Interpretation and Further Development of the Hypnodensity Representation of Sleep Structure

Iris A. M. Huijben[1,2], **Lieke W. A. Hermans**[1], **Alessandro C. Rossi**[2], **Sebastiaan Overeem**[1,3], **Merel M. van Gilst**[1,3], **and Ruud J. G. van Sloun**[1]

[1] Dept. of Electrical Engineering, Eindhoven University of Technology, 5612 AP Eindhoven, The Netherlands
[2] Onera Health, 5617 BD Eindhoven, The Netherlands
[3] Sleep Medicine Center Kempenhaeghe, 5591 VE Heeze, The Netherlands

E-mail: `i.a.m.huijben@tue.nl`

April 2022

Abstract. *Objective:* Data acquired during a sleep recording is typically compressed into a *hypnogram*; a visual representation of manually annotated sleep stages over the night. Recently, a richer *hypnodensity* representation was proposed that provides a probability distribution over these stages at each point in time. In this work we investigate how to interpret a hypnodensity plot, and reveal its implicit assumptions. We, moreover, seek alternative representations to acquire additional information about continuities in the sleeping brain. *Approach:* We recap softmax classification theory, and empirically validate the interpretation of a hypnodensity plot. Unsupervised learning and the non-linear softmax activation are studied to find representations that are less dependent on the manual sleep staging decision process. Experiments are performed both in a synthetic setup, and on sleep recordings. *Main results:* A hypnodensity plot, predicted by a supervised classifier, represents the probability with which the sleep expert assigned a label to an epoch. It thus reflects annotator behaviour, and is thereby only indirectly linked to underlying continuous dynamics of the brain. Unsupervised training was shown to result in hypnodensity plots that were less dependent on this annotation process. Moreover, pre-softmax predictions were found to better reflect continuous brain dynamics than the post-softmax counterparts (i.e. the hypnodensity plot). *Significance:* This study provides insights in, and proposes new, representations of sleep that may enhance our comprehension about sleep and sleep disorders.

Keywords: Sleep, Hypnogram, Hypnodensity, Supervised Learning, Contrastive Predictive Coding, Softmax

1. Introduction

Even though we spend a large part of our lives asleep, there is only a marginal understanding about the processes that happen in our brain during the night. Until the late thirties of the previous century, it was widely believed that sleep is a passive state of the body [1], opposed to the active state of wakefulness. The discovery of the electroencephalogram (EEG), a recording that measures electrical activity of the brain via electrodes on the scalp, has been a fundamental step. Quickly after its discovery, patterns in the sleeping brain were described, which still form the basis for the way we describe sleep nowadays.

The current standard for sleep analysis comes from the American Academy of Sleep Medicine (AASM) [2], which recommends a polysomnography (PSG) measurement that comprises among others EEG, electromyography (EMG), and electrooculography (EOG). Five different states have been distinguished, through which a sleeping brain transitions (multiple times) during the night: rapid eye movement (REM) sleep, non-REM sleep (subdivided into N1, N2, and N3), and wakefulness (W). Given a PSG recording, a sleep expert manually labels each non-overlapping 30-second window with one of the five discrete states to create a *hypnogram*; a visual representation of assigned sleep stages over the full night.

While a hypnogram has proven clinical utility, it is a strongly compressed representation, which can only represent abrupt sleep stage switches. In reality, it is, however, to be expected that transitioning between two states yields a more gradual pattern and does not happen at boundaries of pre-defined data windows. As such, an alternative representation for sleep data was recently proposed, which the authors called a *hypnodensity* plot [3]. Rather than selecting one sleep stage for each window, this hypnodensity representation reveals a probability distribution over the five AASM stages (see fig. 8 for an example), which may be provided at any temporal resolution. A hypnodensity plot could give insights in (yet) unexplained phenomena, and therefore has the potential to induce a paradigm shift in sleep medicine. It was, for example, already shown to contain discriminative patterns for patients with narcolepsy [3].

Despite the clinical possibilities, hypnodensity plots have not yet been used in clinical practice. We suspect one important aspect to be the major reason: while a hypnogram is typically created by a sleep expert that follows the AASM guidelines, a hypnodensity plot is generally predicted using a computer model. As a consequence, the exact relation between recorded data and the predicted probability distributions (i.e. the hypnodensity plot) remained unclear so far. Questions that now arise are: *Is a hypnodensity representation, predicted by a supervised classifier, revealing continuous dynamics of the sleeping brain, or does it mainly show model uncertainty?*, and *What are the implicit assumptions and driving factors that play a role in creating such a hypnodensity plot?* In case it reflects model uncertainty, *Can we find alternative ways that reveal the continuous brain dynamics?*

The machine learning model that predicts a hypnodensity plot, as proposed by the original authors [3], is a supervised neural classifier, of which the final softmax-activated outputs are considered the hypnodensity representation. Earlier efforts from the machine learning community have already provided valuable insights about interpretation of such softmax probabilities [4, 5]; these probabilities are known to reflect a probability distribution that coincides with the (expectation of a) decision process of assigning one of the possible labels to a data point. We thus hypothesize that a hypnodensity plot that is predicted by a supervised model, reflects the probability with which the data point (or PSG window in this case) was assigned with a certain label (i.e. AASM sleep stage) by the expert(s) that annotated the data set. We wonder whether this distribution, as well, reflects continuous dynamics of the sleeping brain.

The hypnodensity representation [3] was, among other, suggested to be used for research about disorders that are related to sleep stage dissocations, or local sleep phenoma [6]. It was, e.g., shown to contain more moments with probability mass spread over several sleep stages for narcoleptic patients, as compared to non-narcoleptic controls. Narcoleptic patients are indeed known to exhibit sleep/wake dissociations, however, given what is known about softmax-activated predictions of supervised models, it can be questioned whether the difference in hypnodensity plots between the two groups was truly reflecting underlying differences in brain dynamics, or whether it mainly reflected a difference in decision processes for manual sleep stage scoring between both groups [7].

This study will contribute in finding answers to the aforementioned questions. The contributions can be summarized as follows:

- We formulate a signal- and an annotation model (Section 2) to study the relation between recordings and annotated AASM stages (i.e. the hypnogram), and its implications for the hypnodensity representation.
- We generate a PSG-inspired synthetic dataset to experimentally investigate the relation between recordings and predictions of hypnodensity plots, as a function of training strategy (supervised vs unsupervised) and final non-linearity (pre- vs post-softmax predictions) (Section 4).
- We validate drawn conclusions from the synthetic experiments on real PSG recordings of healthy sleepers. To this end, we compare hypnodensity plots predicted under the same varying circumstances as in the synthetic experiments, and validate that the effects of these factors are similar on real data (Section 5).

2. Problem formulation and modelling

The authors of [3] propose to use the softmax-activated outputs of a supervised neural classifier that is trained on PSG data with expert's annotations, to acquire a hypnodensity plot. In this section we discuss the interpretation of predictions from such a classifier in a synthetic setup that is inspired by PSG recordings. To this end, we first introduce a generative signal model that generates (heavily simplified) PSG-like data (Section 2.1), and an annotation model (Section 2.2) that can capture the conversion from PSG data to an expert's annotation (i.e. a selected sleep stage). The signal and annotation model are visually summarized in fig. 1. Section 2.3 subsequently provides information on the hypnodensity-predicting model and its optimization.

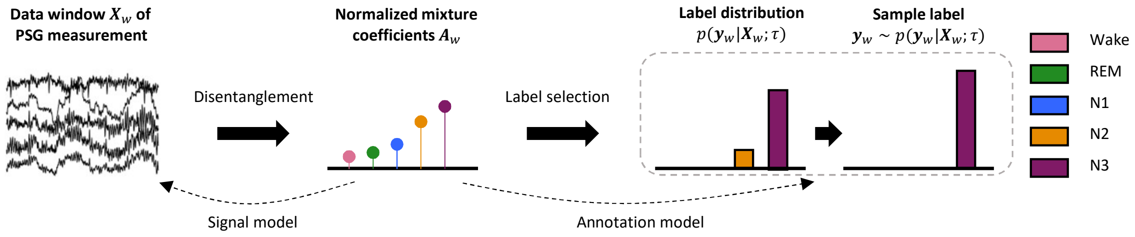

Data window X_w of PSG measurement
Disentanglement
Normalized mixture coefficients A_w
Label selection
Label distribution $p(y_w|X_w;\tau)$
Sample label $y_w \sim p(y_w|X_w;\tau)$
Wake
REM
N1
N2
N3
Signal model
Annotation model

Figure 1: We distinguish a signal and annotation model. The signal model assumes that all channels are a non-linear mixture of some implicit (or hidden) classes that each characterize one of the sleep stages. The annotation model converts the contribution of each of these classes to one selected sleep stage. Due to uncertainty in expert annotations, we know that scoring is a stochastic process, which may be modelled as a conditional label distribution per window.

2.1. Signal model

A typical PSG recording $\mathbf{X}^{(k)} \in \mathbb{R}^{ch \times W \times l}$, with index k, contains time series of $L = W \times l$ samples (W number of 30-second windows, each of l samples) from ch number of channels (e.g. multiple EEG channels, EMG, and EOG). We model the data generation/measurement as a non-linear generative mixing process of C latent signals $\boldsymbol{s}_c^{(k)}$, with $1 \leq c \leq C$, where each signal aggregates typical characteristics associated with a specific sleep stage (or class).

In other words, each PSG window is assumed to contain data which are a non-linear spatial and temporal (over 30 s) accumulation of characteristics that are typical for certain sleep stages. Given five AASM-defined sleep stages, we may hypothesize that these data are thus generated from $C = 5$ latent signals that each represent one sleep stage. When manually scoring a PSG window, an expert (implicitly) determines how much of the characteristics belong to either of the five stages, and based on some rules (i.e. the AASM standard), determines the final sleep stage (more on this in Section 2.2).

The amplitudes $\tilde{\boldsymbol{a}}_c^{(k)}$ of the latent signals are modelled to vary over time, i.e. characteristics belonging to a certain sleep stage can be fully absent in some moments, while present (with a certain amount) at other moments. The resulting signal model yields:

$$\mathbf{X}^{(k)} = h\Big(\tilde{\mathbf{A}}^{(k)} * \mathbf{S}^{(k)}\Big), \tag{1}$$

where $\mathbf{S}^{(k)} \in \mathbb{R}^{C \times W \times l}$ contains the signals of all classes, $\tilde{\mathbf{A}}^{(k)} \in \mathbb{R}_{\geq 0}^{C \times W \times l}$ contains the corresponding time-varying amplitudes, $h : \mathbb{R}^{C \times W \times l} \rightarrow \mathbb{R}^{ch \times W \times l}$ is a non-linear spatial mixing function, and $*$ denotes an element-wise multiplication. The normalized amplitudes, that sum to one over the C classes at every moment in time, are denoted with $\mathbf{A}^{(k)}$. In the context of non-linear mixing, these normalized amplitudes are also called *mixture coefficients*. Figure 1 depicts the described signal model, and table A1 in Appendix A provides a summary of all introduced notations and symbols.

2.2. Annotation model

Sleep stage annotations are in clinical practice assigned to 30 seconds of data. We denote the w^{th} 30-second data window with $\mathbf{X}_w^{(k)} \in \mathbb{R}^{ch \times l}$, which is of length $l = 30 \times f_s$, with f_s the sampling frequency in Hz. Analogously, we define $\tilde{\mathbf{A}}_w^{(k)} \in \mathbb{R}_{\geq 0}^{C \times l}$ and $\mathbf{A}_w^{(k)} \in \{\mathbb{R}_{\geq 0}^{C \times l} : \sum_c \mathbf{A}_w^{(k)} = 1\}$, being the unnormalized, respectively normalized, amplitudes of the mixed signals in the window with index w.

A sleep expert assigns a label $\boldsymbol{y}_w^{(k)}$ to a PSG window by means of an (internal) decision process. Despite the aim of the AASM rules to standardize this process, both inter- and intra-rater variability exist [8], which can be explained by the stochastic nature of human decision making. To model this stochastic decision process, we model each label as a sample from a probability distribution over sleep stages, that is conditioned upon the mixture coefficients of the characteristics belonging to these stages. Omitting the (k)-superscript for readability, this conditional distribution - serving as a *generative label distribution* - yields:

$$p(\boldsymbol{y}_w|\mathbf{X}_w;\tau) = \sigma_\tau\{\log \operatorname{avg}_l(\mathbf{A}_w)\} \propto \exp\{\frac{\log \operatorname{avg}_l(\mathbf{A}_w)}{\tau}\}, \tag{2}$$

where σ_τ denotes a tempered softmax function with temperature parameter $\tau \in \mathbb{R}_{\geq 0}$, and $\operatorname{avg}_l(\cdot)$ returns the average over l samples. In the following, we use the one-hot embedding of labels, and therefore redefine the domain of a label to: $\boldsymbol{y}_w^{(k)} \in \{0,1\}^C$, with $|\boldsymbol{y}_w^{(k)}| = 1$. Figure 1 depicts the described signal and annotation models.

For $\tau = 1$, the selection of a class is linearly related to the mixture coefficients of each class. On the other hand, when $\tau \to 0^+$, the distribution becomes degenerate (i.e. one-hot) and the 'sampling' process becomes fully deterministic. This models the (unrealistic) scenario where experts would always make the same decision, and inter- and intra-rater variability does not exist.

For $0 < \tau < 1$, the distribution's entropy is lowered (compared to $\tau = 1$), and classes with a high mixture coefficient are selected with a higher probability than denoted by their contribution to the mixture, while classes with lower mixture coefficients are selected with a lower probability.

This latter setting (i.e. $0 < \tau < 1$) models sleep staging according to the AASM standard, in which non-linear decision boundaries are used. For example, when a K-complex is detected, the window should in any case be classified as N2, even if only, say, 60% of the window shows characteristics that belong to N2. Similarly, if at least half of the window shows Wake-like characteristics, the window should be assigned the Wake label.

Note that in practice, an expert selects a sleep stage directly given the raw data. Though, the processes of *disentangling* the raw data into characteristics that describe various sleep stages and *selecting* the most appropriate sleep stage, can be considered an implicit processes that takes place during decision making.

2.3. Hypnodensity-predicting neural network

The authors of [3] propose to use a feedforward supervised neural classifier to predict a hypnodensity plot from PSG data. To this end, the classifier model p_{m}, parameterized by θ, makes a conditional prediction of class probabilities: $\hat{\boldsymbol{y}}_w^{(k)} \in \{\mathbb{R}_{\geq 0}^C : |\hat{\boldsymbol{y}}_w^{(k)}| = 1\}$, given some input data $\mathbf{X}_w^{(k)}$. Model parameters θ are optimized by maximizing the log-likelihood of the expert labels, using a training set of $(\mathbf{X}_w^{(k)}, \boldsymbol{y}_w^{(k)})$-pairs that approximate the data-generating distribution $p_{\mathrm{d}}(\mathbf{X}, \boldsymbol{y})$. The optimization problem yields (omitting all k- and w-super/subscripts for clarity):

$$\theta^* = \underset{\theta}{argmax}\{\mathbb{E}_{\hat{p}_{\mathrm{d}}(\mathbf{X},\boldsymbol{y})} \log p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\theta)\}, \tag{3}$$

where $\hat{p}_{\mathrm{d}}(\mathbf{X}, \boldsymbol{y})$ is the approximation of the true data-generating distribution.

We design the hypnodensity-predicting model as a feedforward neural network that comprises a convolutional encoder, and a a non-linear classifier, similar to the model proposed by [3]. The convolutional encoder converts a data window $\mathbf{X}_w^{(k)}$ to a latent representation: $\boldsymbol{z}_w^{(k)} = \mathrm{Enc}\left(\mathbf{X}_w^{(k)}\right) \in \mathbb{R}^F$, with F the number of features in the resulting embedding.

A standard multi-class classification model subsequently maps each embedding to class predictions between 0 and 1, with a total sum of 1 over the classes. It takes the form $\hat{\boldsymbol{y}}_w^{(k)} = \sigma\big(\mathbf{W}\boldsymbol{z}_w^{(k)} + \boldsymbol{b}\big)$, with trainable parameters $\mathbf{W} \in \mathbb{R}^{C\times F}$, and $\boldsymbol{b} \in \mathbb{R}^C$, and σ the softmax function. In case of having a classification goal (i.e. when aiming for an automated sleep stage classifier), the largest entry of the softmax outputs is conventionally selected. In contrast, the authors of [3] propose to omit this last step, and directly use the softmax output $\hat{\boldsymbol{y}}_w^{(k)}$, being the predicted hypnodensity plot of recording k for window w (i.e. $\hat{\boldsymbol{y}}^{(k)}$ entails the full hypnodensity plot belonging to recording k). Appendix B provides more details regarding the model architecture and training procedure.

3. Theoretical background

In this section we provide theoretical background on likelihood maximization of supervised neural classifiers. The optimization problem, as given in eq. (3), can be rewritten using the monotonicity and translation invariance of the argmax and argmin-functions:

$$\begin{aligned}\theta^* = \; & \underset{\theta}{argmax}\{\mathbb{E}_{\hat{p}_{\mathrm{d}}(\mathbf{X},\boldsymbol{y})} \log p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\theta)\} = \\ & \underset{\theta}{argmin}\big\{ -\{\mathbb{E}_{\hat{p}_{\mathrm{d}}(\mathbf{X},\boldsymbol{y})}[\log p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\theta)]\big\} = \\ & \underset{\theta}{argmin}\big\{\mathbb{E}_{\hat{p}_{\mathrm{d}}(\mathbf{X},\boldsymbol{y})}[\log \hat{p}_{\mathrm{d}}(\boldsymbol{y}|\mathbf{X}) - \log p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\theta)]\big\} = \\ & \underset{\theta}{argmin}\big\{\,\mathrm{D}_{\mathrm{KL}}\left(\hat{p}_{\mathrm{d}}(\boldsymbol{y}|\mathbf{X})||p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\theta)\right)\big\}. \end{aligned} \tag{4}$$

From the above equalities it can be seen that training a supervised classifier by maximizing the log-likelihood of the expert annotations, is equivalent to minimizing the Kullback-Leibler (KL)-divergence between the empirical conditional data distribution $\hat{p}_{\mathrm{d}}(\boldsymbol{y}|\mathbf{X})$ and the conditional distribution as trained by the model $p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\theta)$ [5, ch. 5].

The KL-divergence between two discrete probability distributions P and Q, both with C classes, is defined as follows:

$$\mathrm{D}_{\mathrm{KL}}\left(P||Q\right) = \sum_{c=1}^{C} P_c \log \frac{P_c}{Q_c}, \tag{5}$$

and is minimized when both distributions perfectly match. In other words, the probabilistic predictions of the supervised model mimic the conditional probability over the classes, as defined in the data set used for training the model.

The above statement only holds under the assumptions of having independent data points, and using a model that has enough capacity to minimize the aforementioned KL-divergence. On the other hand, when designing a model with too much capacity, overfitting happens and the KL-divergence is perfectly minimized, at the cost of generalizability to unseen data.

4. Synthetic experiments

This section describes the experiments on synthetically-generated PSG-like data. Their generation is described in Section 4.1. Section 4.2 covers the used methodologies, and results are discussed in Section 4.3.

4.1. Data generation

We created a synthetic dataset according to the signal model as introduced in eq. (1), and generated each channel in $\mathbf{X}^{(k)}$ as a non-linear combination of a set of ($C = 3$) independent classes, where each class represents a (fictitious) sleep stage. The signal corresponding to each class was generated as a (discretized) sinusoidal signal, with a class-dependent frequency, a random phase, and an amplitude that is described by a smoothened square wave, such that it smoothly varies between 0 and 1 over time. The varying amplitude thus represents the presence (with a certain amount) or absence of characteristics belonging to a class. We generated $K = 200$ random 'recordings', which

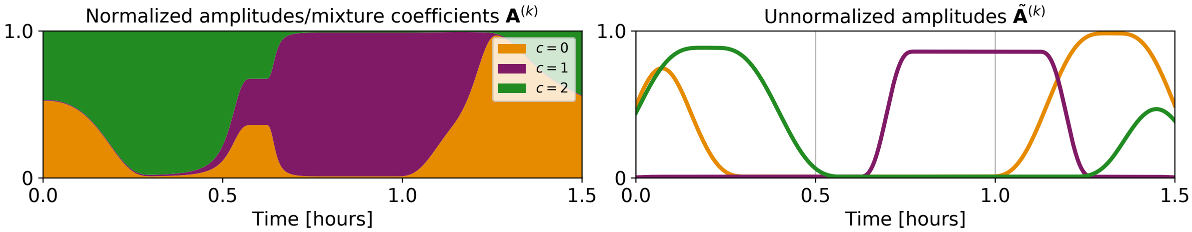


Figure 2: A sequence from the synthetic test set. Left: the normalized amplitudes that sum to one and serve as the mixture coefficients of the signals belonging to the three classes. Right: the corresponding unnormalized amplitudes of the three generated signals over time.

were split into a training, validation and a hold-out test set. Figure 2 shows an example from the test set, with normalized amplitudes (or mixture coefficients) on the left, and the corresponding unnormalized amplitudes on the right. Annotations were generated by sampling from the label distribution, as provided in eq. (2). Appendix C.1 provides more details about the generation of this dataset.

4.2. Methodology

In order to investigate the factors and assumptions that drive a hypnodensity representation, we compare pre- and post-softmax predictions, and unsupervised vs supervised training, for which the methodologies are discussed in Section 4.2.1 and 4.2.2, respectively.

4.2.1. Pre- vs post-softmax predictions The final activation function used in the hypnodensity-predicting network is the softmax function σ, which converts unconstrained predictions $\hat{\hat{\boldsymbol{y}}}_w \in \mathbb{R}^C$ to normalized probabilities $\hat{\boldsymbol{y}}_w \in \{\mathbb{R}^C_{\geq 0} : |\hat{\boldsymbol{y}}_w| = 1\}$:

$$\hat{\boldsymbol{y}}_w = \sigma(\hat{\hat{\boldsymbol{y}}}_w) = \frac{\exp \hat{\hat{\boldsymbol{y}}}_w}{\sum_c \exp \hat{\hat{\boldsymbol{y}}}_w}.$$

We investigate the effect of the non-linearity as introduced by using a softmax function as a final activation, by comparing pre-softmax predictions, to post-softmax (i.e. hypnodensity) predictions. If the (implicit) annotation model of real PSG data indeed follows a distribution close to the one as given in eq. (2), we deduce that the pre-softmax predictions would have more tendency than the post-softmax counterparts, to reveal the (unnormalized) contributions/amplitudes of characteristics in a window that belong to the different sleep stages.

Since the post-softmax predictions yield a (normalized) probability vector for each window w, we may use the KL-divergence (see eq. (5)) as a metric to compare this distribution with both the normalized amplitudes $\mathbf{A}^{(k)}_w$, and the label distribution $\hat{p}_\text{d}(\boldsymbol{y}^{(k)}_w | \mathbf{X}^{(k)}_w)$ used to generate corresponding labels for supervised training. In our synthetic setup, we explicitly defined this conditional label distribution according to eq. (2), thus $\hat{p}_{\mathrm{d}}(\boldsymbol{y}_w^{(k)}|\mathbf{X}_w^{(k)}) := p(\boldsymbol{y}_w^{(k)}|\mathbf{X}_w^{(k)};\tau)$. To ease notation in the results section, we define the following two metrics:

$$\mathrm{D}_{\mathrm{KL}}(\hat{p}_{\mathrm{d}}||\hat{\boldsymbol{y}}) := \frac{1}{K}\sum_{k=1}^{K}\mathrm{median}_w\left\{\mathrm{D}_{\mathrm{KL}}\big(\hat{p}_{\mathrm{d}}(\boldsymbol{y}_w^{(k)}|\mathbf{X}_w^{(k)})||\hat{\boldsymbol{y}}_w^{(k)}\big)\right\}, \tag{6}$$

$$\mathrm{D}_{\mathrm{KL}}(\mathbf{A}||\hat{\boldsymbol{y}}) := \frac{1}{K}\sum_{k=1}^{K}\mathrm{median}_w\left\{\mathrm{D}_{\mathrm{KL}}\big(\,\mathrm{avg}_l(\mathbf{A}_w^{(k)})||\hat{\boldsymbol{y}}_w^{(k)}\big)\right\}, \tag{7}$$

where median_w computes the median over the W windows. Due to the unnormalized nature of the pre-softmax predictions, KL-divergences can not be computed on these unnormalized vectors.

4.2.2. Supervised vs unsupervised encoding Additionally to comparing pre- and post-softmax predictions, we compare the fully supervised setting, where the model is trained using input-label pairs, to a setting in which the full encoder is trained in an unsupervised fashion. A supervised classifier (with its design as described in Section 2.3) is subsequently trained on the resulting 'unsupervised embeddings', while freezing the encoder's parameters.

For unsupervised training of the encoder, we leverage Contrastive Predictive Coding (CPC) [9], a recently proposed framework for self-supervised learning, which has already been found useful to model EEG data [10]. CPC is able to model slow features [9], i.e. slowly varying data characteristics, like the normalized amplitudes $\mathbf{A}$ in our signal model.

From a mathematical perspective, predicting the mixture coefficients (or amplitudes) requires solving a non-linear independent component analysis (ICA) problem, which has proven to be non-identifiable [11]. However, recent advances showed that the problem becomes identifiable under the assumed presence of an auxiliary variable [12]. The contrastive learning paradigm has shown to conform to this assumption [12, 13], and is able to invert the signal model, or data-generating process.

As such we hypothesize that a classifier trained on the unsupervised embeddings will have more tendency to make predictions that are related to the mixture coefficients, than a fully supervised model, which has more tendency to depend on the expert's annotations.

CPC leverages contrastive learning, which builds upon the idea to teach the model that 'similar data points' should be embedded closely together, while 'dissimilar data points' should be repelled. In the framework of CPC, a similar data point (or positive sample) is defined as a future embedding, with respect to a current causal embedding (i.e. incorporating past information as well). Negative samples, on the other hand, are drawn from a random moment within or between (i.e. from a different) recordings. We use within-subject sampling, and randomly draw three negative samples per positive sample. Set $\mathcal{Z}_p'^{(k)}$ contains the embeddings of these three negative samples, and is renewed for every data point and in every training iteration. We define $\mathcal{Z}_p''^{(k)} := \mathcal{Z}_p'^{(k)} \cup \{\boldsymbol{z}_{w+p}\}$, which contains both the negatives and positive embedding. The unsupervised CPC training objective, yields:

$$\mathcal{L} = \frac{1}{P}\sum_{p=1}^{P}\mathcal{L}_p, \text{ with} \tag{8}$$

$$\mathcal{L}_p = -\mathbb{E}_{\hat{p}_\mathrm{d}(\mathbf{x})}\Big[\log\frac{\exp(\boldsymbol{z}_{w+p}^T\mathbf{V}_p\boldsymbol{z}_w)}{\sum_{\boldsymbol{z}\in\mathcal{Z}_p''^{(k)}}\exp(\boldsymbol{z}^T\mathbf{V}_p\boldsymbol{z}_w)}\Big],$$

with $P = 10$ the number of future windows, $\boldsymbol{z}_w$ the current embedding, $\boldsymbol{z}_{w+p}$ the future embedding at index $w+p$, and $\mathbf{V}_p \in \mathbb{R}^{F\times F}$ a trainable mapping between both embeddings. In the following, we refer to the model which's encoder is trained using CPC, and a subsequent classifier is trained supervised, as the *unsupervised* or CPC model.

4.3. Results

4.3.1. Post-softmax predictions We start with an empirical investigation of the post-softmax predictions of the supervised model. To this end, four supervised models were trained with labels that have been generated from label distributions as given in eq. (2), with varying values of $\tau = \{0^+, \frac{1}{4}, \frac{1}{2}, 1\}$. This parameter can be seen as a slider for the amount of uncertainty that is present during labelling the dataset (high τ implies high uncertainty). Table 1 shows the two KL-divergence metrics, as introduced in eq. (6) and eq. (7) (one model per row). Note that the label distribution equals the normalized amplitudes for $\tau = 1$.

For all values of τ, it can be seen that the KL-divergence with the label distribution is lower (or equal, for $\tau = 1$) than with the normalized amplitudes, implying that the softmax outputs have more tendency to reflect the label distribution, than the data-characteristic as captured in the normalized amplitudes of the signals belonging to the different classes. Figure 4 visually compares the prediction of a random test case example, to the normalized amplitudes, for $\tau \to 0^+$ (a), and $\tau = \frac{1}{4}$ (b). Again it can clearly be seen how the model aims to mimic the label distribution. The difference between the label distribution and the normalized amplitudes is most apparent for values of τ close to zero (fig. 4a).

We additionally cross-compare the model predictions with label distributions with varying values for τ. Figure 3 shows a heat map of these results, in which the x-axis denotes the value of τ of the distribution from which labels were drawn during training, and the y-axis indicates this value during evaluation. The heat maps shows that the model predictions indeed aligns best with the label distribution that was used during training (seen from the dark green diagonal). The results in this section are all in line with the proof in eq. (4), which showed that a supervised classifier that is trained by likelihood maximization aims to mimic the conditional label distribution.

Table 1: KL-divergence between model predictions $\hat{\boldsymbol{y}}$ and the normalized amplitudes $\mathbf{A}$, and conditional label distribution $\hat{p}_{\mathrm{d}}$, respectively, for models trained with labels drawn from $p(\boldsymbol{y}|\mathbf{X};\tau)$ for varying τ (each row is one model). The KL-divergence with the label distribution is lower than with the normalized amplitudes, a better match of the former with the model prediction.

τ	$\mathrm{D}_{\mathrm{KL}}(\hat{p}_{\mathrm{d}}\|\|\hat{\boldsymbol{y}})$	$\mathrm{D}_{\mathrm{KL}}(\mathbf{A}\|\|\hat{\boldsymbol{y}})$
0^+	6.4e-2	2.1
1/4	8.3e-3	9.8e-2
1/2	6.9e-3	4.1e-2
1	9.5e-3	9.5e-3

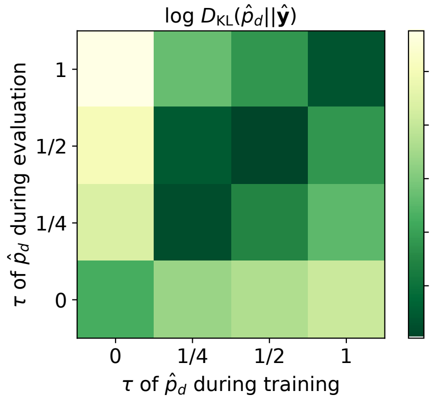
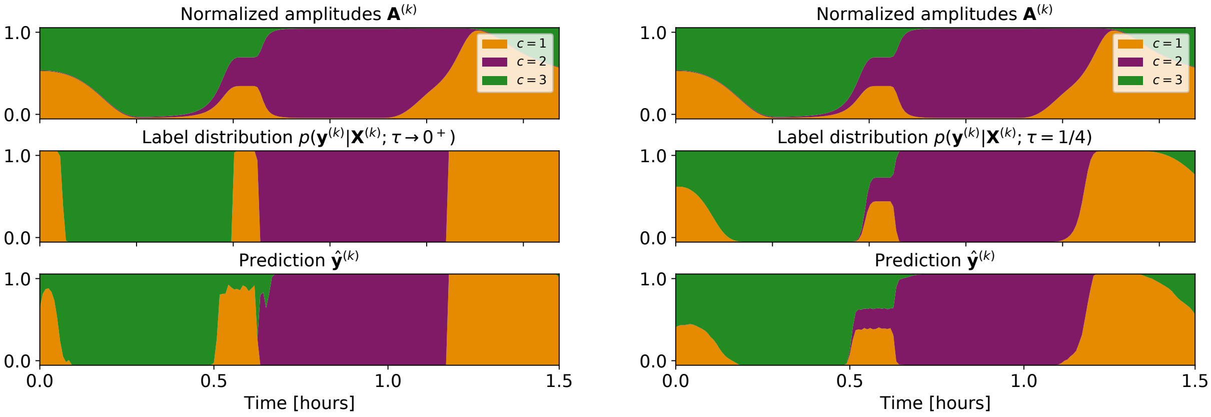


Figure 3: Cross-comparison of KL-divergences (in log-scale) between label distributions $p(\boldsymbol{y}|\mathbf{X};\tau)$ with varying τ during evaluation, and models trained with different values of τ. Lower is better.

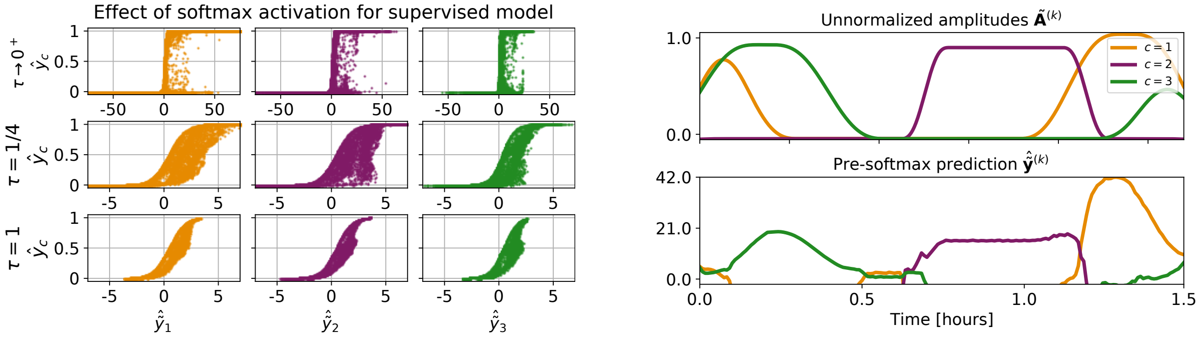


(a) (b)

Figure 4: The normalized amplitudes (top), label distribution (middle) and supervised model prediction (bottom) of a representative sample from the synthetic test set. Predictions of models, trained with label distributions with a) $\tau \to 0^+$, and b) $\tau = \frac{1}{4}$ are shown. It can clearly be seen that the predictions depend on the value of τ, and have the tendency to follow the corresponding label distribution.

4.3.2. Pre- vs post-softmax predictions Figure 5a shows the non-linear effect of the final softmax activation in a supervised neural classifier, trained with label distributions with varying values of τ. The x-axis denotes the pre-softmax predictions per class, while the y-axis denotes the corresponding post-softmax prediction. Each dot represents one window of one recording from the test set.

It can be seen that the pre-softmax input range, and therewith the softmax non-linearity increased for lower values of τ used during training the model. Mainly in case of deterministic label selection (i.e. for $\tau \to 0^+$, when the label distribution has zero entropy; top row), the softmax tended to push the class probabilities to zero or



(a) (b)

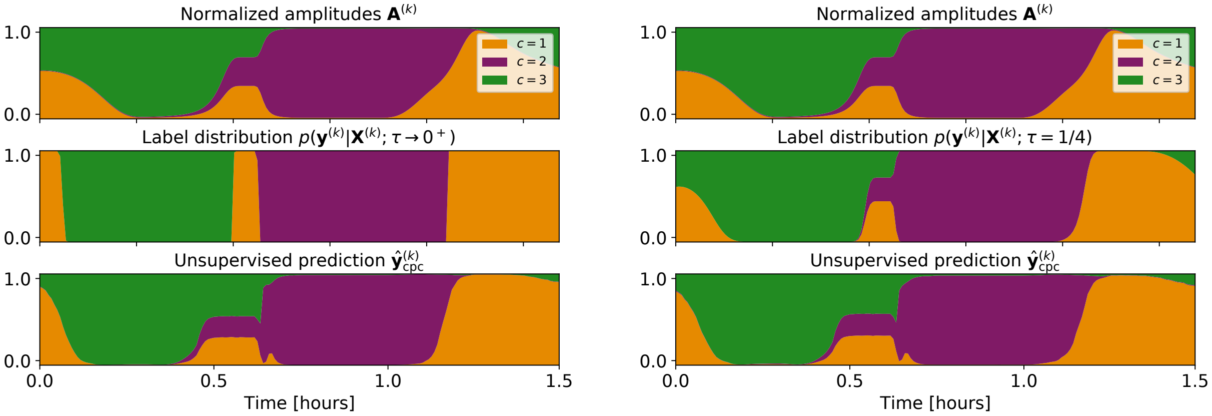
Figure 5: a) The pre-softmax predictions (x-axis) are highly non-linearly related to the post-softmax predictions (y-axis) for a supervised model with $\tau \to 0^+$ in the label distributions used for training (top); the softmax pushes most class predictions towards 0 or 1. When training with a non-degenerate label distribution (i.e. $\tau > 0$), the softmax outputs are less discrete (middle, bottom). b) The pre-softmax predictions (for $\tau \to 0^+$) (bottom) show dynamics that moderately resemble the unnormalized amplitudes (top).

Table 2: KL-divergence with the normalized amplitudes $\mathbf{A}$ (i.e. mixture coefficients) and the label distribution $p(\boldsymbol{y}|\mathbf{X};\tau)$ for encoders trained using Contrastive Predictive Coding, and classifiers trained with labels drawn from label distributions with varying temperature values τ (each row is one model). All models show a lower KL-divergence with the mixture coefficients than with the label generating distribution. For $\tau = 1$, the two are equivalent, hence the equivalent KL-divergences.

τ	$\mathrm{D_{KL}}(\hat{p}_{\mathrm{d}}\|\|\hat{\boldsymbol{y}})$	$\mathrm{D_{KL}}(\mathbf{A}\|\|\hat{\boldsymbol{y}})$
0^+	.21	6.3e-2
1/4	5.2e-2	4.7e-2
1/2	5.1e-2	4.0e-2
1	4.2e-2	4.2e-2

one. For $\tau > 0$, i.e. when sampling from the label distribution is a stochastic process due to its non-zero entropy, the softmax outputs are not anymore pushed towards such binary decisions (middle & bottom row). The effect of the softmax activation in a supervised classifier thus depends on the entropy of the generative label distribution. In real PSG data, this entropy can be seen as the amount of uncertainty the expert had when selecting labels for annotating the dataset.

To illustrate that mainly the softmax activation has a large influence on mimicking the label distribution with the post-softmax predictions, the pre-softmax predictions for one test set example are plotted in fig. 5b, for the model trained with label distribution $p(\boldsymbol{y}|\mathbf{X};\tau \to 0^+)$. Indeed, even though the post-softmax predictions were mimicking the label distribution (as was seen from fig. 4a), the pre-softmax predictions $\hat{\tilde{\boldsymbol{y}}}^{(k)}$ are (moderately) resembling the unnormalized amplitudes.

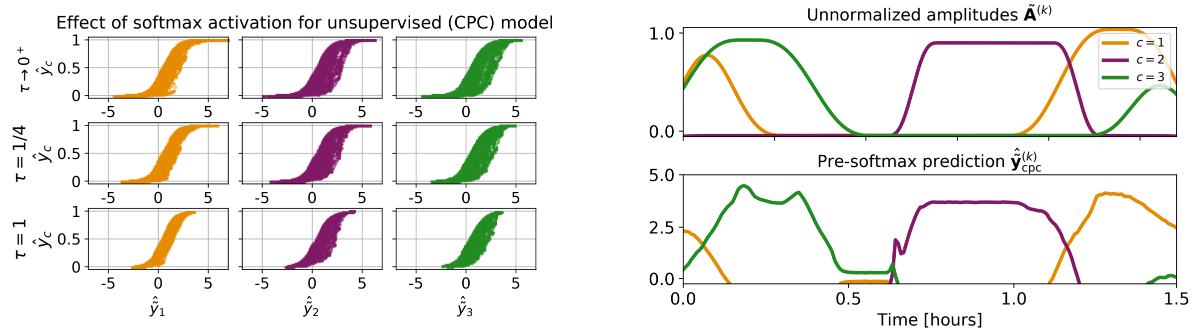


(a) (b)

Figure 6: The normalized amplitudes (top), label distribution (middle) and CPC model prediction (bottom) of a representative sample from the test set of the synthetic data set. We display predictions for which the classifiers are trained with label distributions with a) $\tau \to 0^+$, and b) $\tau = \frac{1}{4}$. The predictions are much less influenced by the value of τ, as compared to the supervised model, displayed in fig. 4, and therefore show more tendency to predict the normalized amplitudes.

4.3.3. Supervised vs unsupervised encoding Table 2 shows the KL-divergence metrics of the (post-softmax) predictions of the unsupervised model. This KL-divergence is, for all values of τ, lower with respect to the normalized amplitudes $\mathbf{A}$, than with the conditional label distribution $\hat{p}_d$. This implies that the unsupervised model, in contrast to the supervised model for which the results were exactly opposite (see table 1), makes a prediction that is closer to the normalized amplitudes than to the label distribution. Figure 6 shows the predictions for $\tau \to 0^+$ (a) and $\tau = \frac{1}{4}$ (b) for the same test set example as for which the supervised predictions were shown in fig. 4. It is clearly visible that the unsupervised model's post-softmax prediction is rather independent of the value of τ used for training the classifier. This can be explained by the fact that the classifier is of such low capacity (only a linear mapping with a softmax activation), that it is unable to fit the label distribution. As a result, the predictions are closer to data-characteristics, rather than label-charactersitics, which in this case results in predictions that are close to the normalized amplitudes.

4.3.4. Interaction effect between softmax and (un)supervised training In fig. 5a it was already shown that the final softmax activation of the supervised model operated in a different regime, dependent on the value of τ used during training the model. Figure 7a shows that this effect was almost fully omitted when training the full encoder unsupervised. Note that the range of the x-axis in the top-row is now equivalent to the this range in the middle and bottom row, while these ranges highly differed for the supervised model (see fig. 5a). The pre-softmax predictions of the unsupervised model seem to be slightly closer to the true unnormalized amplitudes, as seen from fig. 7b, compared to this prediction of the supervised model, as seen in fig. 5b.



(a)

(b)

Figure 7: a) Using unsupervised encoding, the operating range, and thus the non-linear effect of the softmax function is almost independent of the value of τ used in the label distribution during training the classifier. b) The pre-softmax predictions (for $\tau \to 0^+$) (bottom) show dynamics that slightly better resemble the unnormalized amplitudes (top) than the pre-softmax predictions of the supervised model (shown in fig. 5b).

5. Experiments on real PSG data

Results on synthetic data showed that the predictions of a supervised neural classifier revealed label-characteristics, i.e. the generative label distribution. In the context of sleep recordings this can be seen as predicting the probability that a certain data window would have been labelled as one of the different sleep stages, by the expert(s) that labelled the dataset used for training the model. An unsupervisedly-trained model was shown to yield similar post-softmax predictions, while being less reliant on the quality of the labels (modelled as τ). Moreover, pre-softmax predictions were shown to correspond to data-characteristics, i.e. the unnormalized amplitudes of the different mixture signals. In this section we perform similar experiments on real PSG data, and compare the effects to the aforementioned conclusions from the synthetic setup.

5.1. Polysomnography data

We used a dataset of nocturnal video-PSG recordings of 96 healthy sleepers, that were recorded according to the AASM recommendations [2] in Sleep Medicine Center Kempenhaeghe Heeze, the Netherlands. Annotations were created by visual sleep staging on windows of 30 seconds, performed by an experienced and certified sleep technician from Sleep Medicine Center Kempenhaeghe. From the full PSG recordings, we selected EEG (F3/F4, C3/C4, O1/O2), chin EMG (Chin1/Chin2), and EOG (E1/E2) derivations, since these are typically used for manual AASM scoring as well. We randomly generated a training ($K = 150$), validation ($K = 20$) and hold-out test set ($K = 22$). Appendix C.2 provides more details about the dataset and the applied preprocessing.

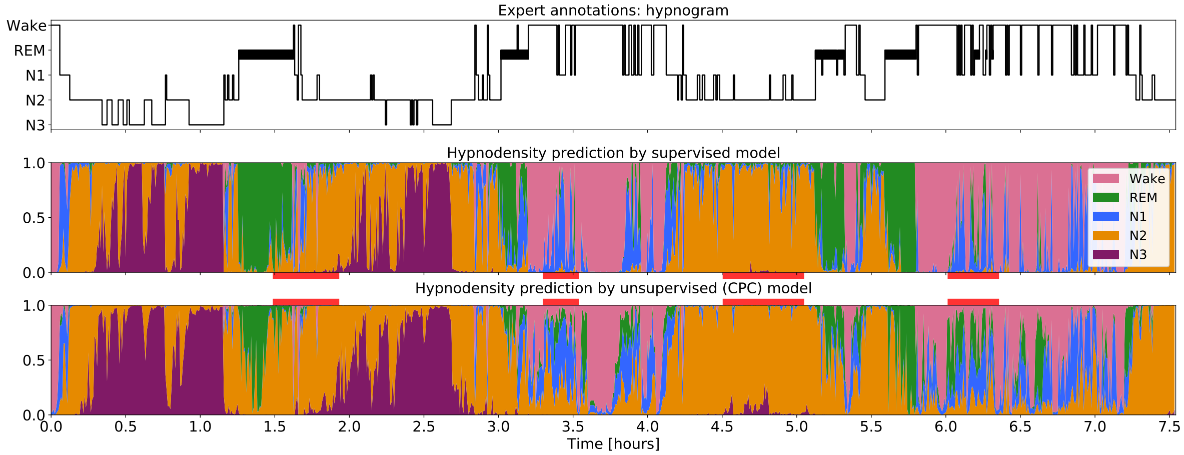
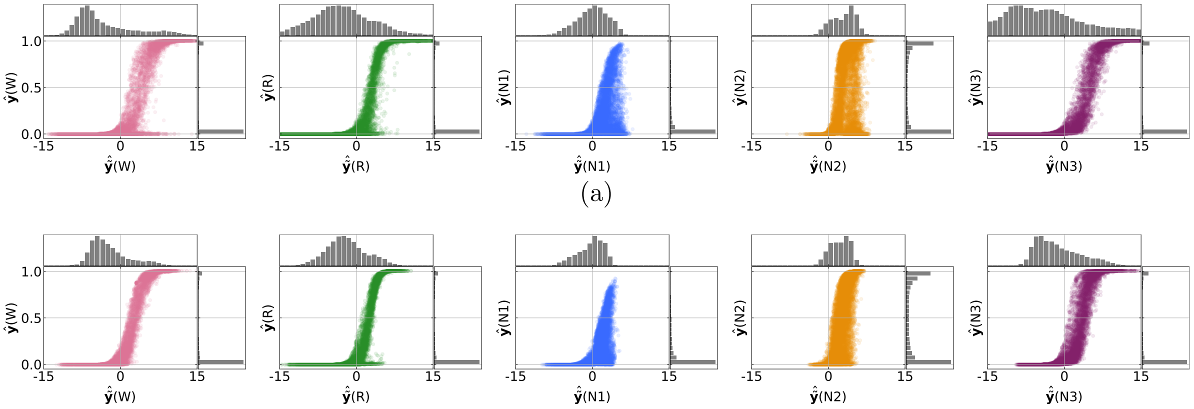


Figure 8: The hypnogram as annotated by a sleep technician (top), and predicted hypnodensity plots by the supervised (middle) and unsupervised model (bottom). The general trend looks similar, but differences are visible (indicated with the red bars), e.g. the unsupervised model in general shows smoother transitions. Note that the unsupervised model does not just predict a smoothened version of the supervised prediction: hard transitions (e.g. at 1.2 hours) can still be predicted as well.



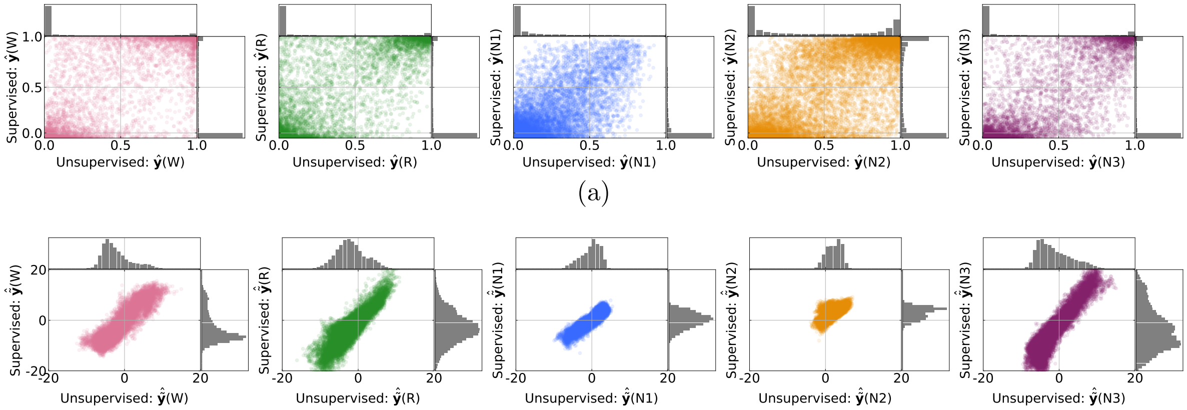
(a)

(b)

Figure 9: Scatter plots that plot the pre-softmax against the post-softmax predictions for the supervised (a) and unsupervised (b) model. Each dot is a window from the full test set. Given the smooth shapes, also for the supervised model, the manual sleep staging process must exhibit stochasticity (i.e. $\tau > 0$).

5.2. Results

5.2.1. Supervised vs unsupervised hypnodensity plot Figure 8 shows the (post-softmax) predictions (i.e. hypnodensity plots) from both the supervised (middle row) and unsupervised (bottom row) model for one representative recording of the test set. For reference, the top row shows the hypnogram as annotated by the sleep technician. The general trend of both predictions looks similar, but differences can be noted (some are indicated by the red bars between both plots). For example, low amounts of N2 or N3 were sometimes predicted by the unsupervised model, while the supervised counterpart



(a)

(b)

Figure 10: Scatter plots that compare the post-softmax (a), and pre-softmax (b) predictions of the supervised (y-axes) and unsupervised model (x-axes). Each dot is a window from the full test set. Given the more linear relation between both models' pre-softmax predictions (as opposed to post-softmax predictions), it can be implied that the non-linear softmax has a different effect on both models.

did not show these low contributions. This difference in spread of predicted probabilities over the classes was also reflected in average entropy of predictions of all windows in the test set, which was found to be $\mathrm{H} = 0.30 \pm 0.06$ for the supervised, and $\mathrm{H} = 0.43 \pm 0.1$ for the unsupervised model.

Occasions where the hypnogram showed rapid transitioning behavior (e.g. around 6.2 hours), were characterized by high entropy predictions from the unsupervised model, while the supervised model predicted a lower-entropy but more time-varying distribution over sleep stages. These rapid changes in predictions of the supervised model possibly reflect the fact that the model was trained using the annotated hypnogram that also contains these (discrete) switches between sleep stages. A similar effect was visible in the synthetic setup, where the supervised model had more tendency to result in more abrupt transitions (fig. 4 vs fig. 6). Despite the more smooth prediction by the unsupervised model, note that it is still able to predict abrupt transitions as well (e.g. at 1.2 hours), so it can not simply be considered a smoothened version of the supervised prediction.

5.2.2. Pre- vs post-softmax predictions Figure 9 plots the pre- versus post-softmax predictions of all sleep stages, for both the supervised (a) and unsupervised (b) model (each dot is one window of one recording from the test set). The softmax effect of the supervised model was found to be less non-linear than was seen for the synthetic case where $\tau \rightarrow 0^+$ (see fig. 5a), implying that the implicit label distribution in our training dataset was non-degenerate (i.e. $\tau > 0$). In other words, the expert labels were assigned with a certain form of stochasticity, which is in line with the known imperfectness of manual sleep stage scoring. *It should thus be realized that thanks to the presence of inter- and intra-rater disagreement in manual sleep staging, a supervised hypnodensity plot exhibits a smooth pattern over time.*

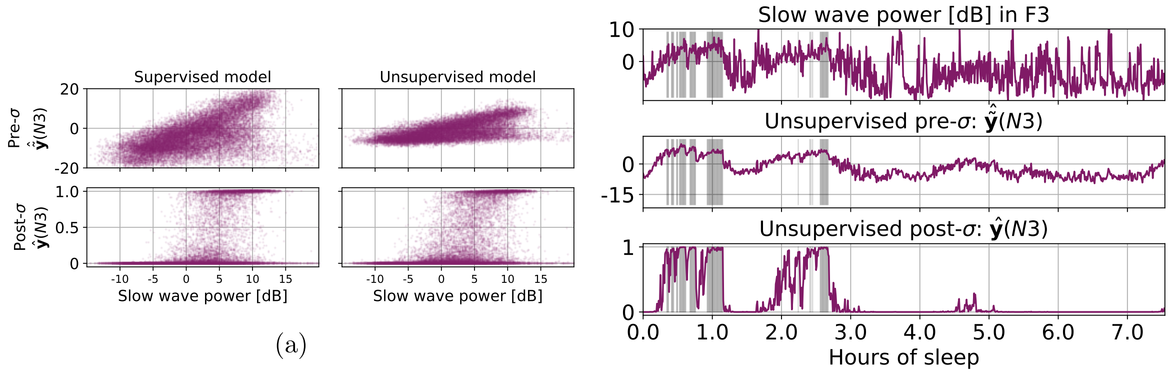
Interestingly, the softmax effect seems different across sleep stages, which can best be seen from the vertical histograms that show the post-softmax distributions (fig. 9). For example, N1 is by neither of the models predicted with a 100% probability, and the slope of the scatter plots is clearly steeper for N2 as compared to W, REM and N3 sleep.

5.2.3. Interaction effect between softmax and (un)supervised training To investigate further in which way the supervised and unsupervised model differed, we plot their predictions against each other in fig. 10a (post-softmax), and fig. 10b (pre-softmax). From this visualization, it can be seen that the softmax activation has a different effect on the supervised model, as compared to the unsupervised model; the pre-softmax predictions of both models seem linearly related with only small deviations from this linear trend, while the relation between the post-softmax predictions was more spread out. So even though the general trend of the predicted hypnodensity plots looked similar (see fig. 8), when zooming in on window level (as done here, since each dot is one window), differences in post-softmax predictions do exist between both models, which might exhibit clinically relevant information.

5.2.4. N3 prediction vs slow wave power Given the fact that both the underlying signal model and the generative label distribution are evidently unknown in real PSG data, we seek an additional approach to draw conclusions on the two different type of models and their pre- and post-softmax predictions. It is known that slow waves (positively) relate to the depth of sleep [1], and the AASM selection criterion for scoring N3 is based upon the amplitude of these slow waves [2]. As such, we can use the slow wave power as a surrogate for the contribution of the deepest sleep phase N3, to the total mixture of characteristics belonging to different stages. To this end, we compare the four predictions (i.e. (un)supervised and pre- vs post-softmax) for N3, to the amount of slow wave (0.5-2 Hz) power in the frontal EEG lead (F3 or F4).

Figure 11a plots both pre- and post-softmax predictions for N3 of both models against the slow wave power for all windows in the test set. It can clearly be seen that, for both models, the pre-softmax prediction better follows a linear relation with the slow wave power, than its post-softmax counterpart. Figure 11b depicts the pre/post-softmax predictions for N3 from the unsupervised model, and slow wave power over time, for the same recording as depicted in fig. 8. This figure clearly shows how the softmax outputs of the unsupervised model, despite being more continuous/smooth than the supervised predictions, still tended to follow the N3 annotations (in grey), whereas the pre-softmax outputs better captured the continuity of deep sleep.

Note the tails in fig. 11a-top, where a low value for N3 was predicted, while high slow wave power was computed. A recheck confirmed that these tails were not caused by a low-quality measurement for one of the patients in the test set, but was present for multiple patients. A possible explanation can be that low-frequency content, slightly above 0.5 Hz (i.e. included in the slow wave range), entered the spectrum during wake



(a)

(b)

Figure 11: (a) The frontal EEG slow wave power against the predicted pre-softmax values (top), and post-softmax probabilities (bottom), over the full test set. The pre-softmax predictions correspond much better with slow wave power for both models. (b) Illustrative example that shows the pre- and post-softmax predictions over time for the unsupervised model. The grey lines indicate windows that were annotated as N3. The pre-softmax prediction better follows the continuity of slow wave power.

episodes as a consequence of movement artifacts. Figure 11b indeed showed this effect at 3.7 hours, where high slow wave power was present, but the data were annotated as Wake (seen from the hypnogram in fig. 8-top).

6. Discussion

In this work, we investigated the interpretation of the recently proposed hypnodensity representation of a PSG recording [3], being a probability distribution over sleep stages throughout the night. Great potential is foreseen for this representation, in addition to the conventional hypnogram, as it, e.g., opens up a legion of research directions about sleep disorders that are known to be related to sleep stage dissociations or local sleep phenomena [6]. In order to shed light on the interpretation of a hypnodensity representation, we proposed a PSG-inspired synthetic dataset that comprised non-linearly mixed measurements of signals belonging to different classes, analogous to the sleep stages, labelled with the most prevailing class (see Section 2.1 and 2.2). Of course these data are simplified and subject to design choices, possibly hampering full generalizability to real PSG data. Nevertheless, similarities between the results on the synthetic case and real PSG data were found, validating our proposed signal and annotation models. In the following, we will discuss the answers to the research questions as posed in the introduction.

6.1. Answering the research questions

First, we wondered whether a hypnodensity plot reflects continuous dynamics of the sleeping brain, or whether it shows model uncertainty, and what the implicit assumptions and driving factors of its prediction are.

Both theoretical (section 3) and empirical evidence (section 4.3.1) showed that *a hypnodensity plot, predicted by a supervised classifier, reveals the probabilities with which a window was assigned to either of the classes by the expert(s) that annotated the dataset used for training the model.* This finding is also in line with an observation the authors made in the original hypnodensity work [3], which stated that the hypnodensity representation resembled the inter-rater disagreement across multiple scorers that annotated their dataset. Note that a supervised classifier only mimics the label distribution under the assumption that the model exhibits the 'right amount' of capacity. In other words, a model that is too small has a large amount of inherent (called epistemic) uncertainty, which increases the average entropy of the probability distributions in the hypnodensity plot. On the other hand, a model that has too much capacity has the tendency to over-fit on the training set and becomes over-confident. In the machine learning community, such models are known as uncalibrated models [14,15]. Given the relatively simple model architecture as used in this work, and the fact that overfitting was not observed when comparing the training and validation log-likelihood during training, we assume our model was not uncalibrated.

Given the fact that the supervised model did predict non-zero entropy distributions (see fig. 8-middle), it can be concluded that the (implicit) label distribution in our dataset was non-degenerate (i.e. $\tau > 0$, or in other words; our experienced sleep technician assigned labels in a non-deterministic fashion, i.e. with uncertainty). This conclusion is in line with the fact that manual sleep staging yields disagreement both within and among sleep experts [8]. It can thus be concluded that *a supervisedly-predicted hypnodensity representation is able to exhibit smoothness across sleep stages (and therefore reveals additional information with respect to a hypnogram), thanks to the inter- and intra-rater disagreement of sleep expert(s) that annotated the dataset.* This is interesting, as scorer disagreement is generally considered a negative consequence of manual scoring, while a hypnodensity plot, predicted by a supervised classifier thus actually requires it. It is, however, important to realize that the strong label-dependency of the supervisedly-predicted hypnodensity plot may result in research conclusions that are strictly reliant on the expert annotations that were used in the specific study. In other words, when drawing conclusions about sleep disorders by means of a supervisedly-predicted hypnodensity plot, researchers may not want to rely on one dataset that is annotated by, e.g., inexperienced scorer(s), as it might highly influence the hypnodensity plots.

Predicting a hypnodensity representation using an unsupervisedly-trained encoder, followed by a supervised classifier, on the other hand, was shown to be less dependent on the (un)certainty of the expert annotations (modelled with τ in this work, and shown on synthetic data). This was explained by the fact that only the low-capacity classifier may be influenced by these, while the full encoder was trained without their availability. The unsupervisedly-predicted hypnodensity plots on real PSG data exhibited higher entropy, as compared to these plots predicted by the supervised model. This finding again implies a lower dependency of the unsupervised model on the hard/discrete expert annotations, as opposed to the supervised model. *Unsupervised training thus seems a reliable strategy to acquire a hypnodensity representation of sleep that is less influenced by the quality of the expert annotations, as compared to models that that are trained in a supervised fashion.*

Second, we wondered whether alternative approaches could be found that would better reflect continuous brain dynamics. *Both for the synthetic and real dataset, it was shown that the pre-softmax predictions of both the supervised and unsupervised models revealed continuous data dynamics, which were more smooth over time than their post-softmax counterparts.* Note that the value of a pre-softmax prediction at one point in time, has no direct physical interpretation, nor relative meaning with respect to other stages due to the unnormalized nature. Nevertheless, it may contain clinically relevant information when considering the interplay of these pre-softmax predictions over time or across classes.

The fact that the final softmax activation was found a main contributor to convert data-characteristics (e.g. the the slow wave power) to label characteristics (the hypnodensity plot), validates our annotation model as provided in eq. (2). It also implies that the uncertainty that is present in the labelling process (visualized in the hypnodensity representation) is, at least partly, caused by mixture of sleep stage-dependent data characteristics in the windows. The hypnodensity plot, while displaying the label distribution, is thus not fully independent of the continuities of the sleeping brain, and still provides a (non-linear) reflection of those.

6.2. Future work

Despite the fact that the largest part of the unsupervised model was trained without expert annotations, and the supervised classifier only exhibited low capacity, a difference was still found between pre- and post-softmax predictions of this model (as discussed ealier). This difference thus teaches us that the low-capacity classifier still pushed the post-softmax predictions of the CPC model towards the label distribution. When interested in the continuous dynamics of the sleeping brain, it was thus seen that the pre-softmax predictions would be more suitable to consider, but they were already mentioned to be unnormalized. This finding opens up a new research direction, in which one may investigate how CPC/unsupervised embeddings can be mapped to AASM classes, without relying (again too much) on the labels during classifier training.

For the supervised model, the authors of [3] observed a more smooth hypnodensity representation when using memory, implemented as a Long Short-Term Memory (LSTM) cell, as part of the model. Addition of such memory is expected to have a similar effect on both supervised and unsupervised models, and therefore hypothesized to not change the drawn conclusions. Still, as suggested by [3], it will likely improve smoothness of the hypnodensity plots of both models, which might be desirable in certain circumstances. Note, however, that this smoothing property may hamper visibility of rapidly-changing patterns, which might be a biomarker of certain sleep disorders. Memory should therefore be used with caution.

In this research, we did not investigate influence of design choices like data window length (which was fixed to 30 s), and the number and type of measurement channels used. The former is, however, not expected to change the drawn conclusions regarding interpretability of hypnodensity representations, but using smaller windows might facilitate research to local sleep phenomena as the predictions would suffer less from temporal aggregation of data. Regarding the number of channels, the authors of [16] depicted (supervised) hypnodensity plots, predicted from one EEG channel only, and showed that these plots did not drastically differ dependent on the chosen channel. However, visual inspection revealed that their single-channel hypnodensity plots yielded higher entropy than the presented (supervised) hypnodensity plot in this work and in [3]. Since conventional ICA requires at least a number of measurement channels equal or larger than the number of sources to be revealed, it might be expected that the number of channels fed to a machine learning model that (implicitly) performs (non-linear) ICA under our signal model, may affect the hypnodensity plot as well. Research that compares hypnodensity plots, predicted from one or multiple channels might therefore be useful, especially with an eye upon the raising trend of consumer electronics for measuring sleep that tend to incorporate fewer sensors (i.e. channels) than the conventional PSG recording.

7. Conclusion

In this work, we investigated how to interpret the recently-proposed hypnodensity representation [3] of a PSG recording, when predicted by either a supervisedly- or unsupervisedly-trained (Contrastive Predictive Coding) machine learning model. Moreover, the final softmax-activated outputs, as well as the pre-softmax predictions, of both models were analyzed. The following conclusions could be drawn: A hypnodensity plot, predicted by a supervised model, reveals the label distribution from which sleep stages were (implicitly) drawn/assigned during manual sleep staging by a sleep expert. In other words, it reflects the uncertainty of a human decision process of assigning AASM labels. It, therefore, is a representation of sleep, which is highly dependent on the amount of the expert's scoring uncertainty (i.e. the value of τ in our model). A hypnodensity plot predicted by an unsupervised model, on the other hand, was shown to be less dependent on this uncertainty, and therefore provides a more robust representation of sleep.

Despite the revelation of ample information in the hypnodensity plot (with respect to a hypnogram), potentially relevant information on continues processes in the sleeping brain may have been non-linearly transformed due to the final normalizing softmax activation (both for supervised and unsupervised training). The pre-softmax class predictions, on the other hand, were shown to have a better linear relation with continuous brain dynamics.

Thus, when aiming for a hypnodensity representation that represents AASM label probabilities, using an unsupervisedly-trained model seems most safe to prevent conclusions that are (too much) biased by label characteristics of the used dataset. When continuous dynamics of a specific sleep stage are to be investigated, the (unnormalized) pre-softmax prediction for that stage might be more suitable.

This work opens up new research directions regarding the effect of the number of channels, used window length, and model design choices on both supervised and unsupervised models that predict hypnodensity plots. Moreover, biomarkers in both pre- and post-softmax predictions might be searched for that distinguish different patient groups.

Acknowledgments

This work was supported by Onera Health and the project 'OP-SLEEP'. The project 'OP-SLEEP' is made possible by the European Regional Development Fund, in the context of OPZuid.

Appendix A. Symbols and notations

Table A1 shows the most used symbols and notations, as used in this work.

Symbol	**Domain**	**Meaning**
C	$\mathbb{N}$	Number of classes, indexed with $1 \leq c \leq C$
W	$\mathbb{N}$	Number of non-overlapping 30-second windows in a recording, indexed with $1 \leq w \leq W$
K	$\mathbb{N}$	Number of recordings, indexed with $1 \leq k \leq K$
l	$\mathbb{N}$	Number of samples in one 30-second window
$L := W \times l$	$\mathbb{N}$	Number of samples in one recording
ch	$\mathbb{N}$	Number of recording channels
$\mathbf{X}^{(k)}$	$\mathbb{R}^{ch \times W \times l}$	All data of recording k
$\mathbf{X}_w^{(k)}$	$\mathbb{R}^{ch \times l}$	30-second data window with index w of recording k
$\mathbf{S}^{(k)}$	$\mathbb{R}^{C \times W \times l}$	Signals of C classes of recording k
$\boldsymbol{s}_c^{(k)}$	$\mathbb{R}^{L}$	Signal belonging to class c for recording k
$\tilde{\mathbf{A}}^{(k)}$	$\mathbb{R}_{\geq 0}^{C \times W \times l}$	Unnormalized amplitudes of recording k
$\mathbf{A}^{(k)}$	$\{\mathbb{R}_{\geq 0}^{C \times W \times l} : \sum_c \mathbf{A}^{(k)} = 1\}$	Normalized amplitudes of recording k
$\tilde{\mathbf{A}}_w^{(k)}$	$\mathbb{R}_{\geq 0}^{C \times l}$	Unnormalized amplitudes in window w of recording k
$\mathbf{A}_w^{(k)}$	$\{\mathbb{R}_{\geq 0}^{C \times l} : \sum_c \mathbf{A}_w^{(k)} = 1\}$	Normalized amplitudes in window w of recording k
$\tilde{\boldsymbol{a}}_c^{(k)}$	$\mathbb{R}^{L}$	Unnormalized amplitudes of class c of recording k
$\boldsymbol{a}_c^{(k)}$	$\mathbb{R}^{L}$	Normalized amplitudes of class c of recording k
$\mathbf{Y}^{(k)}$	$\{0,1\}^{C \times W}$	One-hot embeddings of ground-truth class labels of recording k
$\boldsymbol{y}_w^{(k)}$	$\{0,1\}^{C}$	One-hot embedding of ground-truth class label for window w of recording k
$\hat{\boldsymbol{y}}^{(k)}$	$\{\mathbb{R}_{\geq 0}^{C \times W} : \sum_c \hat{\boldsymbol{y}}^{(k)} = 1\}$	'Soft' class predictions of recording k
$\hat{\boldsymbol{y}}_w^{(k)}$	$\{\mathbb{R}_{\geq 0}^{C} : \sum_c \hat{\boldsymbol{y}}_w^{(k)} = 1\}$	'Soft' class prediction for window w of recording k
$\hat{\boldsymbol{y}}_{w;\text{super}}^{(k)} := \hat{\boldsymbol{y}}_w^{(k)}$	$\{\mathbb{R}_{\geq 0}^{C} : \sum_c \hat{\boldsymbol{y}}_{w;\text{super}}^{(k)} = 1\}$	'Soft' class prediction for window w of recording k by the supervised model.
$\hat{\boldsymbol{y}}_{w;\text{cpc}}^{(k)}$	$\{\mathbb{R}_{\geq 0}^{C} : \sum_c \hat{\boldsymbol{y}}_{w;\text{cpc}}^{(k)} = 1\}$	'Soft' class prediction for window w of recording k by the CPC model.
$\hat{\tilde{\boldsymbol{y}}}_{w;\text{super}}^{(k)}$	$\mathbb{R}^{C}$	Pre-softmax class prediction for window w of recording k by the supervised model.
$\hat{\tilde{\boldsymbol{y}}}_{w;\text{cpc}}^{(k)}$	$\mathbb{R}^{C}$	Pre-softmax class prediction for window w of recording k by the CPC model.

Table A1: The meaning and domain of the symbols used in this work that are related to data and their annotations.

Appendix B. Encoder architecture and training details

The architecture of the encoder followed standard practice in supervised classification model design [5]. Enc$(\cdot)$ comprised three consecutive blocks, where each block containd a 1D temporal convolutional layer, activated by a LeakyReLU (negative slope of 0.01), followed by a 1D max pooling layer, and finally a dropout layer ($p = 0.1$). After the third full block, a fourth 1D convolutional layer was added, followed by average pooling that reduced the temporal dimension to size 1, creating a 1D embedding of size F. All convolutional layers had a bias term, and used strides and dilations of 1. The number of channels differed for the real data $(16, 32, 64, 128)$ vs synthetic data $(4, 8, 16, 32)$ setup, to account for the higher complexity of real data. The used kernels were of size $(15, 9, 5, 3)$ for the four convolutions, and the max pooling layers used kernels of size 5 (with stride 5).

In order to make the fairest between supervised and unsupervised training (see Section 4.2.2), we kept both the parameter initializations and the encoder's design equivalent for both strategies (except for the dropout rates in the CPC encoding trained on synthetic data, for which lower values appeared more beneficial: $(0.1, 0.0, 0.0)$).

All supervised models were trained using the categorical cross-entropy (or negative log-likelihood) loss, in batches of 128 training pairs. Unsupervised encodings were trained with the CPC objective as given in eq. (8), and batches of size 64. The Adam optimizer with default settings [17] was used in all experiments, with a learning rate of 1e-4 for most experiments. Only the supervised classifier, and CPC encoding on synthetic data were trained with learning rates of 1e-3 and 5e-4, respectively. All models were maximally trained for 500 epochs, where one epoch defined one push trough of each data window in the training set. The classifiers trained after CPC encoding, were maximally trained for 100 epochs. In each experiment, the model with the lowest validation loss was finally selected. All experiments were run with the same seed for randomization.

Appendix C. Data sets

Appendix C.1. Synthetic data

To create a synthetic dataset, the signal model as introduced in eq. (1) was used. Each channel ($ch = 3$) in $\mathbf{X}$ was modelled as a non-linear combination of a set of ($C = 3$) independent signals, where each signal represented a (fictitious) sleep stage. The data of 'recording k' was defined as $\mathbf{X}^{(k)} = h(\tilde{\mathbf{A}}^{(k)} * \mathbf{S}^{(k)})$, where $\tilde{\mathbf{A}}^{(k)} \in \mathbb{R}_{\geq 0}^{C \times W \times l}$ are the unnormalized amplitudes of recording k, and $\mathbf{S} \in \mathbb{R}^{C \times W \times l}$ the corresponding signals.

Each signal $\boldsymbol{s}_c^{(k)}$ was generated as a (discretized) sinusoidal signal, with a frequency between $f_c - 0.5$ and $f_c + 0.5$ Hz, a random phase, and an amplitude $\tilde{\boldsymbol{a}}_c^{(k)}$ that is described by a smoothened square wave (sw). More specifically, for each k, we defined three independent signals, with $c \in \{1, 2, 3\}$:

$$\boldsymbol{s}_c^{(k)}[n] = \sin\{2\pi(\frac{f_c + u[-\frac{1}{2}, \frac{1}{2}]}{f_s})[n] + 2\pi u[0, 1]\}, \tag{C.1}$$

with $u[a, b]$ being a realization of a uniform random variable between a and b, and $\{f_1, f_2, f_3\} = \{2.5, 6, 11\}$ Hz. Each signal's length was $L = 5.4e5$ samples, sampled at a frequency of $f_s = 100$ Hz, resulting in a 'recording' of 5400 seconds, thereby mimicking the length of one average sleep cycle.

The (unnormalized) amplitude $\tilde{\boldsymbol{a}}_c^{(k)}$ of the c^{th} signal of subject k, was defined as:

$$\tilde{\boldsymbol{a}}_c^{(k)}[n] = \frac{\text{Hanning}_\nu \circledast \text{sw}[\text{n}; \Phi]}{|\,\text{Hanning}_\nu\,|}, \tag{C.2}$$

where $\circledast$ denotes a convolutional operator, $\nu = u[\frac{1}{20}, \frac{1}{4}]Lf_s$ is the length of the applied Hanning window, and the square wave's parameters are given by $\Phi = \{\text{period} = L, \text{sampling\_freq} = f_s, \text{duty\_cycle} = \frac{1}{2}, \text{phase} = 2\pi u[0,1], \text{min\_value} = \frac{1}{100}, \text{max\_value} = u[\frac{1}{2}, 1]\}$.

From the earlier definition of $\mathbf{X}^{(k)}$ it can be seen that mixing function $h(\cdot)$ is independent of k, i.e. 'recording'-independent. This results in a simplified but valid model, since certain sleep stage characteristics are in practice also measured more in certain channels than in others for all subjects (e.g. slow waves are mainly recorded in the frontal EEG electrodes). Only small deviations - resulting from inter-patient differences - are not captured by choosing one shared setting. For brevity, we define $\boldsymbol{as}_c := \tilde{\boldsymbol{a}}_c^{(k)} * \boldsymbol{s}_c^{(k)}$ here. We defined the non-linear mixing in $h(\cdot)$ as:

$$h(\mathbf{A}^{(k)} * \mathbf{S}^{(k)}) = \begin{bmatrix} 0.3\ \boldsymbol{as}_1 * \boldsymbol{as}_1 + 0.7\ \boldsymbol{as}_3 \\ 0.6\ \boldsymbol{as}_1 + 0.4\ \boldsymbol{as}_2 * \boldsymbol{as}_3 \\ 0.4\ \boldsymbol{as}_1 + 0.5\ \boldsymbol{as}_2 + 0.1\ (\boldsymbol{as}_3)^2 \end{bmatrix}.$$

We finally generated $K = 200$ random 'recordings', which were split into a training, validation and a hold-out test set of sizes 75, 25, and 100, respectively. Figure 2 shows an example from the test set, with normalized amplitudes (or mixture coefficients) on the left, and the corresponding unnormalized amplitudes on the right. From the right figure it can be seen that the heights, phases, and the steepnesses of the amplitudes differ per signal, caused by the injected stochasticity in the data generating process. As a result of this stochasticity, we see that at any moment zero to three class signals stages might co-exist. Absence of characteristics belonging to any of the sleep stages, might in practice occur when electrodes become disconnected.

Appendix C.2. Polysomnography data

We used a dataset of nocturnal PSG recordings, collected as part of the Healthbed study, which's main aim was development of technologies for sleep analyses. The study prototcol (W17.128) was approved by the medical ethics committee of Maxima Medical Center, Veldhoven, the Netherlands. The dataset includes one clinical video-PSG recording for each subject, made according to the AASM recommendations in Sleep Medicine Center Kempenhaeghe. The data analysis protocol for our study (CSG_2021_007_00) was approved by the medical ethics committee of Sleep Medicine Center Kempenhaeghe (11/11/2019).

The study included 96 (60 females) healthy subjects, with an age between 18 and 64. The exclusion criteria were: 1) any diagnosed sleep disorder, 2) a Pittsburgh Sleep Quality Index [18] ≥ 6, or Insomnia Severity Index [19] > 7, 3) indication of depression or anxiety disorder measured with the Hospital Anxiety and Depression Scale [20] (score $>$ 8), 4) pregnancy, shift work, use of any medication except for birth control medicine, and 5) presence of clinically relevant neurological or psychiatric disorders or other somatic disorders that could influence sleep.

Visual sleep staging on windows of 30 seconds was performed according to AASM criteria [2] by an experienced and certified sleep technician. from Sleep Medicine Center Kempenhaeghe. In a previous institutional sleep scoring reliability check, inter-scorer reliability of this technician, compared to other experts was assessed at 85.6% on average (range 83-88%).

From the full PSG recordings, we selected EEG (F4, C4, O2, F3, C3, O1), chin EMG (Chin2, Chin1), and EOG (E2, E1) derivations, since these are typically used for manual AASM scoring as well. Since the EEG and EMG derivations contain redundancy among the left and right hemisphere, the odd and even measurements of all subjects were added as separate recordings to the final dataset ‡. For simplicity, the two EOG recordings were split in a similar fashion, even though these recordings can not be considered fully redundant. As an example; channel data $\mathbf{X}^{(k)} \in \mathbb{R}^{5\times W\times l}$, where k, e.g., refers to the even recording of one of the subjects, thus contained the F4, C4, O2, E2, and Chin2 derivations.

Following [3], all derivations were filtered with a zero-phase (i.e. two-directional) 5th order Butterworth band-pass filter, with cut-off frequencies of 0.2 and 49 Hz. It was followed by another zero-phase 5th order Butterworth notch filter between 49 and 51 Hz, to better suppress powerline interference. All channels were originally recorded with a sampling rate of 512 Hz, but (after filtering) down-sampled to 128 Hz to reduce computational complexity. Channels were normalized within-patient and per channel, yielding mean subtraction, followed by normalization such that amplitudes of 95% of the samples were mapped between -1 and +1.

Finally, the data were randomly split in a training, validation and hold-out test set, comprising respectively $K = 150$, $K = 20$, and $K = 22$ recordings (each recording being either even or odd). Even and odd recordings from the same subject were in all cases assigned to the same subset.

[1] Kryger M H, Roth T and Dement W C 2011 Principles and practice of sleep medicine fifth edition

[2] Berry R B, Brooks R, Gamaldo C E, Harding S M, Marcus C, Vaughn B V *et al.* 2012 *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine* **176** 2012

[3] Stephansen J B, Olesen A N, Olsen M, Ambati A, Leary E B, Moore H E, Carrillo O, Lin L, Han F, Yan H, Sun Y L, Dauvilliers Y, Scholz S, Barateau L, Hogl B, Stefani A, Hong S C, Kim T W, Pizza F, Plazzi G, Vandi S, Antelmi E, Perrin D, Kuna S T, Schweitzer P K, Kushida C, Peppard P E, Sorensen H B, Jennum P and Mignot E 2018 *Nature Communications* **9** 1–15 ISSN 20411723 URL `http://dx.doi.org/10.1038/s41467-018-07229-3`

[4] Niculescu-Mizil A and Caruana R 2005 Predicting good probabilities with supervised learning *Proceedings of the 22nd international conference on Machine learning* pp 625–632

---

‡ EEG recordings of the left and right hemispheres are denoted with odd, respectively, even numbers in the international 10-20 electrode positioning [1].

[5] Goodfellow I 2018 *Deep Learning* vol 12

[6] Nobili L, De Gennaro L, Proserpio P, Moroni F, Sarasso S, Pigorini A, De Carli F and Ferrara M 2012 *Progress in Brain Research* **199** 219–232 ISSN 18757855

[7] Zhang X, Dong X, Kantelhardt J W, Li J, Zhao L, Garcia C, Glos M, Penzel T and Han F 2015 *Sleep and Breathing* **19** 191–195

[8] Rosenberg R S and Van Hout S 2013 *Journal of clinical sleep medicine* **9** 81–87

[9] Oord A v d, Li Y and Vinyals O 2019 *arXiv preprint arXiv:1807.03748* URL `http://arxiv.org/abs/1807.03748`

[10] Banville H, Chehab O, Hyvärinen A, Engemann D A and Gramfort A 2020 *Journal of Neural Engineering* 1–32 ISSN 23318422

[11] Hyvärinen A and Pajunen P 1999 *Neural networks* **12** 429–439

[12] Hyvarinen A, Sasaki H and Turner R E 2018 Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning *The 22nd International Conference on Artificial Intelligence and Statistics* vol 89 pp 859–868 URL `http://arxiv.org/abs/1805.08651`

[13] Zimmermann R S, Sharma Y, Schneider S, Bethge M and Brendel W 2021 *arxiv* URL `http://arxiv.org/abs/2102.08850`

[14] Guo C, Pleiss G, Sun Y and Weinberger K Q 2017 On calibration of modern neural networks *ICML* pp 1321–1330 ISSN 23318422

[15] Ulmer D and Cinà G 2020 Know your limits: Monotonicity & softmax make neural classifiers overconfident on OOD data *UAI* ISSN 23318422

[16] Krauss P, Metzner C, Joshi N, Schulze H, Traxdorf M, Maier A and Schilling A 2021 *Neurobiology of sleep and circadian rhythms* **10** 100064 URL `https://doi.org/10.1101/2020.06.25.170464`

[17] Kingma D P and Ba J 2014 *arXiv preprint arXiv:1412.6980*

[18] Buysse D J, Reynolds III C F, Monk T H, Berman S R and Kupfer D J 1989 *Psychiatry research* **28** 193–213

[19] Morin C M, Belleville G, Bélanger L and Ivers H 2011 *Sleep* **34** 601–608

[20] Snaith R P 2003 *Health and quality of life outcomes* **1** 1–4