Unsupervised Online Classifier in Sleep Scoring for Sleep Deprivation Studies

Paul-Antoine Libourel, Alexandra Corneyllie, Pierre-Hervé Luppi, Guy Chouvet, Damien Gervasoni

Unsupervised Online Classifier in Sleep Scoring for Sleep Deprivation Studies

Paul-Antoine Libourel, MSc[1,2,3,4]; Alexandra Corneyllie, MSc[1,2,3,4,5]; Pierre-Hervé Luppi, PhD[1,2,3,4]; Guy Chouvet, PhD[1,2,3,4]; Damien Gervasoni, PhD[1,2,3,4]

[1]Centre de Recherche en Neurosciences de Lyon (CRNL), Lyon, France; [2]Institut National de la Santé et de la Recherche Médicale (INSERM), Lyon, France; [3]Centre National de la Recherche Scientifique (CNRS), Lyon, France; [4]Université Claude Bernard Lyon 1 (UCBL1), Lyon, France; [5]Institut National des Sciences Appliquées (INSA), Lyon, France

DISCLOSURE STATEMENT: This is a government supported work (Agence Nationale de la Recherche, Centre National de la Recherche Scientifique). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Author contributions: Dr. Gervasoni, Dr. Chouvet, and Dr. Luppi conceived and designed the algorithm and experiments; Dr. Gervasoni, Ms. Corneyllie, Dr. Chouvet, and Dr. Luppi performed the experiments and analyzed the data; Dr. Gervasoni wrote the paper. The authors have indicated no financial conflicts of interest.

Short title: Unsupervised Sleep Scoring for Sleep Deprivation

Submitted for publication May, 2014

Submitted in final revised form August, 2014

Accepted for publication September, 2014

Address correspondence to: Damien Gervasoni, Centre de Recherche en Neurosciences de Lyon, INSERM U1028 - CNRS U5292 - Université Claude Bernard Lyon 1, 50 Avenue Tony Garnier, 69366 Lyon, cedex 07, France; Tel: 33 (0)4 37 28 74 98; Fax: 33 (0)4 37 28 76 01; E-mail: damien.gervasoni@univ-lyon1.fr

Study Objective: This study was designed to evaluate an unsupervised adaptive algorithm for real-time detection of sleep and wake states in rodents.

Design: We designed a Bayesian classifier that automatically extracts electroencephalogram (EEG) and electromyogram (EMG) features and categorizes non-overlapping 5-s epochs into one of the three major sleep and wake states without any human supervision. This sleep-scoring algorithm is coupled online with a new device to perform selective paradoxical sleep deprivation (PSD).

Settings: Controlled laboratory settings for chronic polygraphic sleep recordings and selective PSD.

Participants: Ten adult Sprague-Dawley rats instrumented for chronic polysomnographic recordings

Measurements: The performance of the algorithm is evaluated by comparison with the score obtained by a human expert reader. Online detection of PS is then validated with a PSD protocol with duration of 72 hours.

Results: Our algorithm gave a high concordance with human scoring with an average κ coefficient >70%. Notably, the specificity to detect PS reached 92%. Selective PSD using real-time detection of PS strongly reduced PS amounts, leaving only brief PS bouts necessary for the detection of PS in EEG and EMG signals (4.7±0.7% over 72 h, versus 8.9±0.5% in baseline), and was followed by a significant PS rebound (23.3±3.3% over 150 minutes).

Conclusions: Our fully unsupervised data-driven algorithm overcomes some limitations of the other automated methods such as the selection of representative descriptors or threshold settings. When used online and coupled with our sleep deprivation device, it represents a better option for selective PSD than other methods like the tedious gentle handling or the platform method.

Keywords: sleep staging, automatic scoring, unsupervised algorithm, paradoxical sleep deprivation

INTRODUCTION

Behavioral states are the expression of the large-scale dynamics of the central nervous system. As assessed by electroencephalogram (EEG) recordings, brain activity is continuously changing. Attempting to define or characterize brain states is nothing but imposing segmentation to ever-changing electrical brain signals by dividing recordings into categorical homogenous bouts, with the risk of neglecting their graded dynamics. The interpretation of an animal's vigilance states from polygraphic recordings takes advantage of a series of invariants contained in the EEG and other signals that are readily observable visually, either on a succession of epochs of equal length or on freely delimited intervals. Still being performed manually in many laboratories, sleep scoring is a tedious task that has motivated the development of a variety of automated methods, and the literature continuously describes new ones which can satisfy most experts in the field.[1]

Whether they use a single EEG channel or combine EEG with electromyography (EMG) and electro-oculograms (EOG), all automated methods presume that information about the behavioral state is contained within the signal(s).[2-5] They thus share common features in their design, such as the extraction of indices or descriptors from the raw signals, and a decision process to assign a state to an epoch. After extracting one or more characteristic features such as delta (1-4 Hz) or theta (5-9 Hz) power,[6-8] a logical paradigm is then used to compare incoming epochs to predefined templates of each state to identify the sleep state. This pattern recognition process often reproduces what an experimenter would do when visually inspecting recording charts. However, in contrast to human scorers, automated methods can objectively and infallibly apply the same rigorous criteria over multiple recordings. Such recognition of sleep and wake states supposes a consensual definition of the vigilance states that need to be identified.

Even with the well-known electrophysiological invariants, sleep scoring in rodents remains subject to a high variability and suffers from poor standardization: While a gold-standard classification method exists for human sleep recordings,[9] there are no standardized

guidelines for rodent sleep.[1] Since pioneering reports,[6] individual laboratories have often refined their categories and corresponding criteria on their own. While most studies consider three major waking and sleep states, up to seven stages can be identified,[10] not to mention studies distinguishing states and the transitions between them.[11] The lack of consensus on the characteristic elements of sleep stages to consider and the lack of standard rules to build a hypnogram accounts for interrater variability and leads to difficulties in the quantification and rigorous comparison of sleep studies among laboratories.[1,12-14]

Automated sleep staging methods also include a decision process in charge of assigning a score as close as possible to the one a human expert would assign. This process is often based on sequential logic rules with a series of dual choices leading to individual classes (decision tree).[15-17] Because of inter-individual variability, the decision rules frequently include thresholds which need to be adjusted for each animal.[18] Most often these methods require human intervention either to set thresholds for one or more parameters, or to select representative templates of each vigilance state. While the use of threshold allows a high flexibility, for instance, to adapt the detection if the signal quality changes over time,[16,19] it also has the disadvantage of introducing subjectivity and bias. Similarly, manual selection of representative templates by two distinct experimenters might lead to two distinct interpretation results.

In order to minimize the subjectivity introduced by human supervision and allow selective sleep deprivation, we developed an unsupervised algorithm inspired by a previous study in rodents.[20] Our program is aimed at unambiguously defining the signature of each state (self-training) for each animal, and is able to automatically categorize 5-s epochs with a decision process based on the probability of them belonging to a given sleep stage. At present, the method implements a Bayesian classifier with a series of five objective EEG and EMG indices to evaluate the probability of observing one of three vigilance states based on *a priori* knowledge of the values of the indices for each state.[20] The decision, instead of being sequential, is a one-step process that combines the five indices in a single product of factors and uses the maximum probability (or likelihood) to assign a state to an epoch.

We further wanted the method to allow a real-time detection of SWS and/ or PS in order to trigger external devices such as mechanical sleep deprivation or optogenetic stimulators in a state-dependent manner. Accordingly, we tested our algorithm in a long (72 h) selective PS deprivation paradigm with an innovative experimental device. In addition, we wanted our software and its mathematical functions to use the smallest amount of computer resources in order to maximize the number of animals simultaneously recorded. Since more than half of the automated systems distinguish three main vigilance states,[1] we restricted our algorithm to the sole detection of waking, slow wave (NREM) sleep, and paradoxical (REM) sleep. It is important to note, however, that the method can easily be generalized to n states, provided that appropriate criteria are defined for the supplemental classes. Considering the classical criteria used to identify sleep and wake states, our software uses only two physiological signals—EEG and EMG.

MATERIALS AND METHODS

We first present the architecture of our sleep classifier and then the methods used for *in vivo* data collection, including sleep recordings and selective PS deprivation in chronically implanted rats.

Software Architecture

The software has been designed to class 5-s epochs. For an *a priori* defined number of 3 sleep and wake states, we designed a 3-class classifier in which the decision is probabilistic by using a likelihood function. Initially developed with Spike 2 (Cambridge Electronic Design), the algorithm was later translated into Matlab programming language (Mathworks). As illustrated in Figure 1, our algorithm is a 2-phase program:

- Self-training phase: extraction and normalization of the indices (step 1), self-training (step 2)
- Scoring phase: extraction and normalization of the indices (step 1), scoring (step 2).

The training phase is executed offline on 8-h baseline recordings, and the obtained template values are then used online for scoring phase.

Data Preprocessing

Occasional artifacts are present in the EEG and EMG signals, mostly occurring during the awake state. Prior to any kind of treatment, epochs showing signal saturation or movement artifacts are automatically removed if they contain >10 saturated samples on the EEG channel (ADC at 512 Hz). Typically, <2% of the epochs are excluded from an 8-h recording session. The remaining epochs are considered as valid and further processed for measurements of EEG and EMG features.

Training Phase

STEP 1: EXTRACTION OF INDICES AND NORMALIZATION

As in all classifiers,[1,20,21] the first step consists in the extraction of a set of indices identifying each polygraphic epoch by a point in a multidimensional space. Here, 4 EEG parameters and 1 EMG parameter are extracted from each 5-s epoch: the standard deviation of the rectified EEG (SD-EEG), the number of sign inversions of the filtered EEG (Zero-crossings), theta (5-9 Hz) to delta (0.5-4.5 Hz) power ratio (hereafter named EEG Ratio 1), and the 0.5-20 Hz/ 0.5-55 Hz power ratio (EEG Ratio 2).[22,23] The values of the spectral power in selected bands result from a fast Fourier transform (FFT) of the filtered EEG with 0.5 Hz resolution. A Hanning windowing procedure is applied before FFT. The EMG signal is subjected to a simple rectification and its median amplitude calculated. The median was chosen instead of the mean because of its lower sensitivity to extreme values. These 5 indices were selected by trial and error according to their relatively high ability to discriminate at least one state from the 2 others (Figure 2Figure 2). For each of them, we assign easily and consistently a level to either high or low, based on physiological observations. The 5 indices and their level in each state (templates) constituted the *a priori* elements. Since 4 indices are derived from a single EEG channel, it is likely that they co-vary or are correlated.[15] However,

because of the subsequent use of the law of total probability (see supplementary material), we assume that they are relatively independent from each other. Because of the peculiar distribution of the indices and the subsequent use of a likelihood function, the feature extraction is followed by a nonlinear normalization. In this process we use the quantiles 0, 0.1, 0.5, 0.9, and 1 of the distribution function to generate a cumulative density transfer function based on cubic spline (Piecewise Cubic Hermite Polynomial [PCHIP] fitting).[24] This function is by consequence centered on the median and bounded between 0 and 1 and permits to evaluate for each index its normalized value (Figure 3 and Supplementary material). At the end of step 1, each valid epoch is therefore represented by a series of 5 normalized indices.

STEP 2: SELF-TRAINING PHASE AND EXTRACTION OF STATE TEMPLATES

At the beginning of the training phase, we assume that the distribution of each index conditioned on states is Gaussian. We therefore consider for each state, 5 Gaussian distributions characterized by a mean set arbitrarily to 0.9 or 0.1 based on the *a priori* knowledge of their high or low level respectively, and a standard deviation set arbitrarily to 1. For instance, the mean of the index "EMG median" is set to 0.9±1 for WK, and to 0.1±1 for both SWS and PS. The Gaussians are used to compare the normalized indices for all incoming epochs of a baseline (reference) file. The comparisons are computed using the complementary error function (1-erf(x)), where erf is the gauss error function (see supplementary materials). With the assumption that the factors are independent, the probabilities obtained for the 5 indices of a given epoch are then combined in a single product of factors for each state, equivalent to a likelihood function in order to estimate the probability for the epoch to belong to WK, SWS, and PS. Among these 3 probabilities, the algorithm identifies the maximal one (maximum likelihood); if superior to 0.1 and ≥10 times superior to the 2 others, a decision is made to update the corresponding state template. Both conditions for updating the templates are implemented to obtain templates from unequivocal epochs, excluding transitional epochs (see supplementary material). State templates are updated by calculating the running average and standard deviation of the 5 indices with the following formula:

$$\bar{x}_{S,n} = \frac{(n-1).\bar{x}_{S,n-1} + x_i}{n}$$

Where $\bar{x}_{S,n}$ is the new template value of index X for the state S built from n epochs updated with the value x of the individual epoch i.

The standard deviation is simultaneously updated by using the classical formula.

$$\bar{\sigma}_{S,n} = \sqrt{\frac{(n-1) \times \left(\bar{x}^2_{S,n-1} + \bar{\sigma}^2_{S,n-1}\right) + {x_i}^2}{\mathrm{n}} - \bar{x}^2_{S,n}}$$

For a given index, the formula used ensured that all epochs selected with the maximum likelihood equally contribute to the overall average. With initial values of 0.9 or 0.1, respectively for high or low level of an index X for the state S, the new average asymptotically and often rapidly (2-4 h of recording) converges towards a unique value that is specific to the current data file, and thus an animal.

At the end of this training phase, all valid 5-s epochs of the baseline file are scanned, and those presenting a high likelihood to belong to one state contributed to the template representative of that state. The resulting state templates values are saved in a text file and subsequently used for real-time scoring (or score the baseline).

Scoring Phase

Real-time scoring is carried out essentially with the same procedure described for the training phase (Figure 1). For each animal, scoring is done using the templates and transfer functions previously determined from the 8-h baseline recording obtained the day before. No change is applied to the recording setup, such as amplification gain or filtering of EEG and EMG channels. Scoring is performed every second on the current 5-s epoch; this sliding window procedure has been introduced to maximize the reactivity of our sleep deprivation system.

STEP 1: EXTRACTION OF INDICES AND NORMALIZATION

The algorithm extracts from the last valid 5-s epochs the EMG and EEG indices. The values are then normalized using the transfer function from the training phase. Whenever an incoming epoch shows signal saturation, it is labeled as artifact and not further processed.

STEP 2: SCORING

Using the complementary erf function with the templates obtained from the training phase, the algorithm computes the probabilities to belong to each of the 3 states and scores the corresponding epoch with the one with the maximum likelihood. All incoming epochs are scored, even a probability level as low as 0.01. We chose to ignore the level of uncertainty during the real-time detection to minimize the risk of missing an occurrence of PS. For the same reason, and to optimize the reactivity of the system, no automated error checking or retroactive correction was applied during the real-time analysis.

Surgical Procedures and Sleep Recordings

All experiments were conducted according to the National Charter on the ethics of animal experimentation, the European Union Directives (86/609/EEC and 2010/63/UE) and procedures were approved by our local Animal Care and Use Committee (Comité d'Ethique en Expérimentation Animale – Université Claude Bernard – Lyon 1) under the references BH-2006-09 and BH-2006-10.

Under general anesthesia (Ketamine 90 mg/kg, Xylazine 10 mg/kg, IP), male Sprague-Dawley rats weighing 280-320 g were implanted for chronic EEG and EMG recordings. The animals were placed in a stereotaxic apparatus with ear and nose bars and their body temperature maintained at 37±1°C with a heating pad and a rectal temperature probe. The skin covering the skull was rubbed with iodine, sectioned longitudinally, and reclined to expose the bone. Four trephine holes were made to insert extradural stainless steel EEG electrodes over the frontal, parietal, and occipital cortices, and over the cerebellum (reference). Two stainless steel wires were inserted into the neck muscle to record the EMG. All electrodes wires were then soldered to a single 6-pin connector (Plastics-One) fixed to the skull with dental

acrylic (Paladur, Heraeus Kuzler). The skin was then sutured and the animals left for recovery after an injection of Carprofen (Rimadyl, 5 mg/kg, SC). Complete postsurgical recovery was observed after 24 h, and a delay of 7 to 10 days was necessary for a complete healing of the wound around the implant. During this period, weight, overt behavior, and eating and drinking abilities were monitored, and appropriate analgesia was provided when needed.

Continuous sleep recordings were conducted after complete recovery from the surgical procedure. Placed in individual barrels, the rats were acclimated to the recording chamber for 2 consecutive days, with a tethered recording cable connecting the implant to a swivel connector. Room temperature was set at 21±1°C, and a 12-h light-dark cycle was maintained throughout the experiment (lights on at 07:00, light off at 19:00). Frontal, parietal, and occipital monopolar EEG and bipolar EMG signals were collected via an amplifier (AlphaOmega, Inc), filtered (bandwidth 1-100 Hz for the EEG, 10-100 Hz for the EMG), digitized (Micro1401, Cambridge Electronic Design, sampling rate 512 Hz), and stored on a computer using CED Spike 2. Only one EEG channel, usually parietal, among the 3 EEG recorded were used together with the EMG for the identification of sleep and wake states.

An automated PS deprivation (APSD, n = 10) of 72 h was performed to evaluate the efficiency of the real-time detection of PS. Among the 10 animals, 3 went first through a 72-h PS deprivation with the platform technique (PPSD),[25-27] and after a minimal delay of 2 weeks, through a second long PSD (APSD) with a new homemade device (redesigned by ViewPoint Life Sciences) coupled to our automated scoring algorithm. This device was developed to sleep deprive the animal in its home cage without the need of human intervention. For this purpose, a small solenoid was placed underneath the floor of the barrel and driven by a TTL pulse sent by the CED recording interface at each PS occurrence detected by our algorithm. The solenoid briskly lifts (TTL pulse 50 ms) up the floor of the barrel 1 centimeter up and lets it return to its initial level, causing a global waking stimulus. The animal was placed into the device during the control, the APSD, and the recovery period. Prior to the 72-h automatic PSD, a 24-h baseline recording was done for each animal, as 3 consecutive 8-h files. The 8-h baseline files encompassing the daylight period (with presumably higher PS amounts) were processed

offline to score them and extract the state templates and transfer functions for the online detection.

Manual Sleep Staging and Evaluation of the Performance of the Algorithm

To evaluate the performance of the algorithm, visual sleep staging was performed *a posteriori* on all recordings using classical descriptions of sleep and waking states.[6-8,22,23] Using a homemade script displaying EEG and EMG signals and EEG spectral density, 5-s epochs were manually scored as WK, SWS, or PS according to the following criteria: high and variable amplitude EMG, and a low voltage, fast activity EEG with theta rhythm during exploratory behavior for WK[7,28,29]; low amplitude EMG with no phasic events, and a high voltage EEG with slow waves (1-4 Hz) and spindles (10-14 Hz) for SWS[30]; and concomitant very low voltage EMG, and a low voltage EEG with a marked periodicity in the theta band (5-9 Hz) for PS. Sleep scores obtained from the algorithm and from a human expert were compared by computing confusion matrices and Cohen κ coefficient of agreement.[31,32] The performance of the algorithm was further assessed by calculating the sensitivity and specificity to detect WK, SWS, and PS, as well as the corresponding positive and negative predictive values (PPV and NPV) (see supplementary material). Unless otherwise noted, descriptive statistics use mean ± SEM. The median and interquartile range (IQR) are sometimes used to better describe the distribution of values in small samples.

RESULTS

Efficiency of the Self-Learning Paradigm

We first evaluated the ability of our algorithm to establish the signatures of sleep and waking states for each animal formed by the 5 indices extracted from individual recordings. The database consisted of 24-h baseline recordings of 7 rats by contiguous sections of 8 hours (02:00-10:00; 10:00-18:00; 18:00-02:00). Each 8-h data file was analyzed in 5-s epochs, representing a total of 5760 epochs per file. On average, <2% of the epochs were discarded because of artifacts in 1 of the 2 signals. At the end of step 1, 5 indices—EEG Ratio 1 and 2,

zero crossing, EEG SD, and EMG median amplitude—were concomitantly extracted from all artifact-free epochs of a given data file and normalized. As illustrated in Figure 3 (see also supplementary material), the polynomial normalization process resulted in a linearization of the distributions of the indices and bounded them between 0 and 1. This process was imposed by the subsequent use of probabilities, i.e., real positive values that cannot exceed 1.

At the beginning of the training phase, indices were arbitrarily set to 0.9 for high or 0.1 for low values. Throughout the training phase the 5 indices forming a state template converged towards distinct values that were unique to each animal (Figure 4). Initially, when the maximum likelihood function corresponded to a state that appeared inaccurate *a posteriori*, its template values would be erroneously updated. However, because of the large number of 5-s epochs used for each template (>200) and of the convergent functions used to compute the new average and standard deviation, such sporadic errors had limited consequences on the final convergence.

Since epochs were selected on a probabilistic basis, the templates were built from a proportion of valid epochs that varied between recording files: the average proportion was 42.4±2.1% of all valid epochs (range 24.4–57.6; n = 21). This proportion of epochs contributing to the templates was also expected to be linked to the prevalence of each state. To further test the influence of the amount of the states in the size of the learning sets, we compared the number of epochs selected for templates between the recording files encompassing the daylight period (10:00-18:00) and the flanking files covering the dark period (02:00-10:00; 18:00-02:00). We found that the number of epochs selected to form the SWS and PS templates was slightly higher for the recording files encompassing the daylight period (10:00-18:00) than for the flanking files covering the dark period (02:00-10:00; 18:00-02:00), respectively 1658±193 and 1394±114 (29.2±3.2% versus 24.7±1.9% of valid epochs) for SWS templates and 386±54 and 252±31 (6.9±0.9% versus 4.5±0.6%of valid epochs) for PS templates. Reciprocally, fewer epochs contributed to the WK template in the “day” recording than in the “night” recordings—583±63 versus 628±57, respectively (10.3±1.1 versus 11.1±1.1% of valid epochs). A Mann-Whitney U-test with an α threshold of 0.1% showed no significant difference

in the size of the learning sets between “day” and “night” files, but its statistical power (P = 59%) was too low to draw any firm conclusions.

Once template values were calculated for each state, they were used in the final scoring phase without being further updated. When scoring was performed online, although it was technically possible, templates were not updated. This constraint was added to maintain equivalent processing of incoming epochs over time and to prevent any error that would have changed the templates and impaired automated detection of PS for the algorithm-driven deprivation.

Scoring of each epoch was done by computing the likelihood of each state, i.e., the probability of each epoch to belong to WK, SWS, and PS classes. Based on the product of probabilities calculated for the 5 indices, likelihood values reflected the degree of similarity of an epoch to the WK, SWS, and PS templates. The maximal value among the probability of WK, SWS, and PS was selected to assign a state to the epoch. As illustrated in Figure 5, likelihood values for each state varied according to the evolution of EEG and EMG indices and their similarity to the average template values in each state. Overall, for the great majority of epochs, one value of probability among the 3 calculated for each epoch was always much higher than the others, so that a state could be assigned with a high degree of confidence. When the probabilities of individual epochs were plotted in a 3D space with the state assigned manually (panel B in Figure 5), we observed occasional mismatches, i.e., epochs that were mistakenly scored by the algorithm. In such cases, the values of probabilities for WK, SWS, and PS were <0.02 with only slight differences between them.

Performance of the Algorithm To Identify Sleep and Wake States

Each file was first scored by the algorithm with the templates built from the self-training procedure and also *a posteriori* analyzed manually by an expert to evaluate the quality of labeling. The hypnograms obtained were then compared pair-wise. For this purpose, confusion matrices were built for each file (see example in Table 1), allowing the calculation of a joint probability of agreement representing the proportion of epochs similarly classed by the

algorithm and the expert. The performance of the algorithm was further assessed by calculating the sensitivity and specificity of the algorithm to detect WK, SWS, and PS and the corresponding positive and negative predictive values. Figure 6 illustrates the comparison of a score obtained with the algorithm and the one obtained from a human experimenter. Sporadic differences between human and computer scores are visible on the color-coded hypnograms. Importantly, these mismatches were observed whenever the maximal probability was very low. In the example shown in Figure 6, the algorithm detected 2 PS occurrences (vertical arrows) that were interpreted as WK or SWS by the experimenter. For such occurrences, though the probability of PS was the highest compared to the probability of WK and SWS, its absolute value was <0.01.

The median joint probability of agreement calculated from the confusion matrices was 0.83 (IQR = 0.22, n = 21), with a maximum observed of 0.93; the median κ was 0.72 (IQR = 0.30; n = 21).

Since we had previously observed that daytime and nighttime recording files presented slight differences in the size of the learning sets for the 3 states, we asked whether these differences had an influence on the overall performance of the algorithm. For data files recorded during 10:00-18:00, the median probability of agreement and median κ were 0.85 (IQR = 0.08) and 0.73 (IQR = 0.10, n = 7). For nighttime recordings, these values were slightly inferior—respectively 0.71 (IQR = 0.23) and 0.50 (IQR = 0.42, n = 14). A Mann-Whitney U-test showed no significant difference (α = 0.1%) but statistical power was too low to conclude with confidence on the existence of an actual difference of accuracy between “day” and “night” files.

To test the flexibility of the algorithm, we used the state templates built from the files recorded between 10:00 and 18:00 to analyze the 2 flanking files mostly recorded during the dark phase (02:00-10:00; 18:00-02:00). In this configuration, the median joint probability of agreement reached 0.91 (IQR = 0.14) with a maximum of 0.97, and the median κ coefficient reached 0.78 (IQR = 0.31, n = 14), corresponding to a substantial agreement. Overall, using the templates from recordings made during the “day” improved sensitivity and specificity to detect WK, SWS, and PS (Table 2). Taking into account the actual prevalence of PS in these

“night” files (6.7±1.1%), the median PPV (probability that a positive detection of PS corresponds to its presence) was 76.2% (IQR = 34.5, n = 14). The median NPV (probability that a negative PS detection corresponds to its absence) was 98.8% (IQR = 95.0, n = 14). In other words, our algorithm presented a slight bias towards a detection of false positives. Despite a slight decrease of sensitivity and the risk of false PS detection, we used the templates issued from the day recordings for our subsequent experiments. Since we aimed to use our algorithm online for selective PS deprivations, i.e., to maximize PS detection we thus favored specificity over sensitivity to maximize PS detection.

Performance of the Algorithm for the Online Detection of PS

The ability of the algorithm to detect PS was also evaluated online during selective PS deprivation for 72 h (APSD) and compared to 72-h classic platform PSD (PPSD). For this purpose, baseline recordings were collected and scored beforehand with the algorithm to establish individual templates of indices of the 3 states. These templates were extracted from the 8-h files recorded between 10:00 and 18:00, and directly used in the online detection program. At the end of the experiments, all recordings were visually scored offline to evaluate *a posteriori* the quality of detection.

APSD was performed in a group of 10 animals for 72 hours. Baseline recordings made the day before the deprivation did not reveal any abnormality. The amounts of sleep and wake states in this group of animals (Table 3) were in accordance with the values routinely observed in our laboratory. A subgroup of 3 rats was first submitted to a 72-h PSD with the 3-platform method (PPSD) and then after ≥2 weeks to automated PSD (APSD). In agreement with previous studies,[26,27,33] PPSD efficiently suppressed PS, with <1% of PS remaining over 72 h (0.93±0.03) compared to 9.5±1.1% in baseline condition (n = 3). PPSD was accompanied by a slight decrease in the amount of SWS (23.7±1.3% versus 39.7±2.4% in baseline). During the recovery period that followed the PPSD, a PS rebound was observed with an average percentage of PS of 28.0±3.2% over 150 min, representing 47.7±6.9 minutes of PS. The duration of 150 min corresponds to the recovery period routinely used in the laboratory for

functional neuroanatomical studies using the early gene c-Fos.[26,27,33] The average latency to the first PS episode during the recovery period was 41.2±5.6 min (n = 3).

For all animals, APSD was performed for 72 h starting at 10:00. Every second, incoming 5-sec epochs were analyzed by the algorithm, and a TTL pulse was sent to the solenoid whenever an epoch was identified as PS. The brisk shake applied to the cage's floor by the solenoid resulted in a prompt awakening of each animal. Two to 3 additional TTL pulses were sent until the PS signs faded out in the subsequent incoming epochs. Consequently, a residual amount of PS was observable during APSD as short isolated PS bouts corresponding to the short amount of data necessary to identify the initiation of a PS episode and to apply the waking stimulus. Nonetheless, when deprived with this automated method, a drastic reduction of PS occured over 72 h (Table 3), with no occurrence lasting more than 7.9±0.5s, for an average number of 1447±175 attempts (n = 10). With a PS pressure building up throughout the PSD, we observed an increasing number of PS attempts (data not shown). When APSD was stopped, PS rebound was promptly observed in all animals to the detriment of WK (Table 3). The first PS episode was observed with an average latency of 1.7±1.0 min after the end of APSD. During the recovery period, PS episodes had a longer duration, with an average duration of 2.7±0.2 min versus 1.2±0.1 min in baseline, and recurred on average every 9.6±1.1 min (12.1±1.4 min in baseline).

DISCUSSION

Advantages and Drawbacks of the Use of a Probabilistic Approach

We introduced in this study an unsupervised algorithm that uses a probabilistic approach for sleep scoring. While it is not the first program based on a Bayes classifier,[1,34] to our knowledge it is the first that is completely unsupervised and used in real time to perform selective PS deprivation. The choice of a probabilistic classifier was motivated by our attempt to limit human intervention, while relying on *a priori* knowledge of the characteristics of sleep and waking states, such as the well-described spectral features of the EEG and EMG

amplitude.[2,7,8,20,22,23] During the training phase, five EEG and EMG indices are extracted and normalized; state templates are initialized in concordance with physiological observations, and automatically adjusted by a self-training procedure to form final templates specific to each animal. The scoring phase assigns a state to an epoch based on the probabilistic distance between the indices extracted and the templates, with the intuitive assumption that the shorter the distance is, the higher the likelihood is of an epoch to belong to a class ("birds of a feather flock together").

Previous studies that have also implemented a Bayes classifier in a sleep scoring program were successful in terms of performance, with an agreement between a human rater and software at least equivalent to the interrater agreement of 92%.[1,34] Overall performance based on the global accuracy reported in the literature for sleep classifiers is usually between 70% and 95%.[21] With a median coefficient of agreement of 83%, our algorithm also gave a high concordance with human ratings. If global accuracy could be used as an absolute comparison criterion, then our program would be ranged among the most satisfactory classifiers. There is, however, such a wide variety of classifiers that comparisons of performance solely based on global accuracy are in fact very difficult and do not reflect the actual reliability nor its convenience for a daily use. Other measures, such as sensitivity and specificity could be used as well, and still would not bring relevant information as to the advantages of such a method over another one.

When developing this classifier, our main objective was to remove any human supervision, in both the training phase and in the pattern recognition process. The first requirement in our project was thus for the program to be able to create representative templates of each state for each animal with the closest match possible to physiological observations. With the use of the likelihood function, our self-training procedure was able to select in each recording a series of epochs with the maximal resemblance to the ideal state templates in which the indices were set to 0.9 or 0.1 for high or low, and to automatically adapt these signatures to each animal. Indeed, a convergence of the five indices towards values specific to each animal was always observed. Our preliminary tests in which high and low initial

values were set to 0.8 and 0.2 instead of 0.9 and 0.1 gave similar results. Therefore the values used to initialize the high and low levels of the indices can in fact be arbitrarily set. Unlike the naïve Bayes classifier developed by Rytkonen and colleagues and other automated methods,[1,34,17,18] our self-adaptive procedure is thus able to determine for a given animal the actual level of a low and a high index, and therefore does not require a manual selection of epochs representative of each state. As previously demonstrated,[17] normalization of the input variables is essential to eliminate the need of animal-specific thresholds, and to overcome the problem of signal instability over time. In line with these results, our five indices and the original process used to normalize them allowed a training phase and compilation of state templates that did not depend on the gain of the signal, did not require any threshold, and still allowed a certain degree of flexibility (see below). An advantage of the absence of supervision in this training phase is that state signatures can reproducibly be obtained from a given sleep recording, with resulting templates that will invariably be the same.

A consequence of the use of a probabilistic routine to build the templates of the three states is that the actual number of epochs selected in each file can hardly be predicted. Indeed, our results show that the number of epochs contributing to the templates was variable from one file to the other, and related to the prevalence of the sleep and wake states. This was particularly true for PS, which was the least represented state, especially in recordings encompassing the dark period when rodents are mostly awake. Our results further show that for the data files recorded at night, our algorithm performed better with templates based on daylight recordings. This observation illustrates the intuitive assumption that for a given animal, a template of a state will be the most representative if it is obtained with the largest sample of epochs. It further confirms the importance of the data selected manually or, as here, automatically to build representative templates.[1,20]

While a variable size of the training set for each state may appear as a major drawback for experimenters wishing to control every single parameter of an experiment, others might appreciate the comfort of not having to impose their own selection of representative epochs, particularly when running experiments in multiple animals at once. Even so, our sleep scoring

program leaves the possibility to manually score an entire data file, and to establish the state templates from all epochs manually scored. It is possible therefore to compensate for the low prevalence of PS by selecting as many WK and SWS epochs as there are PS epochs.

Another advantage offered by the probabilistic approach is the flexibility of the program, i.e., its ability to use predefined signatures for scoring new data (forward procedure). As pointed out by Robert and colleagues, an inflexible classifier might not be able to operate suitably on new data.[1] The results obtained in our sleep deprivation experiments show that without applying any correction to the state templates built from the baseline recordings, our algorithm efficiently detects all PS occurrences. This satisfying result could be linked to the fact that the context of selective PS deprivation induces only marginal changes in the EEG waveform,[35] and that the combination of indices used, notably the EEG power ratios, is flexible enough to adapt to the modifications reported.[36] The efficiency of our algorithm, however, might not be the same in the context of pharmacological studies in which EEG waveform is strongly altered by drugs or in genetically modified animal models in which EEG descriptors might also significantly differ between animal strains.[37,38] For instance, the dopamine-transporter KO mouse presents hyperactivity and significant spectral alterations of the EEG when placed in a novel environment compared to its wild-type littermate.[39,40] Our procedure extracts information from the frequency domain (EEG power ratios) and the time domain of the EEG (standard deviation of the EEG), both having advantages and drawbacks.[1] Although it has not been tested in the present study, we believe that with these EEG descriptors and the EMG information, and its ability to adapt the signature of each state, our classifier might still operate correctly in the context of a pharmacological study or in another strain. Our algorithm was developed for sleep recordings in rats, and a preliminary study done in mice without any change to the code shows a global accuracy that is similar to the one observed in the present study.[41] Additionally it would be interesting to test it in the context of a total sleep deprivation (SWS and PS) during which sleep pressure induced by prolonged wakefulness elicits a substantial changes in EEG features.[42-44]

Window Size and Contextual Information

From one laboratory to another, sleep recordings are scored differently, mostly on an epoch-by-epoch basis but with distinct epoch widths, ranging from 2 to 30 seconds.[1] Here we used 5-s epochs,[17] a duration routinely used in the laboratory. With a sampling rate of 512 Hz for EMG and EEG signals, a 5-s window provided enough data points for the extraction of relevant EEG and EMG features and allowed a fair placement of the limits of sleep/ wake bouts and thus precision in the calculation of sleep and wake amounts. Keeping in mind that the use of short epochs could induce an overestimation of short duration episodes,[2,45] we also chose 5-s epochs to limit the number of mixed epochs, i.e., epochs in which at least two states are present. It is important to note that the algorithm can easily be adapted to work on smaller or even larger epochs, provided that when using small ones there are enough data points for spectral calculations. Such possibility is permitted by the use of five EEG and EMG indices that measure tonic features of sleep and wake states as opposed to phasic signs such as ocular movements, which might not be tracked accurately on long epochs.

Efficiency of the Probabilistic Detection of PS for Automated Sleep Deprivation

Our sleep deprivation method is inspired by the pioneering experiments in which the information on the sleep state of an animal is fed back through a mechanical device known as the disk apparatus to wake up an animal.[46] Similar devices use EEG/ EMG monitoring to ensure the bar rotates only when the animal enters a sleep-like state and control a stir bar to simulate gentle handling (e.g., Pinnacle's sleep deprivation system [Pinnacle Technology]). Such feedback systems have been successfully used to perform unsupervised total sleep deprivation and selective PS deprivation in rats and mice. These techniques, however, require human intervention to establish thresholds for each animal. In the present study, state templates were automatically built from baseline recordings without input from an experimenter, and our algorithm efficiently detected PS occurrences without missing any of them. Our results show that our algorithm presents a slight bias toward a detection of false

positive, i.e., WK or SWS epochs mistakenly labeled as PS. This is the consequence of the choice we made to maximize PS detection and thus to favor specificity over sensitivity. In the context of sleep deprivation experiments, this over-detection had limited consequences. Approximately two-thirds of false detections occurred during WK, so a waking stimulus would not change the actual state; as to the false detection occurring during SWS, because of the peculiar sleep dynamics observed during sleep deprivations,[43,47] it is difficult to determine whether those stimulations were systematically shortening SWS episodes or anticipating an imminent PS occurrence. A possible way to prevent this bias would be to apply a smoothing and/ or conditional correction (for instance, do not score PS if previous epoch is WK). Other alternatives would be to add supplementary state templates such as quiet waking, drowsiness, or intermediate states, or to condition the decision based on the actual level of probabilities and prevent any stimulation in a situation of high uncertainty.

Our 72-h sleep deprivation paradigm was efficient and strongly decreased PS. This result demonstrates that our algorithm is able to maintain an accurate detection and deprivation of PS for up to 72 h, despite the cumulating debt of PS and the increasing homeostatic pressure for PS. It further shows the ability of the program to overcome the marginal changes of the EEG features that can be caused by the deprivation.[35] Because of the necessary presence of a few seconds of PS for detection, a residual amount of PS is present throughout the APSD in the form of short aborted PS episodes. These residual amounts of PS might account for the lesser amount of PS observed during the recovery period of 72-h APSD compared to PS rebound after PPSD.[26,27,33] APSD also differed from the PPSD by the time course of the PS rebound. Indeed the latency to the first PS episode from the end of APSD was much shorter. This result is likely due to the change of the environment of the animal with the platform method. Indeed, in PPSD experiments rats are transferred to a clean cage at the end of the privation. With our automated method, no such change was made, so that the animals could readily start recovering their PS debt. In terms of comfort for an animal, we therefore believe that our automated sleep deprivation system is a valuable alternative to the classical platform technique for long-term sleep deprivation. It is also worth mentioning that in our sleep

deprivation system, waking is evoked by a sudden shake of the cage floor, a stimulus from which an animal can hardly escape. Besides, the stimulation is not applied by the contact of an actuator, and it does not force the animal to move. Using the same device in another series of experiments,[48] we found that the PS amounts and latencies observed during the recovery after 6-h automated PSD were similar to those reported after gentle handling PSD of the same duration.[49] Our sleep deprivation system therefore constitutes a valuable alternative to the gentle handling technique for short selective PS deprivation. Its performance remains to be determined in the context of total sleep deprivation as well as sleep fragmentation.

To date, our algorithm has been able to detect the three major sleep and wake states usually scored in sleep laboratories with a high degree of confidence. Still, improvements and adaptations are possible, such as adding more states to the classifier. The underlying probabilistic core can easily be extended to “n” states, or even sub-states. Provided that objective measures are found to separate them, one can imagine a distinction between light and deep slow wave sleep. When performing automatic selective PS deprivation, it would be particularly interesting to detect the intermediate state.[10,50,51] This transition state that mostly occurs between SWS and PS is characterized by short period (3-5 sec) during which the EEG presents a mix of theta activity and SWS spindles.[22,23,50] The detection of this transitional state would allow an anticipation of PS occurrences and a complete suppression of PS. The counterpart, however, could be an increase of inadequate stimulation since, in normal conditions, bouts of intermediate state are not always followed by actual PS episodes.

Conclusion

In agreement with previous studies, our results show that sleep scoring algorithms based on Bayes classifiers can reach high performance in terms of global accuracy. They can thus considerably alleviate the tedious and time-consuming task of offline analysis of sleep recordings. With our self-training paradigm, users are no longer required to adjust thresholds or decision rules for each animal. The ability of our program to detect sleep and wake states in real-time allows long unsupervised sleep deprivation or any kind of state- dependent

stimulation, such as optogenetic light pulses. As a totally unsupervised classifier, it completely suppresses human bias and therefore increases reproducibility of sleep analyses. Finally, we believe that such a classifier can easily be generalized and adapted to others species like cats, primates, and even humans.

Acknowledgment: The authors wish to gratefully thank Jérémy Nigri and Sébastien Arthaud for their help in the conduct of the animal experiments and Dr. Anthony Herrel for carefully revising the manuscript.

Table 1. Confusion matrix for an 8-hour file recorded during the daylight period

		Automated score			
		WK	SWS	PS	sum
Human score	WK	1097	246	247	1590
	SWS	15	2836	132	2983
	PS	65	267	845	1177
	sum	1177	3349	1224	5750

Table 2. Comparison of the median sensitivity and median specificity to detect WK, SWS, and PS in "night" recorded files (18:00-02:00 and 02:00-10:00) with templates extracted from files recorded during the day (10:00-18:00) or with templates built from the nighttime files themselves.

	WK		SWS		PS	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
"day" template	91.7 (26.6)	95.7 (7.4)	95.0 (11.0)	93.6 (6.1)	79.4 (11.3)	98.1 (4.9)
"within-file" template	62.0 (27.3)	99.3 (0.7)	97.9 (4.3)	88.2 (11.2)	90.9 (10.6)	84.6 (19.0)

Values are given as percentages, number in brackets are the interquartile ranges.

Table 3. Average waking and sleep amounts during and after long automated (72-h) selective PS deprivation (n = 10).

	WK	SWS	PS
Baseline (24h)	53.4±1.4% (760.7±18.8)	37.6±1.3% (537.1±19.5)	8.9±0.5% (127.4±7.7)
Automated PSD (72h)	64.3±1.3% (2763.9 ± 56.1)	31.0±1.1% (1331.8 ± 48.9)	4.7±0.7% (203.5±28.2)
Baseline (150 min)	37.3±3.0% (56.0±4.5)	52.0±3.1% (78.1±4.6)	10.7±1.0% 16.0±.6
PS recovery (150 min)	18.4±2.7% (27.7±4.0)	50.4±3.1% (75.7±4.6)	22.8±1.7% (34.3±2.5)

Values are given as percentages ± SEM; values between brackets are the corresponding amounts in minutes. Baseline values are measured over 24 h or over 150 min at the time of the day corresponding to the 72-h APSD (starting at 10:00) or to the recovery period (starting at 16:00), respectively.

REFERENCES

1. Robert C, Guilpin C, Limoge A. Automated sleep staging systems in rats. J Neurosci Methods 1999;88:111-22.
2. Karasinski P, Stinus L, Robert C, Limoge A. Real-time sleep-wake scoring in the rat using a single EEG channel. Sleep 1994;17:113-9.
3. Koley B, Dey D. An ensemble system for automatic sleep stage classification using single channel EEG signal. Comput Biol Med 2012;42:1186-95.
4. Mendelson WB, Vaughn WJ, Walsh MJ, Wyatt RJ. A signal analysis approach to rat sleep scoring instrumentation. Waking Sleeping 1980;4:1-8.
5. Gandolfo G, Glin L, Lacoste G, Rodi M, Gottesmann G. Automatic sleep-wake scoring in the rat on microcomputer APPLE II. Int J Biomed Comput 1988;23:83-95.
6. Timo-Iaria C, Negrao N, Schmidek WR, Hoshino K, Lobato de Menezes CE, Leme da Rocha T. Phases and states of sleep in the rat. Physiol Behav 1970;5:1057-62.
7. Winson J. Patterns of hippocampal theta rhythm in the freely moving rat. Electroencephalogr Clin Neurophysiol 1974;36:291-301.
8. Monmaur P. Phasic hippocampal activity during paradoxical sleep in the rat: selective suppression after diazepam administration. Experientia 1981;37:261-2.
9. Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Washington, DC: US Department of Health, Education and Welfare; 1968.
10. Gottesmann C. Detection of seven sleep-waking stages in the rat. Neurosci Biobehav Rev 1992;16:31-8.
11. Cape EG, Jones BE. Effects of glutamate agonist versus procaine microinjections into the basal forebrain cholinergic cell area upon gamma and theta EEG activity and sleep-wake state. Eur J Neurosci 2000;12:2166-84.
12. Penzel T, Conradt R. Computer based sleep recording and analysis. Sleep Med Rev 2000;4:131-48.

13. Neckelmann D, Olsen OE, Fagerland S, Ursin R. The reliability and functional validity of visual and semiautomatic sleep/wake scoring in the Moll-Wistar rat. Sleep 1994;17:120-31.

14. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 2009;18:74-84.

15. Costa-Miserachs D, Portell-Cortes I, Torras-Garcia M, Morgado-Bernal I. Automated sleep staging in rat with a standard spreadsheet. J Neurosci Methods 2003;130:93-101.

16. Louis RP, Lee J, Stephenson R. Design and validation of a computer-based sleep-scoring algorithm. J Neurosci Methods 2004;133:71-80.

17. Stephenson R, Caron AM, Cassel DB, Kostela JC. Automated analysis of sleep-wake state in rats. J Neurosci Methods 2009;184:263-74.

18. Gross BA, Walsh CM, Turakhia AA, Booth V, Mashour GA, Poe GR. Open-source logic-based automated sleep scoring software using electrophysiological recordings in rats. J Neurosci Methods 2009;184:10-8.

19. Mileva-Seitz VR, Louis RP, Stephenson R. A visual aid for computer-based analysis of sleep-wake state in rats. J Neurosci Methods 2005;148:43-8.

20. Chouvet G, Odet P, Valatx JL, Pujol JF. An automatic sleep classifier for laboratory rodents. Waking Sleeping 1980;4:9-31.

21. Charbonnier S, Zoubek L, Lesecq S, Chapotot F. Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging. Comput Biol Med 2011;41:380-9.

22. Gervasoni D, Lin SC, Ribeiro S, Soares ES, Pantoja J, Nicolelis MA. Global forebrain dynamics predict rat behavioral states and their transitions. J Neurosci 2004;24:11137-47.

23. Lin SC, Gervasoni D. Defining global brain states using multielectrode field potential recordings. In: Nicolelis MA, ed. Methods for neural ensemble recordings, 2nd ed. Boca Raton: CRC Press; 2008:145-68.

24. Fritsch FN, Carlson RE. Monotone piecewise cubic interpolation. SIAM J Numer Anal 1980;17:238-46.

25. Mendelson WB, Guthrie RD, Frederick G, Wyatt RJ. The flower pot technique of rapid eye movement (REM) sleep deprivation. Pharmacol Biochem Behav 1974;2:553-6.

26. Verret L, Fort P, Boissard R, Gervasoni D, Goutagny R, Luppi PH. Localization of the GABAergic neurons responsible for the inhibition of locus coeruleus noradrenergic neurons during paradoxical sleep in the rat. In 3rd Forum of the Federation of European Neuroscience Societies (FENS), Paris (France); 2002.

27. Verret L, Leger L, Fort P, Luppi PH. Cholinergic and noncholinergic brainstem neurons expressing Fos after paradoxical (REM) sleep deprivation and recovery. Eur J Neurosci 2005;21:2488-504.

28. Vanderwolf CH. Hippocampal electrical activity and voluntary movement in the rat. Electroencephalogr Clin Neurophysiol 1969;26:407-18.

29. Morales FR, Roig JA, Monti JM, Macadar O, Budelli R. Septal unit activity and hippocampal EEG during the sleep-wakefulness cycle of the rat. Physiol Behav 1971;6:563-7.

30. Bjorvatn B, Fagerland S, Ursin R. EEG power densities (0.5-20 Hz) in different sleep-wake stages in rats. Physiol Behav 1998;63:413-7.

31. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurements 1960;20:37-46.

32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

33. Sapin E, Lapray D, Berod A, et al. Localization of the brainstem GABAergic neurons controlling paradoxical (REM) sleep. PLoS One 2009;4:e4272.

34. Rytkonen KM, Zitting J, Porkka-Heiskanen T. Automated sleep scoring in rats and mice using the naive Bayes classifier. J Neurosci Methods 2011;202:60-4.

35. Franken P, Dijk DJ, Tobler I, Borbely AA. Sleep deprivation in rats: effects on EEG power spectra, vigilance states, and cortical temperature. Am J Physiol 1991;261:R198-208.

36. Corsi-Cabrera M, Ponce-De-Leon M, Juarez J, Ramos J. Effects of paradoxical sleep deprivation and stress on the waking EEG of the rat. Physiol Behav 1994;55:1021-7.

37. Franken P, Malafosse A, Tafti M. Genetic variation in EEG activity during sleep in inbred mice. Am J Physiol 1998;275:R1127-37.

38. Franken P, Malafosse A, Tafti M. Genetic determinants of sleep regulation in inbred mice. Sleep 1999;22:155-69.

39. Dzirasa K, Ramsey AJ, Takahashi DY, et al. Hyperdopaminergia and NMDA receptor hypofunction disrupt neural phase signaling. J Neurosci 2009;29:8215-24.

40. Dzirasa K, Ribeiro S, Costa R, et al. Dopaminergic control of sleep-wake states. J Neurosci 2006;26:10577-89.

41. Arthaud S, Libourel P-A, Gervasoni D, Fort P, Luppi PH. Selective paradoxical (REM) sleep deprivation in mice using a new unsupervised automated method. In 21st Congress of the European Sleep Research Society, Paris, France; 2012:326-27.

42. Ugalde E, Corsi-Cabrera M, Juarez J, Ramos J, Arce C. Waking electroencephalogram activity as a consequence of sleep and total sleep deprivation in the rat. Sleep 1994;17:226-30.

43. Rechtschaffen A, Bergmann BM, Gilliland MA, Bauer K. Effects of method, duration, and sleep stage on rebounds from sleep deprivation in the rat. Sleep 1999;22:11-31.

44. Vyazovskiy VV, Achermann P, Tobler I. Sleep homeostasis in the rat in the light and dark period. Brain Res Bull 2007;74:37-44.

45. Trachsel L, Tobler I, Achermann P, Borbely AA. Sleep continuity and the REM-nonREM cycle in the rat under baseline conditions and after sleep deprivation. Physiol Behav 1991;49:575-80.

46. Rechtschaffen A, Gilliland MA, Bergmann BM, Winter JB. Physiological correlates of prolonged sleep deprivation in rats. Science 1983;221:182-4.

47. Rechtschaffen A, Bergmann BM. Sleep deprivation in the rat: an update of the 1989 paper. Sleep 2002;25:18-24.

48. Libourel PA, Corneyllie A, Chouvet G, Luppi PH, Gervasoni D. Unsupervised paradoxical sleep deprivation using polygraphic signals in rats: a new alternative to the "flower pot" technique. In 40th Annual Meeting of the Society for Neuroscience, San Diego (USA); 2010.

49. Jego S, Salvert D, Renouard L, et al. Tuberal hypothalamic neurons secreting the satiety molecule Nesfatin-1 are critically involved in paradoxical (REM) sleep homeostasis. PLoS One 2012;7:e52525.

50. Gottesmann C. Données sur l'activité corticale au cours du sommeil profond chez le rat. C R Soc Biol 1964;158:1829-34.

51. Benington JH, Kodali SK, Heller HC. Scoring transitions to REM sleep in rats based on the EEG phenomena of pre-REM sleep: an improved analysis of sleep structure. Sleep 1994;17:28-36.

Figure legends

Figure 1: Block diagram of the algorithm. The self-training phase is done on a baseline file recorded between 10:00 18:00 (gray rectangle in the timeline at the bottom). The first step of this phase performs the extraction of the 5 indices over 5 s from the first to the last epoch (Ep_0 to Ep_{max}), establishes their distribution, and computes the normalization of these cumulative distributions by a piecewise cubic Hermite polynomial fitting of the quantiles. The transfer functions generated to normalize the indices are exported in a text file for use in the real time scoring and PS deprivation. The second step, self-training, starts with a set of indices with a mean initially set to either 0.9 (high) or 0.1 (low), based on the *a priori* knowledge of their high or low level for a given state. The standard deviation for each index is initially set to 1. The state templates are incrementally updated with the values of individual epochs after the calculation of a probability of each state by way of a likelihood function. The maximal probability among the probability of WK, SWS, and PS sets the state template to be updated if the maximal probability is higher than 0.1 and at least 10 times higher than the others. At the end of the training phase, state templates are saved and exported in a text file. The scoring phase is done in real time for the PS deprivation (black rectangles in the timeline). During this phase and for every incoming epoch, the same processes as those described in the training phase are carried out. The main differences are that the normalization is done from the transfer function generated during the training phase and the likelihood is evaluated from the templates obtain at the end of the training phase. The final score of the epoch under consideration is then obtained by taking the maximum of likelihood among WK, SWS, and PS probabilities.

Figure 2: Distribution of the five indices extracted from 5-s epochs as a function of state. Each index allows a distinction between one state and the 2 others. EEG power ratios 1 and 2 are obtained from fast Fourier transform (FFT). The standard error of the rectified EEG is an index of dispersion of the EEG that is lower during both WK and PS and high during SWS, and has therefore a distribution across states that is similar to the one of the EEG power ratio 2 (0.5-20Hz)/(0.5-55Hz). What can appear as a redundancy is, however, not detrimental to the distinction between SWS and activated states. Taken individually these 5 indices show distinct discriminative power, inversely related to the overlap of the distribution across states. Importantly, the one-dimensional overlap disappears when considering all indices in a 5-dimensional space.

Figure 3: Distribution of EEG zero-crossing values before and after the multinomial normalization. The histograms show the distribution of the raw (left side) and normalized (right side) zero crossing values observed in all epochs, and separately in PS, SWS, and WK epochs. The number and width of classes in the histograms are the same within a column.

Figure 1

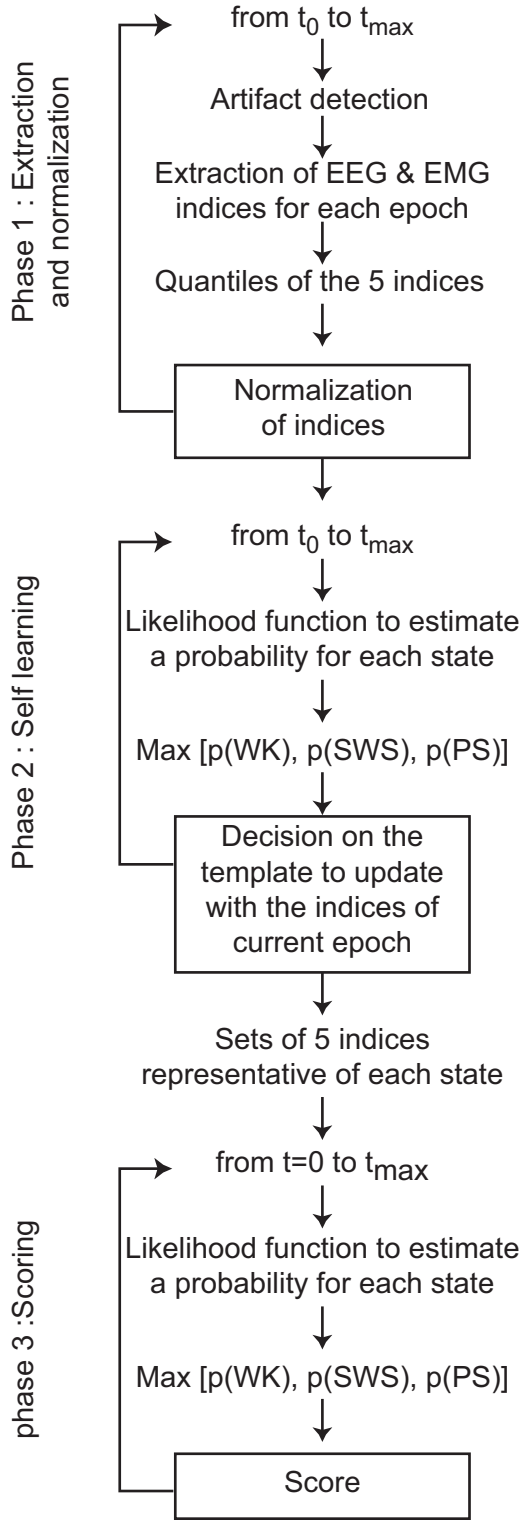

Phase 1 : Extraction and normalization
from t_0 to t_{max}
Artifact detection
Extraction of EEG & EMG indices for each epoch
Quantiles of the 5 indices
Normalization of indices
Phase 2 : Self learning
from t_0 to t_{max}
Likelihood function to estimate a probability for each state
Max [p(WK), p(SWS), p(PS)]
Decision on the template to update with the indices of current epoch
Sets of 5 indices representative of each state
phase 3 :Scoring
from t=0 to t_{max}
Likelihood function to estimate a probability for each state
Max [p(WK), p(SWS), p(PS)]
Score

Figure 2

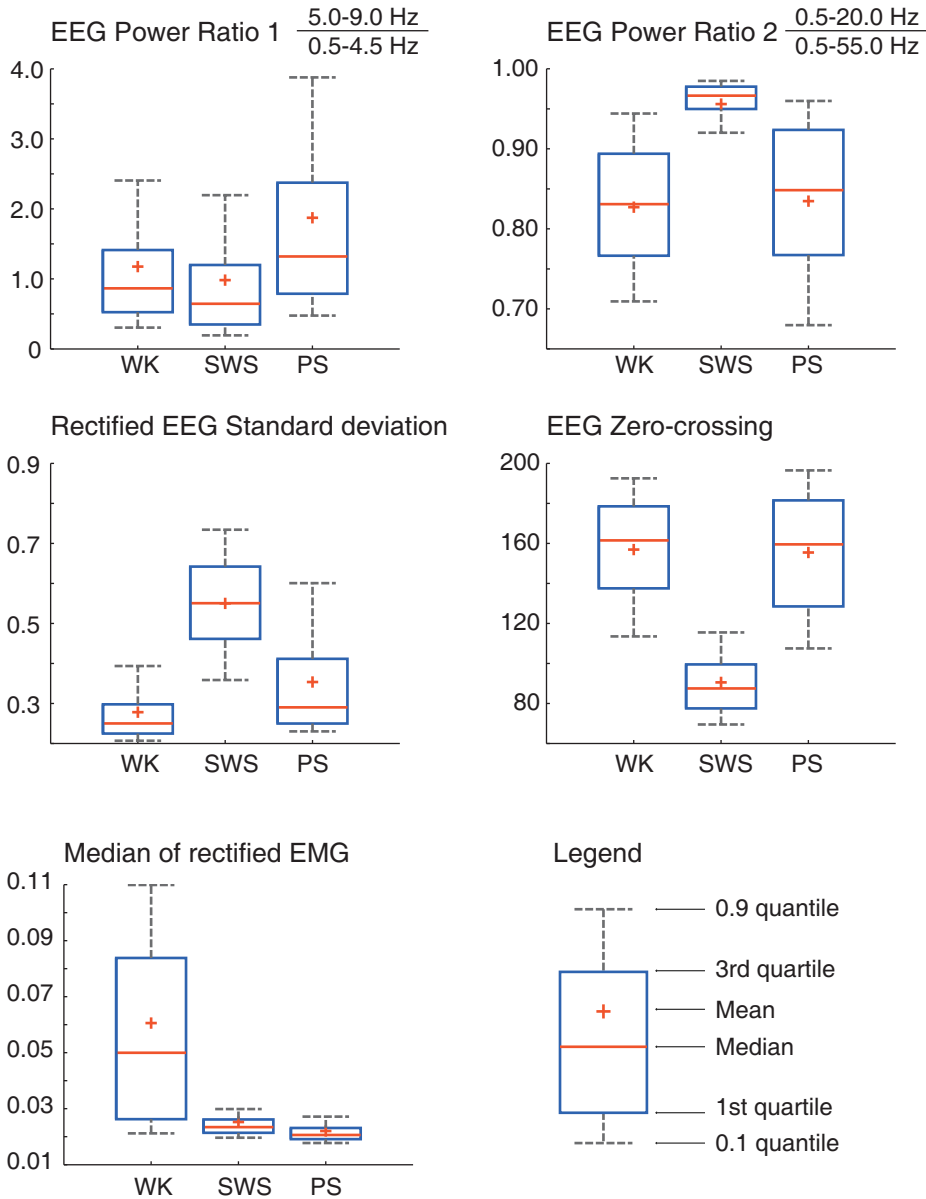

EEG Power Ratio 1
5.0-9.0 Hz
0.5-4.5 Hz
4.0
3.0
2.0
1.0
0
WK
SWS
PS
EEG Power Ratio 2
0.5-20.0 Hz
0.5-55.0 Hz
1.00
0.90
0.80
0.70
WK
SWS
PS
Rectified EEG Standard deviation
0.9
0.7
0.5
0.3
WK
SWS
PS
EEG Zero-crossing
200
160
120
80
WK
SWS
PS
Median of rectified EMG
0.11
0.09
0.07
0.05
0.03
0.01
WK
SWS
PS
Legend
0.9 quantile
3rd quartile
Mean
Median
1st quartile
0.1 quantile

Figure 3

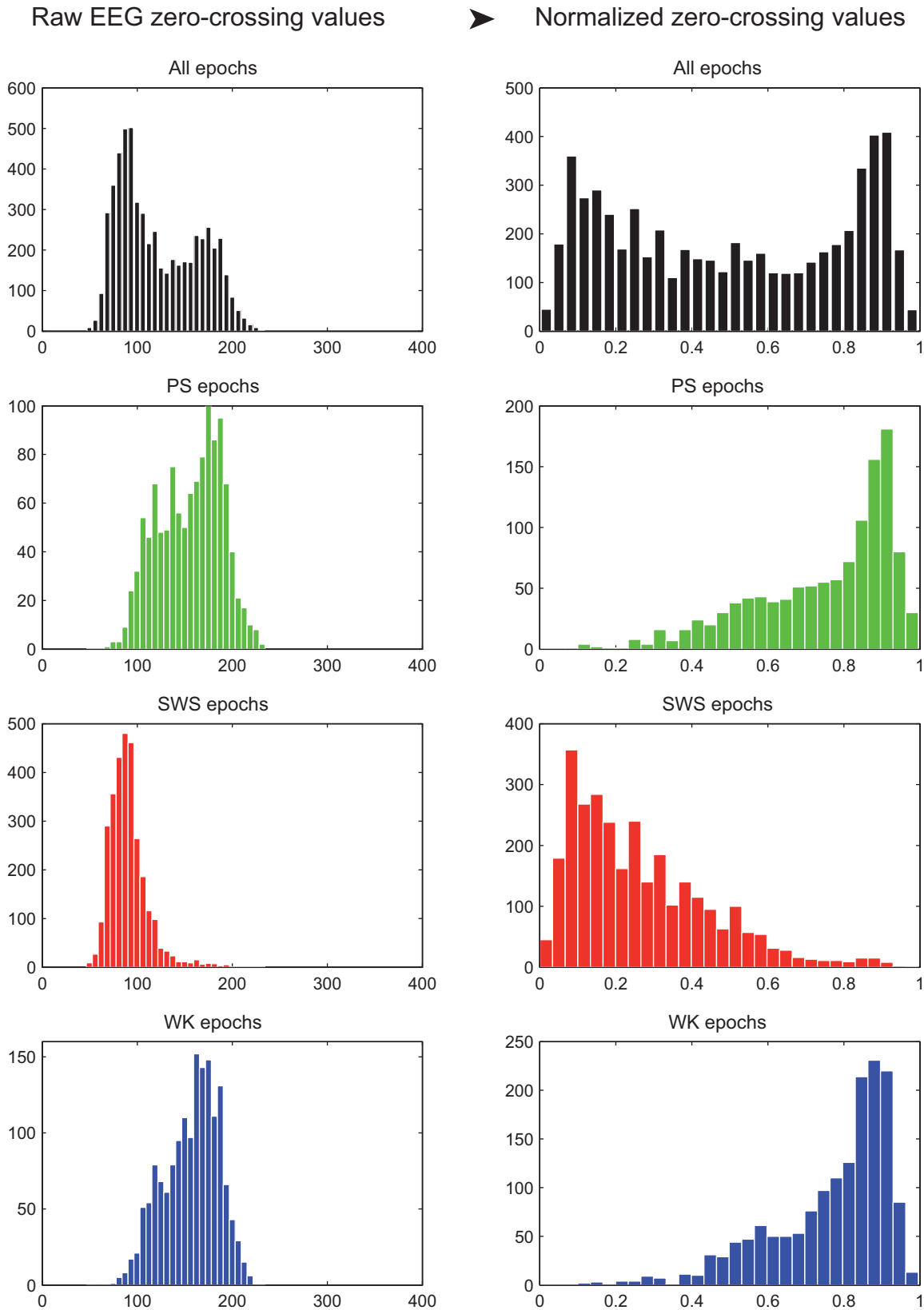

Raw EEG zero-crossing values
Normalized zero-crossing values
All epochs
All epochs
PS epochs
PS epochs
SWS epochs
SWS epochs
WK epochs
WK epochs

Figure 4

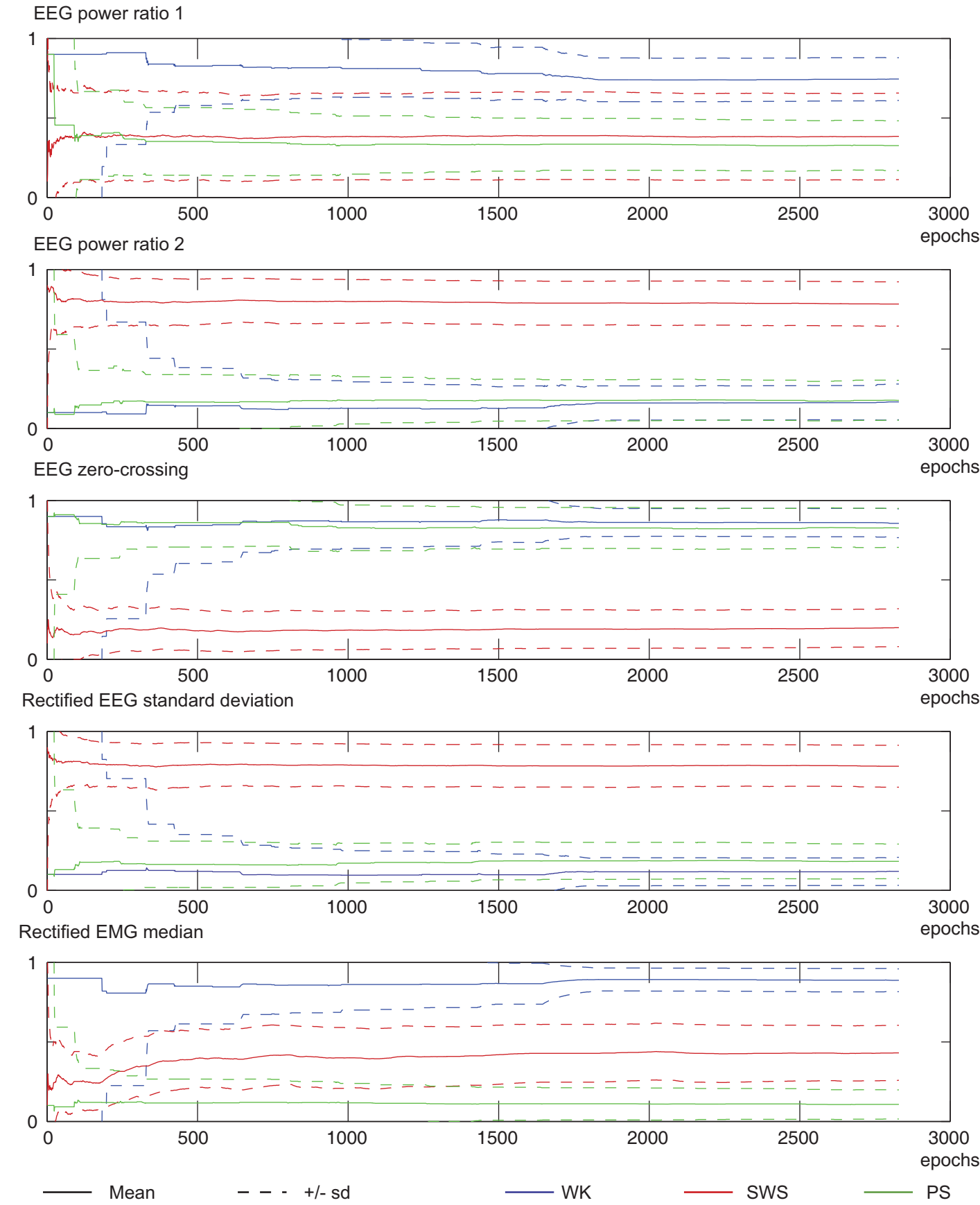

EEG power ratio 1
1
0
0
500
1000
1500
2000
2500
3000
epochs
EEG power ratio 2
EEG zero-crossing
Rectified EEG standard deviation
Rectified EMG median
Mean
+/- sd
WK
SWS
PS

Figure 5

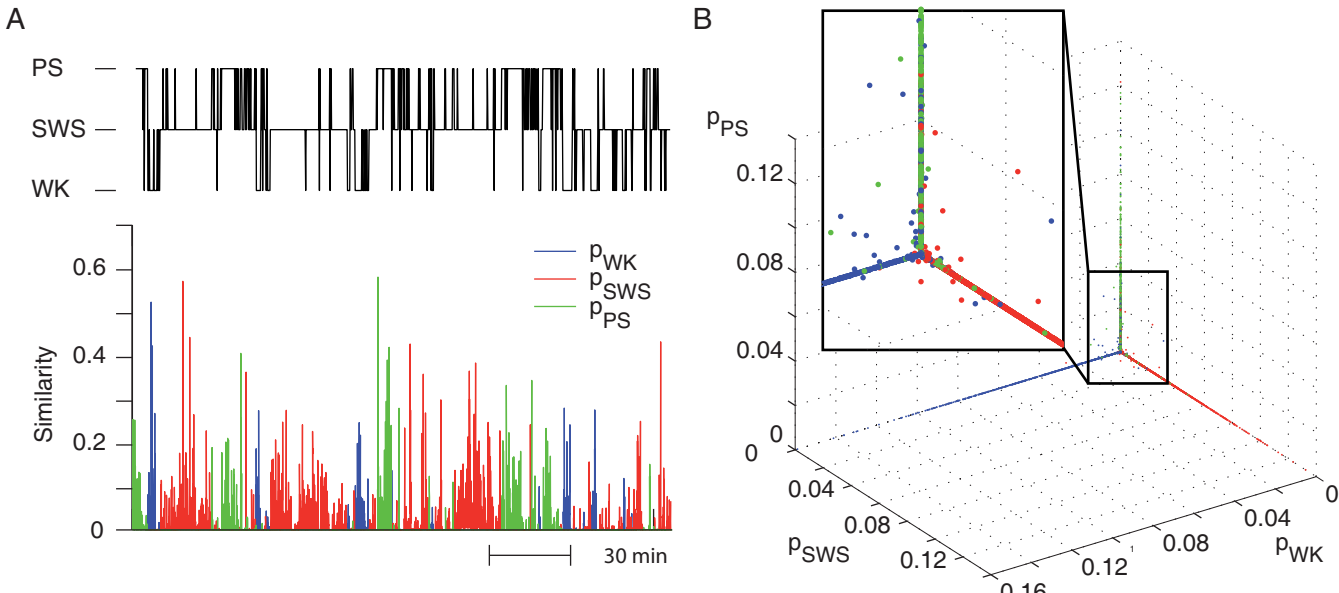
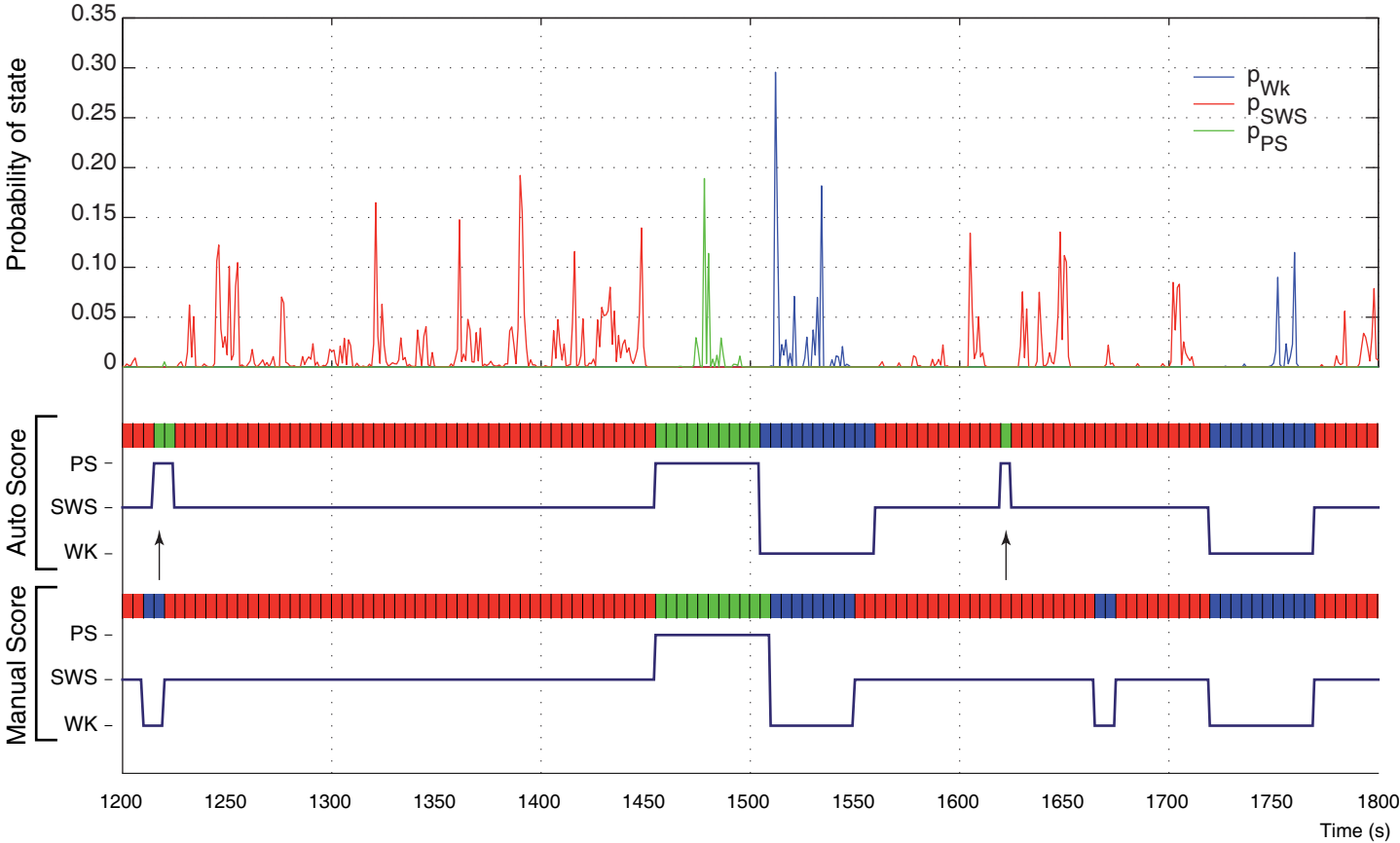

A
PS
SWS
WK
Similarity
0.6
0.4
0.2
0
p_{WK}
p_{SWS}
p_{PS}
30 min
B
p_{PS}
0.12
0.08
0.04
0
0
0.04
0.08
0.12
0.16
p_{SWS}
0
0.04
0.08
0.12
p_{WK}

Figure 6. Comparison of manual and automatic scores



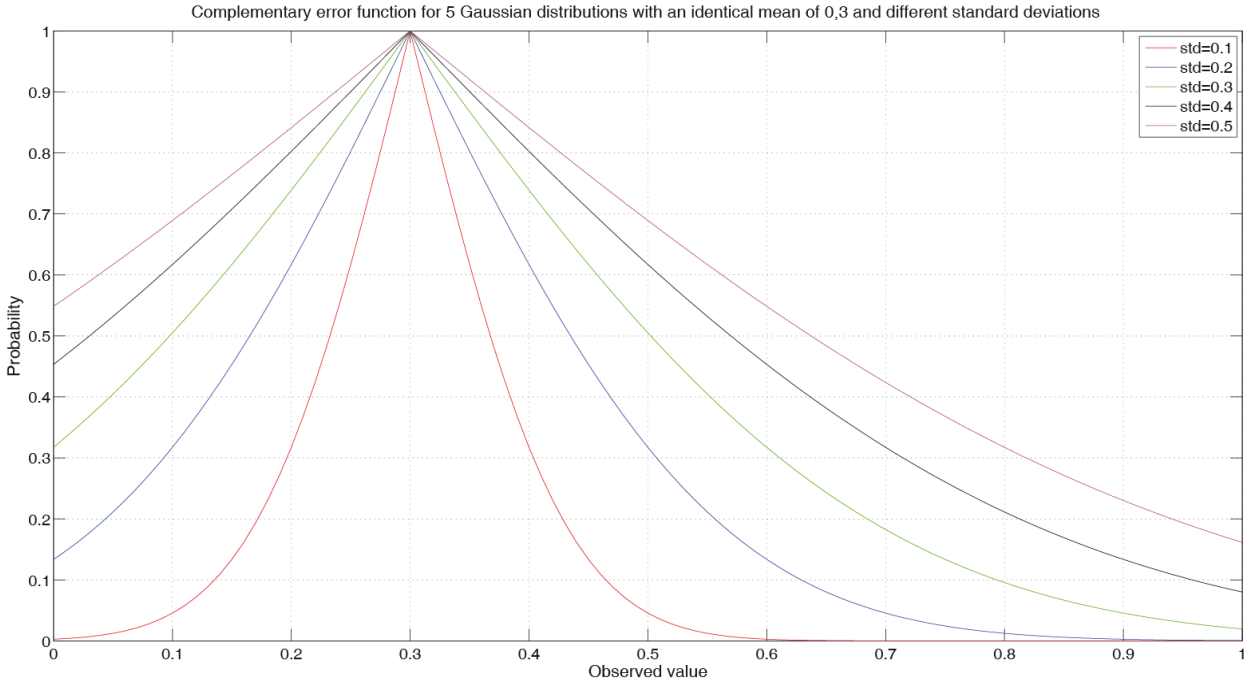
This supplementary material is intended to provide additional details on the probabilistic basis of our Bayesian classifier, the normalization process, and measurements of the performance of the state detection algorithm.

Probabilistic basis our sleep-scoring algorithm

If we consider a standard normal distribution, i.e. a normalized Gaussian function with mean m=0, and standard deviation s=0.5, the complementary Gaussian error function (*erfc*) gives an estimation of the proximity to the mean. By using the function below we estimate for a state *k* the proximity of an observed value x to the mean of a distribution with mean m and standard deviation s.

$$p_k = erfc\left(\frac{|\, x - m_k \,|}{\sqrt{2}.s_k}\right)$$

Below is a graph of this function for five Gaussian distributions sharing the same mean of 0.3, but having distinct standard deviations.



For a single index, for instance the EEG zero crossing, we can estimate the probabilities p_{WK}, p_{SWS}, and p_{PS} for an observed value x to be close to the respective mean (m_{WK}, m_{SWS}, and m_{PS}) of the distributions (with standard deviations s_{WK}, s_{SWS} and s_{PS}) of the index for epochs of WK, SWS, and PS.

The reasoning for single distributions can be extended to the n indices characterizing a state k. Given the law of total probability, we can compute the following product of probabilities:

$$\mathrm{F}(\mathrm{x}, \mathrm{p_k}) = \prod_{\mathrm{j}} (p_{j,k}), \qquad \text{with } \mathrm{j} = 1, \mathrm{n}$$

$$\text{where } p_{j,k} = erfc\left(\frac{|\, \mathrm{x_j} - \mathrm{m_{j,k}} \,|}{\sqrt{2}.s_{j,k}}\right)$$

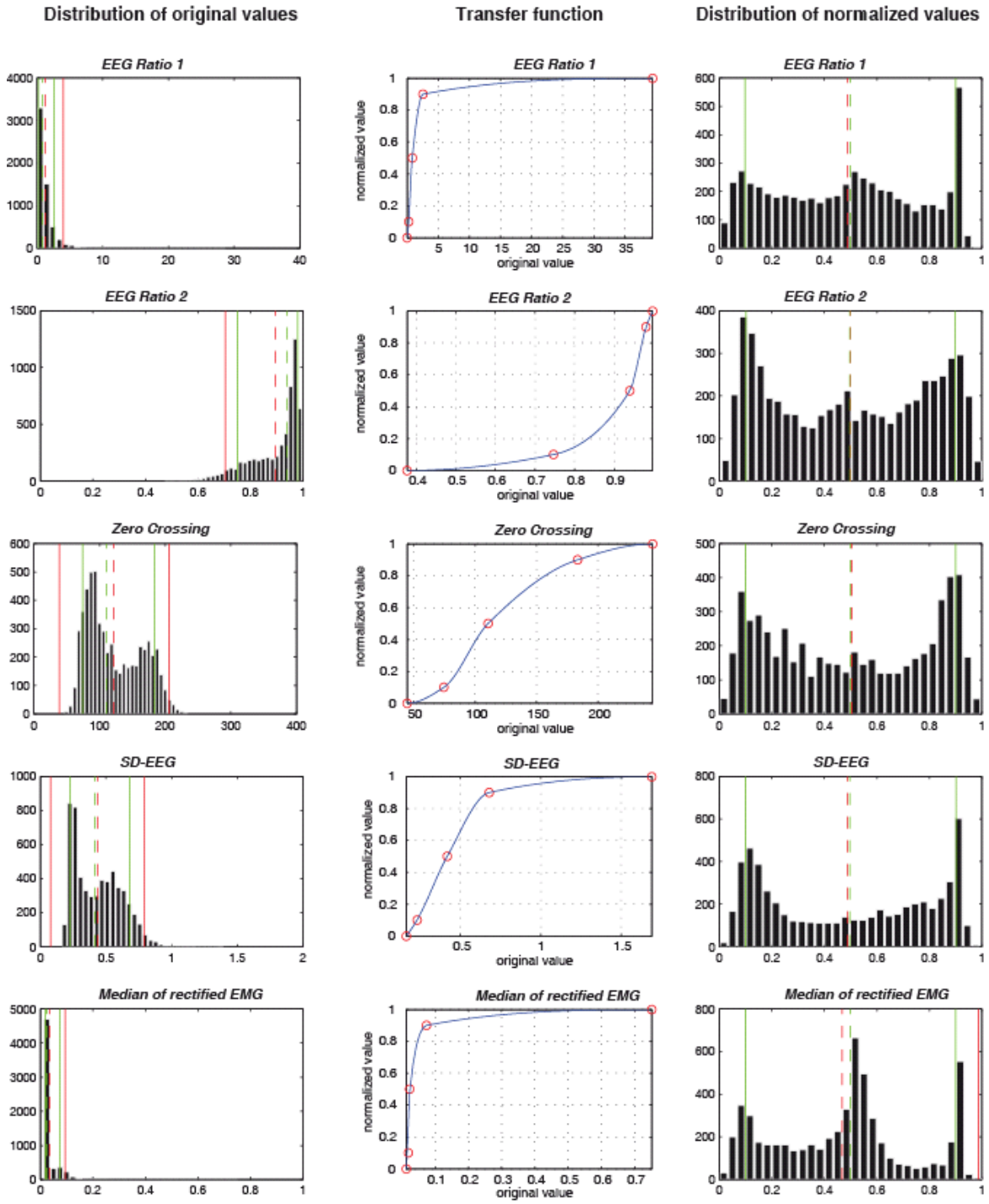
The function $\mathrm{F}(\mathrm{x}, \mathrm{p_k})$ is a likelihood function of the class *k*, given the outcome x of a sample. For a given state the actual value of a likelihood function bears no meaning but it is used in comparison with the values calculated for the other states: with *K* classes characterized by their *n*-dimensional template, and x_j an observation (epoch to score), the maximum of the *K* likelihood functions will give the class the observation *x* likely belongs to.

Table of indices characterizing each state and their levels, known from physiological observations

	WK	SWS	PS
Standard deviation of rectified EEG	Low	High	Low
EEG zero crossing	High	Low	High
EEG Power ratio 1 5-9Hz / 0.5-4.5Hz	High	Low	High
EEG Power ratio 2 0.5-20Hz / 0.5-55Hz	Low	High	Low
Mean amplitude of rectified EMG	High	Low	Low

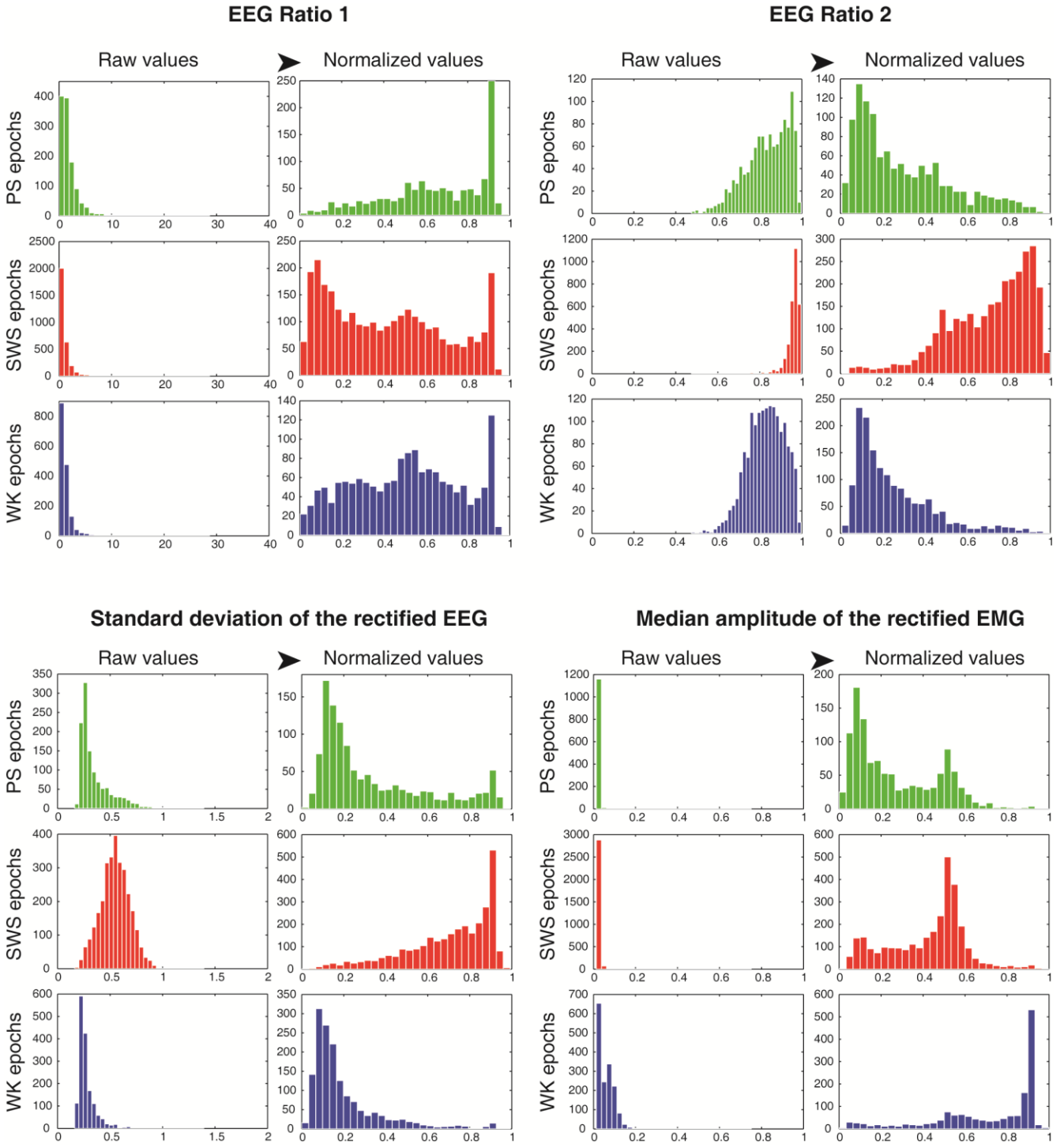
Polynomial normalization process

Our sleep scoring algorithm processes the values of the five indices to bound them between 0 and 1 using a Piecewise Cubic Hermite Polynomial (PCHIP) regression on the quantiles (Fritsch and Carlson, 1980). The following figure illustrates the transfer functions and their effect on the distribution of the indices for all epochs.



Legend : - - - - mean, —— mean ± 2 standard deviation, - - - - median, —— quantiles 0.1 and 0.9 (histograms)
O quantiles 0 (minimum), 0.1, 0.5 (median), 0.9, and 1 (maximum)

The following figure illustrates the effect of the polynomial normalization on the distributions of the EEG Power Ratio 1 (5-9Hz / 0.5-4.5Hz), EEG Power Ratio 2 (0.5-20Hz / 0.5-55Hz), the standard deviation of the EEG, and of the median amplitude of the rectified EMG. Normalization of the EEG zero crossing is illustrated in Figure 3 of the manuscript.



Parameters calculated to evaluate the performance of the algorithm

Confusion matrices are formed by counting the number of epochs classified in the three classes by the Human expert and by the algorithm

		Automated score			
		WK	SWS	PS	
Manual score	WK	n_{11}			$n_{1.}$
	SWS		n_{22}		$n_{2.}$
	PS			n_{33}	$n_{3.}$
		$n_{.1}$	$n_{.2}$	$n_{.3}$	n

The joint probability of agreement is the number of epochs scored in each state both by the Human expert and the algorithm over the total number of epochs $(n_{11}+n_{22}+n_{33})/n$.

Cohen's (unweighted) Kappa coefficient (Cohen, 1960) expresses a relative difference between the observed agreement among raters p_o and the hypothetical probability p_e of chance agreement (under the hypothesis of independence between raters):

$$K = \frac{P_o - P_e}{1 - P_e}$$

where $p_o = \sum_{i=1}^{y} p_{ii} = \frac{1}{n}\sum_{i=1}^{y} n_{ii}$

and $p_e = \sum_{i=1}^{y} p_{.i}p_{i.} = \frac{1}{n^2}\sum_{i=1}^{y} n_{.i}n_{i.}$

with y the number of classes, $p_{.i}$ the column probabilities and $p_{i.}$ the row probabilities.

Below is the qualitative classification of agreement as a function of Kappa values that has been proposed by Landis and Koch (1977). Note that limits between classes are arbitrary.

Agreement	Kappa
Almost perfect	> 0,81
Substantial	0,80 - 0,61
Moderate	0,60 - 0,41
Fair	0,40 - 0,21
Slight	0,20 - 0,0
Poor	< 0,0

References

Cohen J. (1960) A coefficient of agreement for nominal scales., Educ. Psychol. Meas. 20: 27-46.

Landis J.R., Koch G.G. (1977) The measurement of observer agreement for categorical data. Biometrics 33;159-174.

For a given class or state, sensitivity, specificity, positive and negative predictive values (PPV and NPV) are calculated from the binary matrix below derived from the previous matrix.

		State assessed visually		
		Present	Absent	
Algorithm's outcome	Detected	True positive (TP)	False positive (FP)	
	Not detected	False negative (FN)	True negative (TN)	

Sensitivity is defined as the ratio TP / (TP + FN). Also called true positive rate, sensitivity measures the proportion of actual positive detection identified as such. It is equivalent to the probability of a positive detection when the state is present.

Specificity is defined as the ratio TN / (TN + FP). Also called true negative rate, it is equivalent to the probability of a negative detection when the state is absent.

The PPV is defined as the ratio TP / (TP + FP); the NPV is defined as the ratio TN / (TN + FN). The PPV is equivalent to the probability that a detection of a state by the algorithm corresponds to its presence. In other words it is the probability that a state is present when the algorithm positively detects it. Conversely, the NPV is the probability that a non detection of a state by the algorithm corresponds to its absence. Both PPV and NPV depend on the prevalence of the state, i.e. on its proportion in a sleep recording.

Note that numbers on the Y scales are different between states, reflecting their distinct prevalence.

Figure 4: Convergence of the five template values during the training phase. For each index, template values are initially set to either 0.1 (low) or 0.9 (high), with a standard deviation of 1. Subsequently, a probability is calculated for each state and the max likelihood is use to select which state template is updated. The average and standard deviation of the five indices are recalculated accordingly for that state. At the end of the training phase a 5-dimensional set of parameters is obtained for the 3 states. This process is adaptive in the sense that templates are updated with new data sample.

Figure 5: For each epoch and each state the likelihood function combined the probabilities calculated from the five indices in a single value. In terms of statistical inference, the result of the likelihood function is the probability that a given epoch belongs to the class WK, SWS, or PS, and thus reflected the similarity to the representative template of each state. Panel A illustrates the evolution of the 3 probabilities across time, and the hypnogram resulting from the decision rule based on the maximum likelihood. Even when the values of probabilities are low, all epochs are labeled with the state for which the likelihood value is maximal. Panel B shows the 3-dimensional distribution of the values of probabilities for each epoch. Each dot represents a single epoch with its probabilities to resemble to WK, SWS and PS as X, Y, and Z coordinates, and is color-coded with the state visually assigned. The great majority of dots are clustered along the axes, demonstrating the high concordance between the state manually assigned and the result of the selection based on the maximum probability. Note that the rare dots that appear misplaced (insert) have very low probability values.

Figure 6: Comparison of the score obtained with the algorithm and the ones obtained from a human expert. Sporadic differences are clearly visible on these color-coded hypnograms, with most of them coinciding with very low probabilities, i.e., high uncertainty. In this example PS is erroneously attributed to two consecutive epochs, manually scored as WK (first arrow), that correspond to a micro-arousal (activated EEG) without concomitant muscle activity. Similarly another epoch with a low amplitude EEG and very low EMG was detected as PS (second arrow). What can appear as an over detection of PS illustrates the originality of the method aimed at not missing any PS occurrence in the context of a selective PS deprivation.