assessed days (also called documents $d_{1:D}$) share the same set of daily routines (also called topics $\beta_{1:K}$) that are defined as Dirichlet distributions over the observed set of PA descriptors (also called words W or terms of a fixed vocabulary). Observing activities in patients is a difficult task since it is time intensive and intrusive. At the same time, patients are not able to accurately self-report their physical activities [126], and the training of a classifier requires annotations that in a daily life scenario are difficult to obtain. In order to make the methodology fully unsupervised, we assume that the observed words (input of the model) are composed by multimodal measures coming from a body area sensor network. Each assessed day exhibits PA daily routines in different proportion indicated as $\theta_{1:D}$, i.e., each day has a different distribution over the routines that also follows a Dirichlet distribution. In such a model, the N words ($W_{d,n1:N}$) that compose the D documents are the only random variables observed and depend on the per word topic assignment ($Z_{d,n}$) and all the β_k. The daily routines then are composed indirectly by low-level PA measures that belong, with a certain probability distribution, to different thematic areas. Different routines will have different PA measures with different probabilities. Topic modelling algorithms calculate the hidden structure that likely generated the observed collection of documents. In particular, for this analysis, we use variational inference to approximate the intractable posterior distribution over hidden variables defined by LDA. In a nutshell, variational inference posits a parametrized family of distributions over the hidden structure, and then, finds the member of that family that is closest to the posterior according to the Kullback–Leibler divergence. The reader is invited to refer to [139] for an exhaustive explanation.

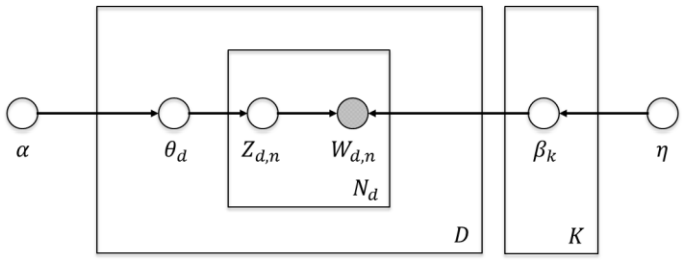

Figure 37 Graphical model for LDA. Each node is a random variable, edges denote possible dependences. The only observed variables (shaded) are the words (W). The distribution of the words in a routine (β) and the distribution of the routine in a document (θ) depend only on the topic hyperparameters η and α that control the mean shape and sparsity of the distributions. Z represents the word topic assignment.

6.4 Methods

In the application of LDA, a word, defined to be an instance of a vocabulary, is considered as the basic unit of discrete data. In this work, multimodal monitoring signals, composed of PA measures and physiological responses, are symbolized first as letters. Our approach in selecting the letters and then the words composing the vocabulary benefits from a methodology that preserves the interpretability of the vocabulary and that allows the generation of words that could also not appear in the current documents. The multidimensional data space is first divided in subspaces, and for each subspace, we use cluster analysis to divide PA measures and physiological responses into discrete clusters levels (letters) that are then combined to form words. We use soft assignments to link an intensity bout (IB) with the words in the vocabulary