sensor measures that belong, with a certain probability distribution, to different thematic areas. Our hypothesis is that nights related to different group of subjects would have a different distribution over the sleep modalities that in turn would be composed by different distributions over symbolic words defined combining discretized sensor measurements. The proposed methodology will first extract all the β_k in a data driven fashion using data from a subset of COPD patients and healthy subjects and then, for each assessed night, it will calculate a probabilistic feature vector ϑ composed by the histogram of activation probabilities of all the sleep modalities $\beta_{1:K}$. These probabilistic features will be used to classify a large cohort of patients.

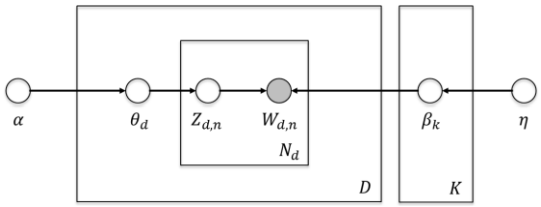

Figure 43 Graphical model for LDA. Each node is a random variable, edges denote possible dependences. The only observed variables (shaded) are the symbols (W). The distribution of the symbols in a sleep modality (β) and the distribution of the sleep modalities during a night (θ) depend only on the sleep modalities hyperparameters η and α that control the mean shape and sparsity of the distributions. Z represents the symbol sleep modality assignment.

7.4 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 such us metabolic equivalent of task (MET) [76], temperature, galvanic skin response and number of steps were first converted into a discrete alphabet of letters and then combined into symbolic words. 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 symbols that actually do not occur in the current documents. The multidimensional data space is first divided in subspaces according to the METs values in order to conveniently define a repertoire of physical activities in which a person may participate [76]. In each subspace, we divided temperature and galvanic skin response data into partitions (letters) that are then combined to form string of symbolic words. We extract the vocabulary in a subset of COPD and healthy patients and show that the constructed vocabulary is able to model the data of a much larger cohort of patients. The dataset and the methodology developed are described in detail in the following sections.

7.4.1 Participants

Data from 1384 patients from ten countries (United Kingdom, Ireland, The Netherlands, Germany, Switzerland, Italy, Spain, The United States of America, Brazil, and Australia) diagnosed with mild to very severe COPD were pooled from previous studies (references can be found in the appendix) and considered for analysis. Participants were included if they had COPD with a post-bronchodilator ratio of forced expiratory volume in the first second (FEV_1) to forced vital