2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340905

Identification of Sleep Patterns via Clustering of Hypnodensities

Joshua R. Mirth, Christopher L. Felton, Clifton R. Haider, Stuart J. McCarter,
Timothy I. Morgenthaler, Erik K. St. Louis, and David R. Holmes, III[1]

***Abstract*— Sleep patterns vary widely between individuals. We explore methods for identifying populations exhibiting similar sleep patterns in an automated fashion using polysomnography data. Our novel approach applies unsupervised machine learning algorithms to hypnodensities graphs generated by a pre-trained neural network. In a population of 100 subjects we identify two stable clusters whose characteristics we visualize graphically and through estimates of total sleep time. We also find that the hypnodensity representation of the sleep stages produces more robust clustering results than the same methods applied to traditional hypnograms.**

I. INTRODUCTION

Sleep is a highly variably characteristic of human physiology. Individuals exhibit needs for different amounts of total daily sleep, as well as different patterns of awakenings and sleep cycles within a period of sleep. Typically sleep is characterized as cycling through four stages: three non-rapid eye movement phases (N1, N2, and N3 or slow-wave sleep), and one rapid eye movement (REM) phase. The cycle from N1 through REM occurs multiple times during a given night of sleep. The number of cycles, their duration, and the duration of each phase vary from individual to individual and across different nights of sleep for any one subject.

In this paper we explore methods for identifying subpopulations which exhibit similar patterns of sleep. Using waveforms from full night polysomnography (PSG) recordings, we apply an unsupervised machine learning approach to locate clusters of subjects whose sleep cycles follow similar sequences. There is a lack of clear metrics for sleep quality, so an unsupervised approach can help identify patterns that affect sleep quality [1]. Polysomnography, which includes electrocardigorams (ECG), electroencephelography (EEG), and other bio-physiological waveforms, provides a rich source of data on nearly all relevant physiological components of sleep. The raw waveforms, however, are challenging to automated clustering techniques due to their extreme high dimensionality: an eight-hour recording at 500Hz provides over $14,400,000$ observations per channel. Unlike supervised classification tasks which can benefit from very high-dimensional inputs, unsupervised learning suffers a "curse of dimensionality."

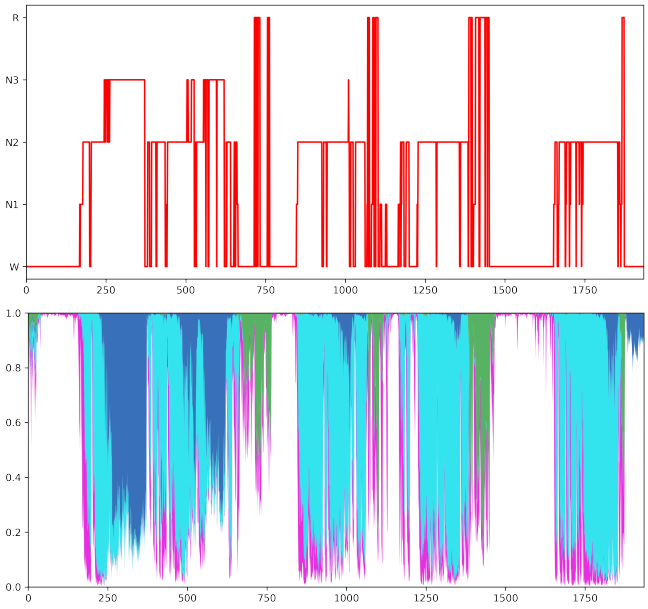

Fig. 1. Example of a hypnogram and hypnodensity. The hypnogram (top) shows the sleep stage for each epoch. The hypnodensity (bottom) represents the probability of each sleep stage at a given epoch. The stages are encoded by the colors: white, wake; pink: N1; light blue, N2; dark blue, N3; green, REM.

Traditionally, each 30-second epoch of PSG data is identified as belonging to one of the four sleep stages, or as a wakeful period, by a trained human scorer. The resulting hypnogram (see Fig. 1, top panel) illustrates the subject's sequence of sleep cycles over the night. Hypnograms provide a much lower-dimensional representation of the sleep period which is amenable to time-series clustering techniques. There are several potential downsides to hypnograms, however. First, they are a discretization of the data. Transitions between sleep stages are a continuous process, so many epochs are intermediate periods. Scoring these stages is to some extent subjective and studies have found that typical agreement between human scorers is only about 83% [2]. There is a growing literature on automating this process using machine learning [3]. An approach that avoids both pitfalls are the *hypnodensities* introduced in [4]. In a hypnodensity each epoch of the PSG recording is represented by a tuple $(p_W, p_R, p_{N1}, p_{N2}, p_{N3})$ indicating the probability that the subject is in sleep stage wake, REM, or non-REM 1, 2, or 3, respectively. This gives a continuous representation of sleep

This material is based upon work supported by the Office of Naval Research under Contract No. N00014-19-C-2017. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

[1]Mayo Clinic, Rochester, Minnesota 55905 USA. mirth.joshua@mayo.edu, felton.christopher@mayo.edu, haider.clifton@mayo.edu, McCarter.Stuart@mayo.edu TMorgenthaler@mayo.edu, StLouis.Erik@mayo.edu holmes.david3@mayo.edu.

stage which is more physiologically appropriate and allows for a more rigorous use of clustering algorithms (which are usually designed for continuous, rather than discrete input data).

Our clustering pipeline consists of applying a pre-trained neural network to convert raw PSGs to hypnodensities. We then cluster the hypnodensities using the k-means (and optionally a PCA dimensionality reduction step). Section II describes the data and pipeline in detail. We compare the effectiveness of clustering sleep patterns using hypnodensities and hypnogram representation in Section III. We then evaluate the characteristics of the identified subpopoulation clusters in Section IV.

II. METHODS

The data consist of 100 full night PSG recordings, collected retrospecively from a population of adults (mean age 46.3 ± 17.2) undergoing overnight sleep studies in the Mayo Clinic Rochester Sleep Lab in accordance with a protocal approved by the IRB of Mayo Clinic. The PSG data include EEG, ECG, electrooculogram (EOG), chin electromyography (EMG), airflow, artertial oxygen saturation, and respiraction waveforms. Sleep stages were manually scored by a trained observer using the PSG data. All recordings were overnight with a typical duration around 8 hours (mean 8.37 ± 0.82).

Individual PSG data were converted to hypnodensity representations of sleep state using the pre-trained neural network provided in [4]. Data quality issues prevented two PSGs from being converted to hypnodensities, resulting in a sample size of 98 hypnodensity files. The resulting 5-dimensional time-series data were resampled to a standard legnth of 960 epochs. Each epoch thus represents a fixed percentage of the sleep period rather than a fixed time length.

Clustering algorithms were applied to both the hypnodensities and the original (human-scored) hypnograms. Except where otherwise specified, all computations were implemented in Python using the Scikit-Learn package [5]. For hypnodensity data the k-means algorithm was used. This requires the researcher to specify the number of clusters k, then clusters S_i are chosen such that the inertia

$$I = \sum_{i=1}^{k} \sum_{x \in S_i} \|\vec{x} - \vec{\mu}_i\|_2^2$$

is minimized. That is, the average distance from a data point $\vec{x}$ to the mean of the cluster $\vec{\mu}_i$, is minimized. For hypnograms, k-means is not a suitable algorithm since the mean of the sleep stage values depends on the numerical encoding and the mean will typically not be a valid hypnogram itself. Instead we use the k-modes algorithm, which is suitable for categorical data types [6]. This is an adaptation of the k-means algorithm where the mean value is replaced by the mode of the cluster, and the Euclidean ℓ^2-norm with the ℓ^0-norm.

To further reduce the dimension of the hypnogram or hypnodensity, we apply principle component analysis (PCA) prior to clustering. The data were projected to dimension

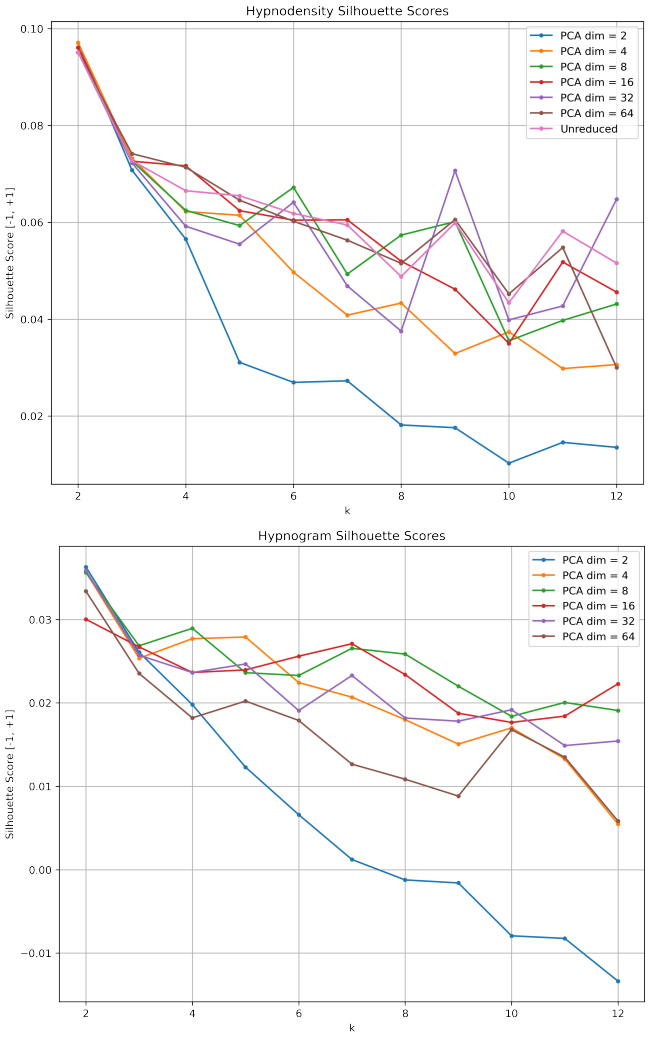

Fig. 2. Silhouette score for k clusters. Regardless of PCA projection, $k = 2$ has by far the highest silhouette score. For hypnograms the unreduced silhouette score is not shown – it is significantly lower than any reduced version.

2^n for $n = 1$ through $n = 6$ prior to clustering. The projected dimensions are weighted linear combinations of different time points, thus the projected data are no longer a time series, nor are they discrete valued in the case of hypnograms. For all PCA data the clustering was performed using k-means.

The number of clusters was chosen using the silhouette coefficient, which quantifies cluster quality based on how many points are near the border between two clusters [7]. Silhouette scores range from -1 to $+1$ with larger scores indicating a more natural clustering. Fig. 2 shows the silhouette scores for different numbers of clusters for the hypnodensity data (unreduced and with various PCA projections) and for the hypnogram data.

Clustering quality was assessed via a stability analysis [8]. Specifically, we resampled 90% of the data points 50 times, then applied the same clustering (and PCA) procedure to the subsets as to the original. The adjusted Rand score,

which measures the similarity between labeling, adjusted for random chance, between the subset clusters and the original clusters (restricted to the 90% subset) was then computed [9]. The stability score for a given k and n is the average adjusted Rand score across all choices of subsamples. A stability score near one indicates that the clustering is robust in the sense that the inclusion of no individual point significantly alters the chosen clusters.

III. RESULTS

For both hypnograms and hypnodensities, with or without PCA, the strongest clustering occurs when $k = 2$ (see Fig. 2). Silhouette scores consistently decrease as the number of clusters k increases. The case $k = 1$ is trivial and so cannot be compared quantitatively, however the qualitative discussion in Section IV suggests that there are reasons to separate the two given clusters. Note that the silhouette score depends on the data metric and hypnograms are measured using the ℓ_0 norm, while hypnodensities use the usual Euclidean ℓ_2 norm. Thus while the hypnodensities typically have larger silhouette scores than the hypnograms, this is not necessarily a meaningful comparison.

The stability scores also indicate that $k = 2$ is the most stable choice of clusters regardless of algorithm or PCA projection (Fig. 3). In particular, hypnodensities with $k = 2$ were the only version of the data with a stability score near one, which is desired for a strong clustering. Stability does not depend on the metric, so we can compare the stability of hypnodensities to hypnograms. The hypnodensity stabilities are significantly higher (mean difference 0.14, $p = 1.6 \times 10^{-14}$). In fact, only four combinations of k and PCA dimension produced a more stable result with the hypnograms. This supports our hypothesis that a continuous measurement (the hypnodensity) produces a more robust clustering than the discretized version.

Stability and silhouette scores both depend upon the embedding, but the optimal $k = 2$ clustering of the hypnodensities was essentially unchanged across different PCA dimensions (Fig. 4). In contrast, the hypnogram clusterings varied significantly, and in general failed to agree with the hypnodensity results. Of particular note, the unreduced hypnodensity returned the same clusters as most of the PCA projected hypnodensities. The unreduced hypnogram clustering (using the k-modes algorithm) was unique, agreeing with neither the PCA hypnograms or the hypnodensities. We take this as further evidence that the hypnodensity representation is more effective for the task of clustering sleep patterns.

To confirm that the two clusters identified in the hypnodensity data are distinct, and not an artifact of the clustering algorithm and unstructured data, we examine the distribution of stability scores. As seen in Fig. 5, for 30% of the samples, the $k = 2$ clustering was perfectly stable on the unreduced hypnodensity data, as many as 50% were stable when a PCA step was included in the pipeline. Stability scores detect unstructured data—the expected number of stability scores equal to one for random data is zero. Thus there is evidence of bimodality within the hypnodensity dataset.

IV. DISCUSSION

Having identified two distinct sleep patterns in the abstract, we attempt to qualitatively describe the differences between these clusters. Since k-means is based on distance to the center of the cluster, it is informative to look at the mean of each cluster in the space of hypnodensities. These means are shown in Fig. 6. These are valid hypnodensity plots, though they do not correspond to the sleep pattern of an actual individual subject. Cluster A contains 20 subjects. It appears to be primarily characterized by a prevalance of wakefulness with little REM sleep. The larger Cluster B, containing the remaining 78 subjcts, shows high probabilities of wakefulness at the beginning and end of the recording period (as expected) but a much lower probability throughout the night. It also exhibits an increasing probability of REM as the night proceeds and a decreasing probabiliy of N3 (slow wave) sleep. Promisingly, these findings reflect the known structure of typical sleep, in which deep sleep (N3) occurs

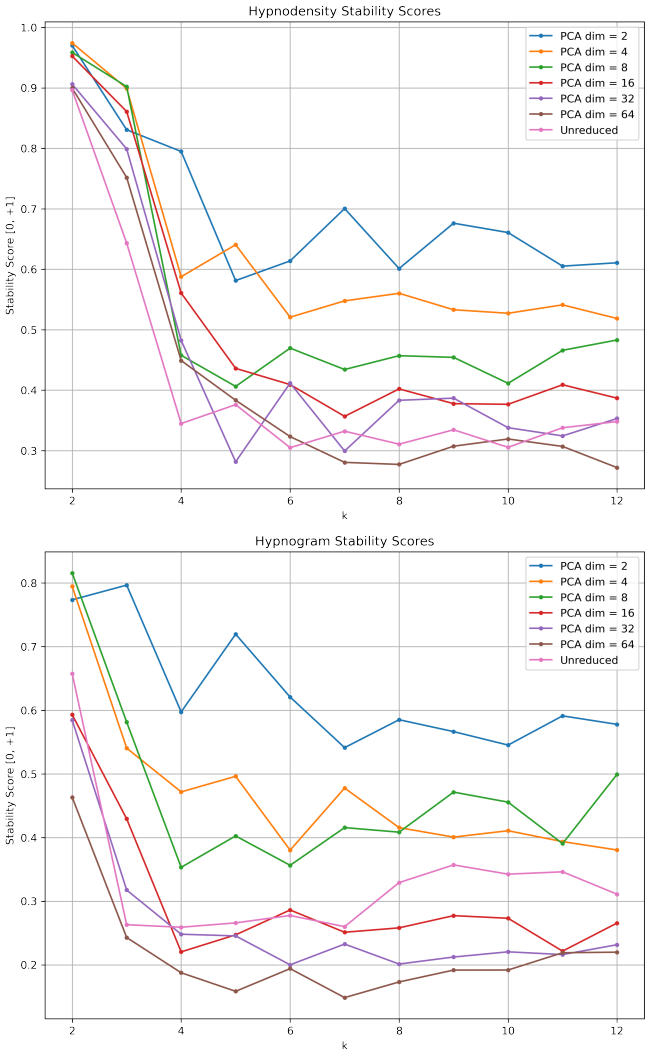

Fig. 3. Stability scores for k clusters for both hypnodensities and hypnograms. The most stable configuration always occurs when $k = 2$. More aggressive dimensionality reduction tends to improve the stability.

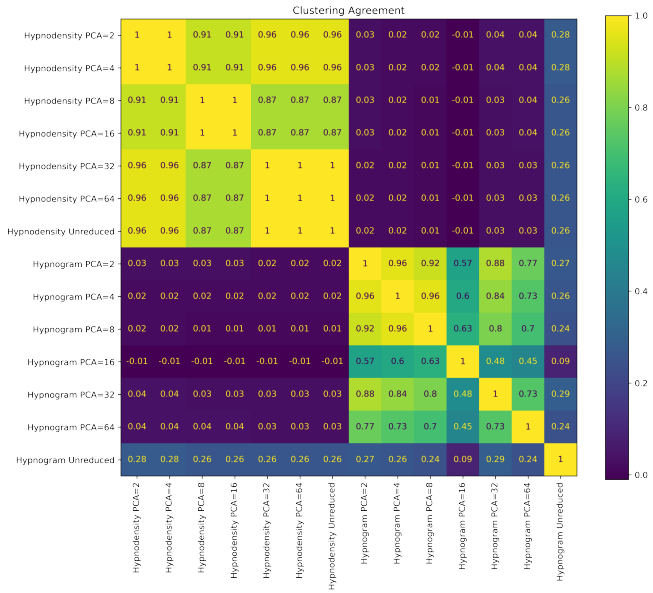

Fig. 4. Adjusted Rand scores between all $k = 2$ clusterings. Different clusterings of hypnodensities registered strong agreement (near one, upper-left block) while clusterings of hypnograms were less consistent (lower-right block). Agreement between hypnodensities and hypnograms was no better than random chance (near zero, upper-right and lower-left).

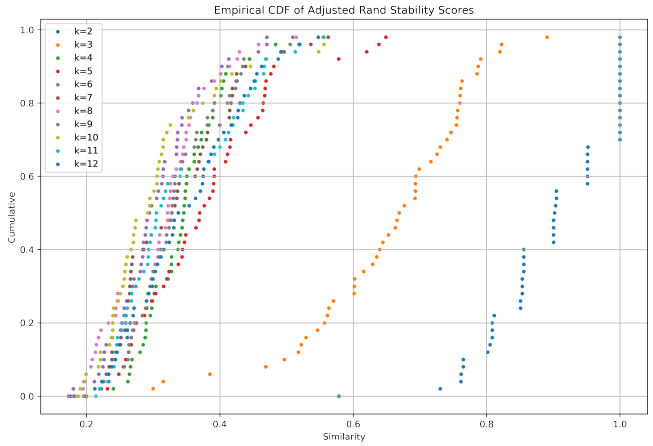

Fig. 5. CDF Plot of stability score for unreduced data. The $k = 2$ case shows 15 stable (Rand score of one) instances, while no $k > 2$ does. The number of stable instances was strictly greater when PCA was included.

more often in the early part of the night, and REM more frequently in the final cycles before awakening.

It appears that we can characterize the two clusters as normal (in Cluster B) and disorded or fitful sleep (in Cluster A). This is confirmed by the total sleep (estimated) sleep time, which averaged 400 minutes in Cluster B, and only 293 in Cluster A. Since the data are observational and the subjects were under study for diagnosis of possible sleep disorders, it is not surprising to identify a large number of instances of poor-quality sleep.

V. CONCLUSIONS

We have demonstrated a fully unsupervised machine learning algorithm for detecting patterns in sleep cycles

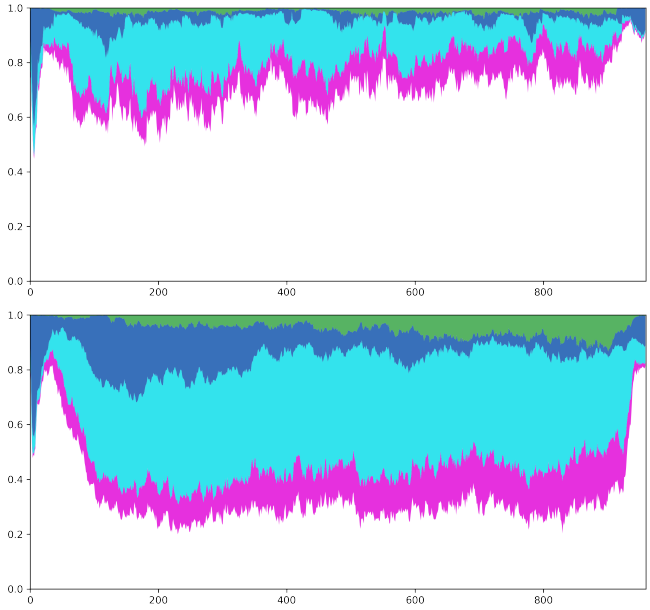

Fig. 6. Mean hypnodensity for Cluster A (top) and Cluster B (bottom).

from polysomnography data. The results show a quantifiable benefit to a continuous (hypnodensity) representation for sleep cycles when applying clustering. The two clusters can be interpreted as normal and fitful or disordered sleepers. With a larger cohort we anticipate that it might be possible to produce finer clusters within these subpopulations. We also anticipate that these methods could be extended in the future to field data collected from wearable devices, allowing for long term studies to detect patterns within larger populations and longitudinally across different nights for the same individual.

REFERENCES

[1] S. J. McCarter et al., "Physiological markers of sleep quality: A scoping review," Sleep Medicine Reviews, vol. 64, p. 101657, Aug. 2022, doi: 10.1016/j.smrv.2022.101657.

[2] R. S. Rosenberg and H. S. Van, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring," Journal of Clinical Sleep Medicine, vol. 09, no. 01, pp. 81-87, doi: 10.5664/jcsm.2350.

[3] L. Fiorillo et al., "Automated sleep scoring: A review of the latest approaches," Sleep Medicine Reviews, vol. 48, p. 101204, Dec. 2019, doi: 10.1016/j.smrv.2019.07.007.

[4] J. B. Stephansen et al., "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," Nature Communications, vol. 9, no. 1, 2018, doi: 10.1038/s41467-018-07229-3.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[6] A. Chaturvedi, P. E. Green, and J. D. Caroll, "K-modes Clustering," Journal of Classification, vol. 18, no. 1, Jan. 2001, doi: 10.1007/s00357-001-0004-3.

[7] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[8] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in Biocomputing 2002, Kauai, Hawaii, USA, Dec. 2001, pp. 6-17. doi: 10.1142/9789812799623_0002.

[9] L. Hubert and P. Arabie, "Comparing partitions," Journal of Classification, vol. 2, no. 1, pp. 193-218, Dec. 1985, doi: 10.1007/BF01908075.