Interpretation and Further Development of the Hypnodensity Representation of Sleep Structure

Interpretation and further development of the hypnodensity representation of sleep structure

IOP Publishing Physiol. Meas. 44 (2023) 015002 https://doi.org/10.1088/1361-6579/aca641

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
7 April 2022

REVISED
28 October 2022

ACCEPTED FOR PUBLICATION
25 November 2022

PUBLISHED
17 January 2023

PAPER

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

Keywords: sleep, hypnogram, hypnodensity, supervised learning, contrastive predictive coding, softmax

Abstract

Objective. The recently-introduced hypnodensity graph provides a probability distribution over sleep stages per data window (i.e. an epoch). This work explored whether this representation reveals continuities that can only be attributed to intra- and inter-rater disagreement of expert scorings, or also to co-occurrence of sleep stage-dependent features within one epoch. *Approach.* We proposed a simplified model for time series like the ones measured during sleep, and a second model to describe the annotation process by an expert. Generating data according to these models, enabled controlled experiments to investigate the interpretation of the hypnodensity graph. Moreover, the influence of both the supervised training strategy, and the used softmax non-linearity were investigated. Polysomnography recordings of 96 healthy sleepers (of which 11 were used as independent test set), were subsequently used to transfer conclusions to real data. *Main results.* A hypnodensity graph, predicted by a supervised neural classifier, represents the probability with which the sleep expert(s) assigned a label to an epoch. It thus reflects annotator behavior, and is thereby only indirectly linked to the ratio of sleep stage-dependent features in the epoch. Unsupervised training was shown to result in hypnodensity graph that were slightly less dependent on this annotation process, resulting in, on average, higher-entropy distributions over sleep stages ($H_{\text{unsupervised}} = 0.41$ versus $H_{\text{supervised}} = 0.29$). Moreover, pre-softmax predictions were, for both training strategies, found to better reflect the ratio of sleep stage-dependent characteristics in an epoch, as compared to the post-softmax counterparts (i.e. the hypnodensity graph). In real data, this was observed from the linear relation between pre-softmax N3 predictions and the amount of delta power. *Significance.* This study provides insights in, and proposes new, representations of sleep that may enhance our comprehension about sleep and sleep disorders.

1. Introduction

Hypnodensity graphs (Stephansen *et al* 2018) have recently been proposed as generalized representations of widely-used hypnograms in sleep medicine. While hypnograms provide one of the five sleep stages as defined by the American Academy of Sleep Medicine (AASM) (Berry *et al* 2012) for each 30 s epoch, a hypnodensity graph reveals a probability distribution over these five stages, possibly at a higher temporal resolution as well. Given that hypnodensity graphs show additional information compared to hypnograms, they could give insights in (yet) unexplained phenomena, and therefore have the potential to induce a paradigm shift in sleep medicine.

Despite the clinical possibilities, hypnodensity graphs 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 graph is generally predicted using a computer model. As a consequence,

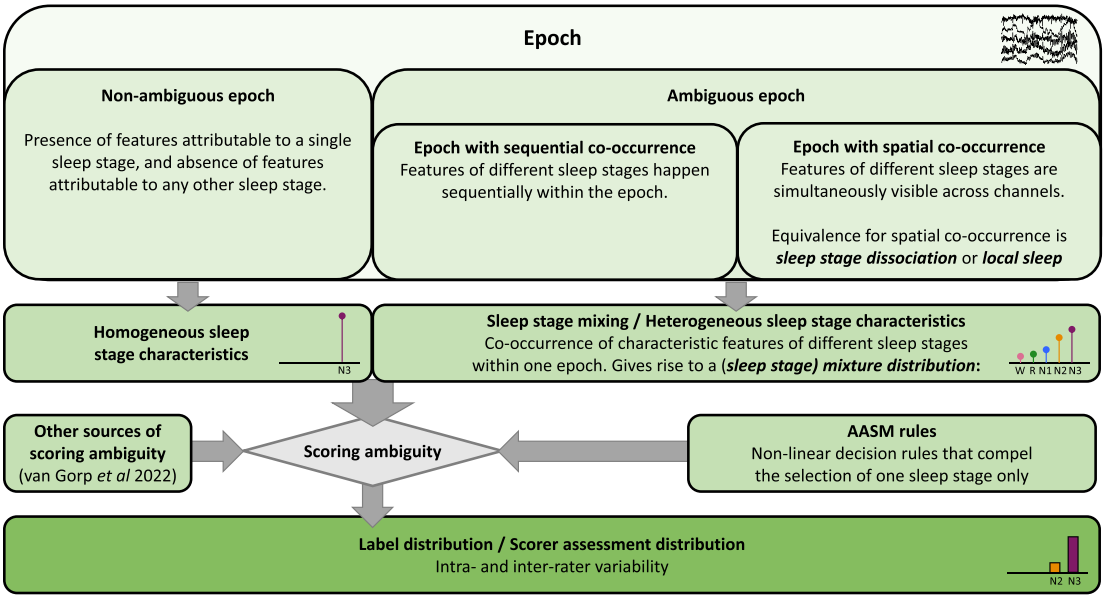

Figure 1. Explanation of the concept of scoring ambiguity. A PSG is scored epoch-based according to the AASM rules. These rules force the scorer to choose one sleep stage, but ambiguity remains within and across scorers due to epochs being ambiguous, or additional sources that lead to scoring ambiguity (van Gorp *et al* 2022). This ambiguity gives rise to a label distribution for each epoch. Icons relate to the ones used in figure 2 as well.

the exact relation between recorded data and the predicted probability distributions for each class (i.e. the hypnodensity graph) remained unclear so far.

The proposed machine learning model that predicts a hypnodensity graph (Stephansen *et al* 2018), is a supervised neural classifier, of which the final softmax-activated outputs are considered the hypnodensity graph. Earlier efforts from the machine learning community have already provided valuable insights about interpretation of such softmax probabilities (Niculescu-Mizil and Caruana 2005, Goodfellow 2018); 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. Indeed, (Bakker *et al* 2022) confirmed that their hypnodensity graphs, predicted by a supervised classifier, coincided with the scoring ambiguity present in the polysomnography (PSG) dataset that was scored by multiple scorers.

The authors that proposed the hypnodensity graph posed a similar conclusion (Stephansen *et al* 2018). However, they also linked the probability for a certain sleep stage (e.g. probability for N3) to the amount of evidence for this sleep stage that is present in the data (e.g. the proportion of slow waves in an epoch) (Stephansen *et al* 2018, page 4). The latter point of view elucidates upon an important implicit assumption that was made by the authors. Namely the assumption that the hypnodensity graph provides an indication about the sleep stage mixture distribution, which we define as the ratio with which characteristic features of different sleep stages are present in the epoch (see figure 1 for an overview about the relation between the mixture distribution and scoring ambiguity). If the aforementioned assumption holds, a hypnodensity graph may provide insights about local sleep phenomena (Nobili *et al* 2012), which exhibit characteristics of different sleep stages simultaneously.

However, given that the hypnodensity graph was also shown to reflect scoring ambiguity, it remains to be questioned whether co-occurrence of sleep stage-dependent features in an epoch is truly reflected in a hypnodensity graph. Moreover, it has not been investigated whether or how the proposed supervised training strategy and the use of the final softmax non-linearity affect the interpretation of the hypnodensity graph. The contributions of this study can be summarized as follows:

- We formulate a simplified signal model 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 graph.
- Given the absence of knowledge about the sleep stage mixture distribution in PSG data, we generate a synthetic dataset according to the proposed signal model and annotation model, to systematically investigate the interpretation of a hypnodensity graph, as a function of training strategy (supervised versus unsupervised) and final non-linearity (pre- versus post-softmax predictions) (section 5).

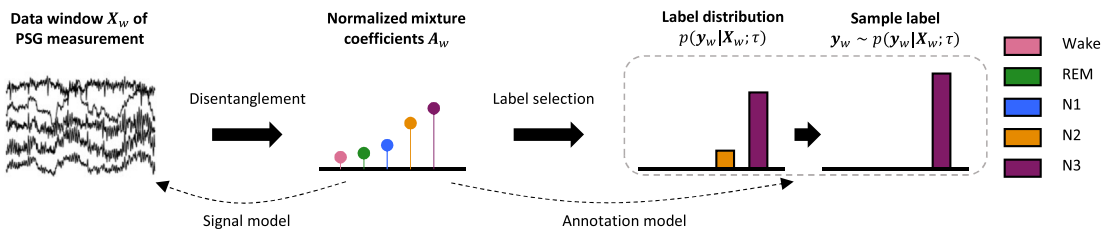

Figure 2. The data are assumed to originate from a signal model, while labels are generated from a generative annotation model. The signal model models all channels as a nonlinear mixture of some implicit (or hidden) classes that each characterize one of the sleep stages. The annotation model converts the contribution of each of these classes to one selected sleep stage. Due to ambiguity in expert annotations, we know that scoring is a stochastic process, which may be modelled as a conditional label distribution per window.

- We validate drawn conclusions from the synthetic experiments on real PSG recordings of healthy sleepers. To this end, we compare hypnodensity graphs 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 6).

2. Problem formulation and modelling

We aim to study the relation between raw PSG data and a predicted hypnodensity graph to get a better understanding of the interpretation of such a graph. A scored PSG recording has an AASM label annotation scored to each epoch, however, no information is generally available about the ratio with which sleep stage dependent-features are present in an epoch, i.e. the mixture distribution. As such, we introduce two (heavily simplified) models that describe time series data and their relation to expert annotations (see figure 2). This signal model (section 2.1) and annotation model (section 2.2) are used to generate synthetic data with which experiments are done in a controlled setup, after which conclusions are linked to real PSG data. Section 2.3 provides information on the hypnodensity-predicting model and its optimization.

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 s 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 nonlinear generative mixing process of C latent signals $\boldsymbol{s}_c^{(k)}$, with $1 \leqslant c \leqslant C$, where each signal aggregates typical characteristics/features associated with a specific sleep stage (or class).

In other words, each epoch is assumed to contain data which are a nonlinear spatial and temporal (over 30 s) accumulation of characteristics that are typical for certain sleep stages. Given five AASM-defined sleep stages, we assume these data to be generated from $C = 5$ latent signals that each represent one sleep stage (see figure 1). When manually scoring a PSG window, an expert (implicitly) determines how much of the visible features belong to either of the five stages (e.g. K-complexes belong to N2, delta waves belong to N3 etc). Based on rules (i.e. the AASM standard), the final sleep stage in subsequently determined (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:

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

where $\mathbf{S}^{(k)} \in \mathbb{R}^{C \times W \times l}$ contains the signals of all classes, $\tilde{\mathbf{A}}^{(k)} \in \mathbb{R}_{\geqslant 0}^{C \times W \times l}$ contains the corresponding time-varying amplitudes, h: $\mathbb{R}^{C \times W \times l} \rightarrow \mathbb{R}^{ch \times W \times l}$ is a nonlinear 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 nonlinear mixing, these normalized amplitudes are also called *mixture coefficients*. Figure 2 depicts the described signal model, and table A1 in appendix A summarizes the introduced notations and symbols.

2.2. Annotation model

Sleep stage annotations are in clinical practice assigned to 30 s epochs of data. We denote the w^{th} 30 s 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 $\tilde{\mathbf{A}}_w^{(k)} \in \mathbb{R}_{\geqslant 0}^{C \times l}$ and $\mathbf{A}_w^{(k)} \in \{\mathbb{R}_{\geqslant 0}^{C \times l}: \sum_c \mathbf{A}_w^{(k)} = 1\}$, 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 (Rosenberg and Van Hout 2013, Younes *et al* 2016). This scoring ambiguity can be caused by inherently ambiguous epochs (see figure 1), and the stochastic nature of human decision making (van Gorp *et al* 2022). To model this stochastic decision process, we model each expert annotation as a sample from a label distribution, being 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 label distribution yields:

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

where σ_τ denotes a tempered softmax function with temperature parameter $\tau \in \mathbb{R}_{\geqslant 0}$, and $\text{avg}_l(\cdot)$ 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 2 depicts the described annotation model.

For $\tau = 1$, the probability of selecting a class is linearly related to the mixture coefficient of that 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 all epochs would be unambiguous since all experts would always assign the same label to a given epoch, and inter- and intra-rater variability would 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 nonlinear decision boundaries are used (see figure 1). 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 features related to Wakefulness, the window should be assigned the Wake label. Parameter τ can thus be seen as a slider for the amount of scoring ambiguity that is present in the labelled dataset (high τ implies high ambiguity).

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

Stephansen *et al* (2018) propose to use a supervised neural classifier to predict a hypnodensity graph from PSG data. To this end, a classifier model p_m, parameterized by θ, makes a conditional prediction of class probabilities: $\hat{\boldsymbol{y}}_w^{(k)} \in \{\mathbb{R}^C_{\geqslant 0}\colon |\hat{\boldsymbol{y}}_w^{(k)}| = 1\}$, 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 $p_\text{d}(\mathbf{X},\,\boldsymbol{y})$. The optimization problem yields (omitting all k- and w-super/subscripts for clarity):

$$\begin{aligned}\theta^* &= \underset{\theta}{\text{argmax}}\{\mathbb{E}_{\hat{p}_\text{d}(\mathbf{X},\boldsymbol{y})}\log p_\text{m}(\hat{\boldsymbol{y}}|\mathbf{X};\,\theta)\}\\ &= \underset{\theta}{\text{argmin}}\,\{\text{D}_\text{KL}(\hat{p}_\text{d}(\boldsymbol{y}|\mathbf{X})||p_\text{m}(\hat{\boldsymbol{y}}|\mathbf{X};\,\theta))\},\end{aligned} \tag{3}$$

where $\hat{p}_\text{d}(\mathbf{X},\,\boldsymbol{y})$ is the approximation of the true data-generating distribution, and D_KL is the Kullback-Leibler (KL)-divergence between the empirical conditional data distribution $\hat{p}_\text{d}(\boldsymbol{y}|\mathbf{X})$ and the conditional distribution as trained by the model $p_\text{m}(\hat{\boldsymbol{y}}|\mathbf{X};\,\theta)$ (Goodfellow 2018, ch. 5). The full derivation of equation (3) can be found in appendix B.

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

$$\text{D}_\text{KL}(P||Q) = \sum_{c=1}^{C} P_c \log \frac{P_c}{Q_c}, \tag{4}$$

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 dataset 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.

We design the hypnodensity-predicting model as a feedforward neural network that comprises a convolutional encoder (including four Leaky ReLU-activated 1D convolutional layers, with pooling and dropout layers in between) and a nonlinear classifier, similar to the model proposed by Stephansen *et al* (2018).

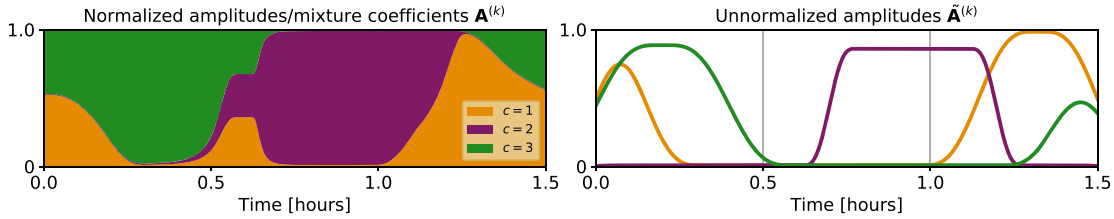

Figure 3. A sequence from the synthetic test set. Left: the normalized amplitudes that sum to one and serve as the mixture coefficients of the signals belonging to the three classes. Right: the corresponding unnormalized amplitudes of the three generated signals over time.

The convolutional encoder converts a data window $\mathbf{X}_w^{(k)}$ to a latent representation: $\boldsymbol{z}_w^{(k)} = \text{Enc}(\mathbf{X}_w^{(k)}) \in \mathbb{R}^F$, 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 $\hat{\boldsymbol{y}}_w^{(k)} = \sigma(\mathbf{W}\boldsymbol{z}_w^{(k)} + \boldsymbol{b})$, 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, Stephansen *et al* (2018) propose to omit this last step, and directly use the softmax output $\hat{\boldsymbol{y}}_w^{(k)}$, being the predicted hypnodensity graph of recording k for window w (i.e. $\hat{\boldsymbol{y}}^{(k)}$ entails the full hypnodensity graph belonging to recording k). Appendix C provides more details regarding the model architecture and training procedure.

3. Datasets

Scored epochs from a PSG measurement come without information regarding the underlying sleep stage mixture distribution in that epoch. As such, directly assessing whether information from the mixture distribution is visible in the hypnodensity graph is hampered. Neither can the effect of model design choices on this aspect be researched. As such, we first generate a synthetic dataset (section 3.1), according to the proposed signal model and annotation model from section 2, which enables controlled experiments. Conclusions drawn in the synthetic setup are, thereafter, validated on real PSG data. These data are described in section 3.2.

3.1. Synthetic data

We created a synthetic dataset according to the signal model as introduced in equation (1), and generated each channel in $\mathbf{X}^{(k)}$ as a nonlinear 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 sinusoidal signal of 90 min (discretized at 100 Hz), with a class-dependent frequency range, 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 were split into a training, validation and a hold-out test set of sizes 75, 25 and 100, respectively. Figure 3 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 equation (2). Appendix D.1 provides additional details about the generation of this dataset.

3.2. Polysomnography data

We used a dataset of nocturnal video-PSG recordings of 96 (60F, age = 36.0 ± 13.6) healthy sleepers (Appendix D.2 provides the definition of 'healthy sleeper' in this work), that were recorded according to the AASM recommendations (Berry *et al* 2012) in Sleep Medicine Center Kempenhaeghe Heeze, The Netherlands. Annotations were created by visual sleep staging on windows of 30 s, 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. 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[4]. For simplicity, the two EOG recordings were split in a similar fashion, even though these

[4] EEG recordings of the left and right hemispheres are denoted with odd, respectively, even numbers in the international 10–20 electrode positioning (Kryger *et al* 2011).

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. We randomly split the $K = 96 \times 2$ recordings in a training ($K = 150$), validation ($K = 20$) and hold-out test set ($K = 22$), while ensuring that the even and odd recording of the same patient were assigned to the same subset. The validation set was used to determine the iteration to stop the training of the model (i.e. at the iteration with the lowest validation loss), and to tune hyperparameters. The hold-out test set was used to evaluate the model's performance on unseen data. appendix D.2 provides more details about the dataset and the applied preprocessing.

4. Methodology

Using the synthetic data, we investigate whether the hypnodensity graph contains information regarding the mixture distribution in an epoch, or only displays the scoring ambiguity (see figure 1 for the difference). To this end, we first introduce two metrics, which are discussed in section 4.1. Second, the effect of the final nonlinear softmax activation, which is defined in section 4.2, is analysed. Third, to investigate the influence of supervised training on the interpretation of the hypnodensity graph, supervised training is compared to unsupervised training of the model. The methodology for unsupervised training is explained in section 4.3.

4.1. Evaluating hypnodensity graphs

A hypnodensity graph yields a (normalized) probability vector for each epoch w. As such, we can (in the synthetic setup) use the KL-divergence (see equation (4)) as a metric to compare this distribution with both the normalized amplitudes $\mathbf{A}_w^{(k)}$, and the label distribution $\hat{p}_{\mathrm{d}}(\boldsymbol{y}_w^{(k)}|\mathbf{X}_w^{(k)})$ used to generate corresponding labels for supervised training. In the synthetic setup, we explicitly defined this conditional label distribution according to equation (2), thus $\hat{p}_{\mathrm{d}}(\boldsymbol{y}_w^{(k)}|\mathbf{X}_w^{(k)}) \coloneqq p(\boldsymbol{y}_w^{(k)}|\mathbf{X}_w^{(k)};\, \tau)$. We define the following two metrics:

$$\mathrm{D}_{\mathrm{KL}}(\hat{p}_{\mathrm{d}}||\hat{\boldsymbol{y}}) \coloneqq \frac{1}{K}\sum_{k=1}^{K} \mathrm{median}_w\{\mathrm{D}_{\mathrm{KL}}(\hat{p}_{\mathrm{d}}(\boldsymbol{y}_w^{(k)}|\mathbf{X}_w^{(k)})||\hat{\boldsymbol{y}}_w^{(k)})\}, \tag{5}$$

$$\mathrm{D}_{\mathrm{KL}}(\mathbf{A}||\hat{\boldsymbol{y}}) \coloneqq \frac{1}{K}\sum_{k=1}^{K} \mathrm{median}_w\{\mathrm{D}_{\mathrm{KL}}(\mathrm{avg}_l(\mathbf{A}_w^{(k)})||\hat{\boldsymbol{y}}_w^{(k)})\}, \tag{6}$$

where median_w computes the median over the W windows.

4.2. Pre- versus 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 $\hat{\boldsymbol{y}}_w \in \{\mathbb{R}^C_{\geqslant 0}\colon |\hat{\boldsymbol{y}}_w| = 1\}$:

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

The effect of the non-linearity as introduced by using a softmax function as a final activation, is investigated by comparing pre-softmax predictions to post-softmax (i.e. hypnodensity) predictions. If the (implicit) model of real PSG data and its annotations indeed resemble the proposed models from sections 2.1 and 2.2, we hypothesize that the pre-softmax predictions have more tendency than the post-softmax counterparts, to reveal the (unnormalized) contributions of sleep stage-dependent characteristics in an epoch, i.e. the mixture distribution.

Due to the unnormalized nature of the pre-softmax predictions, KL-divergences can not be computed on these unnormalized vectors. Moreover, given that the sleep staging mixture distribution is also unknown in real PSG data, we seek an additional approach to draw conclusions about the two different type of models and their pre- and post-softmax predictions. It is known that slow wave power (positively) relates to the depth of sleep (Kryger *et al* 2011), and the AASM selection criterion for scoring N3 is based upon the amplitude of these slow waves (Berry *et al* 2012). 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- versus post-softmax) for N3, to the amount of slow wave (0.5–2 Hz) power in the frontal EEG lead (F3 or F4).

4.3. Supervised versus unsupervised encoding

Lastly, 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) (Oord 2019), a recently proposed framework for self-supervised contrastive learning, which has already been found useful to model EEG data (Banville *et al* 2020). The contrastive learning paradigm has shown to be able to invert the data-generating process, even in case of nonlinearly mixed signals (Hyvärinen *et al* 2018, Zimmermann *et al* 2021), and models slow features (Oord 2019), i.e. slowly varying data characteristics like the normalized amplitudes **A** in our signal model. 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 will more likely depend on the distribution of 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:

$$\begin{aligned} \mathcal{L} &= \frac{1}{P}\sum_{p=1}^{P} \mathcal{L}_p, \text{ with} \\ \mathcal{L}_p &= -\mathbb{E}_{\hat{p}_d(\mathbf{X})}\left[\log \frac{\exp(\boldsymbol{z}_{w+p}^T \mathbf{V}_p \boldsymbol{z}_w)}{\sum_{\boldsymbol{z}\in\mathcal{Z}_p''^{(k)}} \exp(\boldsymbol{z}^T \mathbf{V}_p \boldsymbol{z}_w)}\right], \end{aligned} \tag{7}$$

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.

5. Results on synthetic data

This section describes the results on synthetic data (see section 3.1). The synthetic setup enables analyses on the hypnodensity predictions, while knowing the ground-truth signal and label model, which are absent in real data.

5.1. Hypnodensity graph predictions

We start with an empirical investigation of the post-softmax (i.e. hypnodensity graph) predictions of the supervised model. To this end, four supervised models were trained with labels that had been generated from label distributions as given in equation (2), with varying values of $\tau = \{0^+, \frac{1}{4}, \frac{1}{2}, 1\}$ (higher τ implies higher scoring ambiguity). Table 1 shows the two KL-divergence metrics, as introduced in equations (5) and (6) (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 mixture distribution (i.e. 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 that the model aims to mimic the label distribution. The difference between the label distribution and the normalized amplitudes becomes more apparent for lower values of τ (figure 4(a)), i.e. when the labelling process becomes less ambiguous.

We additionally cross-compare the model predictions with label distributions with varying values for τ. Figure 5 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 optimization problem that is being minimized; equation (3) showed that a supervised classifier that is trained by likelihood maximization mimics the conditional label distribution.

5.2. Pre- versus post-softmax predictions.

Figure 6(a) shows the nonlinear 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

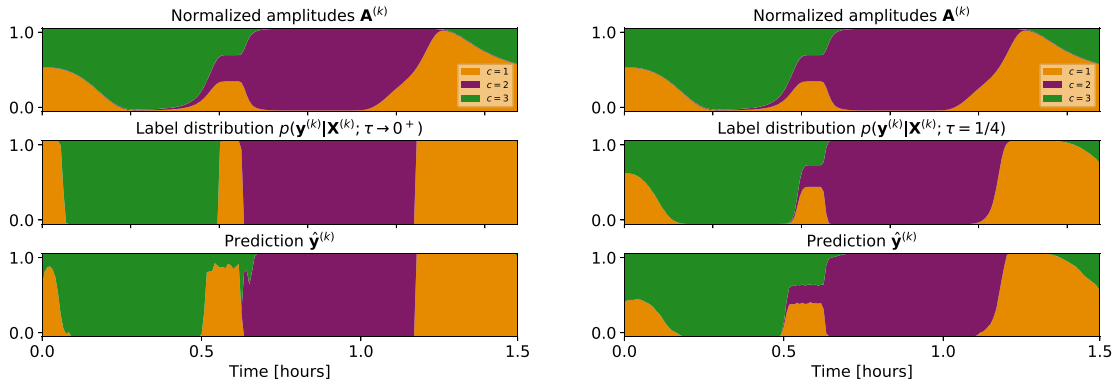

Figure 4. The normalized amplitudes (top), label distribution (middle) and supervised model prediction (bottom) of a representative sample from the synthetic test set. Predictions of models, trained with label distributions with (a) $\tau \to 0^+$, and (b) $\tau = \frac{1}{4}$ are shown. It can clearly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

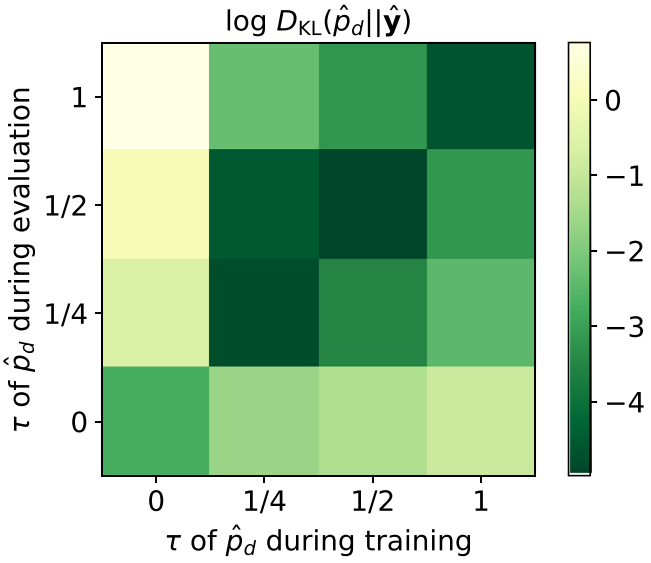

Figure 5. Cross-comparison of KL-divergences (in log-scale) between label distributions $p(\boldsymbol{y}|\mathbf{X}; \tau)$ with varying τ during evaluation, and models trained with different values of τ. Lower is better.

Table 1. KL-divergence between model predictions $\hat{y}$ 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, implying a better match of the former with the model prediction.

τ	$\mathrm{D}_{\mathrm{KL}}(\hat{p}_{\mathrm{d}}\|\|\hat{y})$	$\mathrm{D}_{\mathrm{KL}}(\mathbf{A}\|\|\hat{y})$
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

the y-axis denotes the corresponding post-softmax prediction. Each dot represents one epoch 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/unambiguous 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 one. For $\tau > 0$, i.e. when labelling is ambiguous and follows a stochastic process, the softmax outputs are not anymore pushed towards such binary decisions (middle and bottom row). The effect of

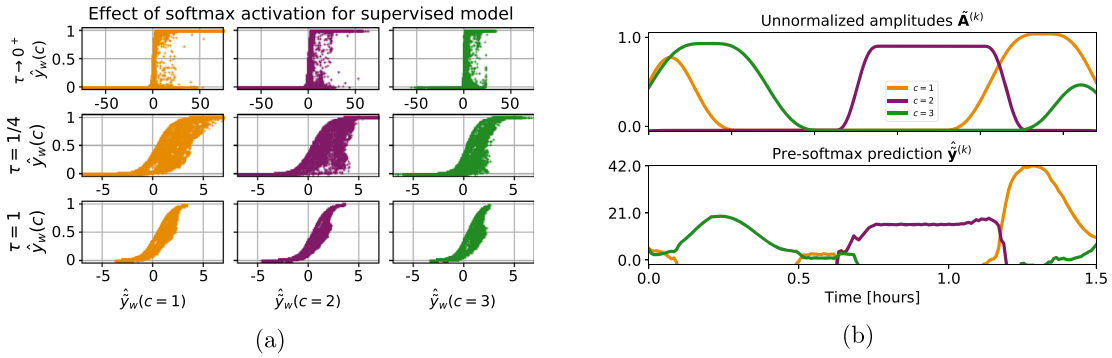

Figure 6. (a) The pre-softmax predictions (x-axis, per class) are highly nonlinearly related to the post-softmax predictions (y-axis) for a supervised model with $\tau \rightarrow 0^+$ in the label distributions used for training (top); the softmax pushes most class predictions towards 0 or 1. When training with a non-degenerate label distribution (i.e. $\tau > 0$), the softmax outputs are less discrete (middle, bottom). (b) The pre-softmax predictions (for $\tau \rightarrow 0^+$) (bottom) show dynamics that moderately resemble the unnormalized amplitudes (top).

the softmax activation in a supervised classifier thus depends on the annotation ambiguity in the dataset, expressed here as the entropy of the generative label distribution. In real PSG data, this entropy corresponds to the amount of sleep staging ambiguity in 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 figure 6(b), for the model trained with label distribution $p(\boldsymbol{y}|\mathbf{X}; \tau \rightarrow 0^+)$. Indeed, even though the post-softmax predictions were mimicking the label distribution (as was seen from figure 5(a)), the pre-softmax predictions $\hat{\tilde{\boldsymbol{y}}}^{(k)}$ are (moderately) resembling the unnormalized amplitudes, and thus might contain information regarding the mixture distribution.

5.3. Supervised versus unsupervised encoding

Table 2 shows the KL-divergence metrics of the (post-softmax) hypnodensity graph 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 opposite, see table 1), makes a prediction that is closer to the normalized amplitudes than to the label distribution. Figure 7 shows the predictions for $\tau \rightarrow 0^+$ (a) and $\tau = \frac{1}{4}$ (b) for the same test set example as for which the supervised predictions were shown in figure 5. It is clearly visible that the unsupervised model's post-softmax prediction is less dependent on the value of τ used for training the classifier, compared to the supervised model. This can be explained by the fact that the classifier is of low capacity (only a linear mapping with a softmax activation), such that it is unable to perfectly fit the label distribution. As a result, the predictions are closer to data-characteristics (i.e. the mixture distribution), rather than label-characteristics.

5.4. Interaction effect between softmax and (un)supervised training

In figure 6(a) 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 8(a) shows that this effect was much less apparent 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 figure 6(a)). The pre-softmax predictions of the unsupervised model seem to be slightly closer to the true unnormalized amplitudes, as seen from figure 8(b), compared to this prediction of the supervised model, as seen in figure 6(b).

6. Results on polysomnography data

Results on synthetic data showed that hypnodensity graphs predicted by a supervised neural classifier revealed label characteristics, i.e. the generative label distribution. In the context of sleep recordings this corresponds to predicting the probability that an epoch would have been labelled as one of the different sleep stages, by the expert(s) that labelled the dataset used for training the model. Mainly the nonlinear softmax activation was shown to play a large role in creating a hypnodensity graph that displays the label distribution. The pre-softmax predictions, on the other hand, seemed to reflect data characteristics, i.e. the unnormalized contributions of the different classes in the mixture distribution, thanks to the lack of nonlinear conversion from the softmax

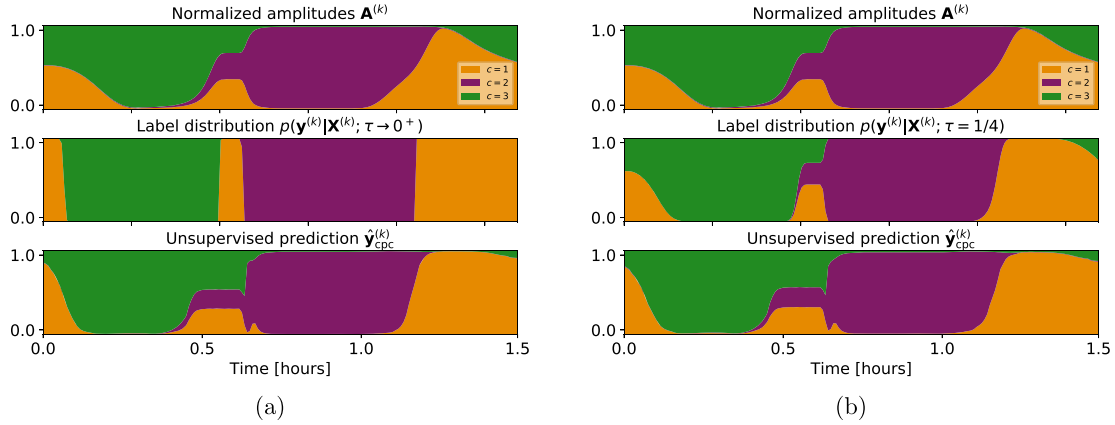

Figure 7. The normalized amplitudes (top), label distribution (middle) and CPC model prediction (bottom) of a representative sample from the test set of the synthetic dataset. Predictions for which the classifier was trained with label distributions with (a) $\tau \to 0^+$, and (b) $\tau = \frac{1}{4}$ are shown. The predictions are much less influenced by the value of τ, as compared to the supervised model, displayed in figure 5, and therefore show more tendency to predict the normalized amplitudes.

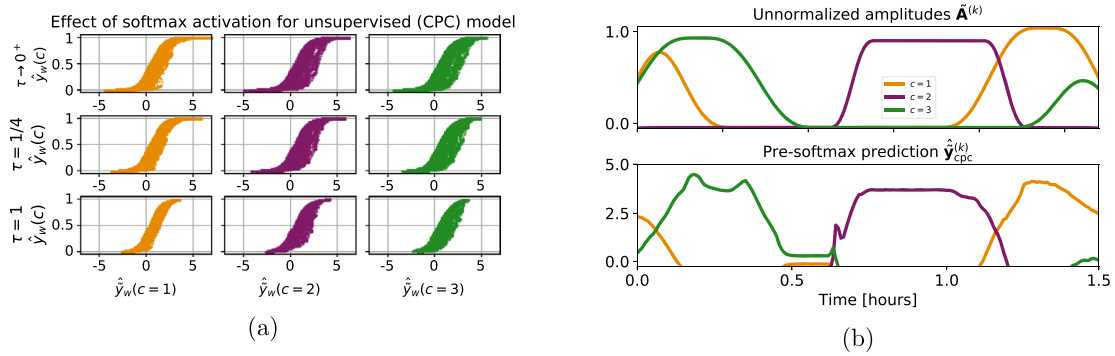

Figure 8. (a) Using unsupervised encoding, the operating range, and thus the nonlinear effect of the softmax function is almost independent of the value of τ used in the label distribution during training the classifier. (b) The pre-softmax predictions (for $\tau \to 0^+$) (bottom) show dynamics that slightly better resemble the unnormalized amplitudes (top) than the pre-softmax predictions of the supervised model (shown in figure 6(b)).

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	$\mathrm{D}_{\mathrm{KL}}(\hat{p}_{\mathrm{d}}\|\|\hat{\boldsymbol{y}})$	$\mathrm{D}_{\mathrm{KL}}(\mathbf{A}\|\|\hat{\boldsymbol{y}})$
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

normalization. This effect seemed even more apparent for the unsupervised model thanks to the lower influence of labels during training. In this section we perform similar experiments on real PSG data, and compare the effects to the aforementioned conclusions from the synthetic setup.

6.1. Hypnodensity graph prediction by a supervised model

Figure 9-middle shows an example of a hypnodensity graph from a representative recording of the test set, predicted by the supervised model. For reference, the top row shows the hypnogram as annotated by the sleep technician. Additionally, figure 10 plots the pre- versus post-softmax predictions for all epochs of all recordings in the test set, separated per sleep stage. From the synthetic case it was seen that for $\tau \to 0^+$, the softmax operated in a highly nonlinear regime for the supervised model (see figure 6(a)), which caused non-smooth boundaries at

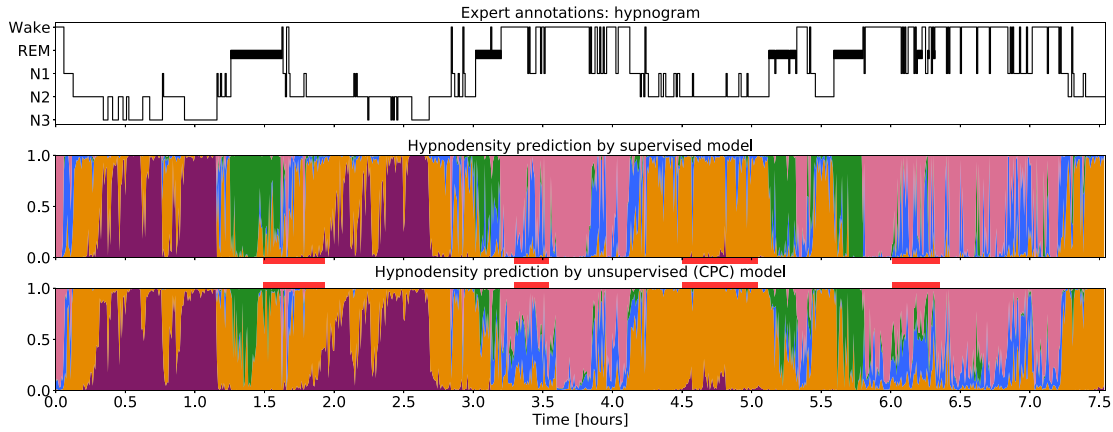

Figure 9. The hypnogram as annotated by a sleep technician (top), and predicted hypnodensity graph by the supervised (middle) and unsupervised model (bottom). The general trend looks similar, but differences are visible (indicated with the red bars), e.g. the unsupervised model in general shows smoother transitions. Note that the unsupervised model does not just predict a smoothened version of the supervised prediction: hard transitions (e.g. at 1.2 h) can still be predicted as well.

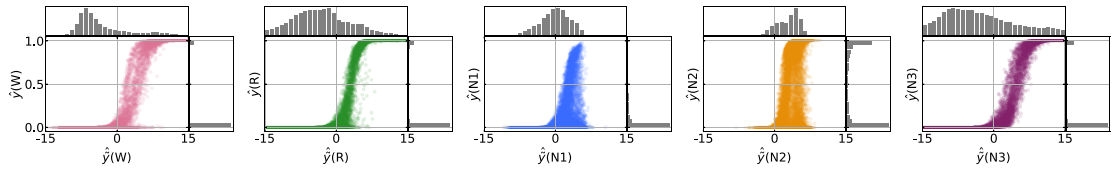

Figure 10. Scatter plots that plot the pre-softmax against the post-softmax predictions for the supervised model. Each dot is an epoch from the full test set. Given the smooth shapes the manual sleep staging process must have exhibited sleep stage scoring ambiguity (i.e. $\tau > 0$).

class transitions (see figure 5(a)). Figure 10, on the other hand, shows a more smooth effect of the softmax non-linearity for real PSG data, which implies 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 amount of ambiguity, which is in line with the known imperfectness of manual sleep stage scoring (Younes *et al* 2016). It should thus be realized that a hypnodensity graph, predicted by a supervised model, only exhibits non-abrupt sleep stage transitions thanks to the fact that sleep staging is ambiguous, and the label distribution (or scorer assessment distribution as called by Stephansen *et al* (2018)) exhibits non-zero entropy thanks to the presence of inter- and intra-rater disagreement.

6.2. Supervised versus unsupervised hypnodensity graphs

In the synthetic setup, unsupervised training, as opposed to supervised training, was found to provide 'hypnodensity graphs' that exhibited more information about the mixture distribution in the epoch (figures 5 versus 7). Figure 11(a) plots the hypnodensity graph probabilities of the supervised model against those of the unsupervised model, sorted per sleep stage, for all epochs in the test set (each dot denotes one epoch). The histograms on the axes of the scatter plots show that the general distributions look very similar, however given the present off-diagonal scatter points it can be seen that differences do exist between the hypnodensity graphs predicted by both models. This difference in predicted probabilities was also reflected in average entropy of probabilities for all epochs in the test set, which was found to be $H = 0.29 \pm 0.04$ for the supervised model, and $H = 0.41 \pm 0.06$ for the unsupervised model.

Qualitative differences between the hypnodensity graph predicted by the unsupervised model (Figure 9-middle), and the one predicted by the supervised model (figure 9-bottom) can indeed be noted, although the general trend looks similar, of which 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 did not show these low contributions. Occasions where the hypnogram showed rapid transitioning behavior (e.g. around 6.2 h), 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

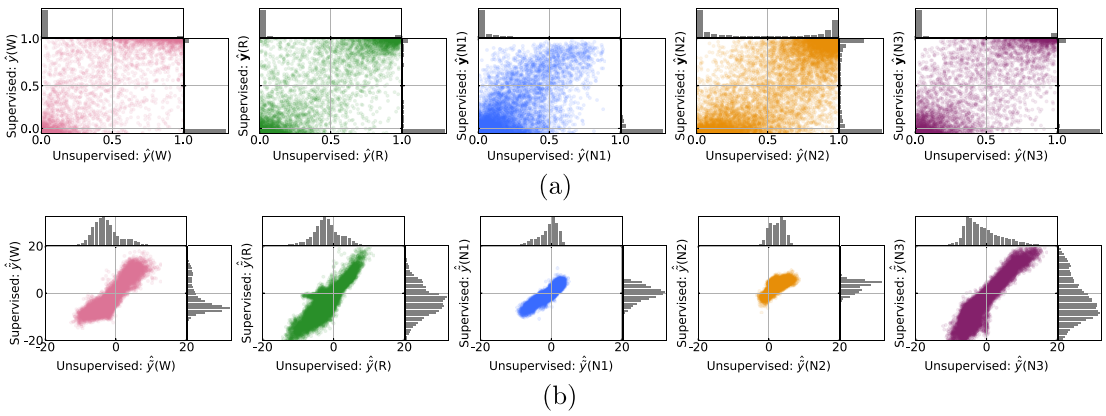

Figure 11. Scatter plots that compare the post-softmax (a), and pre-softmax (b) predictions of the supervised (y-axes) and unsupervised model (x-axes). Each dot is a window from the full test set. The pre-softmax predictions of both models seem better aligned, than the post-softmax hypnodensity graph predictions.

using the annotated hypnogram that also contains these (discrete) switches between sleep stages. 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 h), so it can not simply be considered a smoothened version of the supervised prediction.

Figure E1(b) in appendix E shows the pre- versus post-softmax scatter plots per sleep stage of the unsupervised model, which show similar dynamics as the ones from the supervised model (as visualized in figure 10 and figure E1(a)). In the synthetic setup we found that similar looking plots for both models were only found in case $\tau > 0$ (i.e. when scoring ambiguity is present). As such, given the resemblance of the plots between both models, it confirms our earlier observation that the label distribution in our PSG dataset exhibited scoring ambiguity (i.e. it is non-zero entropy), facilitating non-zero entropy hypnodensity graph predictions by the supervised model. Finally, even though automatic sleep staging using hypnodensity graphs is not the goal of this research, some additional analyses on classification performance of both models are presented in appendix F.

6.3. Interaction effect between softmax and (un)supervised training

To dive further into the effect of supervised training on the hypnodensity graphs, we plot both the post-softmax probabilities and pre-softmax predictions of both the supervised and unsupervised model against each other in figures 11(a), and (b), respectively. We already observed a difference between the probabilities predicted by both models. Interestingly, the pre-softmax predictions showed to be much better linearly aligned between the models, but due to the difference in range, the nonlinear effect of the softmax operated differently, resulting in slightly different hypnodensity graphs. Based on earlier conclusions on synthetic data, the difference between the hypnodensity graphs of both models is expected to be caused by the fact that the unsupervised model is less influenced by the annotations and therefore has more tendency to show attributes of the sleep stage mixture distribution. However, information about this mixture distribution is expected to be even more present in the pre-softmax predictions, which will be investigated in the next paragraph.

6.4. N3 prediction versus slow wave power

Figure 12(a) plots both pre- and post-softmax predictions for N3 of both models against the slow wave power for all epochs 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 12(b) 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 figure 9. 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 figure 12(a)-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 episodes as a consequence of movement artifacts. Figure 12(b) indeed showed this effect at 3.7 h, where high slow wave power was present, but the data were annotated as wake (seen from the hypnogram in figure 9-top).

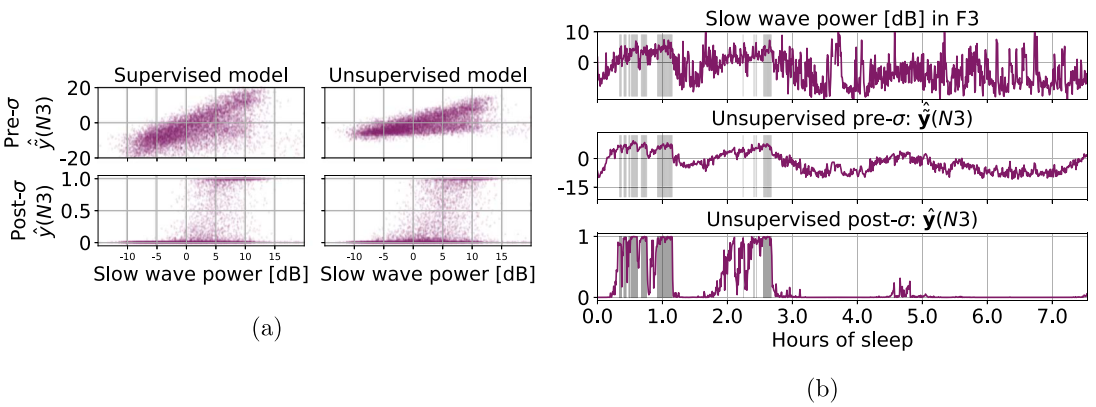

Figure 12. (a) The frontal EEG slow wave power against the predicted pre-softmax values (top), and post-softmax probabilities (bottom), over the full test set. The pre-softmax predictions correspond much better with slow wave power for both models. (b) Illustrative example that shows the pre- and post-softmax 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.

7. Discussion

In this work, we researched the interpretation of the recently proposed hypnodensity graph of a PSG recording (Stephansen *et al* 2018), being a probability distribution over sleep stages throughout the night. We investigated whether such a graph purely reflects scoring ambiguity, or whether it also contains information about the sleep stage mixture distribution in an epoch, being the ratio of sleep stage-dependent features that are present. We, moreover, analysed whether the answer to this question depended on training strategy and/or the effect of the nonlinear softmax function.

We proposed a synthetic dataset that comprised nonlinearly mixed measurements of signals belonging to different classes, analogous to fictitious sleep stages, labelled according to a conditional label distribution (see sections 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 conclusions of this work and future work. We, moreover, provide a future perspective on the use of hypnodensity graphs in the sleep clinic.

7.1. Discussion of results

Both theoretical analyses (section 2.3 and appendix B) and empirical evidence (section 5.1) showed that a hypnodensity graph, predicted by a supervised classifier, reveals the probabilities with which an epoch was assigned to either of the classes by the expert(s) that annotated the dataset used for training the model. This finding is in line with earlier observations that stated that the hypnodensity representation resembled the inter-rater disagreement across multiple scorers that annotated the PSG dataset (Stephansen *et al* 2018, Bakker *et al* 2022). Nevertheless, one should take into account that a supervised classifier only mimics the label distribution under the assumptions that the the model is evaluated on a test dataset that entails the same statistics as the training set, and that the model exhibits the 'right amount' of capacity. In other words, models that are evaluated on an out-of-distribution dataset, or models that are too small exhibit a large amount of epistemic uncertainty (van Gorp *et al* 2022), which increases the average entropy of the hypnodensity graph. 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, creating a lower-entropy hypnodensity graph. In the machine learning community, such models are known as uncalibrated models (Guo *et al* 2017, Ulmer and Cinà 2020). 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 hypnodensity graphs (see figure 9-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 ambiguity). This conclusion is in line with the finding that manual sleep staging yields intra-rater disagreement (Younes *et al* 2016). It can thus be concluded that a supervisedly-predicted hypnodensity graph 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 graph,

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 graph 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 graph, researchers may not want to rely on one dataset that is annotated by, e.g. inexperienced scorer(s), as it might highly influence the graphs.

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 amount of ambiguity across the expert annotations in the dataset (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 graphs on real PSG data exhibited higher entropy, as compared to these graphs 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 graph that is less influenced by the exact amount of ambiguity in the scorings of the specific dataset, as compared to models that that are trained in a supervised fashion. An additional advantage of unsupervised training, is the fact that it is less label-hungry. The full encoder can be trained on large unlabelled datasets, where after the classifier can be trained on a relative small labelled dataset.

When considering the difference between pre-softmax and post-softmax model predictions, the following finding was made. 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 related to the mixture distribution, which were more smooth over time than their post-softmax counterparts. On real PSG data the pre-softmax predictions for deep sleep N3 were shown to corresponds much better to the slow wave power, than their post-softmax counterparts. This finding informs us that we can not simply assume that the hypnodensity graph displays the amount of evidence present in an epoch for each of the sleep stages, as suggested by Stephansen *et al* (2018). It does, however, display the label distribution, which is thus likely a nonlinear reflection of the sleep stage mixture distribution. As such, if one is interested in the mixture distribution of an epoch, i.e. the ratio of the amount of characteristic features per sleep stage, considering the pre-softmax predictions might be more valuable as it does not include the nonlinear softmax conversion. However, the value of a pre-softmax class 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.

7.2. 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. This difference taught us that the low-capacity classifier still pushed the post-softmax predictions of the CPC model towards the label distribution. When interested in the sleep stage mixture distribution, 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, Stephansen *et al* (2018) observed a more smooth hypnodensity representation when using memory, implemented as a long short-term memory 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 Stephansen *et al* (2018), it will likely improve smoothness of the hypnodensity graphs 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 used a convolutional encoder to extract features in a data-driven fashion. Automatic feature extraction, as opposed to manual feature extraction, has the advantage of making a model sensor-agnostic. In other words, when using other type of sensors, a model can be retrained, but does not require new design considerations. We did not investigate the influence of using such automatic feature extraction in this work. Moreover, the influence of design choices like the data window length (which was fixed to 30 s), and the number and type of measurement channels used was also not considered here. 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, Krauss *et al* (2020) depicted (supervised) hypnodensity graphs, 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 graphs yielded higher entropy than the presented (supervised) hypnodensity graph in this work and by Stephansen *et al* (2018). 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 (nonlinear) ICA under our signal model, may affect the hypnodensity graph as well. Research that compares hypnodensity graphs, 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 channels than the conventional PSG recording. When not only the number of channels is compromised, but also the measurement modality is altered, the drawn conclusions should be taken with care. The measured signal(s) should in all cases contain features that allow discrimination between the different sleep stages. If such information is not present in the given sensing modality, we may expect hypnodensity predictions that exhibit near-to uniform probability distributions over the classes, providing low amounts of actual information.

7.3. Clinical implications of hypnodensity graphs

As mentioned in the previous section, certain factors and design choices may be expected to influence the hypnodensity graph. Only once the community fully understands the influences of such choices on a predicted hypnodensity graph, we may start thinking about standardizing this sleep representation into the clinic.

We do foresee several use cases for hypnodensity graphs. First, using them over hypnograms has the advantage of being less dependent on the specific choices made by one sleep expert. Given non-perfect agreement that occurs both within and across human scorers, ambiguity about sleep stages always exists. This ambiguity is, however, not reflected in a hypnogram, while a hypnodensity graph, on the other hand, does show it and may, therefore, provide additional insight about the PSG recording of the patient and its (ambiguous) relation to AASM labels.

Second, a hypnodensity graph could be used as a decision support system for sleep experts that score a hypnogram. In this situation, it is good to realize that the ambiguity that the hypnodensity graph exhibits, may not in all cases be taken away by the expert. This is the case because the probabilities that are visualized actually display the amount of ambiguity the expert(s) (on average) would have about this epoch (at least when the model is evaluated on a PSG recording from a patient that resembles the training population). Nevertheless, the expert is forced to make a choice, which might give us the (possibly false) impression that the final choice is the 'ground-truth' sleep stage with 100% certainty, while in reality the same expert could possible have selected a different sleep stage when you had asked at another moment in time (Younes *et al* 2016). For the interested reader, a further discussion on sources that can be the underlying cause of this effect can be found in van Gorp *et al* (2022).

Third, given the fact that the hypnodensity graph is likely a nonlinear reflection of the mixture of sleep stage-dependent features in an epoch, it may be valuable for research about patient populations that suffer from local sleep phenomena or sleep-wake dissociations (Nobili *et al* 2012). It was already shown to exhibit different patterns for patients with narcolepsy, as compared to healthy controls (Stephansen *et al* 2018). Nevertheless, it should thus be taken into account that there is no one-to-one relation between the actual sleep stage mixture distribution in an epoch, and the predicted hypnodensity graph.

Independent of the final use case of hypnodensity graphs, it is important to note that a model that is aimed to be clinically validated for predicting the graphs, should be trained with a dataset that includes a wide variety of different patients and disorders. In this work we only considered healthy sleepers, and we made a random sub-split into a training, validation and test set, therewith not taking into account equal age ranges or gender balance across the three sets. Moreover, the corresponding hypnograms should be scored by different experts to prevent biased predictions. In this work we used a dataset that was scored by one sleep expert only. We would, thus, like to stress that steps still need to be taken towards clinical validation of hypnodensity-predicting models. This work is solely the start towards a better understanding of this new representation of sleep.

8. Conclusion

In this work, we investigated whether the recently-proposed hypnodensity graph (Stephansen *et al* 2018) of a PSG recording only displays the scorer assessment distribution, or also exhibits information on co-occurrence of sleep stage-dependent characteristic features in an epoch. The following conclusions could be drawn: a hypnodensity graph, predicted by a supervised neural classifier, reveals the label distribution from which sleep stages were (implicitly) drawn/assigned during manual sleep staging by the sleep expert(s) that annotated the dataset. In other words, it reflects the ambiguity 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 scoring ambiguity (i.e. the value of τ in our model). A hypnodensity graph predicted by an unsupervised model, on the other hand, was shown to be less dependent on this amount of ambiguity, and might therefore provide a more robust

hypnodensity graph. Potentially clinically-relevant information on co-occurrence of sleep stage-dependent features in an epoch were shown to be only nonlinearly present in the hypnodensity graph due to the final nonlinear softmax activation (both for supervised and unsupervised training). The pre-softmax class predictions, on the other hand, showed to have a more linear relation with the sleep stage mixture distribution.

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 graphs. 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.

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

Symbol	Domain	Meaning
C	$\mathbb{N}$	Number of classes, indexed with $1 \leqslant c \leqslant C$
W	$\mathbb{N}$	Number of non-overlapping 30 s windows in a recording, indexed with $1 \leqslant w \leqslant W$
K	$\mathbb{N}$	Number of recordings, indexed with $1 \leqslant k \leqslant K$
l	$\mathbb{N}$	Number of samples in one 30 s window
$L \coloneqq W \times l$	$\mathbb{N}$	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 s 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
$\tilde{\mathbf{A}}^{(k)}$	$\mathbb{R}_{\geqslant 0}^{C \times W \times l}$	Unnormalized amplitudes of classes in recording k
$\mathbf{A}^{(k)}$	$\{\mathbb{R}_{\geqslant 0}^{C \times W \times l} \colon \sum_c \mathbf{A}^{(k)} = 1\}$	Normalized amplitudes of classes in recording k
$\tilde{\mathbf{A}}_w^{(k)}$	$\mathbb{R}_{\geqslant 0}^{C \times l}$	Unnormalized amplitudes of classes in window w of recording k
$\mathbf{A}_w^{(k)}$	$\{\mathbb{R}_{\geqslant 0}^{C \times l} \colon \sum_c \mathbf{A}_w^{(k)} = 1\}$	Normalized amplitudes of classes in window w of recording k
$\tilde{\boldsymbol{a}}_c^{(k)}$	$\mathbb{R}^{L}$	Unnormalized amplitudes of class c in recording k
$\boldsymbol{a}_c^{(k)}$	$\mathbb{R}^{L}$	Normalized amplitudes of class c in recording k
$\mathbf{Y}^{(k)}$	$\{0,1\}^{C \times W}$	One-hot embeddings of expert class labels of recording k
$\boldsymbol{y}_w^{(k)}$	$\{0,1\}^{C}$	One-hot embedding of expert class label for window w of recording k
$\hat{\boldsymbol{y}}^{(k)}$	$\{\mathbb{R}_{\geqslant 0}^{C \times W} \colon \sum_c \hat{\boldsymbol{y}}^{(k)} = 1\}$	Post-softmax predictions of recording k
$\hat{\boldsymbol{y}}_w^{(k)}$	$\{\mathbb{R}_{\geqslant 0}^{C} \colon \sum_c \hat{\boldsymbol{y}}_w^{(k)} = 1\}$	Post-softmax prediction for window w of recording k
$\hat{\hat{\boldsymbol{y}}}_w^{(k)}$	$\mathbb{R}^{C}$	Pre-softmax prediction for window w of recording k

Appendix B. Multi-class likelihood optimization

Likelihood maximization of a multi-class classification model, is equivalent to minimizing the 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 $p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\,\theta)$. Here we derive this equivalence.

Using the monotonicity and translation invariance of the argmax and argmin-functions, we can rewrite equation (3) as follows:

$$\begin{aligned}
\theta^* &= \underset{\theta}{\operatorname{argmax}} \{\mathbb{E}_{\hat{p}_{\mathrm{d}}(\mathbf{X},\boldsymbol{y})} \log p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\,\theta)\} \\
&= \underset{\theta}{\operatorname{argmin}} \{-\{\mathbb{E}_{\hat{p}_{\mathrm{d}}(\mathbf{X},\boldsymbol{y})} [\log p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\,\theta)]\} \\
&= \underset{\theta}{\operatorname{argmin}} \{\mathbb{E}_{\hat{p}_{\mathrm{d}}(\mathbf{X},\boldsymbol{y})} [\log \hat{p}_{\mathrm{d}}(\boldsymbol{y}|\mathbf{X}) - \log p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\,\theta)]\} \\
&= \underset{\theta}{\operatorname{argmin}} \{\mathrm{D}_{\mathrm{KL}}(\hat{p}_{\mathrm{d}}(\boldsymbol{y}|\mathbf{X}) || p_{\mathrm{m}}(\hat{\boldsymbol{y}}|\mathbf{X};\,\theta))\}.
\end{aligned} \tag{B1}$$

Appendix C. Encoder architecture and training details

The architecture of the encoder followed standard practice in supervised classification model design (Goodfellow 2018). Enc(·) 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) versus 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 4.3), 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 equation (7), and batches of size 64. The Adam optimizer with default settings (Kingma and Ba 2014) 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 iterations, where one iteration defined one push trough of each data window in the training set. The classifiers trained after CPC encoding, were maximally trained for 100 iteration. In each experiment, the model with the lowest validation loss was finally selected. All experiments were run with the same seed for randomization.

Appendix D. Datasets

D.1. Synthetic data

To create a synthetic dataset, the signal model as introduced in equation (1) was used. Each channel ($ch = 3$) in $\mathbf{X}$ was modelled as a nonlinear 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 $\tilde{\mathbf{A}}^{(k)} \in \mathbb{R}_{\geqslant 0}^{C \times W \times l}$ 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 $c \in \{1, 2, 3\}$:

$$\boldsymbol{s}_c^{(k)}[n] = \sin\left\{2\pi\left(\frac{f_c + u\left[-\frac{1}{2}, \frac{1}{2}\right]}{f_s}\right)[n] + 2\pi u[0,\ 1]\right\}, \tag{D1}$$

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 s, 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:

$$\tilde{\boldsymbol{a}}_c^{(k)}[n] = \frac{\text{Hanning}_\nu \odot \text{sw[n; } \Phi]}{|\text{Hanning}_\nu|}, \tag{D2}$$

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

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 nonlinear 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 3 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.

D.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_2019_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 (mean = 36.0, std. dev. = 13.6). Subjects with the following criteria were considered non-healthy sleepers, and therefore excluded for participation: (1) any diagnosed sleep disorder, (2) a Pittsburgh Sleep Quality Index (Buysse *et al* 1989) $\geqslant 6$, or Insomnia Severity Index (Morin *et al* 2011) > 7, (3) indication of depression or anxiety disorder measured with the Hospital Anxiety and Depression Scale (Snaith 2003) (score > 8), (4) pregnancy, (5) shift work, (6) use of any medication except for birth control medicine, and (7) presence of clinically relevant neurological or psychiatric disorders or other somatic disorders that could influence sleep.

Visual sleep staging on epochs of 30 s was performed according to AASM criteria (Berry *et al* 2012) 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[5]. 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 Stephansen *et al* (2018), 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.

Appendix E. Additional plots for unsupervised training

Figure E1 shows the pre-softmax predictions against the post-softmax predictions for both the supervised (a) and unsupervised (b) model. From the synthetic experiments it was found that, when both models show similar pre-post softmax scatter plots, scoring ambiguity was present (modelled with $\tau > 0$ in the synthetic case). Also, in this situation of present scoring ambiguity, the supervised model showed smooth scatter plots, like the ones we see in figure E1(a). As such, we conclude that the manual sleep staging process in our PSG dataset (that only exhibited one scorer) must have exhibited sleep stage scoring ambiguity, i.e. intra-rater disagreement.

[5] EEG recordings of the left and right hemispheres are denoted with odd, respectively, even numbers in the international 10–20 electrode positioning Kryger *et al* 2011.

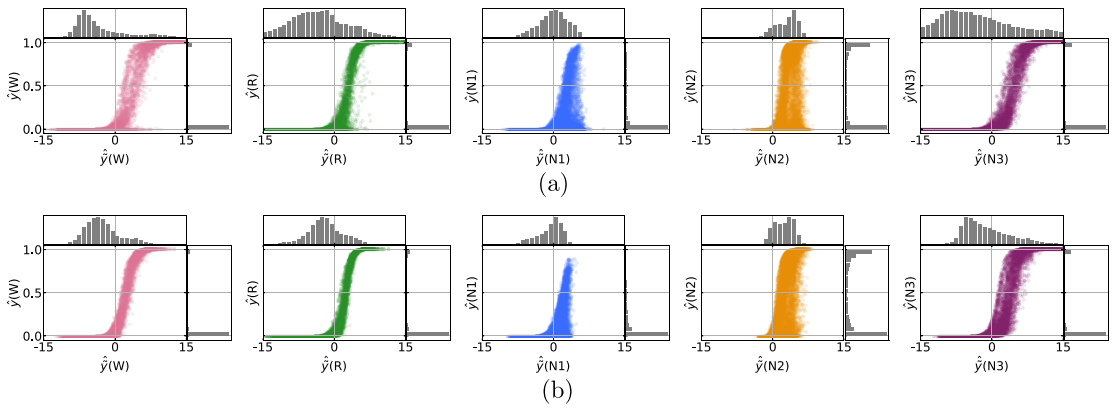

Figure E1. Scatter plots that plot the pre-softmax against the post-softmax predictions for the supervised model (a) and the unsupervised model (b). Each dot is an epoch from the full test set. Given the smooth shapes for the supervised model, and the very similar shapes between both models, the manual sleep staging process must have exhibited sleep stage scoring ambiguity (i.e. $\tau > 0$).

Interestingly, the softmax effect seems different across sleep stages, which can best be seen from the vertical histograms that show the post-softmax distributions (figure E1(a)). For example, N1 is never 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.

Appendix F. Analyses on classification performance

The hypnodensity graph was originally proposed to be a richer representation than the hypnogram, and this research was concerned with investigating its interpretation. Nevertheless, a hypnodensity graph can always be converted again to a hypnogram by selecting the sleep stage with the highest probability for each epoch in the night. Doing this, the predicted hypnogram can be compared to the annotations by means of Cohen's kappa κ Cohen (1960). Note that achieving the highest possible classification performance is thus not the goal of this study. This appendix only serves to provide the reader with a complete overview of all quantitative results.

Classification performance on the full PSG test set for the supervised, respectively unsupervised model, was found to be $\kappa = 0.80 \pm 0.05$ and $\kappa = 0.76 \pm 0.07$ (mean $\pm$ std. dev. across nights). As also mentioned in section 6.2, the average entropy of predictions for all test set epoch was found to be $H = 0.29 \pm 0.04$ for the supervised, and $H = 0.41 \pm 0.06$ for the unsupervised model. We are interested to see whether a relation exists between the probability of the highest class, and whether or not the epoch was classified correctly, compared to the expert annotation. As such, we analyse the highest class probability of epochs that were classified correctly versus epoch that got an incorrect sleep stage prediction.

Figure F1(a) shows the boxplots of the highest class probability for correct versus incorrectly classified epoch for both the supervised (left) and unsupervised (right) model. Indeed, epoch that were classified correctly tended to have a higher maximum class probability in the hypnodensity graph.

We may be interested in using the entropy of the hypnodensity graph as a measure for determining whether a sleep stage prediction may be trusted or not. In this context, we may define a *True postive* as a decision in which the user trusted the model's predicted sleep stage while the model was indeed correct. Similarly, a *False positive* can be defined as a situation in which the model's prediction was trusted, while it was incorrect.

The receiver operating characteristic (ROC) curve for both models is shown in figure F1(b). The area under the curve (AUC) for the supervised, respectively unsupervised model is AUC = 0.82 and AUC = 0.81. The hypnodensity graphs of the supervised model thus have slightly better capability to serve as a thrusthworthiness measure for automated sleep staging or as a decision support system for guided manual sleep staging.

In the situation of serving as a decision support system, it is good to realize that the ambiguity that the hypnodensity graph exhibits, may not in all cases be taken away by the expert, since the hypnodensity actually reflects ambiguity that is apparently present across (and within) experts about the epoch at hand. Only in the case where an epoch is out-of-distribution, with respect to data seen during training, part of the ambiguity is caused by uncertainty due to data being out-of-distribution, and can possibly be taken away by keeping an expert in the loop. The interested reader is referred to (van Gorp *et al* 2022) for a further discussion on uncertainty in human versus model sleep scorers.

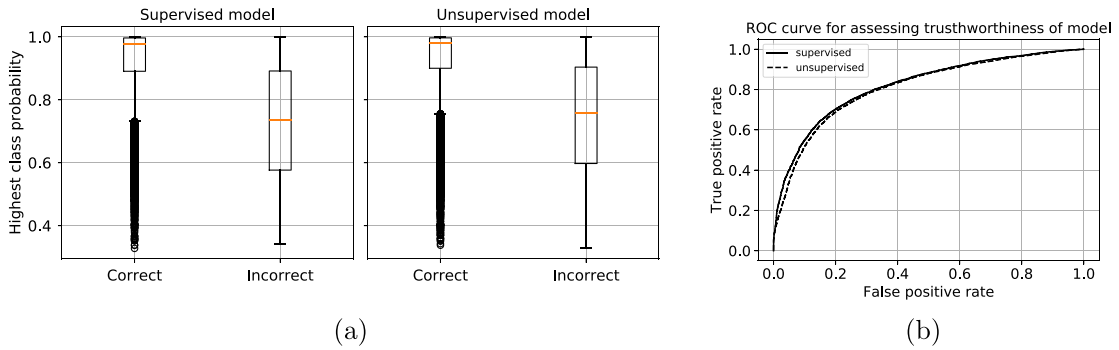

Figure F1. (a) Highest class probability of PSG epoch that were classified correctly and incorrectly, both for the supervised (left) and unsupervised (right) model. Correctly classified epochs exhibit a higher maximum class probability in the hypnodensity graph for both models. (b) ROC curve for assessing the thrustworthiness of the model. *True positive* indicates that the user correctly assessed that the model could be trusted. A *False positive* indicates that the user trusted the model, while the model was incorrect.

References

Bakker J P, Ross M, Cerny A, Vasko R, Shaw E, Kuna S, Magalang U J, Punjabi N M and Anderer P 2022 Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring *Sleep* **45** 1–12

Banville H, Chehab O, Hyvärinen A, Engemann D A and Gramfort A 2020 Uncovering the structure of clinical EEG signals with self-supervised learning *J. Neural Eng.* **18** 1–32

Berry R B *et al* 2012 The AASM manual for the scoring of sleep and associated events *Rules, Terminology Tech. Specifications, Darien, Illinois, Am Acad Sleep Med.* **85** 597–619

Buysse D J, Reynolds C F III, Monk T H, Berman S R and Kupfer D J 1989 The pittsburgh sleep quality index: a new instrument for psychiatric practice and research *Psychiatry Res.* **28** 193–213

Cohen J 1960 A coefficient of agreement for nominal scales *Educ. Psychol. Meas.* **20** 37–46

Goodfellow I 2018 *Deep Learning* vol. 12 (Cambridge, MA: MIT press)

Guo C, Pleiss G, Sun Y and Weinberger K Q 2017 On calibration of modern neural networks *Proceedings of the International Conference on Machine Learning (ICML)* 1321–30

Hyvärinen 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* 89 pp 859–68

Kingma D P and Ba J 2014 *Proceedings of the International Conference on Learning Representations (ICLR)* 2014 Adam: a method for stochastic optimization arXiv:1412.6980

Krauss P, Metzner C, Joshi N, Schulze H, Traxdorf M, Maier A and Schilling A 2020 Analysis and visualization of sleep stages based on deep neural networks *Neurobiol. Sleep Circadian Rhythms* **10** 100064

Kryger M H, Roth T and Dement W C 2011 *Principles and Practice of Sleep Medicine* (Philadelphia: Elsevier Saunderslocation) 5th edn

Morin C M, Belleville G, Bélanger L and Ivers H 2011 The insomnia severity index: psychometric indicators to detect insomnia cases and evaluate treatment response *Sleep* **34** 601–8

Niculescu-Mizil A and Caruana R 2005 Predicting good probabilities with supervised learning *Proceedings of the 22nd International Conference on Machine Learning* pp 625–32

Nobili L, De Gennaro L, Proserpio P, Moroni F, Sarasso S, Pigorini A, De Carli F and Ferrara M 2012 Local aspects of sleep: Observations from intracerebral recordings in humans *Prog. Brain Res.* **199** 219–32

Oord A V D, Li Y and Oriol V 2019 Representation learning with contrastive predictive coding arXiv:1807.03748

Rosenberg R S and Van Hout S 2013 The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring *J. Clin. Sleep Med.* **9** 81–7

Snaith R P 2003 The hospital anxiety and depression scale *Health Quality Life Outcomes* **1** 1–4

Stephansen J B *et al* 2018 Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy *Nat. Commun.* **9** 1–15

Ulmer D and Cinà G 2021 Know your limits: Monotonicity & softmax make neural classifiers overconfident on OOD data *Uncertainty in Artificial Intelligence* (PMLR) 161, 1766–76

van Gorp H, Huijben I A M, Fonseca P, van Sloun R J G, Overeem S and van Gilst M M 2022 Certainty about uncertainty in sleep staging: a theoretical framework *Sleep* **45** zsac134

Younes M, Raneri J and Hanly P 2016 Staging sleep in polysomnograms: analysis of inter-scorer variability *J Clin. Sleep Med.* **12** 885–94

Zimmermann R S, Sharma Y, Schneider S, Bethge M and Brendel W 2021 Contrastive learning inverts the data generating process *International Conference on Machine Learning (ICML)* 12979–12990