tasks, Saria et al. [134] proposed a method that discovers in a time-series recurring sub sequences having similar shape, but that, at the same time, can exhibit significant variability (deformable motifs). Number and length of the motifs need to be specified beforehand and, if applied to multidimensional time series, the method assumes that motifs along all the dimensions are happening synchronously.

6.2.4 Probabilistic Unsupervised Modelling

A probabilistic approach for unsupervised mining of electronic health data has been introduced by Schulam et al. [135]. Time series of clinical markers were clustered taking into account confounding factors that might affect data. In line with the generative modelling of multidimensional time series outlined by the authors and increasing the abstraction level of the activity bouts, we might think of these constructs as primitive descriptors of PA that are latently coupled to each other. Organizing bouts in higher level and coherent structures that co-occur during a day, more informative PA constructs could be described (i.e., PA routines). Topic models suit this goal since they represent a class of algorithms able to discover hidden structures in collections of documents. Due to their pattern discovery capability, they have been explored in the wearable sensor and activity recognition community. Huynh et al. [136] showed that daily routines of activities can be recognized as a probabilistic combination of activity labels, such as walking, discussing at whiteboard, etc. They also addressed the point of avoiding supervised learning approaches by clustering raw sensor data in order to build the vocabulary of activity primitives. Their approach was tested only on data from one single subject, and the vocabulary of primitives had a fixed size chosen a priori. Seiter et al. [137] compared three topic model approaches and analysed three public datasets with different properties affecting the discovery, such as primitive rate, activity composite specificity, primitive sequence similarity, and composite-instance ratio. These authors compared the activity composite discovery performance against the performance of a k-means clustering algorithm providing guidelines for optimal parameter selection. Their results indicated that latent Dirichlet allocation (LDA) shows higher robustness against noise compared to k-means and other topic modelling approaches. The application of a nonparametric framework to create the vocabulary and discover human routines from sensor data was investigated using a Dirichlet process Gaussian mixture model (DPGMM) in [138]. Although this approach does not need to specify the number of unique artificial words, it assumes that data should come from the same distributions used to create the mixture model. In the case of populations with different activity behaviour as healthy subjects and COPD patients at different stages, this might not be true and the mapping of the raw sensor data to vocabulary words might not be correct.

6.3 Background

Topic models are algorithms for discovering the main themes that pervade a large and unstructured collection of documents. Data are treated as observations arising from a generative probabilistic process, in which hidden variables reflect the thematic structure of a collection of documents. The intuition behind using the LDA [139] to discover PA routines is that each day is a mixture of thematically coherent PA measures just as a text document is a mixture of thematically coherent words. The graphical model for LDA is provided in Figure 37. All the