A probabilistic approach to high-resolution sleep analysis

P. Sykacek[1], S. Roberts[1], I. Rezek[1], A. Flexer[2], and G. Dorffner[2]

[1] Robotics Research Group, Dept. Eng. Sci.,
University of Oxford, Parks Road, Oxford OX1 6PJ, UK
psyk@robots.ox.ac.uk,
[2] Austrian Research Institute for Artificial Intelligence (ÖFAI), Schottengasse 3,
A-1010 Vienna, Austria

Abstract. We propose in this paper an entirely probabilistic approach to sleep analysis. The analyser uses features extracted from 6 EEG channels as inputs and predicts the probabilities that the sleeping subject is either awake, in deep sleep or in rapid eye movement (REM) sleep. These probability estimates are provided for different temporal resolutions down to 1 second. The architecture uses a "divide and conquer" strategy, where the decisions of simple experts are fused by what is usually refered to as "naïve Bayes" classification. In order to show that the proposed method provides viable means for sleep analysis, we present some results obtained from recordings of good and bad sleep and the corresponding manual scorings.

1 Introduction

This paper proposes a sleep analyser that maps all night EEG recordings to 3 probability plots. We model the probabilities of being awake, in deep sleep and in REM sleep with temporal resolutions down to one second. Allowing for probabilities and for high temporal resolutions is motivated by prior experiences with a similar method which were reported in [PRTJ97]. The hope is that using such an approach allows us to improve upon classical Rechtschaffen and Kales (R&K) rules [RK68]. Although useful for many purposes, this standard has raised dissatisfaction (e.g. [PSKH91]).

- The rules are based on rather short events in the EEG. In the worst case up to 98% of the EEG signal in a scoring window could be ignored.
- R&K rules use amplitude based criteria and rather large scoring windows, which is valid for healthy young subjects. The rules do however not account for aging effects and they fail to work for some important sleep-disturbances (e.g. sleep apnoea).

The analyser proposed in this paper differs from that in [PRTJ97] by allowing for an entirely probabilistic approach. This has the advantage that all uncertainties (e.g. from noise contamination of the EEG recordings) will lead to less certain decisions about the state of the sleeping subject. The analyser consists of three major building blocks:

- a preprocessing stage,
- a classification stage implemented as a generative model
- and a sensor fusion stage.

The paper also presents some first results obtained from four recordings. Two of the recordings represent "good sleep"[1] (a young male and an elderly female) and two recordings represent "bad sleep" (two middle aged females, one control and a patient with an anxiety disorder). We conclude from this evaluation that the proposed method captures these overall aspects. Contrary to the R&K scoring, which does not find any stage 4 in the recording of the elderly female, the proposed method shows a clear sleep cycles.

2 Methods

The first step in the analyser design was to decide upon an optimal preprocessing technique. We performed an extensive investigation with different feature subset selection (FSS) techniques. Details of this analysis were reported in [SRR+99] and [Syk00b]. The investigation revealed that we should prefer the use of complexity measures and autoregressive (AR) model parameters (without preferring either of these techniques) as opposed to classical FFT based features. We decided for theoretical reasons[2] to use AR models in a lattice filter representation. The second result of this evaluation is that the optimal number of AR coefficients for *classifying* EEG is as small as 3. These lattice filter coefficients are used as inputs in a modular design, which is motivated by several interesting properties.

- We build several simple classifiers (experts) on top of each electrode and hence avoid the curse of dimensionality of fitting one large model.
- The low input dimension allows us to use a fully generative model which is trained using labelled and unlabelled data[3].
- Fusing multiple experts will increase the reliability of the overall system because information from such segments where the experts disagree (e.g. caused by electrode failure) will be downweighted.
- The proposed architecture can be extended very efficiently.

2.1 A Bayesian lattice filter

The lattice filter is a representation of an auto regressive (AR) model. The parameters of the lattice filter are the so called refection coefficients, below denoted as ρ_m. Equation (1) shows an AR model, where $y[t]$ is a sample of a time series at time t, a_m is the m-th AR coefficient and $e[t]$ a sample of an independently

[1] That was judged by human experts.
[2] A probabilistic formulation of complexity measures has so far not been successful.
[3] We use labels only for such segments that were unanimously classified by three human experts. Hence we have plenty of unlabelled data.

identically distributed (i.i.d.) Gaussian innovation sequence with zero mean and precision (inverse variance) β.

$$y[t] = \sum_{m=1}^{M} y[t-m]a_m + e[t] \tag{1}$$

We refer to [Lju99] for details of how to reparameterise Equation (1) in terms of reflection coefficients. Using a lattice filter representation increases the computational efficiency and allows us to use a "stability prior" [4]. The reflection coefficients of stable AR models must lie in the intervall $[-1, 1]$. This stability requirement may be coded by a flat prior $p(\rho_m) = 0.5$. For the precision β we use the prior $p(\beta) = 1/\beta$ as is suggested in [Jef61]. The posterior in Equation (2) is obtained by integrating the AR coefficients and β out of the joint distribution over all coefficients.

$$p(\rho_m|\mathcal{Y}, I_m) \propto p(\mathcal{Y}|I_m)\frac{1}{\sqrt{2\pi}s}\exp\left(-\frac{1}{2s^2}(\rho_m - \hat{\rho}_m)^2\right), \tag{2}$$

Equation (2) denotes the time series used for inference as $\mathcal{Y}$. The model indicator, I_m, represents the hypothesis that an m-th order AR model has generated $\mathcal{Y}$. The multiplicative constant $p(\mathcal{Y}|I_m)$ is the so called *model evidence*. Further aspects of how to calculate and to use the evidence for model selection can be found in [Syk00a].

Equation (3) shows the most probable value of ρ_m and its variance. The equations use N as number of samples in $\mathcal{Y}$. We define $\underline{\epsilon}_m$ as the vector of forward prediction errors of the $(m-1)$-th order AR model in $\mathcal{Y}$ and $\underline{r}_m$ as the vector of the corresponding backward prediction errors[5]

$$\hat{\rho}_m = -\frac{\underline{r}_m^{\mathsf{T}}\underline{\epsilon}_m}{\underline{r}_m^{\mathsf{T}}\underline{r}_m} \text{ and } s^2 = \frac{1-(\hat{\rho}_m)^2}{(N-1)} \tag{3}$$

Inference of reflection coefficients was performed with the following settings:

- Recordings with different sampling rates were resampled to 200 Hz.
- Calculations use a two seconds window that is shifted by one second offsets.
- The data in each window are linearly detrended and normalised to unit standard deviation.
- The time delay operation which is performed in the order recursion is coupled with pre and post windowing (see [Lju99]).

2.2 Classification

As already mentioned, the classifier models the joint density over inputs and labels. We use a fully generative model with K classes, where each of the class

[4] Sleep EEG is necessarily generated by stable AR models.

[5] The backward prediction errors are the errors when precicting one time step into the past. That is with model order $m-1$, we predict $y[t-m]$.

conditional densities are modelled by a mixture model with D Gaussian kernels. Equation (4) abreviates the set of reflection coefficients which acts as input into the classifier as $\underline{x} = \{\rho_1, \rho_2, ..., \rho_m\}$. Hence the classifier is expressed as

$$p(\underline{x}) = \sum_{k=1}^{K} P_k p(\underline{x}|k), \text{ where } p(\underline{x}|k) = \sum_{d=1}^{D} w_{d,k} p(\underline{x}|\underline{\mu}_d, \underline{\Sigma}_d). \tag{4}$$

In Equation (4) P_k are K class priors and $p(\underline{x}|k)$ are K class conditional densities. Hence we may express the posterior probabilities for classes by Bayes' theorem to give $P(k|\underline{x}) = P_k p(\underline{x}|k)/p(\underline{x})$. The D component densities $p(\underline{x}|\underline{\mu}_d, \underline{\Sigma}_d)$ are normal densities with mean, $\underline{\mu}_d$, and diagonal covariance matrix, $\underline{\Sigma}_d$. The $w_{d,k}$ denote the D class conditional kernel priors. We use this model for two reasons.

- Generative models give us the possibility to solve missing data problems. In particular we exploit their ability to cope with missing target labels.
- Generative models allow for probabilistic cluster assessment and hence provide deeper insight into the problem structure.

The key problem in training this classifier is to choose an appropriate number of Gaussian components, which is implicit when learning is performed in a Bayesian setting ([Bis95]). We use here a variational Bayesian framework ([JGJS99]). Recently [Att99] applied variational learning to Gaussian mixture models, which are similar to the classifier we use here.

Bayesian inference requires that we specify a likelihood function. Using t_n to denote known class labels and $\underline{\varphi}$ for the model coefficients, we obtain the joint likelihood of all labelled, $\mathcal{T}$, and unlabelled data, $\mathcal{X}$.

$$p(\mathcal{T}, \mathcal{X}|\underline{\varphi}) = \prod_{n=1}^{N} P(t_n) \sum_{d=1}^{D} w_{d,t_n} p(\underline{x}_n|\underline{\mu}_d, \underline{\Sigma}_d) \prod_{m=1}^{M} \sum_{k=1}^{K} P(k) \sum_{d=1}^{D} w_{d,k} p(\underline{x}_m|\underline{\mu}_d, \underline{\Sigma}_d) \tag{5}$$

The next step in any Bayesian analysis is to specify priors over all model coefficients. We adopt here a setting that was proposed in [RG97]:

- Each component mean, $\mu_{d,i}$, is given a Gaussian prior: $\mu_{d,i} \sim \mathcal{N}_1(\xi_i, \kappa_{ii}^{-1})$,
- The inverse variances are given a Gamma prior: $\sigma_{d,i}^{-2} \sim \Gamma(\alpha, \beta_i)$,
- The hyper-parameter, β_i, is given a Gamma prior: $\beta_i \sim \Gamma(g, h_i)$,
- The class conditional kernel prior, $\underline{W}_k$, is given a Dirichlet prior: $\underline{W}_k \sim \mathcal{D}(\delta_W, ..., \delta_W)$,
- The class prior, $\underline{P}$, is given a Dirichlet prior: $\underline{P} \sim \mathcal{D}(\delta_P, ..., \delta_P)$,

in which i denotes a particular input dimensions. The hyper-parameters have to be set a-priori. Values for α are between 1 and 2, g is usually between 0.2 and 1 and h_i is typically between $1/R_i^2$ and $10/R_i^2$, with R_i denoting the range of the i-th input. The mean, μ_i, is given a Gaussian prior centred at the range midpoint, ξ_i, with inverse variance $\kappa_{ii} = 1/R_i^2$. Both the prior counts δ_P and δ_W are set to 1 to give the corresponding probabilities the most uninformative proper Dirichlet prior. We note however that our inference results did never

depend critically on the setting of the hyper-parameters. In fact we used the same setting successfully for a range of different problems.

Having specified a likelihood and priors, we are ready to derive the variational approximation of the posterior. The key idea is to obtain an approximation of the Bayesian posterior by maximising a lower bound of the logarithmic *model evidence*, $\log(\int_{\underline{\varphi}} p(\underline{\varphi})p(\mathcal{T},\mathcal{X}|\underline{\varphi})d\underline{\varphi})$. We approximate the posterior by a mean field expansion, $p(\underline{\varphi}|\mathcal{T},\mathcal{X}) = Q(\underline{\varphi}) \approx \prod_{\forall \underline{\varphi}_l} Q(\underline{\varphi}_l)$, and use Jensens inequality to obtain $\mathcal{F}(Q(\underline{\varphi})) = \int_{\underline{\varphi}} \log(\frac{p(\underline{\varphi})p(\mathcal{T},\mathcal{X}|\underline{\varphi})}{Q(\underline{\varphi})})Q(\underline{\varphi})d\underline{\varphi}$ as a lower bound to the logarithmic model evidence. For our classifier, the natural factorisation is $Q(\underline{\varphi}) = \prod_d (Q(\underline{\mu}_d) \times \prod_i Q(\sigma_{d,i})) \times \prod_k Q(W_k) \times Q(P)$. This finally leads to the functional shown in Equation (6), where we use n to denote labelled samples, m to denote unlabelled samples and I as the number of inputs.

$$
\begin{aligned}
\mathcal{F}(Q) = \int_{\mu,\sigma,P,W,\beta} & Q(\mu)Q(\sigma)Q(P)Q(W)Q(\beta) \\
& \left(\log\left[\frac{P(\mu)P(\sigma)P(P)P(W)P(\beta)}{Q(\mu)Q(\sigma)Q(P)Q(W)Q(\beta)}\right]\right. \\
& + \sum_n \left(\log(P(t_n)) + \sum_{d_n} \left(Q(d_n) \left[\log(W_{t_n,d_n}) - \frac{I}{2}\log(2\pi) \right.\right.\right. \\
& \left.\left.\left. - 0.5\sum_i \log(\sigma^2_{d_n,i}) - 0.5(\underline{x}_n - \underline{\mu}_{d_n})^T \underline{\Sigma}^{-1}_{d_n}(\underline{x}_n - \underline{\mu}_{d_n}) - \log(Q(d_n)) \right] \right) \right) \\
& + \sum_m \sum_{k_m} Q(k_m) \left[\log(P_{k_m}) + \sum_{d_m} \left(Q(d_m) \left(\log(W_{k_m,d_m}) - \frac{I}{2}\log(2\pi) \right.\right.\right. \\
& \left.\left.\left. - 0.5\sum_i \log(\sigma^2_{d_m,i}) - 0.5(\underline{x}_m - \underline{\mu}_{d_m})^T \underline{\Sigma}^{-1}_{d_m}(\underline{x}_m - \underline{\mu}_{d_m}) \right.\right.\right. \\
& \left.\left.\left. - \log(Q(d_m)) \right) \right) - \log(Q(k_m)) \right] \Bigg) d\mu d\sigma d\beta dP dW
\end{aligned}
\tag{6}
$$

The mean field assumption (i.e. the independence assumption among the Q-functions) allows the maximisation of Equation (6) to be done separately w.r.t. every Q-function. Maximising w.r.t. one Q-function involves an E-step, where we take the expectation of the functional $\mathcal{F}(Q)$ w.r.t. all other Q-functions, and an M-step which involves minimising a Kullback-Leibler (KL) divergence. After each iteration we evaluate $\mathcal{F}(Q)$ and test for convergence. Usually we observe convergence after at most 100 iterations. Although $\mathcal{F}(Q)$ is only a lower bound of the logarithm of the model evidence, [Syk00a] contains several examples where model selection gave the correct result. Similar iterations are done for predicting probabilities of test samples. The main difference is that the Q-functions over model coefficients are fixed. The maximum of $\mathcal{F}(Q)$ is then found in between two and five iterations.

2.3 Sensor fusion

Sensor fusion is used to combine predictions of multiple experts and across time to obtain a desired resolution. We use a sliding window and assume that the observations in the window are class conditionally independent. In this case we obtain the resulting probability by a "naïve Bayes" expression, which is identical to one of the techniques advocated in [PHR99].

$$P(t|x_{w1}, x_{w2}, ..., x_{wn}) = \frac{P(t)^{-(n-1)} \prod_{i=1}^{n} P(t|x_{wi})}{\sum_k P(k)^{-(n-1)} \prod_{i=1}^{n} P(k|x_{wi})} \tag{7}$$

The optimal window size depends on the resolution we aim at. In this paper we use 30 seconds, which is equivalent to the R&K window length that was used by the human experts.

3 Experiments

The experiments have been peformed using recordings of 14 healthy subjects otained at at different sleep centres. As already mentioned we use only such labels of R&K stages wake, REM sleep and stage 4 that were unanimously classified by three human scorers, one of them providing a consensus scoring. These labels were further subsampled to get equal class priors[6]. The inputs were obtained by extracting three reflection coefficients from each of the EEG electrodes (Fp1, C3, O1, Fp2, C4 and O2).

Using these data, we trained one classifier for each electrode. Across electrodes we find that the optimal number of Gaussian components is between 10 and 19. The plots shown in Figures 1 and 2 illustrate the information extracted by the proposed approach. The left plots in Figure 1 were obtained from a recording of a young male. The right plots in Figure 1 were obtained from a female aged 65 years. It is evident that the probabilities obtained with our sleep analyser show less aging effects than do the corresponding R&K scorings, where there is no sign that the elderly person reaches stage 4 at all. The plots in figure 2 show examples of bad sleep - mostly because of rather long wake phases during the night. Also for these examples we see that the structure of sleep is preserved in our plots. We would like to emphasise that we can easily distinguish between REM and non-REM, which is known to be difficult from EEG alone.

4 Conclusion

We proposed in this paper an entirely probabilistic approach to sleep analysis. The proposed classifier allows the use of unlabelled data for inference. Model selection is part of the training procedure.

[6] As inference is based both on labelled and unlabelled data this does *not* imply that the optimal priors are equal.

Applying the classifier to some sleep recordings we can certainly assess that the proposed method preserves the structure of sleep that is also found in manually labeled R&K scorings. We see a clear indication of deep sleep also for elderly subjects and with that respect have overcome a known problem of R&K scoring rules.

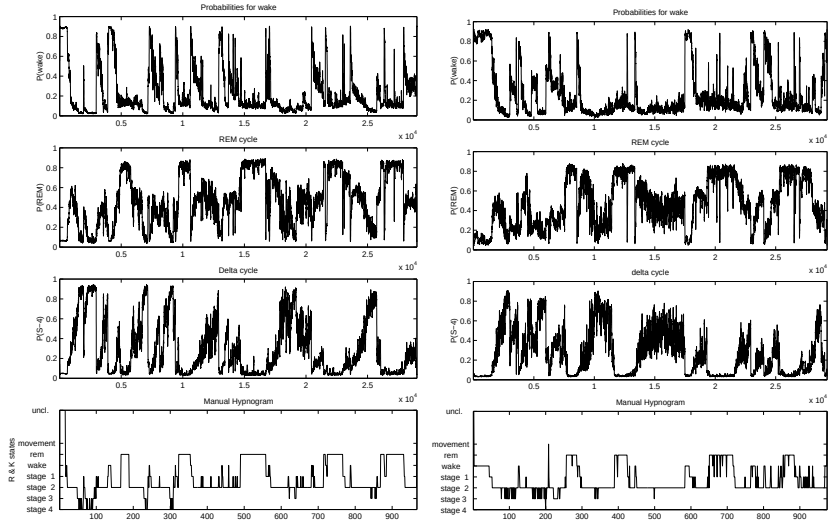

Fig. 1. Probability traces showing good sleep. The probabilities were obtained by combining all EEG channels with a 30 seconds sliding window. An R&K scoring is added to allow for a comparison with the classical scoring technique. Both plots shows a good perfect correspondence between our plots and the R&K labels. The right hand plots were obtained from an elderly subject. They correctly show phases of deep sleep which are missing in the R&K scoring.

Acknowledgements

The authors would like to thank the reviewers of this paper for their valuable suggestions. This work has been done within the project SIESTA, funded by the EC Dg. XII grant BMH4-CT97-2040. Peter Sykacek is currently funded by grant Nr. F46/399 kindly provided by the the BUPA foundation. The authors want to express gratitude to all partners who were involved in this project.

References

[Att99] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conf. on Uncertainty in AI, 1999*, 1999.

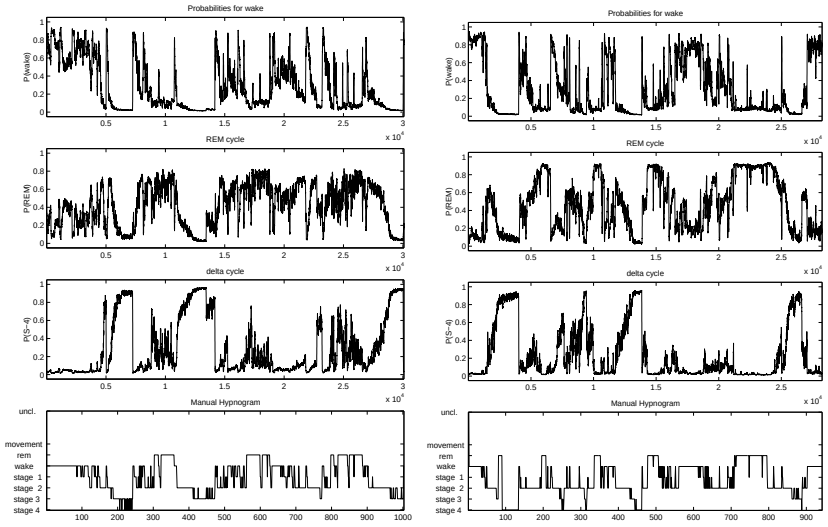

Fig. 2. The traces shown in this figure represent bad sleep. We see that our plots preserve the structure also found in the manual scoring.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[Jef61] H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, third edition, 1961.

[JGJS99] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[Lju99] L. Ljung. *System Identification, Theory for the User*. Prentice-Hall, Englewood Cliffs, New Jersey, 1999.

[PHR99] W. Penny, D. Husmeier, and S. Roberts. Covariance based weighting for optimal combination of model predictions. In *Proceedings of the 8th Conference on Artificial Neural Networks*, pages 826–831, London, 1999. IEE.

[PRTJ97] J. Pardey, S. J. Roberts, L. Tarassenko, and J.Stradling. A new approach to the analysis of the human sleep/wakefulness continuum. *J. of Sleep. Res.*, 5:201–210, 1997.

[PSKH91] T. Penzel, K. Stephan, S. Kubicki, and W.M. Herrmann. Integrated sleep analysis with emphasis on automati methods. *Epilepsy Research Supplement*, 2:177–204, 1991.

[RG97] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal Royal Stat. Soc. B*, 59:731–792, 1997.

[RK68] A. Rechtschaffen and A. Kales. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. NIH Publication No. 204, US Government Printing Office, Washington, DC., 1968.

[SRR+99] P. Sykacek, S. J. Roberts, I. Rezek, A. Flexer, and G. Dorffner. Bayesian wrappers versus conventional filters: Feature subset selection in the Siesta

project. In *Proceedings of the European Medical & Biomedical Engineerinmg Conference*, pages 1652–1653, 1999.

[Syk00a] P. Sykacek. *Bayesian inference for reliable biomedical signal processing*. PhD thesis, Technical University, Vienna, 2000.

[Syk00b] P. Sykacek. On input selection with reversible jump Markov chain Monte Carlo sampling. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 638–644, Boston, MA, 2000. MIT Press.