Naive Scoring of Human Sleep Based on a Hidden Markov Model of the Electroencephalogram

Farid Yaghouby, *Member IEEE-EMBS*, Pradeep Modur and Sridhar Sunderam*, *Member IEEE-EMBS*

***Abstract*— Clinical sleep scoring involves tedious visual review of overnight polysomnograms by a human expert. Many attempts have been made to automate the process by training computer algorithms such as support vector machines and hidden Markov models (HMMs) to replicate human scoring. Such supervised classifiers are typically trained on scored data and then validated on scored out-of-sample data. Here we describe a methodology based on HMMs for scoring an overnight sleep recording without the benefit of a trained initial model. The number of states in the data is not known a priori and is optimized using a Bayes information criterion. When tested on a 22-subject database, this unsupervised classifier agreed well with human scores (mean of Cohen's kappa > 0.7). The HMM also outperformed other unsupervised classifiers (Gaussian mixture models, k-means, and linkage trees), that are capable of naive classification but do not model dynamics, by a significant margin ($p < 0.05$).**

I. Introduction

Sleep quality is a critical determinant of human health and performance. Clinical evaluation of disordered sleep involves overnight polysomnography (PSG) following specific guidelines [1]. A PSG recording includes electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), and other measurements, and is scored by an expert in 30 s epochs into discrete vigilance states, namely wakefulness (Wake), rapid eye movement (REM) sleep, and non-REM (NREM, stages 1-3) sleep [2]. Scoring sleep is difficult and tedious. Many statistical classifiers have been developed to automate this process and replicate human performance [3], sometimes from a single EEG channel alone [4-5]; most require supervision in the form of expert heuristics or a statistical model derived from expert-scored training data to stage sleep; and all are used in essentially the same manner: *i.e.*, by fitting a model to scored data from one set of subjects and validating it on out-of-sample data from another set [3-5]. This gives confidence that the model will work reliably on future subjects.

Supervised classifiers are constrained by the need for (and subjectivity/variability of) human scoring of training data. No method to date generates a reasonable first-pass hypnogram from a sleep recording without supervision: *i.e.*, without previous training. Even hidden Markov models (HMMs), which, strictly speaking, are unsupervised classifiers, are first fitted to training data in which all vigilance states are known to occur, and then used to score test data [6-8]. But in the naive scenario, no initial model is available; nor may all vigilance states occur. Here, we propose a method for using HMMs to score overnight sleep without the benefit of a trained classifier. While supervised classifiers need labeled training data, unsupervised classifiers like the HMM find natural partitions in data that could map signal features onto distinct hidden states. In principle, PSG epochs can be mapped onto vigilance states without prior training—which a supervised classifier cannot do. This could yield a useful first-pass score for a new patient, to be refined by an expert if reasonably accurate.

Implicit in HMMs is the notion of dynamics, that the state follows a trajectory whose likelihood depends on the previous state at any instant. In contrast, most classifiers are "static", *i.e.*, they do not incorporate context when determining state, unless subsequent steps filter classifier output: for instance, a minimum duration criterion, median filtering, exponential updating, and so on. Research on sleep dynamics suggests that human sleep is fairly well represented by a Markov chain model [9]. Since HMMs are built on Markov chains, this may explain their popularity in sleep scoring. However, other unsupervised but "static" classifiers (e.g., Gaussian mixture models or GMMs, *k*-means, *k* nearest neighbors, linkage trees, etc.) that cluster the feature space to score sleep from PSG features have been investigated in the past [10-11]. Whether the assumption of Markov dynamics in HMMs truly translates into better predictive performance compared to other unsupervised static classifiers has not been verified. Here, we test a methodology for naive scoring of human sleep using HMMs. We also compare HMM performance with three unsupervised static classifiers to see if the added computational burden imposed by Markov dynamics is justified by classification performance.

II. Methods

Signal features extracted from 30s epochs of overnight PSGs were modeled using four unsupervised classifiers: an HMM, a GMM, a *k*-means classifier, and a linkage tree. The number of states in each was optimized by an information criterion. Classification accuracy was assessed against expert-scored hypnograms.

A. Data source and feature extraction

This analysis is based on a Physionet database of 22 overnight expert-scored PSGs (6-9 h each; 100 Hz sampling)

This work was supported in part by NIH grant NS065451.

FY and SS are with the Department of Biomedical Engineering, University of Kentucky, Lexington, KY, USA. PM is with the Department of Neurology, University of Texas Southwestern-Austin Program, Austin, TX, USA.

*Address correspondence to: Sridhar Sunderam (Phone: 859-257-5796; Fax: 859-257-1856; e-mail: ssu223@uky.edu).

of healthy subjects (male/female, 18-79 years old, mean ~40) without medications [12-13]. All analysis was performed using Matlab™ (Mathworks, Natick, MA). The hypnograms, which mapped 30s epochs of data onto six states (NREM 1-4, REM, and Wake) were relabeled per the current guidelines of the American Academy of Sleep Medicine [2] by combining NREM stages 3 and 4. Hence, each hypnogram contained up to five labels: *N1*, *N2*, *N3* for NREM, *R* for REM, and *W* for Wake. The Fpz-Cz signal from each subject was bandpass-filtered into seven distinct frequency bands, specifically: δ_L (0.5-2 Hz), δ_H (2-4Hz), $\boldsymbol{\theta}$ (4-9Hz), $\boldsymbol{\alpha}$ (9-12Hz), $\boldsymbol{\sigma}$ (12-16Hz), $\boldsymbol{\beta}$ (16-30Hz) and $\boldsymbol{\gamma}$ (30-45Hz) using 3rd order Butterworth IIR filters. The mean power in these bands was estimated in 30s epochs and combined into "sleep variable" ratios:

$$S_1 = (\boldsymbol{\delta_L} + \boldsymbol{\delta_H}) / \boldsymbol{\theta} \quad (1)$$

$$S_2 = (\boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\sigma} + \boldsymbol{\gamma}) / (\boldsymbol{\delta_L} + \boldsymbol{\delta_H} + \boldsymbol{\theta}) \quad (2)$$

$$S_3 = \boldsymbol{\delta_L} / \boldsymbol{\sigma} \quad (3)$$

Each variable is designed to emphasize contrast between EEG rhythms observed in different states of vigilance: S_1 captures differences between N_3 (strong delta) and *R* (strong theta), S_2 distinguishes N_3 (low frequency) from *W* (broadband activity), and S_3 discriminates N_2 (spindle activity). This three-dimensional vector of features was expressed on a logarithmic scale, which makes the observation distribution approximately Gaussian, and used as the input to the unsupervised classifiers to be evaluated.

B. *Modeling the data using unsupervised classifiers*

The main aims of this analysis are: 1. To perform unsupervised sleep scoring using an HMM; and 2. To compare HMMs, which incorporate dynamics as Markov state transitions, with other unsupervised but static classifiers (GMMs, *k*-means clustering, and linkage trees) that do not have dynamics. In effect, GMMs and HMMs are *parametric* since they are based on a probability model, while *k*-means and linkage trees are *nonparametric* since they are based solely on proximity in the feature space.

Gaussian mixture models. A GMM expresses the distribution of $\boldsymbol{S} = [S_1\ S_2\ S_3]^T$ as a linear mixture of Gaussians: $p(\boldsymbol{S} \mid \Theta) = \sum \alpha_k\, p(\boldsymbol{S} \mid \theta_k)$. Each component *k* corresponds to one of *ns* model states, and θ_k is parameterized by a mean vector and covariance matrix; α_k is a mixing coefficient. Once *ns* is fixed, model parameters are determined from sample PSG data using maximum likelihood estimation. Assuming samples are independent and identically distributed, optimal parameters are those that maximize the function $L(\Theta \mid \boldsymbol{S}_{1:N}) = \prod p(\boldsymbol{S}_i \mid \Theta)$, which expresses the joint likelihood of all samples $i = 1{:}N$. *L* (or more commonly, $\log L$) is optimized via an Expectation-Maximization (E-M) algorithm [14], in which an initial parameter guess is iteratively refined in a way that local convergence is guaranteed. For each subject, we used multiple randomized seeds and selected the solution with largest $\log L$. Then, we labeled each epoch by the GMM component that maximized its probability density.

Hidden Markov models. An HMM is a dynamical model of a sequence or time series [15] that assumes each observation $\boldsymbol{S}_k$ in a sequence to be randomly drawn from a probability distribution conditioned on an underlying nominal state Q_k. $\boldsymbol{S}_k$ is conditionally independent of $\boldsymbol{S}_{k-1}$ given Q_k. The evolution of state Q_k over time follows the Markov property: *i.e.*, given Q_k, the distribution of Q_{k+1} is independent of Q_{k-1}, Q_{k-2}, and so on [16]. Here, we model the observation density $p(\boldsymbol{S} \mid Q)$ as a Gaussian distribution where *Q* is one of *ns* discrete model states that relate to the different states of vigilance. To model a PSG recording using an HMM, its parameters must be fixed: namely, a set of priors π and emission models $p(\boldsymbol{S} \mid Q)$, one for each of the *ns* states; and a matrix of transition probabilities P_{tr} between any two states. Algorithms are available for statistical inference using HMMs [16] that generally involves the recursive application of Bayes rule to compute the probability of a sequence of emissions from an arbitrary sequence of states, and for decoding the most likely sequence of states given an arbitrary sequence of emissions (the Viterbi algorithm). An E-M variant known as the Baum-Welch algorithm is used to estimate HMM parameters for a sample observation sequence $\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_N$ [14]; since the source states are not known a priori, the HMM is an unsupervised model. Since we have chosen a Gaussian emission model for each state, we used the GMMs described in the previous section as the initial guesses of the priors and observation densities of the HMM. Once the model is determined, the Viterbi algorithm is used to decode the sequence of hidden states $Q_0, Q_1, \ldots, Q_N$ most likely to have generated the sequence of emissions. As for GMMs, the likelihood *L* associated with the model can be computed for a sequence of observations.

k-means clustering. This is a well-known unsupervised algorithm, used here to cluster sample vectors of sleep variables into different states. The algorithm starts with *k* randomly selected prototypes or centroids (for *k* states), and then associates each data sample with a centroid based on the Euclidean distance between them in the feature space. The centroids are then recomputed based on the newly determined membership of each state. State labels and centroids are recursively updated until convergence [17].

Hierarchical clustering. A linkage tree is a clustering

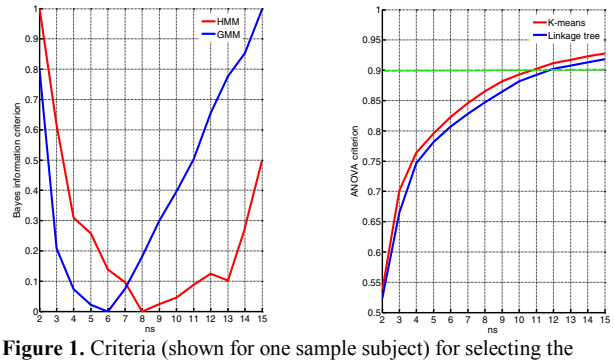

Figure 1. Criteria (shown for one sample subject) for selecting the number of model states *ns* that best fits the data. **Left:** Bayes information criterion (scaled by the dynamic range) passes through a minimum that determines *ns* for GMMs and HMMs. **Right:** Optimal ns for k-means and linkage tree classifiers is chosen as the lowest value for which an F-statistic representing relative variance between states

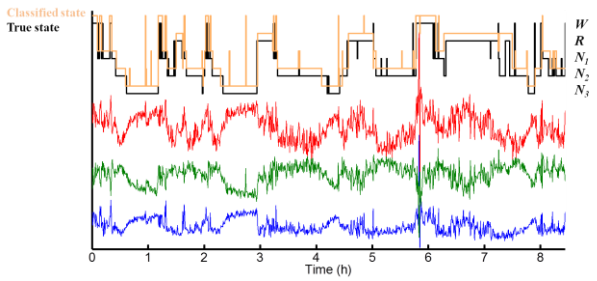

Figure 2. HMM classifier output for a sample overnight sleep recording. Input features S_1, S_2, and S_3 are shown below the model-generated (black) and true (beige) hypnograms for the data (Cohen's

technique that builds a hierarchy of clusters using a "bottom up" approach. It starts with each observation forming its own cluster and then merges clusters based on their proximity to each other to move up the tree [18]. The tree therefore contains successively smaller numbers of clusters (states) at each level until there is only one cluster encompassing all the data at the top. The level at which the tree is "cut" or terminated determines the number of states, and their descendants on the tree inherit their labels.

C. Optimization of the number of classifier states

For an unsupervised classifier, the number of model states *ns* must first be specified. Since the optimal number of states is not known *a priori*, a criterion is needed for the value of *ns* that best predicts the scatter observed in the data. While a large *ns* may give a better fit, the parameter space needs to be kept manageable and overfitting avoided. Also, *ns* should be close—but not necessarily equal—to the actual number *ms* of vigilance states in the sample: some model states may be sub-states of one vigilance state that together determine its distribution in the feature space.

For the parametric classifiers (GMM and HMM), we constructed models with *ns* varying from 2 to 15 Gaussian components. Then the optimal model was chosen by using the Bayes information criterion (BIC) [19], which balances conflicting terms representing the goodness-of-fit of the model and the degrees of freedom respectively:

$$BIC = -2 \log L + k \log n \qquad (4)$$

L is the likelihood of the data given the probability model, n is the number of observations (i.e., epochs of data), and k is the model degrees of freedom based on the total number of fitted parameters in the model. Fig. 1a demonstrates how *BIC* varies with *ns* in a GMM fitted to data from an arbitrary subject (blue graph) whose recording contained all five vigilance states ($ms = 5$). A GMM with $ns = 6$ seems optimal for this subject. For an HMM of the same subject's data a choice of eight model states ($ns = 8$) is deemed optimal. The excess model states turn out to be subcomponents of vigilance states. For the nonparametric classifiers (*k*-means clustering and linkage trees) there is no probabilistic model, so a likelihood measure cannot be defined. Instead, we specify a criterion inspired by the F-statistic typically used in analysis of variance. We selected the optimal *ns* as the smallest value for which the ratio R of the variance between clusters to the total variance crossed 90%. For the sample subject in Fig. 1b, R monotonically increases with *ns* for the *k*-means algorithm and crosses 90% at $ns = 11$. Similarly, $ns = 12$ is optimal for a linkage tree classifier extracted from the same subject's data.

D. Mapping the model states to vigilance states

For each sleep record, dynamic (HMM) and static (GMM, *k*-means and linkage tree) unsupervised classifiers with *ns* optimized by *BIC* or R were constructed. The mapping between model states and vigilance states is not known a priori. In fact, multiple model states may form sub-states of a particular vigilance state; and not all vigilance states may occur in a sleep record (e.g., subject never reaches *N3*, or the recording does not include *W*). Whichever the case, we assume that a sleep physician could quickly inspect a few samples of each model state and fix the true vigilance state, based on which the hypnogram can easily be relabeled. In our analysis, we determine the mapping from model states to vigilance states by computing Cohen's kappa [20], which is a widely used statistical measure of inter-rater agreement. Since kappa takes chance agreement between the nominal states into account, it is a more reliable measure than just the overall proportion of agreement between labels. We applied the mapping that optimized Cohen's kappa for each subject before assessing the performance of each classifier.

E. Assessment of classifier performance

Classifier performance was assessed by comparing model-predicted labels against true hypnogram labels using conventional metrics of detection sensitivity and specificity. The sensitivity (expected true positive rate) of a specific vigilance state reflects the proportion of actual sample epochs of that state correctly identified by the classifier. Conversely, the specificity (expected true negative rate) for a particular state is the proportion of other states *not* wrongly classified as the state of interest. Overall model performance was gauged by kappa while the ability to detect specific states was assessed using sensitivity and specificity.

III. RESULTS

Fig. 3 gives the performance of optimal static and dynamic classifiers on a 22-subject database in terms of Cohen's kappa. The static classifiers appeared to have similar performance with kappa of about 50%, which is considered moderate agreement with expert sleep scores. GMMs and linkage trees performed slightly but not significantly better than *k*-means. HMMs significantly outperformed the static classifiers ($p < 0.05$ by ANOVA), with a median kappa of over 70% (substantial agreement).

Trends in classifier performance in terms of sensitivity and specificity for each vigilance state (Table I) mirrored overall agreement (kappa), with some differences. Linkage trees and k-means gave very similar sensitivity and specificity for all five states. GMMs performed significantly better overall, except for lower sensitivity and higher specificity to *N2*, than the other static classifiers. HMMs gave comparable or significantly higher sensitivity and specificity for all states than any of the static classifiers.

Table I. Performance of unsupervised classifiers by vigilance state.

	Sensitivity (mean ± s.e.m.)				
	N_1	N_2	N_3	R	W
K-means	27.8±3.4	81.3±2.4	61.3±5.6	52.8±4.7	32.9±3.8
Linkage tree	32.2±4.1	82.4±2.5	59.9±6.4	52.1±5.7	35.6±4.8
GMM	39.8±5.5	68.4±3.8	75.9±5.5	63.4±4.6	57.1±7.3
HMM	41.8±4.9	84.1±2.2	68.9±7.4	86.7±1.8	73.3±4.7
	Specificity (mean ± s.e.m.)				
K-means	93.3±1.2	79.1±2.3	93.9±0.9	91.6±1.1	93.6±0.9
Linkage tree	94.4±1	78.8±2.9	93.2±1.6	92.9±1	94.4±1.3
GMM	91.2±1.1	89.5±1.9	91.5±1.7	91.7±1.9	91.9±1.4
HMM	95.3±0.6	89.2±1.2	96.1±0.8	95.6±0.9	96.7±0.5

IV. Discussion

In this work, we compared HMMs with multiple static classifiers for clinical sleep scoring. The presumptive advantage gained by the empirical Markov chain representation of the dynamical sleep state transitions in the HMM has never been verified, but are now clear. Our other goal, to propose and test a means for obtaining reasonable initial sleep scores for an overnight recording without a previously trained model, also appears feasible. In this regard, we proposed a criterion for optimizing the number of states modeled by the classifier from the data without a priori information. This approach improved classification performance compared with similar studies [6-8], which are few in number and presume without justification that all stages of sleep are presented in each recording. Since the purpose of our HMM is to generate a first-pass segmentation, a human expert can quickly match up the model states with conventional vigilance states by reviewing a random sample of each model state. Moreover, our use of three simple power spectral features rather than a wide range of spectral /nonlinear EEG features [3-8] or auxiliary EMG/EOG features [8] results in a simple but more efficient automated sleep scoring technique. Use of the initial band power variables did not improve classifier performance despite the greater dimensionality.

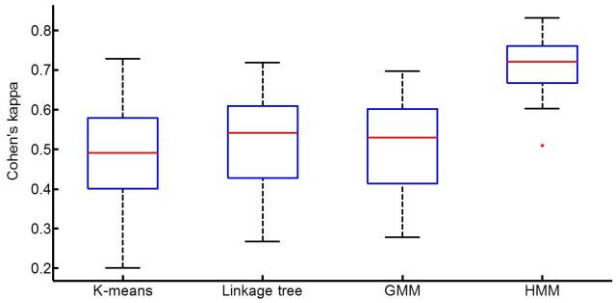

Figure 3. Overall performance of sleep classifiers assessed using Cohen's kappa (n = 22 subjects). HMM performance is significantly better than GMM, linkage tree and k-means classifiers ($p < 0.05$).

References

[1] A. Rechtschaffen , A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects", *CA: BI/BR, Los Angeles*, 1968.

[2] C. Iber, S. Ancoli-Israel, A. Chesson, and S.F. Quan, "The AASM manual for the scoring of sleep and associated events", *American Academy of Sleep Medicine*, 2007.

[3] S. Khalighi , T. Sousa, G. Pires and U. Nunes, "Automatic Sleep Staging: A Computer Assisted Approach for Optimal Combination of Features and Polysomnographic Channels", *Expert Syst Appl*, vol. 40, pp.7046–7059, 2013.

[4] B. Koley and D. Dey, " An ensemble system for automatic sleep stage classification using single channel EEG signal", *Comput Biol Med,* vol. 42, pp. 1186–1195, 2012.

[5] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier", *Comput Meth Prog Bio,*vol. 8, pp. 10-19, 2012.

[6] A. Flexer, G. Grubera and G. Dorffner , "A reliable probabilistic sleep stager based on a single EEG signal," *Artif Intell Med*, vol. 33, 199-207, 2005.

[7] L.G. Doroshenkov, V.A. Konyshev and S.V. Selishchev, "Classification of Human Sleep Stages Based on EEG Processing Using Hidden Markov Models", *Biomedical Engineering,* vol. 41, pp. 25-28, 2006.

[8] S. Pan, C. Kuo, J. Zeng and S. Liang, "A transition-constrained discrete hidden Markov model for automatic sleep staging", *BioMed Eng OnLine*., vol. 11, 2012.

[9] J.W. Kim, J.S. Lee, P. A. Robinson, and D.U. Jeong, "Markov Analysis of Sleep Dynamics", *Phys Rev Lett,* vol .102, pp.178104, 2009.

[10] I. Gath, C. Feuerstein, and A. Geva. "Unsupervised classification and adaptive definition of sleep patterns". *Pattern Recognition Lett*, vol. 15, 977-984, 1994.

[11] H. Escola, E. Poiseau, M. Jobert, and P. Gaillard. "Classification using distance-based segmentation—application to the analysis of EEG signals", *Pattern Recognition Lett*, vol. 12, 327-333, 1991.

[12] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark , J Mietus, G. Moody , C. Peng, H. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals", *Circulation*, vol .101, pp. 215-220, 2000.

[13] B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, J.J.L. Oberyé, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG", *IEEE-BME* vol. 47, 1185-1194, 2000.

[14] JA. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models" *International Computer Science Institute*, 1998.

[15] A. Krough, "An introduction to hidden Markov models for biological sequences", *Computational Methods in Molecular Biology*, 1998.

[16] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", P*roc. IEEE*, vol. 77, 257-286, 1989.

[17] J. B. McQueen, "Some Methods for classification and Analysis of Multivariate Observations", *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press*. pp. 281–297, 1967.

[18] G. J. Székely and M. L. Rizzo, "Hierarchical clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method", *J Classif*, vol. 22, pp. 151-183, 2005.

[19] D. Posada and T.R. Buckley, "Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests", *Syst Biol,* vol. 53, pp. 793-808, 2004.

[20] J. Cohen, "A coefficient of agreement for nominal scales". *Educ Psychol Meas*, vol. 20, pp. 37-46, 1960.