Published in final edited form as:

Neurosci Lett. 2010 January 18; 469(1): 97. doi:10.1016/j.neulet.2009.11.052.

Manual rat sleep classification in principal component space

Timothy P. Gilmour,

The Pennsylvania State University, University Park, PA

Jidong Fang,

The Pennsylvania State University College of Medicine, Hershey, PA

Zhiwei Guan, and

The Pennsylvania State University College of Medicine, Hershey, PA

Thyagarajan Subramanian

The Pennsylvania State University College of Medicine, Hershey, PA

Timothy P. Gilmour: ; Jidong Fang: jfang@psu.edu; Zhiwei Guan: zxcg6@psu.edu; Thyagarajan Subramanian: tsubramanian@hmc.psu.edu

Abstract

A simple method is described for using principal component analysis (PCA) to score rat sleep recordings as awake, rapid-eye-movement (REM) sleep, or non-REM (NREM) sleep. PCA was used to reduce the dimensionality of the features extracted from each epoch to three, and the projections were then graphed in a scatterplot where the clusters were visually apparent. The clusters were then directly manually selected, classifying the entire recording at once. The method was tested in a set of ten 24-hour rat sleep electroencephalogram (EEG) and electromyogram (EMG) recordings. Classifications by two human raters performing traditional epoch-by-epoch scoring were blindly compared with classifications by another two human raters using the new PCA method. Overall inter-rater median percent agreements ranged between 93.7% and 94.9%. Median Cohen's kappa coefficient ranged from 0.890 to 0.909. The PCA method on average required about five minutes for classification of each 24-hour recording. The combination of good accuracy and reduced time compared to traditional sleep scoring suggests that the method may be useful for sleep research.

Keywords

Rat; Sleep scoring; Principal Component Analysis; Electroencephalogram; Electromyogram

Introduction

Sleep classification occupies an important role in medical sleep disorder research, basic neuroscience research, and industrial and transportation safety. Rats are the most frequently used species in animal sleep studies. Classification of rodent sleep and wakefulness using

Corresponding author: Gilmour, Timothy P. Neurology Department, 500 University Drive, Hershey PA 17033, timgilmour@psu.edu, (717) 531-0003 x 283961 phone, (717) 531-6916 fax.

Suggested Reviewers: David Rye, drye@emory.edu, Hans Lüders, scolarc@ccf.org, Hagai Bergman, hagaibe@ekmd.huji.ac.il, Imad Najm, najmi@ccf.org, Stella Papa, spapa@emory.edu, Sudhansu Chokroverty, schok@att.net, Ausim Azizi, ausim.azizi@temple.edu, Richard Brenner, james.brenner@bicrad.com, Nancy Foldvary, foldvan@ccf.org

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

electroencephalogram (EEG) and electromyogram (EMG) recordings has traditionally been a time-intensive manual task. The spectrum of rodent sleep is typically divided into three categories: awake, rapid-eye-movement (REM) sleep, and non-rapid-eye-movement (NREM) sleep [11,13]. These stages can be distinguished by the EMG power and by the EEG power in particular frequency bands. Wakefulness is characterized by high and variable EMG activity and medium EEG activity with power concentrated in beta (15–30Hz) and gamma (30–100Hz) ranges. NREM sleep is characterized by low EMG activity and high EEG activity with power concentrated in the delta (1–4Hz) band. REM sleep is characterized by very low EMG activity and medium EEG activity with power concentrated in the theta (6–10Hz) band.

Many systems have been created in the past two decades to automate this classification, in order to reduce the scoring time and to reduce variability between raters and across time. A recent review of these systems was published by Robert and colleagues [13]. These systems have achieved varying degrees of correspondence with the human rater “gold-standard” scorings.

The automated classification systems fall into two general types: heuristic classifiers or machine-learning classifiers. Both types first extract features from the data, such as the EEG spectral power between 0.5Hz – 5Hz within a particular epoch, or the average EMG amplitude within the epoch. The heuristic classifiers then impose manually-adjustable thresholds on the extracted features to classify each epoch. The machine-learning classifiers use a set of training data to automatically learn classification thresholds, and then likewise impose those thresholds on the data to classify each epoch. The earliest automated systems used manually-adjusted thresholds [9]. Newer systems have tended toward using machine learning algorithms such as neural networks [14] or support vector machines [6]. Many automated classifiers also typically include a post-processing step which takes into account the sequential location of an epoch compared with its temporal neighbors [5,15].

Both heuristic and machine-learning automated approaches, while saving some time over fully manual classification, still require manual adjustment of parameters or manual selection of a training set before the classification. An alternate way to classify the data is to explicitly transform the data into a lower dimensional state space, followed by manual or automatic classification. If the dimensionality is reduced to three using factor analysis such as principal component analysis (PCA), the clusters may be shown visually in a three-dimensional (3-D) scatterplot and the classification may be performed by hand. Such scatterplots have been used in the past as visualization tools [2,3,7,8,11]. For example, Low and colleagues used 3-D EEG PCA scatterplots to compare mammalian and bird sleep, showing distinct clusters corresponding to wakefulness and to NREM, REM, and intermediate sleep (IS). Gervasoni and colleagues used 2-D PCA scatterplots to show the EEG and intracranial local field potentials, showing clusters corresponding to NREM, REM, and IS, active and quiet wakefulness, and even whisker twitching.

However, to our knowledge the idea of using these cluster plots to directly score the recording has not previously been demonstrated. As described below, this approach to manual sleep-wake classification can be faster than traditional epoch-by-epoch manual scoring while eliminating the need for manual selection of training sets or manual adjustment of time-domain heuristic thresholds.

Methods

Theory

The basic principle behind the classification method is to reduce the dimensionality of the feature vector to three dimensions and then ask the human scorer to manually classify the data

based on visually-apparent clusters of similar epochs. We used principal component analysis (PCA) for the dimensionality reduction, although in principle other decompositions could be used.

PCA is a standard technique that computes a set of vectors that point in the directions of greatest (and sequentially decreasing) variance of the original data. For dimensionality reduction, the eigenvectors corresponding to the k largest eigenvalues of the covariance matrix, C , are computed. The full covariance matrix is $C = XX^T = WSS^TW^T$, where X is the data matrix, W is the full matrix of eigenvectors (“principal components”), and S is the full diagonal matrix of eigenvalues. Using only the k largest eigenvectors gives the reduced-dimension data: $Y = W_k^TX$, where W_k is the cropped eigenvector matrix.

For sleep classification, the waveform data is first partitioned into epochs and the features computed for each epoch. The dimensionality of the feature matrix is then reduced to three ($k = 3$) by PCA. The projections of each epoch onto the three largest eigenvectors are drawn onto a three-dimensional scatterplot. The human scorer then rotates the plot for the best view of the clusters, and classifies awake, NREM-sleep and REM-sleep by drawing a box around the visually apparent clusters. Finally, the scoring was refined by classifying as awake any transitional epochs containing high EMG activity but clustered as “NREM” due to high delta activity or artifact. This refinement was accomplished using a second scatterplot with axes of EMG power, EEG high beta, and EEG delta. Figure 1 depicts the scoring process.

Experimental Validation

To test the method, separate 24-hour EEG/EMG recordings from 10 adult male Sprague-Dawley rats were scored by four human raters: two using traditional epoch-by-epoch manual scoring and two using the new PCA-space scoring. The rats were implanted with stainless steel screws over the frontal and parietal cortices for EEG recordings and with wire electrodes in the neck muscles for EMG recordings. Rats were kept on a 12:12 hour light-dark cycle. 24-hour sleep recording started after 2 weeks recovery and additional 4 days of adaptation to the recording cable. Standard rat chow and water were made freely available. Experiments were carried out in accordance with the guidelines in the National Institute of Health Guide for the Care and Use of Laboratory Animals, 1996 revision, and all experimental protocols used in the experiment were approved by the Institutional Animal Care and Use Committee (IACUC).

After recovery and adaption, one channel of EEG and one channel of EMG were recorded from each rat for a 24-hour period. Each signal was amplified and digitized at 128 Hz. The signals were filtered offline in Matlab (Mathworks). After filtering, the signals were segmented into 10 second epochs.

Raters T1 and T2 performed traditional epoch-by-epoch classification on the recordings using the commercial SleepWave program. Raters P1 and P2 performed PCA-space classification. All raters were kept blind to the others' scoring.

For the PCA scoring, seven features were computed from each epoch based on our observations and reports from previous literature [6,13]. These were: absolute EEG power in the 1–4Hz (delta), 5–9Hz (theta), 10–20Hz (low beta), and 30–40Hz (high beta) bands, absolute EMG power in the 1–10Hz band, theta-to-delta ratio, and beta-to-delta ratio (Figure 2).

These seven features were collected into a feature matrix for each 24-hour recording. The three largest principal components were then computed and used to reduce the dimensionality of the feature matrix to three, and the projections were plotted in a three-dimensional scatterplot. The visual clusters were manually classified and the classification of the epochs was graphed as depicted in Figure 1. The secondary scatterplot with axes of the EMG, EEG high beta, and

EEG delta was made available to raters P1 and P2 to aid in refining the scoring based on absolute EMG score as described above.

Results

Overall, the PCA-scoring method closely matched the traditional epoch-by-epoch scoring. Figure 3 shows the agreements between the classifications of all pairwise combinations of raters. The agreement between the two methods was calculated separately for each class (wake, NREM, or REM) as the number of epochs that were classified as that type by both the traditional scoring and the PCA scoring, divided by the total number of epochs classified as that type by the traditional scoring. Figure 3 also shows the values of Cohen's kappa for all pairs of raters. Values of kappa between 0.8 and 1.0 are generally considered to be "very good agreement" [4]. All of the scoring pair kappas fell within this range. Additional data such as a representative confusion matrix, table of the Figure 3 median values, and final cluster figures for all ten rats may be found online in the supplementary material.

Table 1 lists the average classification time for each 24-hour recording. In addition to this time, approximately 1.2 minutes was used per recording used for filtering and processing the data on a Pentium Core 2 T7200 2.00Ghz computer running Matlab 7.2 on Windows XP.

Two methods were used to quantify the cluster separation. First, a one-way multivariate analysis of variance (MANOVA) was performed on the PCA projections to test the null hypotheses that the cluster means were identical (zero-dimensional test) or were on the same line (one-dimensional test). Both null hypotheses were rejected for every rat's clusters at the $\alpha = 0.05$ level of significance, as shown in Table 1.

Second, linear discriminant analysis was used to test the cluster separation. Since the classifications themselves had been performed using manually-drawn lines, simple LDA validation using the original training set would be meaningless, as it would always give "perfect" separation. Instead, 10-fold cross-validation was performed as follows. The temporal order of the epochs was randomized, and a linear discriminant classifier was trained on 90% of each recording and tested on the remaining 10%. The misclassification error was defined as 100% minus the percent of epoch classifications that matched the testing set. This procedure of randomization, training, and classification was repeated 10 times per rat. Table 1 shows the average misclassification results from this cross-validation.

Discussion

The PCA-space classification method is able to classify the sleep/wake state of the rodent to a high degree of agreement with the traditional manual epoch-by-epoch scoring method. The percentage of agreement is comparable or higher than the human inter-rater agreements reported in the literature ranging from 80% to 95% [1,6,10,12,13,15]. The higher agreement on awake and NREM segments than REM segments also matches previous literature reports [6,10,12,13]. REM sleep tends to be harder to classify because its EEG waveforms look similar to wakefulness, and its EMG waveforms show a mixture of hypoactivity interspersed with occasional muscle twitches.

The PCA scores from rater P1 were closer to the traditional scores than from P2, probably reflecting a higher degree of training of rater P1 on the new paradigm. Rater P1 had previous experience of more than one hundred scoring sessions of one hour or more each, while rater P2 had previous experience of only two scoring sessions lasting less than one hour each. The difference in experience between the PCA-space raters combined with the general agreement between their scorings highlights the ease of learning the new method. The REM sleep showed

the poorest classification agreement (especially from rater P2), as expected because of the difficulty of scoring REM sleep in general.

In a previous study, Gervasoni and colleagues reported mediocre sleep classification accuracies using EEG alone in a linear discriminant classifier, suggesting that surface recordings are not enough to produce clear clusters and accurately classify the full range of sleep [7]. However, they did not record EMG in their paradigm, and it is well known that the EMG signal is useful in separating REM sleep from wakefulness. Thus the higher accuracies we report here without using any intracranial recordings are likely attributable to our inclusion of the EMG signal.

In addition to accuracy, the PCA-space method provides a significant speed improvement over the traditional epoch-by-epoch scoring method because it allows classification of all the epochs at once. Assuming thirty minutes as a typical required time for scoring 24 hours of sleep data using traditional methods, Table 2 shows that the PCA method was approximately six times faster than traditional scoring.

Comparison of the PCA-space method with previously published fully-automated sleep classification methods suggest both advantages and disadvantages. The main advantages are the elimination of the need for training data or threshold adjustments, and the unique perspective provided by the state-space representation. The main disadvantages are the classification subjectivity and the difficulty of generalization for extremely large datasets.

All published sleep classification methods require human interaction at some level. The methods vary in at what point in the classification the human interaction is required, and to what extent it is required. The two main types of automated classifiers are currently the heuristic and the machine-learning types. Typical heuristic classifiers require manual adjustment of thresholds before running the algorithm for best performance [2,10,13]. Early classifiers required adjustment of the thresholds separately for each rat and each recording. Some newer heuristic classifiers such as that reported by Stephenson and colleagues require only global dataset parameter adjustments, but this classifier achieved only 89% global agreement with human raters (92% wake, 88% NREM, 80% REM) [15]. This adjustment of parameters typically requires initial human analysis of 1 or 2 hours of rat recordings prior to the automated analysis of the remainder of the recordings. By contrast, the PCA-space method has the advantage of not requiring this separate pre-analysis since the whole recording is directly classified at once. Alternatively, one could say that the PCA-space manually-drawn cluster boundaries are themselves the “thresholds”, imposed in PCA-space, which allows faster classification of the entire recording at once.

Machine-learning sleep classifiers also require human involvement in the selection of training data prior to the automated classification. One study using a neural network as the classifier required 40 minutes of sleep data to be manually classified as a training set [14]. Another more recent study using a support vector machine used 2 hours of recorded sleep data for training, requiring approximately 30 minutes of human rating [6]. In contrast, the PCA-space method has the advantage of not requiring any training data, since it is a direct manual approach.

Another advantage of the PCA approach is the unique perspective of being able to graphically visualize the relationship between the sleep and wake states in each data record. Unlike the “black box” approach of the automated classifiers, the PCA approach keeps the human rater involved and alert to the presence of outliers or systematic biases. Meanwhile, the PCA approach is faster than the traditional epoch-by-epoch sleep scoring.

The flip side of retaining human scoring is that the PCA classification is not as objective as completely automatic classification. Different raters or the same rater on different days might draw the boundaries between the clusters differently, resulting in different scoring.

Another disadvantage is that while the time savings in using the PCA method is considerable for up to a certain amount of data, at some point it may be more efficient to invest several hours in procuring a high quality training set for training an automated classifier. This “break-even point” will depend on how stationary are the recordings across sessions and across rats. We did not investigate using the cluster boundaries drawn for one particular 24-hour recording to classify other recordings from the same rat or different rats. This would be an interesting direction for further research. The same problem is faced by the automated classifiers: recordings that are widely disparate in feature content will always need separate training or parameter adjustment for best results.

A potentially fruitful direction for future research would be to use the PCA-space method as a first stage in a multi-stage classification. For example, the PCA method could provide an easy way to pick “representative” epochs from visually apparent cluster centers which could then be used to train machine-learning algorithms for subsequent high-throughput classification.

Alternatively, since the visually-apparent boundaries of the PCA clusters provide an indication of epochs potentially containing transitions between awake and asleep, this unclear subset of the data could be easily flagged in PCA space for subsequent human epoch-by-epoch analysis.

Another potential refinement of the basic PCA method would be to include a postprocessing step to use sequential epoch information, as used in many automated systems [15]. The classification of each epoch could be adjusted depending on the classifications of its antecedent and following neighbors.

Further refinements of the method might include adjusting the features used, and using normalized rather than absolute features to increase inter-subject generalization.

Besides the rat model, the PCA method is anticipated to be useful for other models, such as mouse, an increasingly used sleep model. An important future question is the utility of the PCA method for classifying human sleep. Initial unpublished results from our laboratories suggest that the method has promise for human sleep classification, although full validation is still in progress. This is an important research question because of the thousands of hours spent on sleep classification in clinics worldwide in the diagnosis of sleep apneas and other disorders.

Conclusion

In summary, the PCA space method is a novel way to quickly classify sleep and wakefulness in rat recordings. It provides accuracy that is comparable to or better than inter-rater agreement, while requiring less time than traditional epoch-by-epoch scoring. In contrast to fully automated classification systems that use heuristics or machine-learning algorithms, it does not require any prior adjustment of parameters or training data selection. The PCA method opens future possibilities for hybrid sleep classification systems that combine the best of all types of currently available classifiers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the NIH NINDS RO1NS42402, NIH R41HL084990, and PA Tobacco Settlement Funds Biomedical Research Grant. The authors would like to thank Thomas Wichmann, Ken Jenkins, and the reviewers for their helpful suggestions.

References

1. Ambrogio C, Koebnick J, Quan SF, Ranieri M, Parthasarathy S. Assessment of sleep in ventilator-supported critically ill patients. *Sleep* 2008;31:1559–1568. [PubMed: 19014076]
2. 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. [PubMed: 8191200]
3. Bergmann BM, Winter JB, Rosenberg RS, Rechtschaffen A. NREM sleep with low-voltage EEG in the rat. *Sleep* 1987;10:1–11. [PubMed: 3563244]
4. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 1960;20:37–46.
5. 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. [PubMed: 14583408]
6. Crisler S, Morrissey MJ, Anch AM, Barnett DW. Sleep-stage scoring in the rat using a support vector machine. *J Neurosci Methods* 2008;168:524–534. [PubMed: 18093659]
7. 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;24:11137–11147. [PubMed: 15590930]
8. Karasinski P, Stinus L, Robert C, Limoge A. Real-time sleep-wake scoring in the rat using a single EEG channel. *Sleep* 1994;17:113–119. [PubMed: 8036365]
9. Kohn M, Litchfield D, Branchey M, Brebbia DR. An automatic hybrid analyzer of sleep stages in the rat. *Electroencephalogr Clin Neurophysiol* 1974;37:518–520. [PubMed: 4138909]
10. Louis RP, Lee J, Stephenson R. Design and validation of a computer-based sleep-scoring algorithm. *Journal of Neuroscience Methods* 2004;133:71–80. [PubMed: 14757347]
11. Low PS, Shank SS, Sejnowski TJ, Margoliash D. Mammalian-like features of sleep structure in zebra finches. *Proc Natl Acad Sci U S A* 2008;105:9081–9086. [PubMed: 18579776]
12. 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;17:120–131. [PubMed: 8036366]
13. Robert C, Guilpin C, Limoge A. Automated sleep staging systems in rats. *Journal of Neuroscience Methods* 1999;88:111–122. [PubMed: 10389657]
14. Robert C, Karasinski P, Natowicz R, Limoge A. Adult rat vigilance states discrimination by artificial neural networks using a single EEG channel. *Physiol Behav* 1996;59:1051–1060. [PubMed: 8737892]
15. Stephenson R, Caron AM, Cassel DB, Kostela JC. Automated analysis of sleep-wake state in rats. *J Neurosci Methods*. 2009

Figure 1.

Scatterplots (from Rat 7) illustrating the classification process. Plots A and B show the initial unclassified epochs, plots C and D show the epochs after the initial PCA classification, and plots E and F show the final classification after refinement using the EMG/EEG-delta scatterplot to classify any high-EMG epochs as awake. Dotted lines in plot C and F depict sample manually-adjusted discriminant lines used for the classification. Blue dots indicate NREM sleep epochs, red dots indicate awake epochs, and orange dots indicate REM sleep epochs.

Figure 2.

Sample 24-hr recording showing the original and PCA-reduced features in ascending order. From bottom: features 1–7 (black, in order starting from feature #1 EEG 1–4Hz at the bottom), principal components 3, 2, and 1 (blue, green, and red respectively), and at top the hypnogram resulting from the PCA classification (from rater P1).

Figure 3.

Pairwise comparison of raters, shown by (A) the inter-rater percent agreement and by (B) Cohen's unweighted kappa coefficient. Raters P1 and P2 used the PCA-space method, and raters T1 and T2 used the traditional epoch-by-epoch method. Percent agreement was calculated as the number of concordant epochs of a particular class divided by the total number of epochs of that class. Boxes show the interquartile range, whiskers show the full extent of the data, and center lines show the median.

Table 1

Classification time and cluster quality statistics from both PCA-space raters. The percent of misclassification errors for each rat's PCA clusters was computed from a linear discriminant classifier trained on 90% of each recording and tested on the remaining 10%. Mean values are listed for each rat from 10 randomized repetitions of each classification, and the mean across all rats is listed. MANOVA results test whether the cluster means were the same (zero-dimensional) and whether the means lay on the same line (one-dimensional).

Rater	Rat											
	1	2	3	4	5	6	7	8	9	10	Mean $\pm$ SD	
P1	Time	4.3	3.9	4.3	4.1	5.1	4.7	3.3	3.7	2.9	4.1 $\pm$ 0.7 min.	
	LDA error	5.46	2.81	4.33	3.68	2.35	2.42	1.45	2.33	3.11	3.12 $\pm$ 1.15 %	
	MANOVA	p < 0.00005 for all rats, both zero- and one-dimensional tests										
P2	Time	4.9	5.5	6.9	7.0	9.7	4.8	6.1	3.5	4.1	3.4	5.6 $\pm$ 1.9 min.
	LDA error	5.61	4.55	4.32	4.01	2.95	4.21	2.20	1.64	2.16	3.50	3.52 $\pm$ 1.26 %
	MANOVA	p < 0.00005 for all rats, both zero- and one-dimensional tests										