce. However, this approach comes with two challenges. Firstly, it does not account for the known non-linear and complex relationships of acceleration with energy expenditure and activity types. Secondly, the relationship between acceleration, energy expenditure, and activity types may differ between populations, between acceleration sensor brands, accelerometer attachment locations, and between data processing metrics.

A possible way to better explore the information in the accelerometer data is the inclusion of extra data metrics in our analyses in addition to the magnitude of acceleration already used. For example, metrics can be added that indicate the orientation of the accelerometer under static conditions. Orientation has been used in related work, first for the detection of human posture [4], but more recently also for the detection of sleep [5], and in the Sedentary Sphere method for detecting sitting behavior and visualizing the data [6].

Further, the data can also be explored using automated methods such as machine learning. Machine learning methods that use labelled data, referred to as supervised machine learning, have previously been used for activity type classification and energy expenditure estimation [7–10]. Although supervised machine learning has shown potential, it has disadvantages similar to

the cut-points approach in that the trained classifier may overfit to the specific experimental conditions under which it was trained, including: characteristics of study participants, the activity types and experimental attributes (e.g. specific chair, bed, or stair case) used in the laboratory setting.

Unsupervised machine learning on the other hand has received less attention in relation to physical activity assessment. Unsupervised learning methods are data-driven, allow identification of the characteristic states in the data, and can be applied to free-living data directly. As a result, they do not require time consuming and expensive calibration studies. In addition, they can combine data features from acceleration and orientation to attempt to better discriminate between physical behaviors.

There are several methods for unsupervised segmentation of time series data. Hidden Markov Models (HMMs) as used by others for accelerometer data [11,12] do not model the duration of the state. The related Hidden Semi-Markov Models (HSMM) have the advantage of modelling time distribution for the different behavioral states. HSMM have proven to be valuable for similar segmentation tasks in ubiquitous computing [13], [14] .

In this study, we explore how an unsupervised method can be used to segment and cluster accelerometer data into informative descriptors of physical behavior using data from 500 fourteen-year-old participants. We apply HSMM to acceleration data with and without addition of the angles of the three accelerometer axis (x, y, z) relative to the horizontal plane and compare this approach with the cut-points approach in terms of data variability explained with principal component analysis. We also examine the plausibility of the agreement between HSMM-defined states, cut-points categories and time-use diary records.

2 Methods

2.1 Participants

Members of the Millennium Cohort Study (MCS), a longitudinal study managed by the Centre for Longitudinal Studies at the UCL Institute of Education participated. The MCS follows over 19,000 young people in the UK born in 2000/1, conducting home interviews every 2 to 4 years. Its sixth survey, at age 14, included the collection of physical activity data using wrist-worn activity monitors [15–17]. The study obtained ethical approval from the National Research Ethics Service (NRES) Research Ethics Committee (REC) – London Central (ref: 13/LO/1786). A random subset of 500 participants was selected for the present analysis.

2.2 Study protocol

Interviewers placed tri-axial accelerometers (GENEActiv) with respondents during home visits and requested them to wear the device on their non-dominant wrist for two complete days; one during the week and one at the weekend, randomly selected at time of placement. Each day lasted for 24 hours: from 4 am in the morning to 4am the following morning. Participants received text messages reminding them to complete the tasks on the selected days. Upon completion of the second day of activity data collection respondents were required to return the accelerometer in a pre-paid envelope. The accelerometer sample frequency was set at 40 Hertz and the dynamic range was ±8*g*. The orientation of the acceleration axes, seen from the anatomical position, is as follows: the x-axis points in medio-lateral direction (direction of thumb), the y-axis in longitudinal direction (direction of middle finger), and the z-axis in dorsal-ventral direction (perpendicular to skin). For further details on the study design and protocol see [16,17]. Additionally, participants were asked to record their categories of behavior in a time use diary[15]. Participants were asked to provide a full record of what they did on the two days

(activities), from 4am to 4am the next day, as well as where they were, who they were with, and how much they liked each activity – using pre-coded lists. Participants were offered the choice between an app and online version (with a paper version available for those unable or unwilling to use one of the other two).

2.3 Accelerometer data pre-processing

The raw data from the accelerometer is processed with R package GGIR [18] which extracts the two days on which the accelerometer was worn, performs an acceleration sensor calibration check, aggregates the data into features which are then used to clean the data for accelerometer non-wear time and to describe the movement kinematics[19–21]. The main features extracted from the data are five second average of the *vector magnitude of body acceleration* and the *orientation of each of the three acceleration sensors relative to the horizontal plane*. The vector magnitude of acceleration is calculated as Euclidian Norm Minus One (ENMO), which in formula corresponds to $max\{\sqrt{acc_x{}^2 + acc_y{}^2 + acc_z{}^2} - 1{,}0\}$, with acc_x, acc_y, and acc_z referring to the three orthogonal acceleration axes pointing in the lateral, distal, and ventral directions, respectively. The orientation angles of the acceleration axis relative to the horizontal plane are calculated as $angle_z = \left(\tan^{-1}\frac{acc_z}{\sqrt{acc_x{}^2+acc_y{}^2}}\right)\cdot\frac{180}{\pi}$, $angle_y = \left(\tan^{-1}\frac{acc_y}{\sqrt{acc_x{}^2+acc_z{}^2}}\right)\cdot\frac{180}{\pi}$,

$and\ angle_x = \left(\tan^{-1}\frac{acc_x}{\sqrt{acc_y{}^2+acc_z{}^2}}\right)\cdot\frac{180}{\pi}$.

For all continuous time periods with no z-angle change of more than 5 degrees lasting at least 5 minutes, the acceleration values were set to zero to take out the possible influence of increased calibration error during sustained inactivity periods, e.g. as a result of temperature [19].

To account for variation in sign of the signal as a result of wearing the accelerometer upside down, the angles were corrected as follows: If the value for the *x-angle* has a positive median during all time periods detected as active (calculation described in the next section) then the device is considered to be worn incorrectly, in that case, the *x-angle* and *y-angle* (flipped around zero) are negated to mirror the orientation.

2.4 Conventional cut-points approach

The acceleration magnitude and angle-z feature are used to assign each of the 5-second epochs to one of the following ten categories:

- Sustained inactivity: all continuous time periods with no z-angle change of more than 5 degrees lasting at least 5 minutes [5];
- Inactivity: defined as acceleration below 40 m*g*, divided into periods of inactivity lasting less than 10 minutes, bouts of inactivity lasting between 10 and 30 minutes, and bouts of inactivity for at least 30 minutes;
- Light physical activity (LPA): defined as acceleration between 40 and 120 m*g*, subdivided into spontaneous LPA lasting less than 1 minute, bouts LPA lasting between 1 and 10 minutes, and bouts LPA lasting at least 10 minutes;
- Moderate or vigorous physical activity (MVPA): defined as acceleration above 100 m*g*, subdivided into spontaneous MVPA lasting less than 1 minute, bouts MVPA lasting between 1 and 10 minutes, and bouts MVPA lasting at least 10 minutes.

In the absence of validated cut-points for our population, we used the reported acceleration values across activity types in children and adults from a study by Hildebrand and colleagues [22]. The choice for three intensity levels, inactivity, light and MVPA, is widely used in the

physical activity research community. The sub-classification of these levels in bout durations, was guided by the common practice to look for bouts of at least 10 minute of MVPA [1], and the common practice of looking for bouts of at least 30 minutes inactivity or sedentary behavior [23]. The sustained inactivity category is a proxy for sleeping time [5].

2.5 Hidden semi-Markov models

The goal of using an unsupervised method is to segment the data in time periods that can be clustered. In a Hidden semi-Markov model (HSMM) clustering is performed into *hidden states.* The word *hidden* is used because the states are not directly observed, but found by the algorithm. Further, the abstract word *states* is used for the data clusters because we do not know (yet) what physical activity class they represent. However, in practice a state can be interpreted as a physical activity class, with its own characteristic distributions of orientation and acceleration values (the observations that are not hidden) and a characteristic distribution of duration. A graphical representation of the HSMM is visualized in Fig 1.

A more detailed explanation of (Bayesian) HSMMs and the chosen configuration is given in the supplement of this paper. For the implementation of the Bayesian HSMMs the python package **pyhsmm (**https://github.com/mattjj/pyhsmm) is used, the code is available on Github ([24] and [25]).

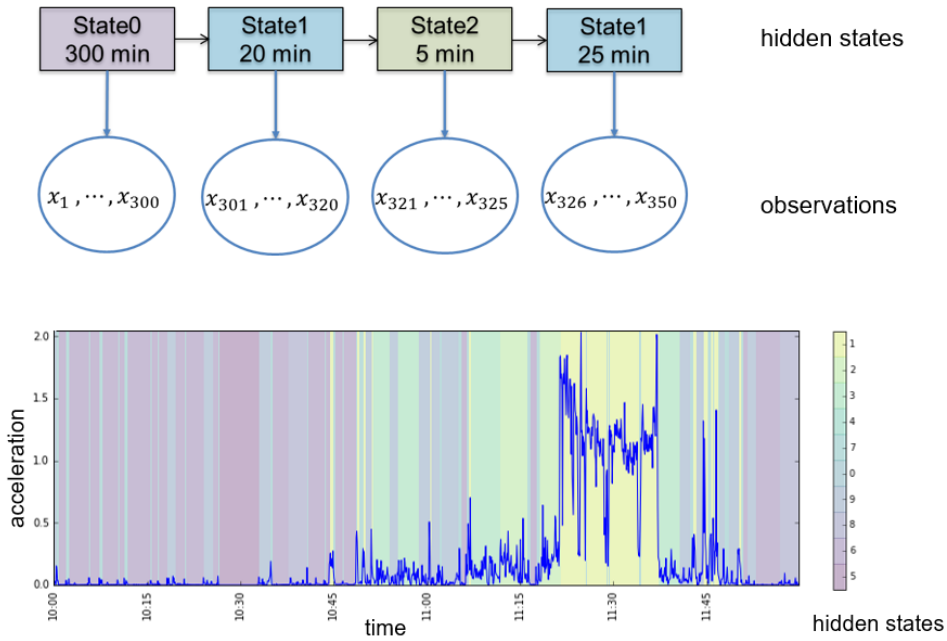

Fig 1. Hidden Semi-Markov model, explained schematically (above) and an example of the resulting segmentation (below). The observations (accelerometer features), denoted by x, are segmented into states of variable length.

The maximum number of states can be given as user input to the model, we choose 10 comparable to the number of cut-points categories. We experimented with two models with different observations as input. The first model used only acceleration, the second model used acceleration together with $angle_x$, $angle_y$ and $angle_z$. We will further refer to the resulting models as the *acceleration* model and the *acceleration+angles* model, respectively. The information used *acceleration* model is most comparable to the cut-points approach, since the cut-points approach only uses angle for one out of the ten categories. It is not possible in HSMM to instruct the model to use specific variables for only a single state.

2.6 Evaluation

Time spent in each state per day and cut-points category was calculated for participants with full 24 hours of data. Principal component analysis was used to quantify the number of principal components needed to explain at least 95% of the variance in the time use variables, as an indicator for the information dimensionality produced by the cut-points approach, HSMM approach with only *acceleration*, and the HSMM approach with *acceleration+angles*, respectively. The PCA statistics give an indication of collinearity between the variables: the more components are needed to explain the variance of the data, the less colinear the variables are.

Next, to assess comparability of the cut-points and HSMM approaches, correlation coefficients were calculated between time spent in states and cut-points categories individually, as well as between groups of states and categories grouped by acceleration level. Further, to establish that the states came with an expected plausible variation in conventional method output, a descriptive comparison was done of HSMM states, acceleration values, angle values, cut-points categories, and time use diary records. To ease interpretation, only days of measurement with full 24 hours of valid data were considered for the descriptive comparison to ease interpretation. Here, valid data was considered measurements with no accelerometer non-wear time [21,26]. We will focus mostly on the HSMM models using *acceleration+angles* since we expect the addition of angle variables to give extra insights, *acceleration* only results will be reported in the supplement.

The random sample of 500 participants was motivated by computational constraints and considering that the cut-points are typically derived in laboratory studies with less than 100 participants. To evaluate that a subsample can generalize to a larger population we tested the reproducibility. A HSMM model was trained on a random subset of 250 participants out of the

total set of 500, Nmax = 10, using only the *acceleration* variable since this makes it easier to compare state distributions. The states of the original model trained on 500 participants and the model trained on the subset were sorted on acceleration mean so that a match between corresponding states of both models could be made. The model parameters (mean and sigma for the observation distributions, lambda for the duration distributions) of the corresponding states of the two models were compared. Evaluation were based on the closeness of distributions for each state using the Kullback-Leibler divergence, which is an information-theoretic measure to calculate the distance between two statistical distributions [27]. The Kullback-Leibler divergence is denoted with KL(P|Q), where P is the distribution of the observed values as learned by the original model on 500 participants, and Q the corresponding distribution for the subset model.

3 Results

A total of 9122 participants accepted to wear the accelerometer, 4970 participants returned the accelerometer and time use diary, out of which a random subsample of data from 500 participants was used for the present study. The demographics of this sample are: 42.4% males; 21.3 ± 4.2 kg/m^2 BMI; 59.8%, 10.9%, 12.5%, and 16.8% from England, Northern Ireland, Scotland, and Wales, respectively, and; Equivalized household income quintiles from lowest to highest are 10.1%, 13.1%, 20.4%, 29.1%, and 27.3%. Note that the data of all 500 participants, regardless of whether the data was complete, could be used for training the HSMM model. Among these observations there was on average 21.0 (Inter quartile range: 20.2-24.0) hours of valid data per day, 38.8% of participants spent at least 60 minutes in MVPA on at least one day, and daily acceleration was 31 ± 13 m*g*. Out of the participants who wore the accelerometer for 24 hours on both days: the accelerometer was worn for 40% of a day; 36.5% spent 60 minutes in MVPA on at least one of those two days, and the acceleration was 34±13 mg [21]. The demographics of the 246 participants who wore the accelerometer for 24 hours on both days was: 41.5% male; 21.2 ± 3.9 kg/m^2 BMI; 61.5%, 8.5%, 12.0%, and 18.0% from England, Northern Ireland, Scotland, and Wales respectively, and; Equivalized income quintiles from lowest to highest were 7.0%, 13.5%, 21.5%, 25.5%, and 32.5%.

3.1 Correlation analyses and PCA

Ten different states were found by the HSMM training for both the models using *acceleration* only and using *acceleration+angles*. The percentage of explained variance for the number of principal components of the derived time use variables is plotted in Fig 2. The lower the curve is, the more dimensions of information the dataset has. Four, five, and eight principle components

were needed to explain 95% of the variance of the variables for the cut-points categories, *acceleration* HSMM model, and *acceleration+angle* HSMM model, respectively.

The average absolute correlation between variables of time spent in the 10 different cut-points defined physical activity categories was 0.27±0.24, for the states of the *acceleration* model it was 0.25±0.18 and for the *acceleration+angles* model it was 0.20±0.13, as illustrated in Fig 3. For the cut-points categories, there were 8 combinations of variables (out of 45 combinations) that had an absolute correlation of more than 0.5, for the *acceleration* model and the *acceleration+angles* model there were 6 and 1 combinations respectively.

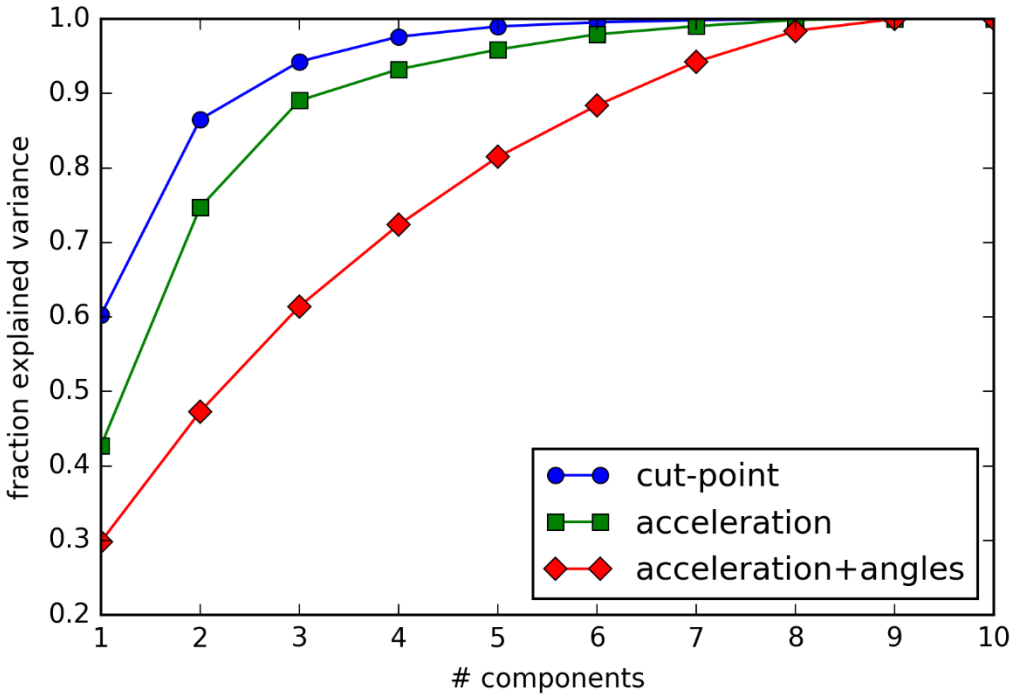

Fig 2: Scree plot showing the explained variance for principal components of time spent in cut-points categories and HSMM states for the HSMM based on acceleration and the HSMM based on acceleration+angles

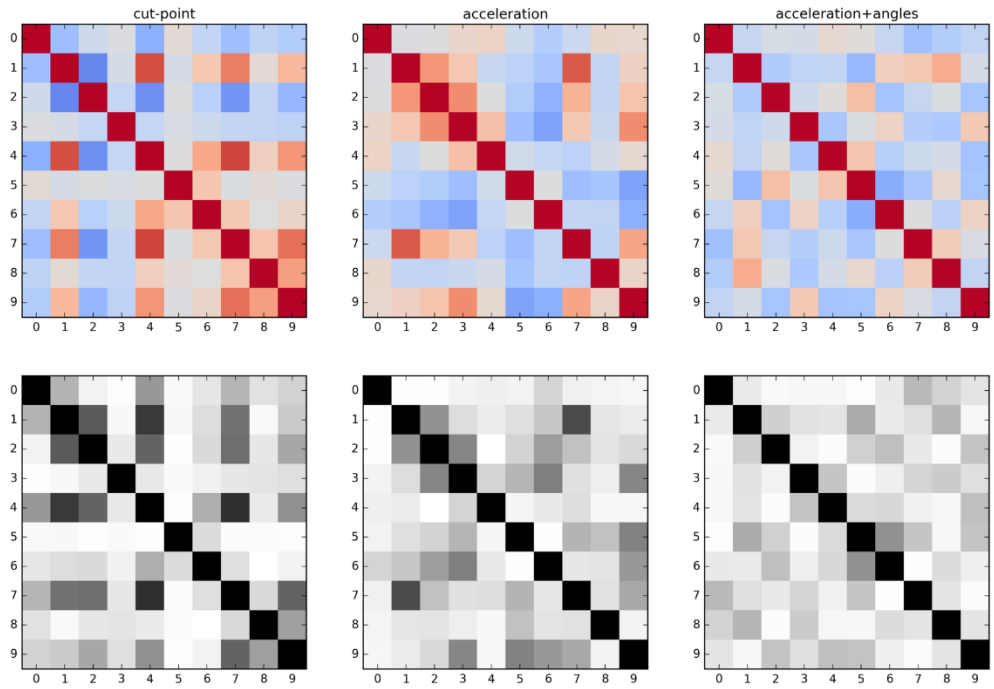

Fig 3: Correlation (top) and absolute value of correlation (bottom) matrices for cut-points categories and model states

3.2 Comparison of acceleration and angles to cut-points states

Distributions in acceleration values, angle values, and durations varied by state and threshold categories, see Fig 4. In Table 1 we see how much time the participants spent on average in each cut-points category and each state. A detailed description of the states in the *acceleration+angles* model is given in the supplement. We also compare the values in Table 1 for parallels between states and cut-points categories. The same can be done for the *acceleration* model (available in table S3 of the supplement).

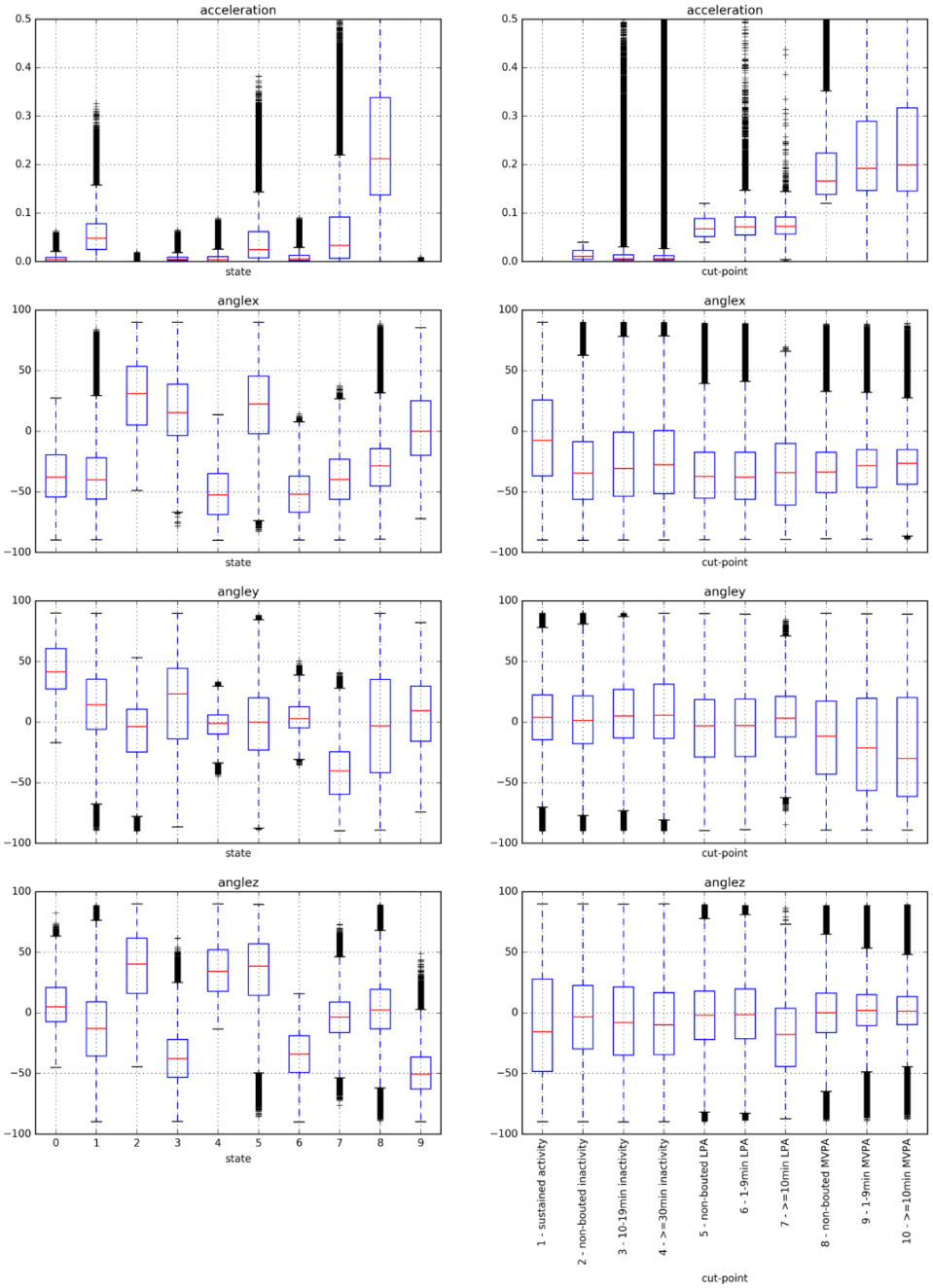

Fig 4: Boxplots for acceleration and angle values per HSMM state (acceleration+angles model, left column) and cut-points category (right column)

The model states and cut-points categories can both be grouped in combinations of respectively states and categories that have similar acceleration levels, corresponding to sustained activity, inactivity, LPA and MVPA, see Table 1. States were grouped as follows: the group {<0>, <2>} corresponds to *sustained activity*, states {<0>, <3>, <4>,<6>} to *inactivity*, states {<5>, <1>, <7>} to *LPA* and state {<8>} to *MVPA*. The Pearson's correlation coefficient between the time spent per grouped cut-points category and the time spent per grouped states were 0.69, 0.56, 0.72 and 0.88, respectively for each level of activity defined.

Table 1: Average time spent (minutes) per participant per day in each state from the acceleration+angle method and cut-points category. The states are sorted on mean acceleration, resulting in higher numbers around the diagonal.

cut-points category	state: 9	2	0	3	4	6	5	1	7	8	total	Avg acc
1 - sustained activity	197	105	29	0	43	37	2	5	13	0	432	0.0
2 - non-bouted inactivity	2	6	32	30	44	52	28	41	47	2	284	14.5
3 - 10-19min inactivity	3	7	23	26	23	33	12	15	15	1	156	13.3
4 - >=30min inactivity	4	12	46	50	34	59	18	27	28	2	280	13.3
5 - non-bouted LPA	0	0	0	0	2	3	19	66	43	7	142	71.6
6 - 1-9min LPA	0	0	0	0	0	0	2	6	5	1	15	75.1
7 - >=10min LPA	0	0	0	0	0	0	0	0	0	0	0	77.3
8 - non-bouted MVPA	0	0	0	0	0	0	5	10	19	23	57	220.4
9 - 1-9min MVPA	0	0	0	0	0	0	1	3	9	15	28	286.7
10 - >=10min MVPA	0	0	0	0	0	0	1	2	6	9	17	310.5
total	206	131	130	106	146	185	88	175	186	60	1412	
avg acc	0.0	0.6	6.5	7.4	7.8	9.6	42.5	55.9	58.2	305.2		

The cut-points categories differed in duration by design (standard deviation of duration means=15.9 minutes), while less difference was observed in durations between states (Fig 5) as there were all shorter than 17 minutes and standard deviation of the duration means was 4.3 minutes for the *acceleration* model and 2.9 minutes for the *acceleration+angles* model. In contrast, the distribution of angle values is different for the states with similar acceleration levels. The average values in the cut-points categories for the x, y and z angles have a standard deviation of respectively 8.1, 9.3 and 6.6 degrees over all cut-points categories, whereas over the states this is 31.7, 21.6 and 31.7 degrees.

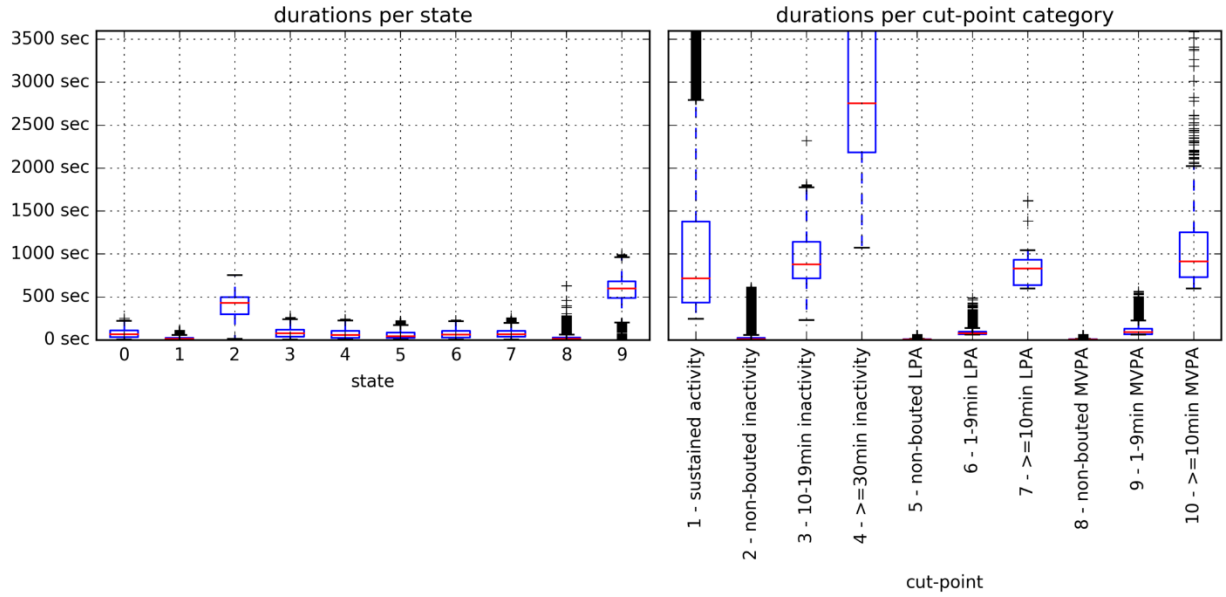

Fig 5: Durations for each HSMM state (acceleration+angles model) (left) and each cut-points category (right)

3.3 Comparison against time use diary

To assess whether there is a relation between the states and activities reported in the time use diary, we focused on the 10 most common activities reported in the time use diary, see Table 2 for a comparison with the *acceleration+angles* model (the same table for the *acceleration* model is available as table S4 in the supplement). The sustained activity states {<9>, <2>} are mostly present during sleep. Strongly sedentary activities, such as "Watching tv .." and "Playing electronic games.." have most time steps in the inactivate states {<0>,<3>,<4>,<6>}, which is consistent with our comparisons against the cut-points approach. States 5, 1, 7 present more time in activities such as speaking or eating a meal. State 8 appears as a mix of short time in different activities.

Table 2: Average time spent (minutes) per participant per day in each state (acceleration+angles model) and the top 10 activities

	state										
	9	**2**	**0**	**3**	**4**	**6**	**5**	**1**	**7**	**8**	**total**
Sleeping and resting (including sick in bed)	132.2	81.3	27.2	30.8	41.7	45.8	17.9	21.8	26.6	4.7	429.9
In class	0.9	1.5	8.0	6.2	10.0	20.5	8.0	19.9	18.6	6.4	99.9
Watch TV, DVDs, downloaded videos	4.4	4.4	12.7	11.3	8.8	12.8	7.2	12.6	11.7	3.4	89.4
Speaking, socialising face-to-face	1.5	1.0	7.7	4.1	6.8	7.0	4.0	10.5	10.4	3.6	56.4
Personal care (including taking a shower/bath, grooming, getting dressed etc.)	2.9	2.7	5.2	3.6	5.3	5.8	4.3	8.3	8.4	2.9	49.4
Eating a meal	1.0	1.2	5.6	4.1	5.9	6.7	4.0	8.5	8.6	3.0	48.6
Playing electronic games and Apps	1.7	1.4	6.6	4.6	6.3	8.4	3.4	6.8	6.7	1.7	47.6
Other activities not listed	1.2	1.0	3.0	2.0	3.5	3.7	2.8	6.0	6.8	2.3	32.2
Homework	1.2	1.0	3.8	2.5	3.6	4.8	2.3	4.8	4.3	1.3	29.6
Travel by car, van (including vehicles owned by friends and family)	0.5	0.6	2.5	1.6	3.4	3.9	2.7	5.7	5.6	2.1	28.6

3.4 Reproducibility

The states of the *acceleration* model trained on a subset of 250 participants were sorted on acceleration mean and then matched with the states from the model trained on the full data set of 500 participants. The model parameters (acceleration mean and variance, duration mean and variance) are plotted in Figs 6 and 7. The Kullback-Leibler (KL) divergence for the acceleration distributions is below 1.0 for all state combinations except the two states with small durations. The duration distributions are less consistent over the two models, with 4 out of 10 state combinations having a KL divergence of larger than 1. The calculated KL divergences are listed in the supplement in Table S5.

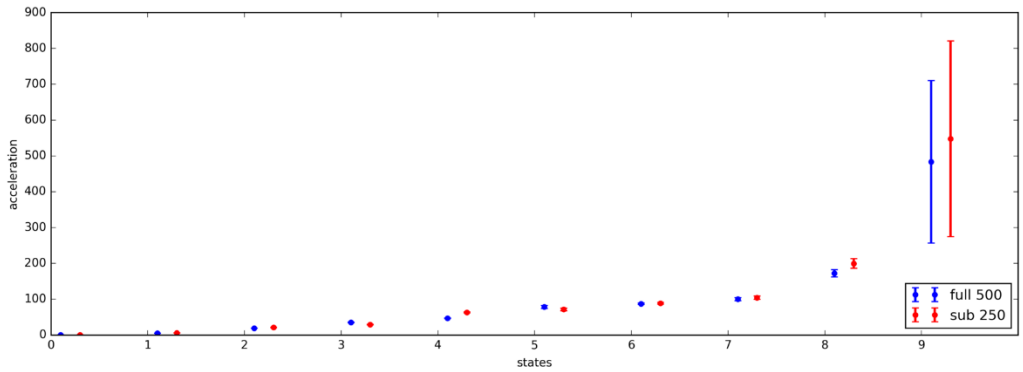

Fig 6: Mean and variance parameters for acceleration

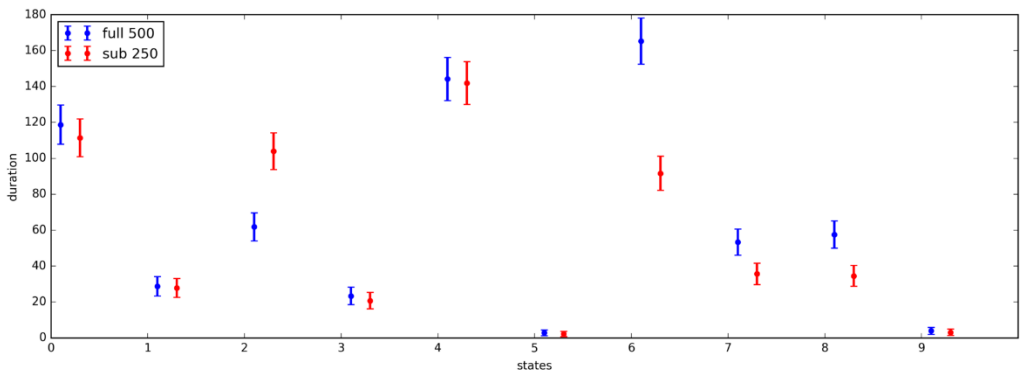

Fig 7: Lambda and variance of parameters of the duration distribution

4 Discussion

This study proposes an unsupervised Hidden semi-Markov Model (HSMM) to segment and cluster data from wearable accelerometers. The HSMM is a possible alternative for the widely-used cut-points approach as it does not require resource expensive calibration studies. The increased use of raw data accelerometry in recent years has resulted also in the need for a new calibration study for every population and every new acceleration processing metric to be used when relying on the cut-points approach [28–33]. Furthermore, these calibration studies are typically limited to small numbers of participants and activity types, thus generalization to different populations and conditions is a known problem [34]. The HSMM approach, examined in the present study, learns 'states' from the data, which are described by the mean and variance of the observations (accelerometer derived time series) and by the lambda of their Poisson distributed duration.

Our findings show that the HSMM derived states were related to cut-points categories. For example, the HSMM found short lasting states with high acceleration and long lasting states with low acceleration, which is consistent with data derived from cut-points approaches. Further, when states and cut-points categories are grouped by acceleration level, correlations of 0.56 and higher are observed between the two approaches. The mean acceleration for some of the states was close to the threshold value used in the cut-points approach, which indicates disagreement on the range or distribution of acceleration for typical behaviors. However, this may well be explained by the fact that the HSMM is driven by the distributions in observations (accelerometer recording) from data collected outside a laboratory in the daily life of British teenagers, while the cut-points approach in our case was driven by cut-points derived from Norwegian children and adults performing a very specific set of activity types in a laboratory

setting. Consequently, it may not be surprising that no perfect agreement is found. More importantly, the HSMM covers a plausible range of acceleration levels (low, medium, and high), durations (from less than a minute to more than 30 minutes), and to some extent, although difficult to interpret, angle ranges. Further, it was reassuring to observe that the principal component analyses applied to time spent in states has a less steep scree plot compared to time use variables based on the cut-points approach. An important strength of the HSMM is that it can account for multi-variate input, even if no prior theory exists on why or how the additional input variables could contribute. In contrast, the cut-points approach needs such a theory [35]. Our PCA results indicates that adding the angles offers a higher dimensional description of physical activity compared to using acceleration only.

The HSMM approach allows us to move towards a description of physical behavior based on, and driven by, the accelerometer data that can feasibly be collected in large scale populations. The HSMM is not biased by the subjective nature of self-report methods, avoids the complexities of accounting for inter-individual variation in body composition in energy expenditure estimation and the variation in the relationship between body composition and energy expenditure between activity types, and avoids the difficulties with generalizability of supervised learning techniques that rely on training data composed of small numbers of participants and/or activity types. The HSMM approach will not directly fit into the research framework aimed at providing public recommendation on easy to grasp constructs like steps or minutes of moderate to vigorous physical activity per day. However, HSMM may speed up and facilitate a data driven approach that could help to understand how variations in observed acceleration and angle relate to health and disease.

Strengths and limitations

For a few activities in the time use diary we saw an agreement with the states. However, the interpretation of a possible relationship was complicated by the broad description of the activity classes in the time use diary. Further studies might investigate this relationship with direct observation or more specific time use diaries. Studies that compare HSMM states to observed or reported activities or energy levels can guide interpretation of the states and provide validation for the hypothesis of a relationship between states and activity types. The states found by the HSMM are based on patterns in the data and are not specific for the research question of a user. To use the HSMM states as physical behavior descriptors in further research, it might be good practice to undertake post-processing of the data. This includes, for example, grouping states that have similar characteristics regarding the research question (e.g. similar acceleration level), or using the majority state for a larger window to adjust for the desired time granularity. In practice, it seems not to be feasible to let the model converge to very consistent state assignment (e.g. < 1% Hamming distance). It is not clear from theory how much of this is attributed to variance in the data, and how much could be gained with more training iterations. Therefore, future research (both empirical and theoretical) is needed to investigate the relationship between data size, population characteristics, and convergence.

Future research is needed to better understand the application of HSMMs on physical activity data. For example, the number of states can be varied to optimize face validity, while retaining interpretability and feasibility in terms of training time. The same holds for the size of the input time frame: instead of 5-second time frames, smaller or larger time frames can be used as input. It is also possible to include more input metrics in the model. The use of different distributions to

represent the data in the HSMM model could be investigated, such as a log-normal distribution for the acceleration metric.

The use of metrics that describe the orientation of the device imposes challenges. Firstly, the interpretation of the distributions of the angle values is difficult, asking for visualization and comparison to specific activities to distinguish between states with different angle distributions. Secondly, the correct wear position of the device becomes crucial. In this work, we took a heuristic approach at correcting for improper worn devices. Future research is needed to build a reliable classifier that determines the wear position if signal features are used that are body side dependent [36].

Since computation time increases with the amount of data used to train the HSMM method, we chose a random subset of 500 participants in this study. The reproducibility experiment suggests that the model for a smaller subset of 250 approaches the model trained on the data of all 500 participants, except for the rarest states. This provides us with confidence that the model generalizes well to more data from the same target group. However, the question remains how much the model generalizes to other populations, e.g. age groups or countries. If enough data are available, it is also possible to train the model for a specific person; this would however make it more difficult to relate the resulting states among the participants. This is a problem of finding the right balance between a model optimized in specific population, and a model that representative for the general population.

We conclude that applying Hidden Semi-Markov Models results in informative states, based on the data from a real-world setting. It is possible to relate the states to conventional cut-points categories, to interpret the meaning of the states. The unsupervised model can easily incorporate

multiple input metrics, so that the states provide a higher dimensional description of physical behavior.

5 References

1. Menai M, van Hees VT, Elbaz A, Kivimaki M, Singh-Manoux A, Sabia S. Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: results from the Whitehall II study. Sci Rep. 2017 Apr;8:45772.
2. Diaz KM, Krupka DJ, Chang MJ, Kronish IM, Moise N, Goldsmith J, et al. Wrist-based cut-points for moderate- and vigorous-intensity physical activity for the Actical accelerometer in adults. J Sports Sci. 2017 Feb;1–7.
3. McGarty AM, Penpraze V, Melville CA. Calibration and Cross-Validation of the ActiGraph wGT3X+ Accelerometer for the Estimation of Physical Activity Intensity in Children with Intellectual Disabilities. PLoS One. 2016;11(10):e0164928.
4. Veltink PH, Bussmann HJ, de Vries W, Martens WJ, Van Lummel RC. Detection of static and dynamic activities using uniaxial\naccelerometers. IEEE Trans Rehabil Eng [Internet]. 1996;4(4):375–85. Available from: http://ieeexplore.ieee.org/ielx4/86/11962/00547939.pdf?tp=&arnumber=547939&isnumber=11962
5. van Hees VT, Sabia S, Anderson KN, Denton SJ, Oliver J, Catt M, et al. A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. PLoS One. 2015;10(11):e0142533.
6. Rowlands A V, Yates T, Olds TS, Davies M, Khunti K, Edwardson CL. Sedentary Sphere: Wrist-Worn Accelerometer-Brand Independent Posture Classification. Med Sci Sports Exerc [Internet]. 2016 Apr [cited 2017 Mar 7];48(4):748–54. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/26559451

7. Rosenberg D, Godbole S, Ellis K, Lacroix A, Natarajan L, Kerr J. Classifiers for Accelerometer-Measured Behaviors in Older Women. Med Sci Sport Exerc. 2017;49(3):0–0.

8. Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas. 2014 Nov;35(11):2191–203.

9. Lyden K, Keadle SK, Staudenmayer J, Freedson PS. A method to estimate free-living active and sedentary behavior from an accelerometer. Med Sci Sports Exerc. 2014 Feb;46(2):386–97.

10. Freedson PS, Lyden K, Kozey-Keadle S, Staudenmayer J. Evaluation of artificial neural network algorithms for predicting METs and activity type from accelerometer data: validation on an independent sample. J Appl Physiol. 2011 Dec;111(6):1804–12.

11. Lin JF-S, Kulić D. Automatic Human Motion Segmentation and Identification using Feature Guided HMM for Physical Rehabilitation Exercises. Work Robot Neurol Rehabil IEEE Int Conf Intell Robot Syst. 2011;33–6.

12. Pober DM, Staudenmayer J, Raphael C, Freedson PS. Development of novel techniques to classify physical activity mode using accelerometers. Med Sci Sport Exerc. 2006;38(9):1626–34.

13. Duong T V., Bui HH, Phung DQ, Venkatesh S. Activity Recognition and Abnormality Detection with. Proc 2005 IEEE Comput Soc Conf Comput Vis Pattern Recognition, IEEE, Washington, D C. 2005;838–45.

14. Van Kasteren TLM, Englebienne G, Kröse BJA. Activity recognition using semi-Markov

models on real world smart home datasets. J Ambient Intell Smart Environ. 2010;2(3):311–25.

15. Ipsos MORI; CLS. Millennium Cohort Study Sixth Sweep (MCS6): Time Use Diary Documentation. 2016.

16. Chatzitheochari S, Fisher K, Gilbert E, Calderwood L, Cleary A, Chatzitheochari S, et al. Measuring young people ' s time-use in the UK Millennium Cohort Study : A mixed-mode time diary approach A mixed-mode time diary approach. 2015;

17. Ipsos MORI, CLS. Millennium Cohort Study Sixth Sweep: Technical report (Version2). Vol. 2017. 2017.

18. van Hees V, Fang Z, Zhao JH, Heywood J, Sabia S. package "GGIR" [Internet]. 2017 [cited 2017 Mar 24]. Available from: https://cran.r-project.org/web/packages/GGIR/GGIR.pdf

19. van Hees VT, Fang Z, Langford J, Assah F, Mohammad A, da Silva IC, et al. Auto-calibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents. J Appl Physiol. 2014 Aug;

20. Sabia S, van Hees VT, Shipley MJ, Trenell MI, Hagger-Johnson G, Elbaz A, et al. Association Between Questionnaire- and Accelerometer-Assessed Physical Activity: The Role of Sociodemographic Factors. Am J Epidemiol. 2014 Feb;

21. van Hees VT, Gorzelniak L, Dean León EC, Eder M, Pias M, Taherian S, et al. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. PLoS One. 2013 Jan;8(4):e61691.

22. Hildebrand M, Van Hees VT, Hansen BH, Ekelund U. Age-Group Comparability of Raw Accelerometer Output from Wrist- and Hip-Worn Monitors. Med Sci Sport Exerc. 2014

Feb;(accepted for publication 2014):1.

23. Bellettiere J, Winkler EAH, Chastin SFM, Kerr J, Owen N, Dunstan DW, et al. Associations of sitting accumulation patterns with cardio-metabolic risk biomarkers in Australian adults. PLoS One. 2017;12(6):e0180119.

24. van Kuppevelt D, van Hees V. hsmm4acc [Internet]. 2017. Available from: https://github.com/wadpac/hsmm4acc

25. van Hees V, van Kuppevelt D. millenniumcohort-acc. 2017; Available from: https://github.com/wadpac/millenniumcohort-acc

26. van Hees VT, Renström F, Wright A, Gradmark A, Catt M, Chen KY, et al. Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. PLoS One. 2011 Jan;6(7):e22922.

27. Kullback S, Leibler RA. On Information and Sufficiency. Ann Math Stat. 1951 Mar;22(1):79–86.

28. Bakrania K, Yates T, Rowlands A V, Esliger DW, Bunnewell S, Sanders J, et al. Intensity Thresholds on Raw Acceleration Data: Euclidean Norm Minus One (ENMO) and Mean Amplitude Deviation (MAD) Approaches. PLoS One. 2016;11(10):e0164045.

29. Vähä-Ypyä H, Vasankari T, Husu P, Mänttäri A, Vuorimaa T, Suni J, et al. Validation of Cut-Points for Evaluating the Intensity of Physical Activity with Accelerometry-Based Mean Amplitude Deviation (MAD). PLoS One. 2015;10(8):e0134813.

30. Esliger DW, Rowlands A V, Hurst TL, Catt M, Murray P, Eston RG. Validation of the GENEA Accelerometer. Med Sci Sports Exerc. 2011 Jun;43(6):1085–93.

31. Hildebrand M, VAN Hees VT, Hansen BH, Ekelund U. Age group comparability of raw accelerometer output from wrist- and hip-worn monitors. Med Sci Sports Exerc. 2014

Sep;46(9):1816–24.

32. Hildebrand M, Hansen BH, van Hees VT, Ekelund U. Evaluation of raw acceleration sedentary thresholds in children and adults. Scand J Med Sci Sports. 2016 Nov;

33. Phillips LRS, Parfitt G, Rowlands A V. Calibration of the GENEA accelerometer for assessment of physical activity intensity in children. J Sci Med Sport. 2013 Mar;16(2):124–8.

34. van Hees VT, Golubic R, Ekelund U, Brage S. Impact of study design on development and evaluation of an activity-type classifier. J Appl Physiol. 2013 Apr;114(8):1042–51.

35. Rowlands A V, Olds TS, Hillsdon M, Pulsford R, Hurst TL, Eston RG, et al. Assessing Sedentary Behavior with the GENEActiv: Introducing the Sedentary Sphere. Med Sci Sports Exerc. 2013 Nov;

36. Rowlands A V, Olds TS, Bakrania K, Stanley RM, Parfitt G, Eston RG, et al. Accelerometer wear-site detection: When one site does not suit all, all of the time. J Sci Med Sport. 2017 Apr;20(4):368–72.

6 List of supplements

S1: Appendix. Description of HSMM

S2: Appendix. Supplementary results

S3: Table. Average time spent (minutes) per participant per day in each state from the acceleration method (sorted by mean acceleration) and cut-points category

S4: Table. Average time spent (minutes) per participant per day in each state (acceleration method) and the top 10 activities, acceleration model

S5: Table. Kullback-Leibler divergences for the distributions on the full model and a subset model

S6: Table. Distribution parameters of the HSMM (acceleration) model, for each state

S7: Table. Distribution parameters of the HSMM (acceleration+angles) model for each state