Magisterarbeit

zum Thema

Continuous analysis of sleep EEG data using Gaussian Mixture Models

ausgeführt am

Institut für Medizinische Kybernetik und Artificial Intelligence
Zentrum für Hirnforschung
der Medizinischen Universität Wien

unter der Anleitung von

A.o. Univ.Prof. Dipl.-Ing. Dr. Georg Dorffner

durch

Bakk.techn. Stefan Neubauer
Matrikelnummer: 9306196
Schwadorferstraße 32, 3100 St. Pölten

Wien, 19. Februar 2006

Abstract

Knowledge gained from studying sleep profiles is an important tool used in clinical practice of sleep medicine. Due to known limitations of and dissatisfaction with the golden standard of sleep scoring – the traditional Rechtschaffen and Kales (R&K) rules (1968) – we describe and validate three alternative models of sleep. The models rely solely on information of electroencephalogram (EEG) data and describe sleep as a continuous process with a high temporal resolution. The first model considers the relatively unambiguous wake, rapid eye movement (REM) sleep, and non-REM sleep states. In the most detailed second model, all stages defined by R&K are considered as continuous processes. Finally, in the third model REM is seen as a discrete process and the substructure of wake and NREM is emphasized. For this model, information from an external classifier is incorporated. The output of all models is validated based on data from both healthy subjects and patients.

From the polysomnographic sleep recordings, data from a central lead EEG channel is used for the analysis. After the application of standard signal processing tools to unify recordings, those are cut into segments of a few seconds. Hierarchical mixtures are used to model the sleep processes by capturing the distribution of the data, where the data is represented by a compact spectral representation of the individual segments using autoregressive modeling. The structure at the top level of the hierarchy is determined by the choice of the processes considered. Gaussian Mixture Models are used in the lower part of the hierarchy. Inference about the model parameters is done in a semi-supervised manner, where information from R&K sleep profiles is used for model selection and parameter initialization. In a second step, unlabeled data is used to allow the models to adapt more freely to the data.

Probability vectors are obtained from the models with high temporal resolution (down to 10 ms) that constitute a continuous description of sleep. Visualization and comparison to R&K sleep profiles showed the usefulness of this representation, where continuous transitions between sleep processes and events of short duration are to be mentioned. To compare the amount of information inherent in the new descriptions, correlations between features derived from the profiles and several external criteria of sleep have been compared. Furthermore, the discriminate power between healthy subjects and several patient groups have been explored. Overall results show that the features derived from the continuous profiles hold about the same amount of information, being superior in some aspects and complementary in others.

Keywords: biomedical signal processing, sleep analysis, sleep quality, human sleep EEG, autoregressive modeling, Gaussian mixture models

Acknowledgement

First, I want to thank Georg Dorffner for his marvelous supervision, and for being a source of inspiration throughout my Master studies. It was great to work within the group of sleep researchers from the Institute of Medical Cybernetics and Artificial Intelligence, the Austrian Research Institute for Artificial Intelligence, and The Siesta Group. Special thanks to Roman Rosipal for the delightful cooperation, to Michael Woertz for helpful comments and feedback, to Silvia Parapatics, Peter Anderer, and Georg Gruber for bringing in their invaluable expertise and to Arthur Flexer for contribution of source code regarding R&K feature extraction.

I am grateful to my former employer Erich Hackhofer for enabling me to smoothly combine work and studies, resulting in mutual benefit. Of great practical usage throughout this work was the well-featured laptop generously provided by the UNIQA Group Austria. Special thanks to Roman Rotter for organizing additional mass storage to hold the sleep data.

More generally, I want to thank Qian Du and Manfred Anzinger for their companionship throughout the studies, my family for their everlasting love and most importantly my wife Michaela for being an invaluable supporter.

This research was supported by The Siesta Group GmbH in the framework of a research project funded by the Austrian Research Promotion Agency (FFG).

Contents

1 Introduction

Sleep is essential for good health, as well as mental and emotional functioning. In the 1950s, rapid eye movement (REM) sleep was discovered, giving evidence that the brain is highly active during sleep. Furthermore, it was observed that REM periods alternate with non-REM (NREM) sleep, and even in deep sleep, where responsiveness is greatly reduced, roughly 80 % of the brain is activated. Many studies have been performed, mainly based on electroencephalography (EEG) signals, where electrical activity of the brain is measured using electrodes placed on the scalp. Progress has been made in understanding the functions of human sleep, but many issues are still to be solved (Siegel 2005).

The golden standard for assessment of sleep is the discrete sleep staging as standardized in the Rechtschaffen & Kales (R&K) manual (Rechtschaffen and Kales 1968), where epochs of typically 30 seconds are classified as wakefulness, NREM stages 1, 2, 3, 4, or REM sleep. The scoring method is designed for visual rating of paper recordings. The typical chart paper speed of 10 mm/sec results in a single page of chart paper (300 mm wide) displaying 30 seconds of data. Sleep staging for epochs of 20 seconds is sometimes encountered, and also epochs of one-minute have been used. Taking shorter epochs make the procedure more tedious, whereas with longer epochs stage changes of short duration may be overlooked.

While R&K scoring is still useful, major drawbacks are the crude division of the sleep process into a few discrete stages, the low temporal resolution, the low inter-rater reliability (Danker-Hopfe et al. 2004), and the limited validity for patients and elderly. A review of the limitations is given by (Himanen and Hasan 2000). Attempts were made to improve these rules, e.g. by (Hori et al. 2001), and recently the American Academy of Sleep Medicine has appointed task forces to review and update the manual (Iber 2004).

A lot of effort was made to automate the visual scoring process to save costs and time, and reliable and validated methods are now available (Anderer et al. 2005). In fact scorings from this automated classifier called *Somnolyzer 24 × 7* are used in this work.

Uncoupled from the R&K manual, methods can be chosen that are not restricted to waveforms that are also identifiable visually. Most of the work concerning alternative descriptions of sleep was done starting in the 1980s, see (Hasan 1996) for a review. One proposal for the reevaluation of sleep analysis resulted in a model with a discrete REM and a continuous wake/NREM process with a time resolution of 1 sec (Kemp 1993). The importance of the microstructure of slow wave and REM sleep is emphasized in (Kubicki and Herrmann 1996). The previous EU-funded SIESTA project resulted, amongst others, in a continuous model based on the cornerstones wake, NREM, and deep sleep, again with a temporal resolution of 1 sec (Sykacek et al. 2001).

In this thesis, we describe and validate continuous sleep models that are based on the following main principles:

- The model should have a high temporal resolution.
- The model should be capable of describing the sleep process as a continuum as opposed to discrete stages.
- The model should be objective and rely solely on information in the data.
- The model should be applicable automatically, without user interaction and parameter tuning.

The challenge is taken by construction of probabilistic models based on Gaussian Mixture Models (GMMs). The continuous probability outputs of the model are the new description of the sleep/wakefulness continuum. The models are most closely related to the model described in (Flexer et al. 2005). Compared to their work, time information is not implicitly modeled, which in turn allows for even higher temporal resolution of down to 10 ms.

The model is formulated on an abstract level in form of a model framework, where several processes are considered, and each process is based on a conglomeration of R&K sleep stages. From the model framework we derive three models:

- A model of Sleep Cornerstones with the continuous processes wake, NREM, REM.
- A continuous model of R&K sleep stages, where each of the R&K sleep stages is considered a continuous process.
- A model of sleep substructure adapting the model framework to consider REM as a discrete process and to model the substructure of wake/NREM by the continuous processes wake, s1, s2 and deep sleep.

The validation and comparison of the models will be done threefold:

- Comparison to R&K scoring, both visually and by comparison of classification results.
- Correlation analysis with age and external information about sleep quality.
- Analysis of discriminative power between healthy subjects and patients.

For the correlation and discrimination analysis, features similar to those derived from R&K sleep profiles are computed. The analyses are also carried out for the R&K-based features, and results are compared.

In the next chapter, we introduce the database and outline the most important concepts used in this thesis. In chapter 3, the model framework including the preprocessing steps is defined and the three continuous models are derived. The features computed from the sleep profiles are described in chapter 4, and, most importantly, chapter 5 gives the results and discussion of the results.

2 Methods and materials

In this chapter, information about the data used in this work is followed by a short review of the most important concepts and techniques used throughout this work, which will allow for a more compact description in the main part of this thesis.
The order of steps done in the analysis is kept in the structure of the subchapters, where signal filtering is followed by segmentation of the signal and autoregressive modeling. As the basis for the probabilistic model, Gaussian Mixture Models are reviewed and a description, along with practical considerations for parameter estimation using the Expectation Maximization algorithm, is given.

2.1 Database

The sleep data used in this thesis was taken from the SIESTA database (Klosch et al. 2001). As part of the SIESTA recording protocol, the participants had to spend two consecutive nights in the sleep laboratory and the full night polysomnographic recordings include at least

- six electroencephalography (EEG) channels with mastoid as reference point: F_{p1}–M_1, C_3–M_2, O_1–M_2, F_{p2}–M_1, C_4–M_1, O_2–M_1 (see Figure 1),
- two electrooculogram (EOG) channels,
- one submental electromyography (EMG) channel,
- one EMG channel recorded from linked electrodes left and right anterior tibialis, and
- various physiological parameters, e.g. electrocardiogram and O_2 saturation.

In addition, several psychometric tests assessing vigilance, sleepiness, and performance were carried out. We give a description in section 2.2 of the measures used in the validation part of the thesis.

Note that mastoids (M_1, M_2, not drawn in Figure 1), i.e. the hard, bony structure behind the ear, serve as reference points. For whole night sleep recordings those are preferred to ear (A_1, A_2) placements as those can cause unpleasant feelings. In the record descriptions, ears/mastoid reference points are often used interchangeably with A_1/A_2 being more commonly encountered.

For this work, the central EEG lead signals C_3–M_2 and psychometric test results from the SIESTA database (Klosch et al. 2001) are used. For training as well as for validation purposes, R&K sleep labels as delivered by the automatic, validated sleep classification system Somnolyzer 24 × 7 (Anderer et al. 2005) are utilized. The labels are given for epochs of 30 seconds, which is the typical length used for R&K sleep scoring. Code 6 "undefined" is given to periods that are not assignable to any sleep stage, e.g. because of artifacts or before "lights off" in the sleep laboratory.

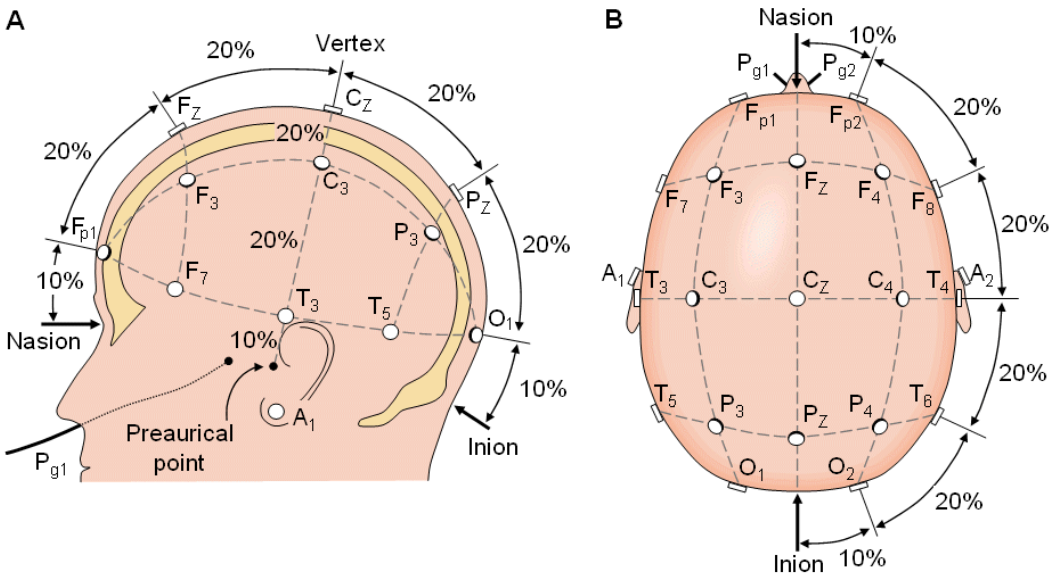

Figure 1: The standardized 10/20 electrode system, seen from left (A) and above (B). Figure from the web version of the book (Malvimvuo and Plonsey 1995), chapter 13.3.

Table 1: R&K codes for sleep staging

R&K Code	Label
0	wake
1	stage 1
2	stage 2
3	stage 3
4	stage 4
5	REM
6	movement
9	undefined

From the SIESTA database, recordings from five healthy controls have been excluded because of missing data, and a further three due to pathological sleep profiles and bad subjective sleep quality. Patient groups consisting of less than 10 subjects were excluded. Table 2 lists the number of controls and patients grouped by sleep disorder as used in this work, with the corresponding ICD-10 (International Statistical Classification of Diseases and Related Health Problems, 10th revised version) (WHO 1992) codes.

Table 2: Number of subjects in the SIESTA database, grouped by sleep disorder.

Group	ICD-10 code	Number of subjects
Healthy Control	-	189
Generalized Anxiety Disorder	F51.0 and F41.1	16
Sleep Apnea Syndrome	G47.3	51
Parkinson's Disease	G20.0	15

The recordings are split into a training and a validation set. The assignment is made for subjects rather than for recordings, i.e. both nights from one subject belong either to the training or the validation set. That way, it is possible to examine differences between two consecutive nights for a subject. As this work

is mainly focused on the sleep of healthy controls, all patients are part of the validation set. For the healthy controls, assignment to one of the sets was made randomly.
The training set was used to tune parameters of the procedures and models, while the validation set was kept for final verification of the results. When appropriate, the training set itself was split into training and testing records.

2.2 Measures of sleep quality

The recording protocol for the SIESTA database (Klosch et al. 2001) comprised a variety of subjective and objective psychophysiological parameters to assess vigilance, sleepiness, and performance. We will use the subjective and objective measures for correlation analysis in section 5.5. A short description of the used measures follows.

The Pittsburgh sleep quality index (PSQI) is a self-rated questionnaire that measures sleep quality and disturbance retrospectively over a 1-month period (Buysse et al. 1989).

The self-rating scale for sleep and awakening quality (SSA) consists of 20 items comprising three subjective categories: quality of sleep, awakening quality, and somatic complaints. Responses are given on an ordinal scale with four possibilities, where lower score means better subjective quality (Saletu et al. 1987). The subscores for the three categories are commonly denoted as SSA1 (quality of sleep), SSA2 (awakening quality), and SSA3 (somatic complaints).

Mood, drive, affectivity, and drowsiness are captured on 100 mm visual analogue scales (ASES), (Folstein and Luria 1973). The subjective well-being in the morning and in the evening is measured by the self-assessment Zerssen Bf-scale (Osgood et al. 1975), where wakefulness, concentration, and extraversion are particularly considered.

Attention and concentration is assessed by the Alphabetical cross-out test (Grunberger 1977). The subject has to mark certain letters in 20 lines consisting of 40 letters each, with a time limit of 10 seconds per line. Several scores are considered: the total score as a measure of attention, the number of errors (absolute and in percentage) for concentration, and the difference between extreme scores for attention variability.

The digit span test (Wechsler 1955) measures numerical memory, by asking the subject to repeat digits in the same and, the second time, in reverse order.

An objective measure of motor activity and drive is given by the test of Fine motor activity (Grunberger 1977). In the test, the subject has to set as many dots as possible in certain squares given on a form within 15 seconds. The performance is measured for both hands individually. Considering also the sum of scores for both hands, we have three parameters originating from the Fine motor activity test.

An overview of the parameters is given in Appendix A2.

2.3 Filtering

We apply a two-step preprocessing procedure to the EEG signal. In the first step, the signal is downsampled to a unified frequency. In the second step, bandpass filtering is done to remove low- and high-frequency components. In the following two subsections, we describe the methods used and refer to alternatives. In the results, section 5.1, we will validate that the methods work as designed.

2.3.1 Downsampling

Resampling to a different frequency is a task commonly encountered in signal processing, and one can choose from various concepts, e.g. nearest neighbor, cubic resampling, spline interpolation, and filter resampling. The filtering method is known to produce good results and is therefore the method of choice. Herewith, the downsampling is done in a general manner by a cascade of three operations, appropriate for performing sampling rate conversion of data by rational ratios p/q (Crochiere 1979):

1. Upsampling the data by a factor of the integer p by inserting zeros.
2. Filtering the upsampled signal by a finite impulse response (FIR) filter.
3. Downsampling the result by keeping every q^{th} sample.

Common EEG recording frequencies encountered are 200 Hz and 256 Hz. For a 200 Hz recording with 100 Hz as target frequency, p=1 and q=2; therefore step 1 is skipped. Note that the lowpass filtering done in step 2 is still essential to prevent aliasing artifacts. Starting with 256 Hz and again taking 100 Hz as target frequency, p=25 and q=64, and all three steps take effect. The design of the FIR filter for step 2 has to result in a lowpass filter, taking into account the upsampling done in step 1 by appropriate choice of filter length. When using the function ‘resample’ from the Matlab Signal Processing Toolbox with the default settings, the filter length is $20\,q + 1$, and the FIR filter used in step 2 is designed as follows:

1. Creation of a lowpass-filter by least-squares error minimization method. For cutoff frequency fc, the desired amplitude of the frequency response is 1 for frequency up to fc, and 0 above. For a given filter length, the integrated squared error is minimized.
2. Multiplication of the lowpass filter coefficients with a Kaiser window, where the sidelobe attenuation of the Fourier transform of the window is controlled by the parameter β, which is set to 5 per default.

Figure 2 illustrates the effect of the windowing procedure using the Kaiser window. Without application of the window, ringing occurs, especially near the cutoff frequency of 50 Hz. This phenomenon is known as “Gibbs effect” and does not vanish with increased filter length. As multiplication in the time domain corresponds to convolution in the frequency domain, the application of the Kaiser window results in a smoother frequency response. While the ringing is greatly reduced, this improvement is at the expense of selectivity, i.e. broadening of the transition band.

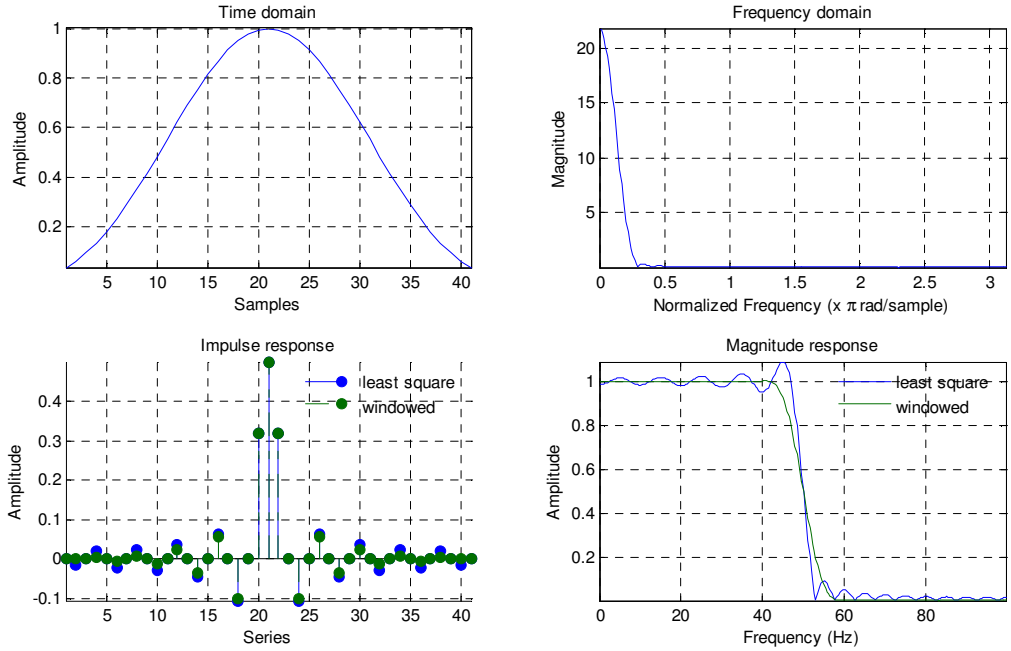

Figure 2: *Top*: Coefficients and frequency magnitude response for Kaiser window of length 41, β=5. *Bottom*: Impulse response and frequency magnitude response for lowpass filter (50 Hz) of step 2, with filter length 41.

For the filter design, available methods are plentiful, e.g. the Butterworth design method or the Chebyshev design method could be used instead of the least squares approach. For the windowing procedure, Blackman and Hamming windows are often used instead of a Kaiser window. As the differences in the outcome are not expected to change results in the following modeling steps, different settings are not compared in this work.

2.3.2 Bandpass filtering

Butterworth filters are infinite impulse response (IIR) filters characterized by a magnitude response that is maximally flat in the passband and monotonic overall (Parks and Burrus 1987). To avoid phase distortion and time shifts, non-causal forward and reverse filtering is done, thus preserving the wave shape for visual inspection of the signal in time domain. The steepened sidelobes in the frequency magnitude response result in slightly shifted cutoff frequencies:

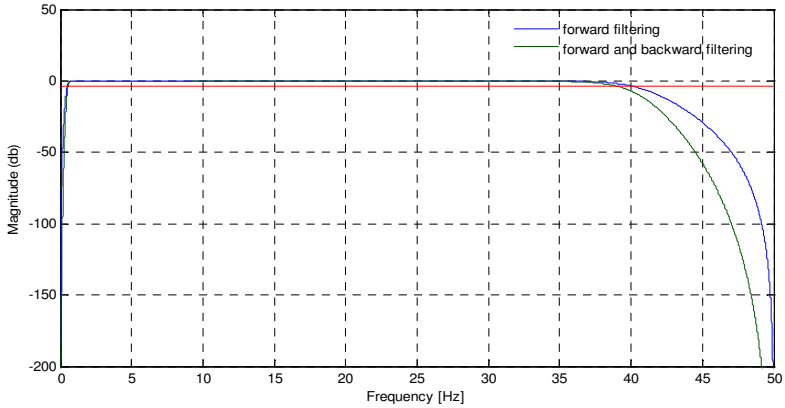

Figure 3: Frequency magnitude response for butter filter of order 8, cutoff frequencies 0.5 and 40 Hz. For the designed filter, the dampening at the cutoff frequencies is 1/√2 (red line).

2.4 Autoregressive modeling

The AR model is a time series model often used in practice which can be characterized as a parametrical spectral analysis method. In contrast to the periodogram with high variance, the AR model has the property of smoothing the spectrum while being able to capture sharp peaks. There are some concerns about misleading results, see for example (Thomson 1990), but comparison with other methods, including non-linear modeling, has shown good performance of the AR model for EEG data (Blinowska and Malinowski 1991). It has been applied to EEG data in at least 160 journal articles (Kaipio and Karjalainen 1997), usually verified experimentally.

We will apply the AR model to short segments of EEG data. In the following section, we give a model description where some details about the spectral interpretation are emphasized, discuss the segmentation and the parameter estimation.

2.4.1 Model description

For an AR model of order P, AR(P), the value of the current signal is predicted by a linear combination of the P previous values:

$$x_t = \sum_{p=1}^{P} a_p x_{t-p} + \varepsilon \tag{1}$$

where ε is zero mean Gaussian noise with constant variance, $\varepsilon \sim N(0,\sigma^2)$, and a_1 ... a_P are called AR coefficients. The values of the time series are denoted by x_1 ... x_T. The forward prediction is therewith given by

$$\tilde{x}_t = \sum_{p=1}^{P} a_p x_{t-p} \tag{2}$$

and the forward prediction error e_t is

$$e_t = x_t - \tilde{x}_t = x_t - \sum_{p=1}^{P} a_p x_{t-p} \tag{3}$$

When the assumptions of the model hold, the predicted values as given by formula (2) can be interpreted as the true signal which is disturbed by white noise. By rearranging equation (3), the AR model can be expressed as an infinite impulse response (IIR) filter where the error value is taken as input, thus leading to a description in the frequency domain:

$$x_t = \sum_{p=1}^{P} a_p x_{t-p} + e_t \tag{4}$$

The sum on the right hand side in equation (3) corresponds to a convolution of the signal [x_1, x_2, ..., x_T] with $[1, a_1, a_2, ..., a_P]$, emerging as multiplication when taking the z-transform:

$$E(z) = A(z)X(z) \quad \text{with} \quad A(z) = 1 + \sum_{p=1}^{P} a_p z^{-p}$$

$$\Leftrightarrow \quad X(z) = E(z)A(z)^{-1} \quad \text{with} \quad A(z)^{-1} = \frac{1}{1 + \sum_{p=1}^{P} a_p z^{-p}} \tag{5}$$

The frequency response can be obtained by evaluation of $H(\omega) = A(e^{-j\omega})^{-1}$. With the assumption of a white noise error process e_t, the spectrum $E(e^{-j\omega})$ will be flat as the frequency components are equally distributed, thus the spectrum of the signal is characterized by $H(\omega)$ alone. Without constraints on the poles, they can lie on the unit z-circle enabling infinite peaks in the frequency domain. Sharp spectral features can be modeled by poles close to the unit z-circle, which is a valuable characteristic of the AR representation. The power spectrum of a signal as estimated by an AR(P) model is given by

$$P_{\mathrm{AR}(P)}(\omega) = \frac{\tilde{\sigma}^2}{\left| 1 + \sum_{p=1}^{P} a_p e^{-jp\omega} \right|} \tag{6}$$

The quality of the model can be evaluated by inspection of the residuals, i.e. the forward prediction error e_t. Those should follow a normal distribution with variance σ^2. Taking the estimate $\tilde{\sigma}^2$, the expected and sample distributions can be compared, for example by examination of a quantile-quantile (QQ) plot. The estimate $\tilde{\sigma}^2$ can not be used as an indicator for the fit of the model, as it is dependent on the amplitude of the signal. This can also be seen as an advantage of the AR model, as e.g. for elderly the amplitude of the EEG signal in slow wave sleep is decreased. Given that the frequency structure is similar to those of adults, the AR model will result in the same AR coefficients for both adults and elderly.

2.4.2 Segmentation

We are not interested in modeling a whole night recording at a glance, rather we want to look at short time intervals, so they can be considered stationary, i.e. the frequency spectrum does not change within this time frame. Choosing too short a time interval, on the other hand, makes it impossible to grasp low frequencies, and estimates are based on a small number of samples. Stationarity of sleep state EEG data is to a large extent fulfilled for segments up to five seconds (Sugimoto et al. 1978). The typical time scale for sleep spindles and K-complexes is one second, and in (Kemp 1993) a time resolution of not more than one second is recommended. Explicit identification of such events is not a requirement within this work, and thus epochs of three seconds were used here.

When splitting up a recording, non-overlapping or overlapping windows can be used. Overlapping windows with small time shifts are especially useful when high time resolution is advantageous. When considering non-overlapping windows, a whole night recording of eight hours is divided into 9600 segments. For a sampling frequency of 100 Hz and the segment length of three seconds, the number of signal values of one segment equals 300, represented by P AR coefficients for an AR(P) model. This can be viewed as a compression of the data by the factor 300 / P. When using overlapping windows with the

smallest time shifts possible, i.e. 10 ms, one gets *P* times as much data. That should illustrate that the choice of the segmentation strategy has a big influence on the amount of data obtained.

2.4.3 Parameter estimation

As the order *P* of the AR model is not known a priori, it can be seen as an additional parameter of the model to be estimated, commonly called model selection in the literature. Given one segment of the signal, various methods exist to select the model order, with the Akaike criterion (Akaike 1973) being widely accepted. An additional constraint in our application is to have to same AR model order for all EEG segments, e.g. independent of sleep stage and subject. The demand on the AR representation within this work is not perfect reconstruction of the spectrum of the signal, but rather the ability to describe different states of brain activity. Setting the model order too low will result in inappropriate representation of the signal, and differentiation would be biased. Higher model order than necessary, on the other hand, would lead us into the curse of dimensionality.
Within this work, the order *P* of the AR model was set at ten, in agreement with the work from (Flexer et al. 2005) and consistent with results from (Vaz et al. 1987). To check the adequacy, the AR spectrum has been compared with the periodogram for single segments, and the mean AR spectrum has been compared between different sleep stages; see the results in section 5.1.3.

Given the model order, parameter estimation is relatively simple because the problem to be solved can be formulated as set of linear equations. The Burg algorithm (Burg 1975) is the preferred estimator for the AR parameters (Broersen 1997). The algorithm minimizes both forward and backward prediction errors, and no assumptions about the signal outside the considered time interval are made. Stability of the AR model is guaranteed by constraining the poles to lie within the unit circle.

2.5 Gaussian mixture model

Gaussian mixture models (GMMs) are widely used and are receiving increasing interest as a natural way of modeling population heterogeneity. An up-to-date monograph is provided by (McLachlan and Peel 2000) and a recent review is given by (Bohning and Seidel 2003). We will use GMMs to model the distribution of AR parameters (AR coefficients), i.e. 10-dimensional vectors where one vector corresponds to an EEG segment of three segments.

With a GMM, a distribution of data vectors $x \in \Re^d$ is approximated by the weighted sum of |*K*| Gaussian components, indexed by $k \in K$:

$$p(x) = \sum_{k \in K} \alpha_k \; p(x \mid \theta_k) \;\text{ with }\; \sum_{k \in K} \alpha_k = 1, \quad \alpha_k \geq 0 \tag{7}$$

The parameters of the Gaussian densities θ_k consist of a mean vector μ_k and a covariance matrix Σ_k, and the density for a given component *k* has the form

$$p(x \mid \theta_k) = (2\pi)^{-d/2} \left|\Sigma_k\right|^{-1/2} e^{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)/2} \tag{8}$$

In the illustrative example shown in Figure 4, samples were drawn from a three-component GMM in two-dimensional space.

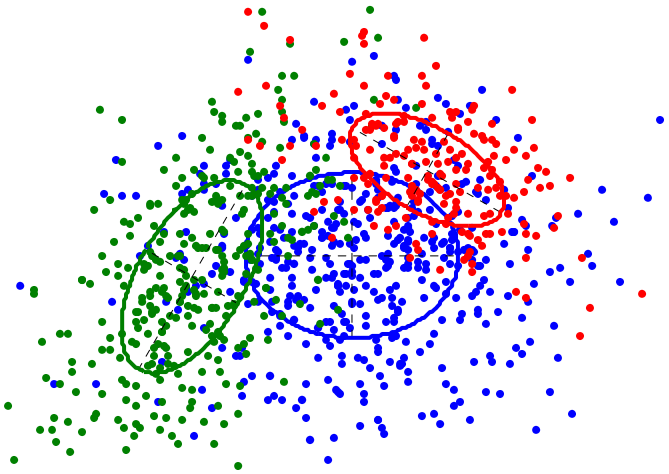

Figure 4: GMM with three components in 2D space, with 1000 samples drawn from the distribution. For each data point, the generating cluster is indicated by color.

By eigenvalue decomposition of the covariance Matrix $\Sigma_k = \lambda_k D_k A_k D_k^T$ (Benfield and Raftery 1993), geometrical properties of the Gaussian densities can be restricted or coupled between components. The parameter λ_k determines the volume, D_k the orientation, and A_k the shape of the k^{th} cluster.

Within the MIXMOD software (Biernacki et al. 2005), three model families with spherical shape, diagonal, or full covariance matrices are considered. An optional equality constraint on the mixing coefficients α_k leads to 28 models.

Estimation is often done by maximum likelihood, which is non-trivial and has some inherent problems. When treating the number of components C as unknown, the likelihood

$$L(\Theta \mid X) = \sum_{n=1}^{N} p(x_n \mid C, \theta, \alpha) \tag{9}$$

is unbounded, which is also true for fixed C with component specific variances.

The Bayesian information criterion (Schwarz 1978) (BIC) is a crude large sample approximation of the logarithm of the Bayes factor (Kass and Raftery 1995), and can be used to choose both the number of components C and the parameterization of the covariance matrix Σ_k. For log-likelihood L, the number p of model parameters and N training samples, BIC is given by

$$\text{BIC} = -2\log(L) + p\log(N) \tag{10}$$

Model selection using BIC is consistent (Keribin 2000) and computationally efficient. Alternatives for choosing the number of components C for a mixture model are Bayesian modeling, e.g. to estimate the posterior distribution of C (Stephens 2000), using cross validation as proposed in (Miloslavsky and van der Laan 2003) or to add one component after another combined with statistical testing (Verbeek et al. 2003).

2.6 Expectation maximization algorithm

A major breakthrough for using GMMs occurred with the appearance of the expectation maximization (EM) algorithm of Dempster et al. (1977), and the associated idea of explicitly representing the mixture components generating each observation by means of latent allocation variables. By using this approach, analytically intractable optimization problems can be simplified and carried out more efficiently. The EM algorithm is designed to locate a maximum likelihood estimate for a broad range of problems. The parameter estimates are updated iteratively, with monotonically increasing likelihood and guarantee of convergence to a local maximum.

For observed data X, it is assumed that a complete data set $Y = (X, Z)$ exists that can be expressed as a joint density function: $p(x, z \mid \Theta) = p(z \mid x, \Theta)\, p(x \mid \Theta)$

The original likelihood as defined in equation (9) is referred to as incomplete-data likelihood. In the E-step of the EM algorithm, the expected value of the complete-data log-likelihood $L(X, Z \mid \Theta)$ is computed with respect to the given data X and the current parameter estimates Θ^{old}. It is assumed that the unknown data Z is a random variable that follows a distribution depending on X and Θ^{old}. In the M-step, the new parameter estimates Θ^{new} are maximized with respect to the old estimates Θ^{old} and the given data X. For convenience, this is commonly noted as

$$\Theta^{new} = \arg\max_{\Theta} Q\left(\Theta, \Theta^{old}\right)$$

The E- and M-steps are repeated as necessary, e.g. until convergence, for a specified number of iterations, or until another criteria is fulfilled. The group of algorithms where the function Q is not maximized but parameters are found to increase, Q is called Generalized EM (GEM). A review of the EM algorithm is e.g. given in (Bishop 1996).

EM for GMM

In the special case of finite mixtures of Gaussians (see section 2.5) without constraints on the mixing proportions and variance matrices, the update procedure is as follows:

$$p^{old}(k \mid x_n) = \frac{\alpha_k{}^{old}\, p^{old}(x_n \mid k)}{\sum_{k' \in K} \alpha_{k'}{}^{old}\, p^{old}(k' \mid x_n)} \tag{11}$$

$$\mu_k{}^{new} = \frac{\sum_{n=1}^{N} p^{old}(k \mid x_n)\, x_n}{\sum_{n=1}^{N} p^{old}(k \mid x_n)} \tag{12}$$

$$\Sigma_k{}^{new} = \frac{\sum_{n=1}^{N} p^{old}\left(k \mid x_n\right)\left(x_n - \mu_k{}^{new}\right)\left(x_n - \mu_k{}^{new}\right)^T}{\sum_{n=1}^{N} p^{old}(k \mid x_n)} \tag{13}$$

$$\alpha_k{}^{new} = \frac{1}{N} \sum_{n=1}^{N} p^{old}(k \mid x_n) \tag{14}$$

There are well-known problems using the EM algorithm. Those encountered during this project were tackled by heuristics, as described below.

Local minimum

The EM algorithm exhibits monotonic convergence and is highly sensitive to the initialization. A review of initialization methods and an empirical comparison is given by (Karlis and Xekalaki 2003). The suggested strategy is to use different initialization values, make a small number of training iterations, and further optimize the best mixture until convergence.

In this work, several models are initialized by randomly choosing training samples as component centers. Furthermore, for some of them a few iterations of k-means clustering (MacQueen 1967) are applied. More concretely, *n* iterations of k-means clustering are done, with *n* drawn uniformly at random from the integers 0 to 3. The model with highest likelihood is retained and further optimized.

Variance collapse

When component *k* is centered at a data point of the training set and the variance is very small, $p(x \mid \theta_k)$ approaches infinity. This problem arises from the fact of unbounded likelihood, as mentioned above. To circumvent this problem, the variances or the mixing proportions have to be constrained. Thus, the components are prevented from collapsing to a small number of data points. The strategies commonly encountered are, to

- add a constant diagonal matrix to the variance matrices after every M step,
- reset the covariance matrix when any of its singular values are too small,
- constrain the mixing proportions, or
- establish ties between the variance matrices.

Within the used toolbox Netlab (Nabney 2002), the resetting strategy is implemented. As an alternative, it would be easy to place an equality constraint on the mixing proportions, resulting in the model labeled p-Lk-Ck in the MIXMOD software.

Outliers

For data vectors with $p(x \mid g)$ approaching zero and numerically evaluating to zero, the log-likelihood of $-\infty$ causes the EM algorithm to fail. To exclude outliers, an iterative procedure was used that was compounded with the initialization strategy:

1) Train several models using different initializations, as outlined above. For performance optimization, just a single starting EM step is done. After that step, for each data vector x_i with posterior close to zero, a counter $errors_i$ is incremented.
2) If the log-likelihood of all models from the first step is $-\infty$, exclude the data points x_j responsible for the breakdown most of the time, i.e. with $errors_j = \max(errors_{1..N})$, and continue with step 1 or stop with an error message if step 1 has already been executed ten times.

3 Models for continuous sleep analysis

In this section, the preprocessing and feature extraction is described, followed by the main part introducing the model framework. Finally, three instances of the model framework are defined: a model of sleep cornerstones, a continuous model of R&K sleep stages, and a model of sleep substructure. The latter stands out in the way that information from an external classifier is incorporated and REM is modeled as a discrete process.

The output of the models is a set of continuous curves representing the probability of each sleep state at a given point in time, where one sleep state is defined as a conglomeration of R&K sleep stages. Single channel EEG data from whole night recordings and R&K labels are used as input. For the training regime, the data is split into two parts, where for the first part the R&K labels are used for supervised training, while the other part is used for an unsupervised *reshuffling* step.

Application of a model to previously unseen recordings yields probability vectors over time. For the model of sleep cornerstones and the continuous model of R&K sleep stages, no more information than the features extracted from the EEG data and the model itself are necessary (see Figure 5). For the model of sleep substructure, the data flow is slightly adapted to include external information. This is described later in section 3.5.

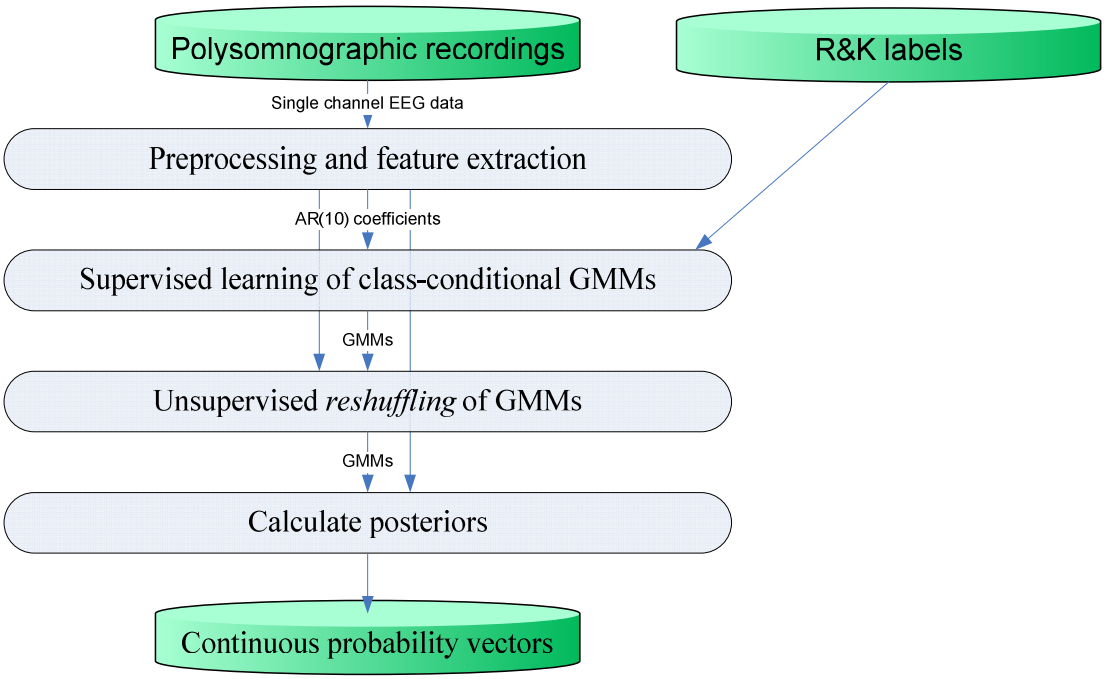

Figure 5: Data flow for model building and usage

3.1 Feature extraction

From preprocessed signals from single channel EEG data, features (autoregressive coefficients) from partitioned recordings are extracted. These features are the basis for the model building as well as the application of the model. See Figure 6 for the sequence of steps involved.

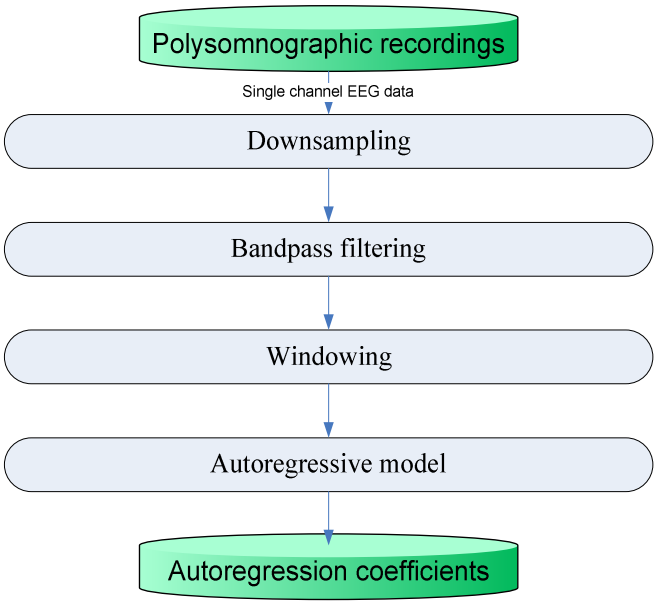

Figure 6: Sequence of steps for data preprocessing and feature selection

The goals of the preprocessing steps, downsampling and bandpass filtering, are to enable uniform processing of the EEG data and to reduce the influence of individual lab measurement settings. To accomplish this, standard signal processing techniques are applied.

The minimum frequency for EEG recording as required by the SIESTA recording protocol is 100 Hz (Klosch et al. 2001). To enable uniform processing, all recordings are resampled to that frequency using the resampling technique described in section 2.3.1.

According to the SIESTA recording protocol, the highpass-cutoff frequency is set between 0.16 and 0.5 Hz. The upper boundary of 0.5 Hz is naturally taken as the cutoff frequency for a highpass filter. To minimize the influence of different amplifier characteristics in the high frequency domain, lowpass filtering with cutoff frequency below 50 Hz is beneficial. In the band below 50 Hz lies the beta band, characteristic of brain activity in the wake state, with the upper limit of 30 Hz. As a compromise, the "golden mean" of 40 Hz is taken as a cutoff frequency for the lowpass filtering.

Both lowpass and highpass filtering can be done in one step by a bandpass filter to pass frequencies from 0.5 to 40 Hz. From the broad range of filters available, a Butterworth filter of order 8 was chosen. See section 2.3.2 for details. As in the downsampling step, the expected differences when using another type of filter or a slightly different filter order are expected to be non-significant, and comparison with different choices is not elaborated.

From the uniformly preprocessed EEG data, the autoregressive coefficients of an AR(10) model are estimated by the Burg algorithm for segments of three seconds. See section 2.4 for a description and discussion of the AR model.

3.2 Model framework

The aim of the framework described in this section is to provide a probabilistic model for describing sleep as a continuous process of sleep/wakefulness. From the model framework, an actual model is derived by specifying classes of sleep stages and possibly by extensions or adaptations, which will be the case in the last model described in this chapter.

For the training of a model, AR vectors from sleep EEG data have to be supplied together with labels for at least a part of the data. As the main result, a finite mixture is created that can be applied to unlabeled data to yield a set of probability vectors over time.

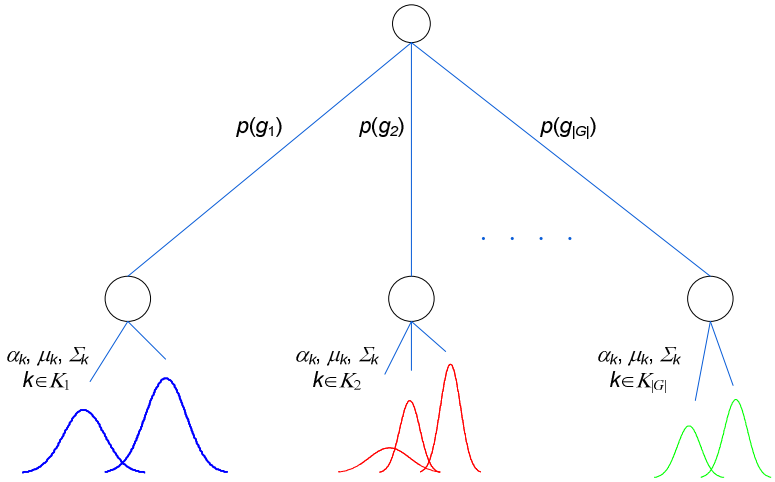

Figure 7: Structure of the hierarchical mixture used in the model framework. The root node represents $p(x)$, that is a weighted sum of the distributions $p(x \mid g)$ on the lower level. Those (bottom) are modeled by mixtures of Gaussian densities.

The basic approach of the model framework is to represent the distribution of the data, i.e. the AR-vectors, by a hierarchical mixture of two levels as shown in Figure 7. The model is independent of time, i.e. the probability of observing a time series $x_1, ..., x_T$ is given by

$$p(x_1, ..., x_T) = \prod_{t=1}^{T} p(x_t)$$

That said, time information and information about the recording can be discarded for the training data, as the ordering does not have any influence. Consequently, we will use the subscript n instead of t to highlight that sample x_n is just a training sample and we do not use time information.

One Gaussian mixture model (GMM) is used to model the class conditional densities (lower level), which are in turn used in the upper level. This can be seen as an extension of Mixture Discriminant Analysis

(MDA) (Hastie and Tibshirani 1996), similar to the model called MclustDA in (Fraley and Raftery 2002). After definition of a-priori class probabilities, Bayes' theorem can be applied to arrive at posterior probabilities for the class memberships.

A step called *reshuffling* is formulated, with the intention of decoupling the model from the class labels and enabling the GMM components to adapt more freely to the true, generating distribution. There are two reasons behind this idea. Firstly, the R&K class labels should just be the foundation of the model, not the one and only criterion. Furthermore, the training labels are partially incorrect, for the following reasons:

- The R&K labels are assigned to segments of 30 seconds, but state changes can happen within this period.
- The R&K rules are ambiguous, resulting in considerable variation of R&K scoring between different experts or systems.

Even with the assumption of perfect and unambiguous rating, overlapping areas in the data for different classes are to be expected, as the R&K rules incorporate mutual dependencies between neighboring segments, while in our model the data is assumed to be time independent. We will further discuss this point later in this section.

This setting can be seen as the problem of handling unreliable data, which is well-known in machine learning literature. By discarding the class labels in the reshuffling step, the problem is stated in the domain of semi-supervised learning. This is treated for a single GMM, e.g. in the MIXMOD software (Biernacki et al. 2005), by using the class information of the labeled data in the EM algorithm. We introduce a simple technique that is computationally efficient and provides viable results within this thesis. We refer to the supervised training using labeled data as step one, and to the reshuffling as step 2. These two steps are not to be confused with the two levels of the mixture. The first step results in a complete model, which is refined in the second step.

The nomenclature used for the model description is as follows:

G	Set of class indices
g	Class identifier, $g \in G$
x_n	Observation vector (AR-coefficients), without class affiliation
z_n	Class affiliation for observation x_n; $z_n \in G$
N_g	Number of training samples for class g. $N_g = \left\lvert \{ z_i \mid z_i = g \} \right\rvert$
N	Total number of training samples assigned to classes. $N = \sum_{g \in G} N_g$
K_g	Set of distinct indices for class-components. The number of components in class g is $\lvert K_g \rvert$. As the indices are distinct, $\bigcap_{g \subseteq G} K_g = \{\}$

K	Indices of all class-components, regardless of class affiliation. $K = \bigcup_{g \in G} K_g$
k	Component-identifier, $k \in K$
group(k)	Group assigned to component k. group$(k) = \{ g \in G \mid k \in K_g \}$

We have $|G|$ predefined classes, denoted by $g \in G$. Depending on the R&K label, each AR vector of the training data is either excluded or assigned to one of the classes. The class-conditional distributions are approximated by fully parameterized GMMs $p(x \mid g)$ as described in 2.5, with parameters and number of components being class-specific. That corresponds to the first level in our mixture hierarchy. In the second level, those GMMs are used as building blocks for a finite mixture of the form

$$p(x) = \sum_{g \in G} p(x, g) = \sum_{g \in G} p(x \mid g)\, p(g) \tag{15}$$

For data with unknown class membership, this can be seen as integrating out the class membership, similar to the components in a single GMM. The a-priori probabilities $p(g)$ have to be non-negative and add up to 1. In our case, they are set according to the observed distribution in the training data:

$$p(g) = \frac{N_g}{N} \tag{16}$$

For the first, supervised, training step class labels z_n are given, but component affiliations within the GMMs are unknown, and the log-likelihood of the model with parameters $\Theta = (\theta_g, \alpha_g ; g \in G)$ is given by

$$\begin{aligned} & L(\Theta \mid x_1, ..., x_N, z_1, ..., z_N) \\ & = \log \prod_{n=1}^{N} \{ p(x_n \mid z_n)\, p(z_n) \} \\ & = \sum_{n=1}^{N} \log p(z_n) + \sum_{n=1}^{N} \log\, p(x_n \mid z_n) \\ & = \sum_{n=1}^{N} \log p(z_n) + \sum_{g \in G} \sum_{n=1}^{N} \{ \delta(z_n, g) \log\, p(x_n \mid g) \} \end{aligned} \tag{17}$$

As the a-priori probabilities are fixed in advance by equation (16), optimizing the likelihood $L(\Theta \mid X, Z)$ given by equation (17) is the same as independent optimization of the GMMs representing the class-conditional distributions, with class log-likelihood

$$L_g(\Theta \mid x_1, ..., x_N, z_1, ..., z_N) = \sum_{n=1}^{N} \delta(z_n, g) \log\, p(x_n \mid g), \quad g \in G \tag{18}$$

Maximum likelihood estimation is done by the EM algorithm, with multiple initializations and outlier treatment as described in section 2.6. Model order selection is done by BIC, with $N^{3/10}$ as the upper limit for the number of components (Bozdogan 1993). The number of components is increased as long as the difference of BIC values gives very strong positive evidence, i.e. as the difference is greater than 10 (Kass and Raftery 1995).

In the reshuffling step, the class labels $z_1, ..., z_N$ of the training data are not used. The same data as in the first step or a different set of data can be used. Both approaches have been tried, with the latter producing

better results. For simplicity, the nomenclature introduced above does not cover the partitioning of the training data for the two steps, but this should not be crucial for the following description. As an alternative to the two consecutive steps, both labeled and unlabeled data could be used in one EM step. With our approach, more emphasis is put on the unlabeled data, as the model can adapt to the data more freely, without being held tightly by labeled data.

The GMMs representing the class-conditional distributions are used to create a compound GMM featuring the union of all class components. Following the same notation as above, but using the identifier $^{(b)}$ to highlight the parameters of the compound GMM, we have:

$$p(x) = \sum_{k \in K} \alpha_k^{(b)} \; p(x \mid \theta_k) \;\text{ with initialization }\; \forall g \in G : \alpha_k^{(b)} = \alpha_k \; p(g), \quad k \in K_g \tag{19}$$

While the equations (15) and (19) are equivalent, the ML estimation of the parameters of the compound GMM has the following consequences:

- Every GMM component is potentially affected by all elements of the training data set.
- The mixing proportions are free to vary, with the limitation that they add up to 1. As a result, the predefined a-priori probabilities (16) would be altered when retransforming the optimized parameters back to the class-specific model.

To overcome the second point, the mixing proportions have to be fixed for groups of components:

$$\forall g \in G : \sum_{k \in K_g} \alpha_{k'}^{(b)} = p(g) \tag{20}$$

That way, only the mixing distributions of the components within the groups are adopted, while the a-priori class probabilities are still defined by equation (16). Incorporation of this constraint into the EM algorithm is shown in Appendix A1. The resulting update formulas are equivalent to those given in equations (11) to (14), but with the update of the mixing proportions (14) replaced by

$$\alpha_k^{new\;(b)} = p(g) \frac{\sum_{n=1}^{N} p^{old}(k \mid x_n)}{\sum_{k' \in K_{group(k)}} \sum_{n=1}^{N} p^{old}(k' \mid x_n)}, \quad k \in K_g \tag{21}$$

When optimizing the compound GMM by the adapted EM algorithm, the parameters θ_k of the component normal distributions are affected directly, while the mixing proportions are transferred to the single class-conditional GMMs by the rule

$$\forall g \in G : \quad \alpha_k = \frac{\alpha_k^{(b)}}{p(g)}, \quad k \in K_g \tag{22}$$

This is just the inversion of equation (19) applied to the initialization of the compound GMM and ensures that the mixing proportions add up to one for each class GMM. For the split-up of the compound GMM into the class GMMs to work, the monotonic characteristic of the EM algorithm is an important assumption. Without this local characteristic, rearrangement of the Gaussian components could occur and

identification of components assigned to groups would be impossible. By taking into account the fact that the EM algorithm is highly dependent on the initial parameter estimates, it is assumed that the components are adjusted to the overall training data set without being distorted. A similar approach was taken in (Flexer et al. 2002), where the Gaussian observation distributions of a Hidden Markov were initialized by a supervised training step. The reshuffling step could also be formulated without the intermediate compound GMM, as from the model perspective the model structure is implicitly kept in the compound GMM.

By applying Bayes' theorem, the posterior probabilities for the classes can be calculated:

$$p(g \mid x) = \frac{p(x \mid g)\, p(g)}{p(x)} \propto p(x \mid g)\, p(g) \tag{23}$$

Note that the prior probabilities $p(g)$ can be easily adjusted when additional information from external sources is available. For example, when we have a model containing a class wake and we do possess information about the occurrence of sleep spindles, the prior probability for class wake can be decreased or set to zero. During the model training, the prior probabilities do not affect the supervised training, but they do in the unsupervised reshuffling-step. Priors deviating severely from the distribution in the training data may cause some of the groups to vanish and others to take over.

The parameters for controlling the optimization procedure are as follows:

- The number of different GMM initializations for step 1.
- The maximum number of iterations for the EM algorithm as used in the initialization step. That is, each of the GMMs is optimized by applying the specified number of EM steps; the one with maximum likelihood is further used.
- The maximum number of EM-iterations applied to the GMM selected in the initialization step as described above
- The maximum number of EM-iterations applied to the compound GMM in the reshuffling step

For the model training done in this thesis, three initializations with 10 EM-iterations are done. The selected model is trained for another 400 cycles. In the reshuffling step, a maximum of 200 iterations are executed. To make the optimization computationally feasible, non-overlapping segments of three seconds are used for both training steps. The training set of 158 recordings (from 79 healthy controls) was randomly split into two sets of equal size. One set is used for the first step and the other in the reshuffling procedure.

An alternative to the constraints on the mixing proportions as defined by equation (20), a simpler model with equal mixing proportions of all components, could be suitable, which is labeled p-Lk-Ck in section 2.5. When class-specific numbers of components are allowed for, a restricted EM algorithm where the mixing proportions are fixed according to the initial values could be applied for both training steps. The main advantage of this model would be that the training procedure would be more stable because component collapses are prevented in a natural way.

In the illustration of the model building process (Figure 8), the densities are sketched in one-dimensional space by their probability density distribution. In the first step, class conditionals are modeled by one GMM each, which are trained on the data assigned to that class (top). The weighted sum of the components yields the class-conditional distribution $p(g \mid x)$. The compound GMM is a weighted combination of the class-conditional GMMs (second row). Based on unlabeled data, this compound GMM is retrained with the sum of the mixing proportions constrained to sum to the class prior for each class. Finally, the reshuffled GMM is split up into the group GMMs, where the mixing proportions have to be normalized to add up to one for each group. The parameters of the component densities (mean and variance matrices) are directly affected by the reshuffling process, and the group priors are not affected at all.

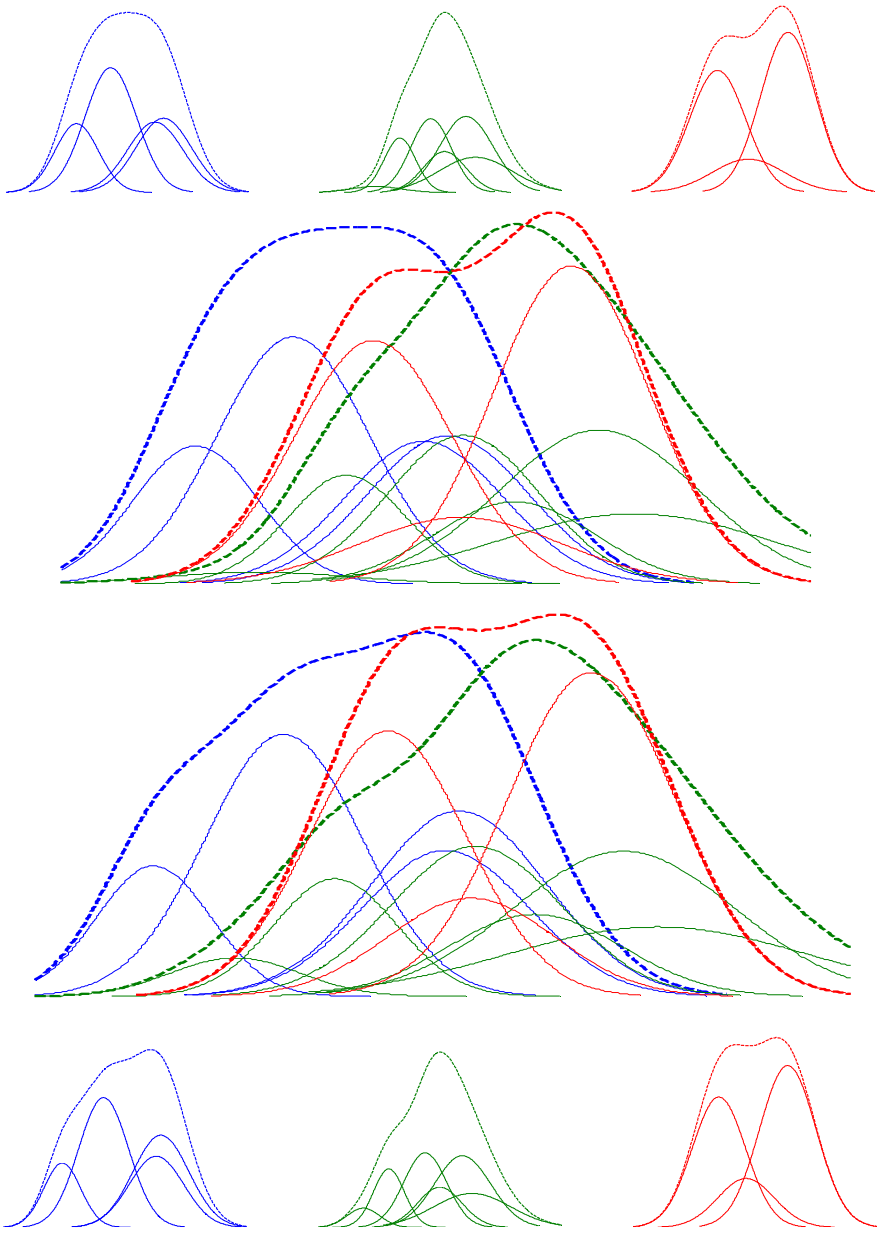

Figure 8: Illustration of the model building process. The densities are taken from a model of Sleep Cornerstones (see chapter 3.3) and are projected to the first principal component of the data. Top: group-conditional densities are trained independently. Second row: composite GMM before reshuffling. Third row: composite GMM after unsupervised reshuffling. Bottom: the reshuffled GMM is split up into the group GMMs.

3.3 Model of Sleep Cornerstones

In the model of sleep cornerstones, the three relatively unambiguous states wake, NREM sleep and REM sleep are considered (see Table 3). In the work done by (Pardey et al. 1996), it was observed that the sleep/wakefulness continuum can be described solely by wake, REM sleep, and the s4 sleep processes. Consequently they developed a model (a multi-layer perceptron) trained on the epochs labeled wake, REM sleep, and sleep stage 4. The output of the model, probability triples over time, were used for assignment to one of the traditional R&K sleep stages, i.e. each probability triple is assigned to one of the six R&K sleep stages.

Gaussian mixture models were used in (Sykacek et al. 1999) to approximate class conditional densities, based on AR parameters (reflection coefficients) and a complexity measure. For inference of the model parameters, periods labeled wake, REM, and s4 were used together with the rest of the data without using class assignments. The model was trained on just one recording and the probability traces were shown for one test recording.

In the continuous sleep analyzer using Hidden markov models (HMM) described in (Flexer et al. 2005), the cornerstones wake, REM, and deep sleep (s4) were used as discrete states of the HMM, and for comparison with R&K scorings, classification to one of the six R&K sleep stages was carried out.

In our model of sleep cornerstones, we use data from wake, REM, and all NREM sleep stages. The benefit of using all NREM sleep stages as opposed to only those of stage 4 is seen in the increased amount of training data. Sleep stage 2 is predominant in NREM sleep, and sleep stage 4 is rarely observed in some subjects. As we expect the probability of NREM sleep to be highest for deep sleep, assignment to one of the six R&K sleep stages should also be possible, similar to that done in the work mentioned above.

Table 3: Groups of sleep stages for the model of sleep cornerstones

Group	R&K codes
wake	wake (0)
NREM	stage 1 (1) stage 2 (2) stage 3 (3) stage 4 (4)
REM	REM (5)

3.4 Model of R&K Sleep Stages

The most detailed level of modeling offered by the model framework is utilized in the model of R&K sleep stages, where each group is defined by exactly one R&K label. All stages except labels "undefined" and "movement" are considered (see Table 4). The main goal of this model is to gain knowledge about which sleep stages are more important than others in respect to sleep quality. Moreover, it will be interesting to investigate whether the NREM sleep states can be distinguished reasonably well or if they are heavily overlapping in the AR(10) representation. The class conditionals for the groups wake and REM are expected to be closely related to those from the model of sleep cornerstones. Minor differences are to be expected because of the different structure of the compound GMM used in the reshuffling step.

Table 4: Groups of sleep stages for the model of R&K sleep stages

Group	R&K codes
wake	wake (0)
s1	stage 1 (1)
s2	stage 2 (2)
s3	stage 3 (3)
s4	stage 4 (4)
REM	REM (5)

3.5 Model of Sleep Substructure

The model of sleep substructure takes an exceptional position because information from an external classifier is incorporated. The requirements of the classifier are to discriminate between the states wake, NREM sleep, REM sleep, movement, and undefined epochs. In the model of sleep substructure, the states wake and NREM are modeled continuously, while REM sleep is seen as a discrete process. NREM is partitioned into the processes s1, s2, and deep sleep. See Table 5 for the list of continuous and discrete states. The aggregation of s3 and s4 as deep sleep is motivated by the predominant delta activity in both stages. Furthermore, restorative theories of sleep ascribe the same functions to s3 and s4 sleep (Kryger et al. 2005).

The motivation behind this approach is that REM detection based solely on EEG data is problematic, and robust, validated automatic classifiers, taking into account additional biosignals, already exist to differentiate the aforementioned states. We use the automated sleep stager Somnolyzer 24 × 7 (Anderer et al. 2005) as an external classifier. As the external information has to be integrated, the data flow is slightly different from that sketched at the beginning of this chapter. See Figure 9 for the adapted version.

In the work done by (Kemp 1993), REM was also modeled as a discrete process, together with a continuous wake/NREM process. Structurally, our model can be seen as a refinement, as the substructure of wake/NREM is emphasized by one process for wake and several for the NREM activity. This should allow extraction of information from the microstructure of s2 and slow wave sleep as proposed by (Kubicki and Herrmann 1996).

We define the model of sleep substructure based on the model framework, with the following refinements:

- In the reshuffling step, information from the external classifier is used to extract wake and NREM periods.
- The output of the model is composed by the four continuous processes wake, s1, s2, deep; and the three discrete traces for REM, movement, and undefined. The discrete traces are fed by the external classifier.
- For the points in time where one of the discrete states is active, the probabilities of the continuous processes are zero by definition.

For the calculation of the posterior group probabilities, no distinction is made between the discrete states wake and NREM. That way, continuous transitions between wake and the NREM groups can be traced, e.g. to capture micro arousals. In the analysis of the model output, however, information from the discrete wake / NREM processes will also be used (see section 4.2.4).

Table 5: Groups of sleep stages for the model of sleep substructure. *:discrete

Group	R&K codes	External classifier
wake	wake (0)	wake
s1	stage 1 (1)	NREM
s2	stage 2 (2)	NREM
deep	stage 3 (3) stage 4 (4)	NREM
REM*	REM (5)	REM
movement*	movement (6)	movement
undefined*	undefined (9)	undefined

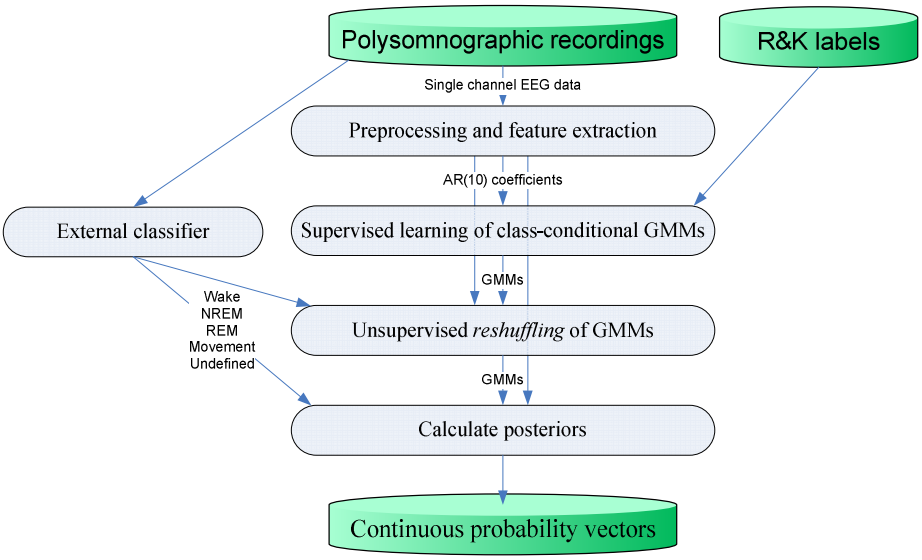

Figure 9: Data flow for the model of sleep substructure. Information from an external wake/REM/NREM classifier is used when applying the model (calculation of the posterior probabilities) to unseen data.

4 Feature calculation

The next step after sleep staging is the calculation of parameters (features). Ideally those features are clinically relevant; e.g. they are correlated to sleep quality or discriminative regarding sleep disorders. We derive such features from the R&K and continuous sleep profiles. Subsequently we use them for correlation analysis and evaluation of discriminative power for different patient groups. In this work, the expressions *R&K features* and *continuous features* are sometimes used in a casual way to refer to features derived from R&K sleep profiles and features derived from continuous sleep profiles.

4.1 R&K features

There is no established consensus to date about the parameters derived from the R&K sleep hypnogram, and terminology and quantities derived vary according to the type of analysis and purpose of the study. However, a number of conventions are fairly common, e.g. to include the times spent in each stage and the time from "lights off" until the first appearance of sleep.

We shortly describe the features used within this work in Table 6, without pointing out alternatives or possible extensions. The variable names start with "rk_" to avoid confusion with similar variables computed from continuous models. All time quantities are given in minutes.

The recognition of sleep cycles was done according to (Feinberg and Floyd 1979). From the extracted begin- and end-times of the sleep cycles, only average length and the number of cycles are kept.

Table 6: Description of features calculated from a R&K sleep hypnogram

Variable name	Description
rk_uns	**Time left unscored** Time not scored (code 9) that is not at the beginning or end of the recording.
rk_tib	**Time in bed computed** Length of the recording without unscored periods (code 9).
rk_tsp	**Total sleep period** Time from first appearance of any sleep stage until final awakening.
rk_tst	**Total sleep time** Time spent in sleep stages 1, 2, 3, 4 or REM.
rk_eff	**Sleep efficiency** Total sleep time (rk_tst) / Time in bed (rk_tib)
rk_sl	**Sleep latency** Time from beginning of the recording until first appearance of S2 or first consecutive three periods of S1.
rk_wtsp	**Wake within Total sleep period (rk_tsp)**
rk_wbb	**Wake before lights-on** Time from final awakening until the end of the recording.
rk_mt	**Time scored as movements (code 8)**
rk_fw	**Number of awakenings during Total sleep time (rk_tst)** Number of coherent periods of wakefulness without taking into account the first and last wake periods.
rk_fs	**Number of state changes during Total sleep period (rk_tsp)**

Variable name	Description
rk_s1	**Time of sleep stage s1**
rk_s2	**Time of sleep stage s2**
rk_s3	**Time of sleep stage s3**
rk_s4	**Time of sleep stage s4**
rk_s4_q*x*	**Time of sleep stage s4 separately for the quarters *n* of the night**
rk_srem	**Time of stage REM**
rk_tst_s1	**Percentage of sleep stage s1 within total sleep time (rk_tst)**
rk_tst_s2	**Percentage of sleep stage s2 within total sleep time (rk_tst)**
rk_tst_s3	**Percentage of sleep stage s3 within total sleep time (rk_tst)**
rk_tst_s4	**Percentage of sleep stage s4 within total sleep time (rk_tst)**
rk_tst_srem	**Percentage of sleep stage REM within total sleep time (rk_tst)**
rk_sc_*xy*	**Number of state changes between stage *x* and stage *y*** As codes for *x* and *y*, 1/2/3/4 are used for sleep stages s1/s2/s3/s4, d for deep sleep (s3 or s4), w for wake. When the stages of two consecutive epochs are identical, rk_sc_*xx* is addressed, e.g. rk_sc_44 corresponds to the number of s4 epochs followed by another s4 epoch.
rk_n_sleep_cycles	**Number of sleep cycles**
rk_n_remc	**Number of REM cycles**
rk_n_nremc	**Number of NREM cycles**
rk_avd_sleep_cycles	**Average duration of sleep cycles**
rk_avd_remc	**Average duration of REM cycles**
rk_avd_nremc	**Average duration of NREM cycles**
rk_rem_latency	**REM latency** Time from first appearance of S2 until beginning of first REM cycle.

4.2 Continuous features

For most of the R&K features described above, similar parameters can be extracted from the continuous models. Possibilities to derive further variables are manifold, and the type of analysis is more complex due to the continuous characteristic. In this chapter, a description of general approaches used for the calculation of features for the continuous models of sleep is followed by the parameter lists for the three continuous models.

4.2.1 General concepts

Two categories of features are derived from the continuous sleep profiles. Features from the first category are based directly on the continuous traces, while in the second category the probabilities are used for classification and the resulting discrete states are used for the feature calculation.

Classification

From the continuous probability traces, similar parameters to the traditional R&K parameters can be derived be choosing a "winner" for each epoch to end up with a discrete sleep hypnogram with high time resolution.

For a continuous model with groups $g \in G$, the winner for the time segment t is determined by maximum a-posteriori estimation:

$$\hat{g}_t = \arg\max_{g \in G} p(g \mid x_t)$$

where the posterior probabilities $p(g \mid x_t)$ are computed by application of the Bayes' theorem as described in the model framework, section 3.2.

Prior to classification, it can be beneficial to smooth the probability traces, to reduce high frequency changes and thus increase the stability of the classification. For all parameters based on classifications, we compute three versions: one without filtering and two with lowpass filtering using moving average filters of length 10 and 100 (denoted by trailing "_f1" for the unfiltered version, "_f10" and "_f100" for the filtered, respectively):

$$p_{\mathrm{f}n}(k \mid x_t) = \frac{1}{n} \sum_{i=t-n}^{t-1} p(k \mid x_i)$$

Number of state changes

As for the discrete R&K labels, the number of state changes can be calculated from the classified epochs. The variables are denoted by "sc_*xy*_f*n*" where *x* and *y* are substituted by group labels. For *x*=*y*, the number of epochs that are followed by another epoch assigned to the same group are counted.

The level of filtering prior to classification is again indicated by f*n*, where *n* is the filter length of the moving average filter.

Area under the curve

The computation of the area under the curve (AUC) of a probability trace poses another possibility of calculating features similar to the times spent in several sleep stages according to R&K sleep staging. In the variable nomenclature, AUC variables are indicated by prefixing the name with "auc_".

The posterior probabilities are given at equidistant times points with a distance of three seconds when using non-overlapping windows. We denote the time indices by $t = 1, ..., T$, and the distance is set to h=3 sec when using non-overlapping windows. Using the composite trapezium rule, linear interpolation between two given points is done, and the AUC is computed by

$$\begin{aligned} AUC_g &= \sum_{t=1}^{T-1} \frac{h}{2} \left\{ p(g \mid x_t) + p(g \mid x_{t+1}) \right\} \\ &= \frac{h}{2} \left\{ \left(2 \sum_{t=1}^{T} p(g \mid x_t) \right) - p(g \mid x_1) - p(g \mid x_T) \right\} \end{aligned}$$

For a big number T and small value for t=1 and t=T, the AUC is approximately

$$AUC_g \approx h \sum_{t=1}^{T} p(g \mid x_t)$$

For a filter of length n, only the sum of the first $n - 1$ posterior values $p(g \mid x_t)$ is affected. For small filter length in comparison to the number of samples of the time series, changes are marginal. For whole night recordings we have around 10,000 windows with lengths of 3 sec, with posterior values between 0 and 1. For that reason, it is sufficient to calculate the AUC from the unsmoothed probability traces only.

To diminish the effect of "noise" in the probability traces, i.e. low probabilities, a cutoff level of 0.1 is set, and values below are set to zero prior to calculation of the AUC. See Figure 10 for an illustrative example.

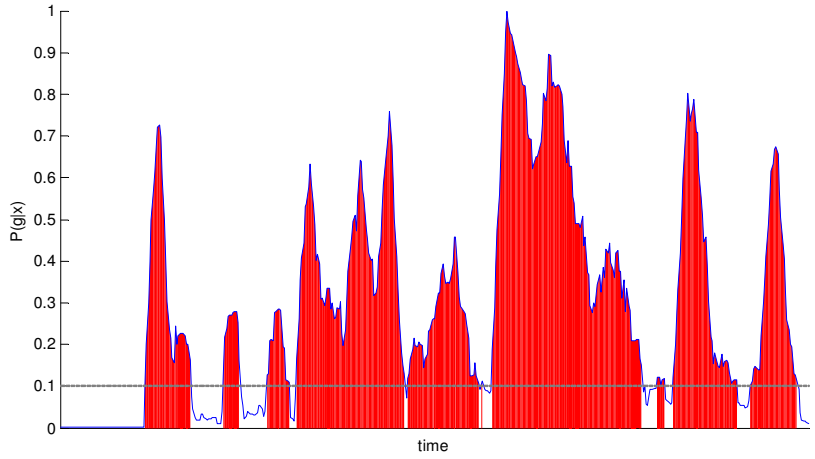

Figure 10: Area under the curve (AUC, red area) with cutoff level of 0.1 (grey horizontal line). Values below 0.1 do not account for the AUC, while the values above are accounted for in full magnitude.

Area under the curve of 1st and 2nd derivative

The AUC of the 1st and 2nd derivative of a probability trace give a measure of complexity. The corresponding parameters are denoted by leading "auc1" and "auc2". Big changes in the probability traces induce high values of the 1st derivative, while for a relatively constant trace the 1st derivative is of low amplitude. We do not want to distinguish between positive and negative changes; therefore the AUC is taken from the absolute values of the derivatives. The 1st derivative is approximated by the central difference

$$\frac{d}{dt} p(g \mid x_t) \approx \frac{p(g \mid x_{t+1}) - p(g \mid x_{t-1})}{2h}$$

The second derivative is computed using the Laplace operator:

$$\frac{d^2}{dt^2} p(g \mid x_t) \approx \frac{p(g \mid x_{t-1}) - 2p(g \mid x_t) + p(g \mid x_{t+1})}{2h^2}$$

In both formulas, *h* equals 3 sec when using non-overlapping windows. Filtering of the posterior values might have a considerable impact, thus variable names end with f*n* as usual.

Entropy

Entropy measures the disorder or randomness of a system. There are different definitions for calculation of the entropy, we use the Burg entropy (Burg 1972):

$$entropy_g^{burg} = \sum_i \log p_i = \log \overline{p}_g + \log \overline{p}_{\sim g} = \log \overline{p}_g + \log(1 - \overline{p}_g)$$
$$= \log\left\{\frac{1}{T}\sum_{t=1}^{T} p(g \mid x_t)\right\} + \log\left\{\frac{1}{T}\sum_{t=1}^{T}\left(1 - p(g \mid x_t)\right)\right\}$$

We distinguish between two probabilities, one for being in the state corresponding to group *g*, and the complementary probability, both expressed as mean probabilities across time. This entropy measure is greatest for a mean probability of 0.5, as shown in Figure 11.

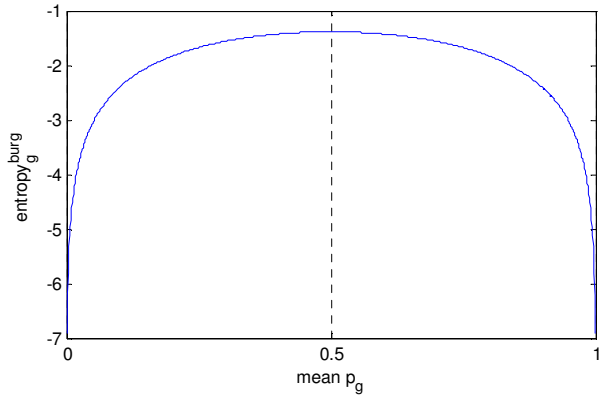

Figure 11: Burg entropy depending on the mean probability for a group

For the whole set of probability vectors, we compute the Shannon entropy

$$entropy^{shannon} = -\sum p_i \log_2 p_i = -\sum_{g \in G} \overline{p}_g \log_2 \overline{p}_g$$
$$= -\sum_{g \in G}\left\{\left(\frac{1}{T}\sum_{t=1}^{T} P(g \mid x_t)\right)\log_2\left(\frac{1}{T}\sum_{t=1}^{T} P(g \mid x_t)\right)\right\}$$

This parameter was also used in (Flexer et al. 2005) to investigate differences in the probability traces depending on sleep laboratories.

Different smoothing levels are not considered, as the distribution of the unsmoothed plots is expected to hold the maximum amount of information. The variables are named "entropy_*group*", where *group* is to be substituted by the group label, and simply "entropy" for the intergroup entropy.

Path length

The path length is another complexity measure, calculated as the mean of the Euclidean distances between consecutive probability vectors $\left(p(g \mid x_t);\, g \in G\right)$ and $\left(p(g \mid x_{t+1});\, g \in G\right)$:

$$pathlength = \frac{1}{T-1}\sum_{t=1}^{T-1}\sqrt{\sum_{g\in G}\left\{p(g \mid x_{t+1}) - p(g \mid x_t)\right\}^2}$$

Note that this quantity does capture the amount of switching between states, but does not take into account an ordering of the states. For example, assume groups wake, s1, s2, and deep sleep; probabilities shifting from s1 to s2 affect the path length in the same way as shifts from s1 to deep sleep or from deep sleep to wake, which reflect more relevant changes of state.

The path length is calculated for different smoothing levels, and the variable names are "path_length_f*n*", where *n* again denotes the filter length of the moving average filter applied preliminarily.

4.2.2 Features for the continuous model of sleep cornerstones

For calculation of the features, only information from the probability traces computed from the EEG data is used. The Total sleep period is worth mentioning, as this is used for the calculation of other features. The idea is to find the first and last appearance of any sleep. By using the probabilities with high temporal resolution, outliers might result in sleep showing up too early. By smoothing the posteriors, this effect might be reduced. For some other parameters, e.g. the AUC of the magnitude of the 1st derivative of wake, just the epochs within the Total sleep period are considered, and different smoothing levels for the parameter calculation are used. In that case, multiple combinations, e.g. smoothing for calculation of the Total sleep time and no smoothing for calculation of the auc1 are conceivable. To keep the numbers of variables reasonably low, these combinations are not implemented. As a result, for the parameters with low filtering, the sleep onset and the final awakening might be biased towards the beginning / end of the recording.

Table 7: Description of features calculated from the output of the model "Sleep cornerstones"

Variable name	Description
cs_tib	**Time in bed** Length of the recording
cs_tsp_f*n*	**Total sleep period** Time from first appearance of any sleep until final awakening, based on classification.
cs_auc_tsp_f*n*_wake	**Area under curve (AUC) for wake within total sleep period (cs_tsp_f*n*)** The filtering is not done for the wake trace, but for calculation of the total sleep period (cs_tsp_f*n*).
cs_auc1_tsp_wake_f*n*	**AUC of magnitude of 1st derivate of wake during total sleep period (cs_tsp_f*n*)**
cs_auc2_tsp_wake_f*n*	**AUC of magnitude of 2nd derivate of wake during total sleep period (cs_tsp_f*n*)**
cs_auc_nrem	**AUC of NREM sleep**
cs_auc_nrem_q*x*	**AUC of NREM sleep, separately for the quarters of the night**
cs_auc1_nrem_f*n*	**AUC of magnitude of 1st derivate of NREM sleep**
cs_auc2_nrem_f*n*	**AUC of magnitude of 2nd derivate of NREM sleep**
cs_auc_rem	**AUC of NREM sleep**

Variable name	Description
cs_auc1_rem_f*n*	**AUC of magnitude of 1st derivate of REM sleep**
cs_auc2_rem_f*n*	**AUC of magnitude of 2nd derivate of REM sleep**
cs_tst_f*n*	**Total sleep time** Time spent in REM or NREM sleep, based on classification.
cs_eff_f*n*	**Sleep efficiency** Total sleep time (cs_tst_f*n*) / Time in bed (cs_tib)
cs_sl_f*n*	**Sleep latency** Time from beginning of the recording until first appearance of REM or NREM sleep, based on classification.
cs_wtsp_f*n*	**Wake within Total sleep period (cs_tsp_f*n*)** Based on classification.
cs_fw_f*n*	**Number of coherent wake periods within Total sleep period (cs_tsp_f*n*)** Based on classification.
cs_wake_f*n*	**Absolute duration of wake** Based on classification, computed for the whole recording.
cs_nrem_f*n*	**Absolute duration of NREM sleep** Based on classification, computed for the whole recording.
cs_nrem_q*x*_f*n*	**Absolute duration of NREM sleep, separately for the quarters of the night**
cs_rem_f*n*	**Absolute duration of REM sleep** Based on classification, computed for the whole recording.
cs_tst_nrem_f*n*	**Percentage of NREM sleep within total sleep time (cs_tst_f*n*)**
cs_tst_rem_f*n*	**Percentage of REM sleep within total sleep time (cs_tst_f*n*)**
cs_entropy_wake	**Entropy of wake trace**
cs_entropy_nrem	**Entropy of NREM trace**
cs_entropy_rem	**Entropy of REM trace**
cs_entropy	**Entropy of wake, NREM, REM**
cs_path_length_f*n*	**Sum of Euclidean distance of posterior vectors**
cs_sc_*xy*_f*n*	**Number of state changes between group *x* and group *y*** Based on classification; the group labels can be wake, nrem, rem.

4.2.3 Features for the continuous model of R&K sleep stages

For the Total sleep period, the same issue as for the model of Sleep cornerstones arises. Analogous to the approach taken there (see above), the same filter length is used for identification of the Total sleep period and the calculation of the features where filtering is taken into account.

Table 8: Description of features calculated from the output of the model "R&K sleep stages"

Variable name	Description
s6_tib	**Time in bed** Length of the recording
s6_tsp_f*n*	**Total sleep period** Time from first appearance of any sleep until final awakening, based on classification.
s6_auc_tsp_f*n*_wake	**Area under curve (AUC) for wake within total sleep period (s6_tsp_f*n*)** The filtering is not done for the wake trace, but for calculation of the total sleep period (s6_tsp_f*n*).
s6_auc1_tsp_wake_f*n*	**AUC of magnitude of 1st derivate of wake during total sleep period**
s6_auc2_tsp_wake_f*n*	**AUC of magnitude of 2nd derivate of wake during total sleep period**
s6_auc_*group*	**AUC of s1, s2, s3, s4, rem**
s6_auc_deep_q*x*	**AUC of deep sleep, separately for the quarters of the night**
s6_auc1_*group*_f*n*	**AUC of magnitude of 1st derivate** The placeholder *group* is replaced by s1, s2, s3, s4, rem
s6_auc2_*group*_f*n*	**AUC of magnitude of 2nd derivate** The placeholder *group* is replaced by s1, s2, s3, s4, rem
s6_tst_f*n*	**Total sleep time** Time spent in REM or in NREM sleep stages; based on classification.
s6_eff_f*n*	**Sleep efficiency** Total sleep time (s6_tst_f*n*) / Time in bed (s6_tib)
s6_sl_f*n*	**Sleep latency** Time from beginning of the recording until first appearance of REM or any NREM sleep, based on classification.
s6_wtsp_f*n*	**Wake within Total sleep period (s6_tsp_f*n*)** Based on classification.
s6_*group*_f*n*	**Absolute duration of sleep stage** Based on classification, computed for the whole recording. The placeholder *group* is replaced by s1, s2, s3, s4, rem.
s6_s4_q*x*_f*n*	**Absolute duration of s4 sleep, separately for the quarters of the night**
s6_tst_*group*_f*n*	**Percentage of sleep stage within total sleep time (s6_tst_f*n*)** Based on classification; replace *group* by s1, s2, s3, s4, rem.
s6_entropy_*group*	**Entropy of trace for particular *group*** the placeholder *group* is replaced by wake, s1, s2, s3, s4, rem.
s6_entropy	**Entropy of wake, s1, s2, s3, s4, REM**
s6_path_length_f*n*	**Sum of Euclidean distance of posterior vectors**
s6_sc_*xy*_f*n*	**Number of state changes between group *x* and group *y*** Based on classification; the group labels are wake, s1, s2, s3, s4, rem.

4.2.4 Features for the continuous model of sleep substructure

The model of sleep substructure deserves extra comments, as discrete states are incorporated into this model. While for the calculation of the posteriors, wake and NREM as given by the external classifier are not used, we use this information to determine sleep onset and the time of final awakening, i.e. the Total sleep period. This is expected to give more robust results especially for the features concerning wake. Furthermore, different smoothing levels do not affect the Total sleep period; therefore more concise comparison of features derived with different filter length is possible.

Table 9: Description of features calculated from the output of the model "Sleep substructure"

Variable name	Descriptions
su_tib	**Time in bed** Length of the recording, without leading and trailing periods marked as "unscored" by the external classifier.
su_tsp	**Total sleep period** Time from first appearance of any sleep until final awakening. External information from wake/NREM/REM classifier is used, where NREM and REM periods are considered as sleep.
su_auc_tsp_f*n*_wake	**Area under curve (AUC) for wake within total sleep period (s6_tsp)** Periods not assigned to wake or NREM by the external classifier are excluded prior to calculation of the AUC.
su_auc1_tsp_wake_f*n*	**AUC of magnitude of 1st derivate of wake during total sleep period (su_tsp)** Periods not assigned to wake or NREM by the external classifier are excluded prior to calculation of the AUC.
su_auc2_tsp_wake_f*n*	**AUC of magnitude of 2nd derivate of wake during total sleep period (su_tsp)** Periods not assigned to wake or NREM by the external classifier are excluded prior to calculation of the AUC.
su_auc_*group*	**AUC** The placeholder *group* is replaced by s1, s2, deep. Periods not assigned to wake or NREM by the external classifier are excluded prior to calculation of the AUC.
su_auc1_*group*_f*n*	**AUC of magnitude of 1st derivate** The placeholder *group* is replaced by s1, s2, deep. Periods not assigned to wake or NREM by the external classifier are excluded prior to calculation of the AUC.
su_auc2_*group*_f*n*	**AUC of magnitude of 2nd derivate** The placeholder *group* is replaced by s1, s2, deep. Periods not assigned to wake or NREM by the external classifier are excluded prior to calculation of the AUC.
su_auc_deep_q*x*	**AUC for deep sleep for quarter *x*** Same as su_auc_deep, but calculated separately for the first, second, third, and last quarter of the night (*x*=1, 2, 3, 4).
su_tst_f*n*	**Total sleep time** Time spent in s1, s2 or deep sleep, based on classification. External information about wake/REM/NREM is not used.
su_eff_f*n*	**Sleep efficiency** Total sleep time (su_tst_f*n*) / Time in bed (su_tib)

Variable name	Descriptions
su_sl_f*n*	**Sleep latency** Time from beginning of the recording until first appearance of s2 or deep sleep, based on classification. External information about wake/REM/NREM is not used.
su_wtsp_f*n*	**Wake within Total sleep period (su_tsp)** Based on classification, external information about wake/REM/NREM is not used.
su_*group*_f*n*	**Absolute duration of sleep stage** Based on classification, computed for the whole recording. The placeholder *group* is replaced by s1, s2, deep, rem. External information about wake/REM/NREM is used for REM only.
su_deep_q*x*_f*n*	**Absolute duration of deep sleep, separately for the quarters of the night**
su_tst_*group*_f*n*	**Percentage of sleep stage within total sleep time (s6_tst_f*n*)** Based on classification; the placeholder *group* is replaced by s1, s2, deep, rem. External information about wake/REM/NREM is used for REM only.
su_entropy_*group*	**Entropy of trace for particular *group*** The placeholder *group* is replaced by wake, s1, s2, deep.
su_entropy	**Entropy of wake, s1, s2, deep, REM**
su_path_length_f*n*	**Sum of Euclidean distance of posterior vectors** Only the continuous groups wake, s1, s2, deep are considered.
su_sc_*xy*_f*n*	**Number of state changes between group *x* and group *y*** Based on classification; the group labels are wake, s1, s2, deep, REM.

5 Results

In the results chapter, we provide outcomes and findings for the three continuous models described in chapter 3. The results given for the EEG preprocessing and feature extraction steps serve as quality control for the chosen methods. For the continuous models, the results regarding model training will focus on the model selection, i.e. the number of components of the class-conditional GMMs, but will also trace changes of the reshuffling step.
Emphasis is placed on the output of the continuous sleep models, i.e. the sleep profiles and features derived thereof. The continuous and the R&K-features are used for correlation analysis with age and external information about sleep quality. Furthermore, the ability of those features to discriminate between healthy controls and several patient groups will be explored.

To ease comparison of the models, the subsections regarding training and analysis treat all models rather than giving results separately for the different models.

5.1 Feature extraction

The cascade of preprocessing and feature extraction from EEG signals was described in section 3.1. For the signal preprocessing steps, we will give examples of the effect on real data signals. That is done as quality control to validate that the methods and techniques applied work as designed. Main emphasis, though, is put on the results from the AR model, where we will inspect results from the parameter estimation, visualize the distribution of the AR coefficients, and look at the spectra reconstructed from the AR models.

5.1.1 Downsampling

The impact of the downsampling procedure on EEG data is shown in Figure 12 for a short time interval, with resampling from 200 Hz to 100 Hz. For the lowpass filtered signal, the effect of smoothing can be seen, e.g. at the beginning of the plot at sec 9000. Without anti-aliasing, the Nyquist theorem is violated and high-frequency components falsify the signal.

Thus, the downsampling procedure is validated and the signal is ready to be passed through to the next preprocessing step.

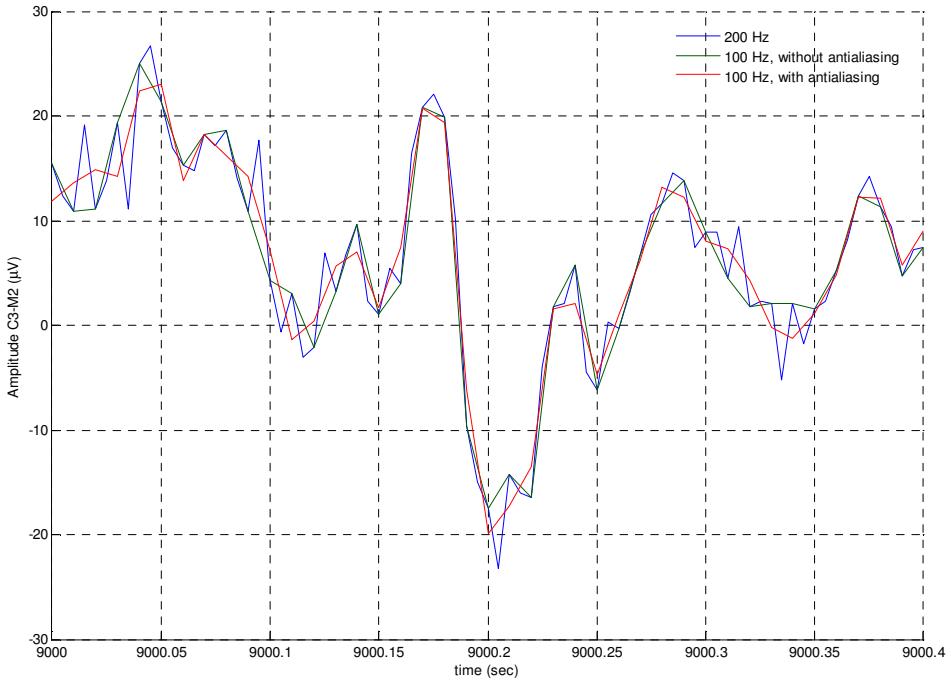

Figure 12: Original and downsampled signal, for 0.4 seconds taken from recording b0001, night 2 (amplifier lowpass filter: 85 Hz)

5.1.2 Bandpass filtering

The effect of removing low-frequency trends can be seen well in the example given in Figure 13. As for the downsampling, we take this random sample as evidence that the chosen implementation works as designed and approve validation of this preprocessing step.

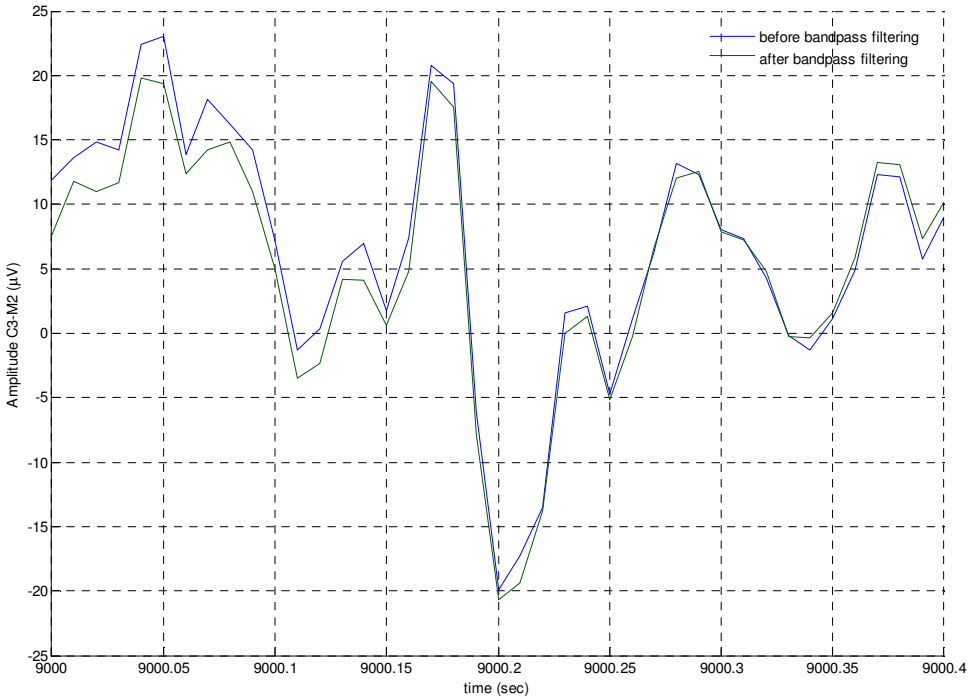

Figure 13: Effect of bandpass filtering, for 0.4 seconds taken from recording b0001, night 2

5.1.3 Autoregressive model

As the AR coefficients are the foundation of the modeling steps to come, a detailed inspection is appropriate. In this section, the goodness of fit, the spectra associated with AR models, and the distribution of the AR vectors are investigated.

In Figure 14, the distribution of the forward prediction error is shown for three segments, and the sample distribution is compared with a zero-mean normal distribution with estimated white noise variance $\tilde{\sigma}^2$ by a quantile-quantile (QQ) plot, that is capable of showing agreement with the expected distribution and gives clues about the type of deviation if the distributions are distinct. For the wake and the s4 examples, agreement with the expected distribution is very high, indicated by the QQ plot approximating a straight line. For s2 the residuals seem to follow a normal distribution, but with lower variance. That means the expected variance is too high, triggered by outliers with high forward prediction error.

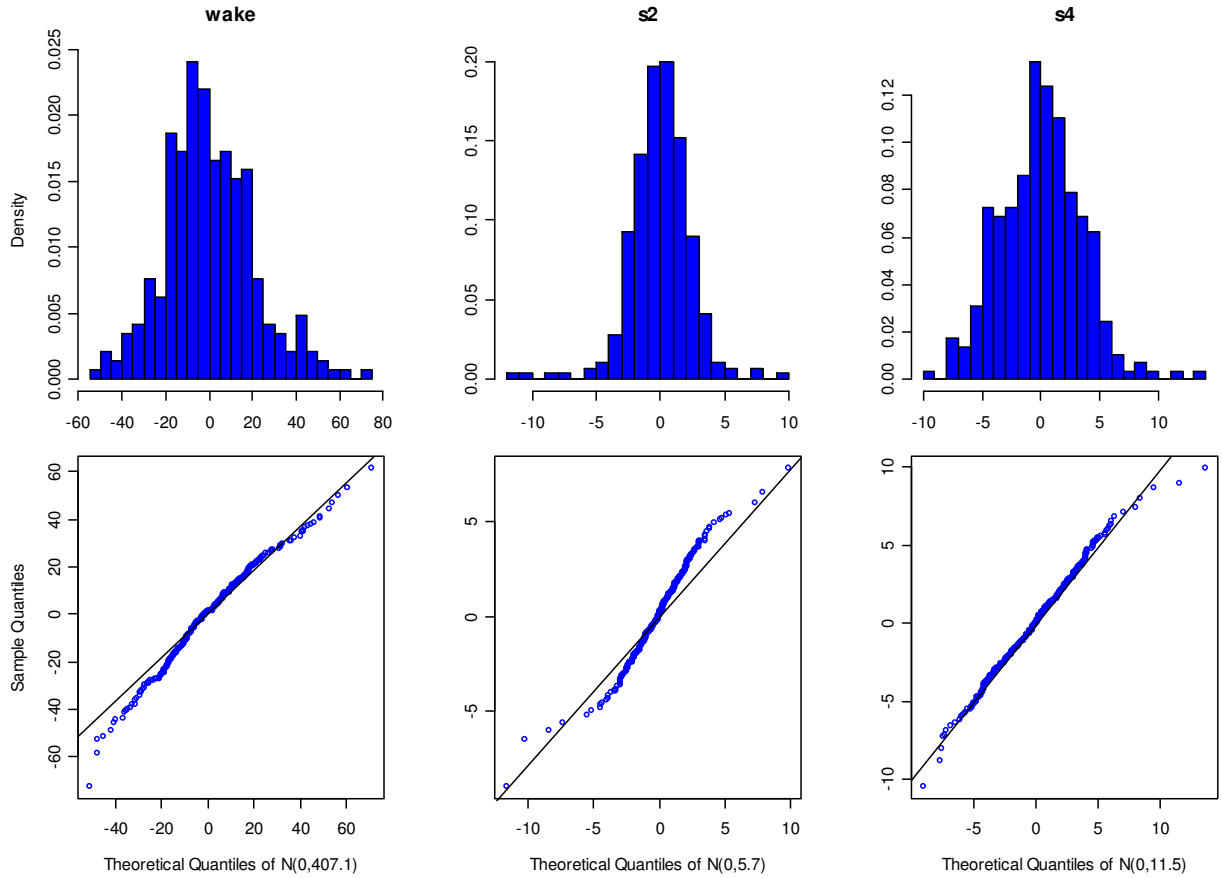

Figure 14: Distributions and quantile-quantile plots of the residuals of the AR(10) models fitted on single segments of recording b0001, night 2

In Figure 15, the spectral estimates for single segments, taken from the same subject at different sleep stages, are shown. As a consequence of the conjugated poles of an AR(P) model, the number of spectral peaks equals $P/2$ for even P, i.e. five peaks for the AR(10) model. The model order of ten seems to be a good choice, as the AR(10) spectrum is smoothed out nicely, while obvious peaks are still existent. As expected, the power in the beta band is a lot higher for wake than in the sleep segments. The slow wave sleep (s4) manifests itself with high power in the low frequency band.

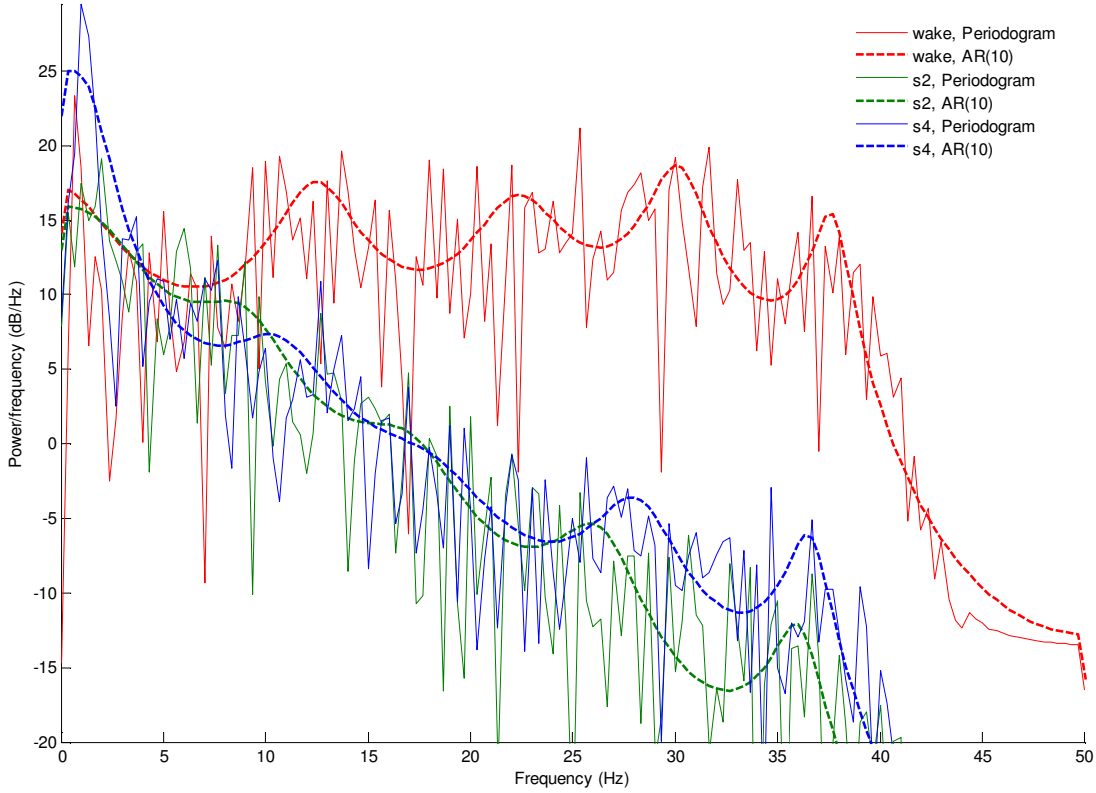

Figure 15: Spectrum of one segment each for states wake, s2, and s4 from recording b0001, night 2 (same segments as used in Figure 14). The periodogram was computed by the discrete Fourier transform without windowing; the AR coefficients were estimated by the Burg algorithm.

Inspection of the mean power spectrum, grouped by the sleep stage as scored by R&K rules, gives an impression of the discriminative power of the AR(10) representation. For the mean power spectrum from one recording as shown in Figure 16, a number of remarks are to be made:

- Wake and s1 have very similar power spectrum, with more power in the high frequency band than s2, s3, or s4.
- Deep sleep (s3 and s4) is characterized by high power in the low-frequency band.
- The mean spectrums for s3 and s4 differ just slightly in the delta-band, but s4 has less power in the high frequency band.
- As expected, the power in high frequencies for s4 is below those for s2. For the segments taken from a single recording in Figure 15, s4 epochs had slightly higher power in the high frequency band than the s2 epochs.

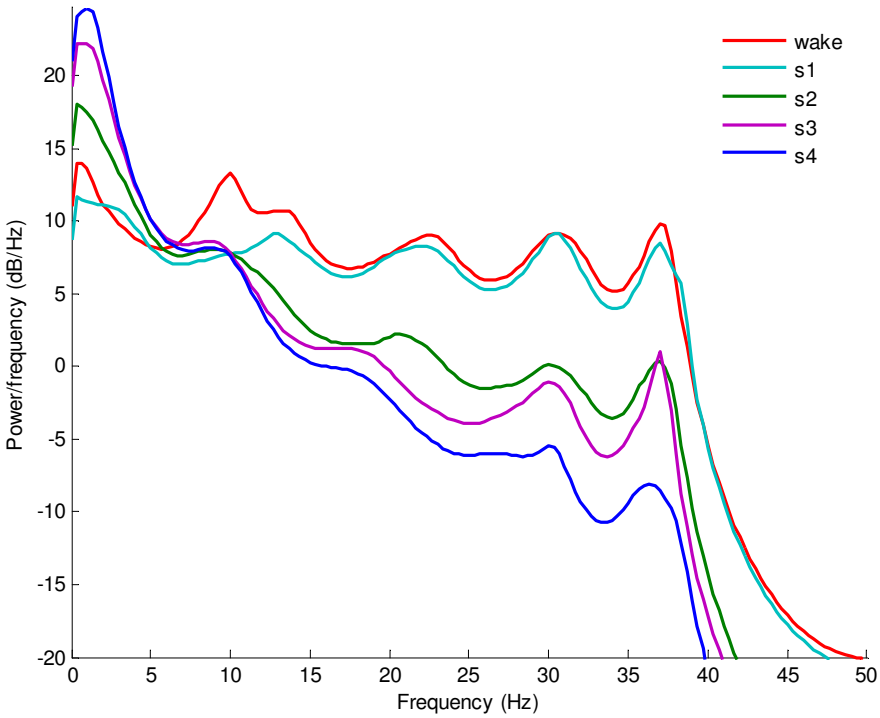

Figure 16: Mean AR(10) power spectrum of recording b0001, night 2. The grouping was done based on R&K labels.

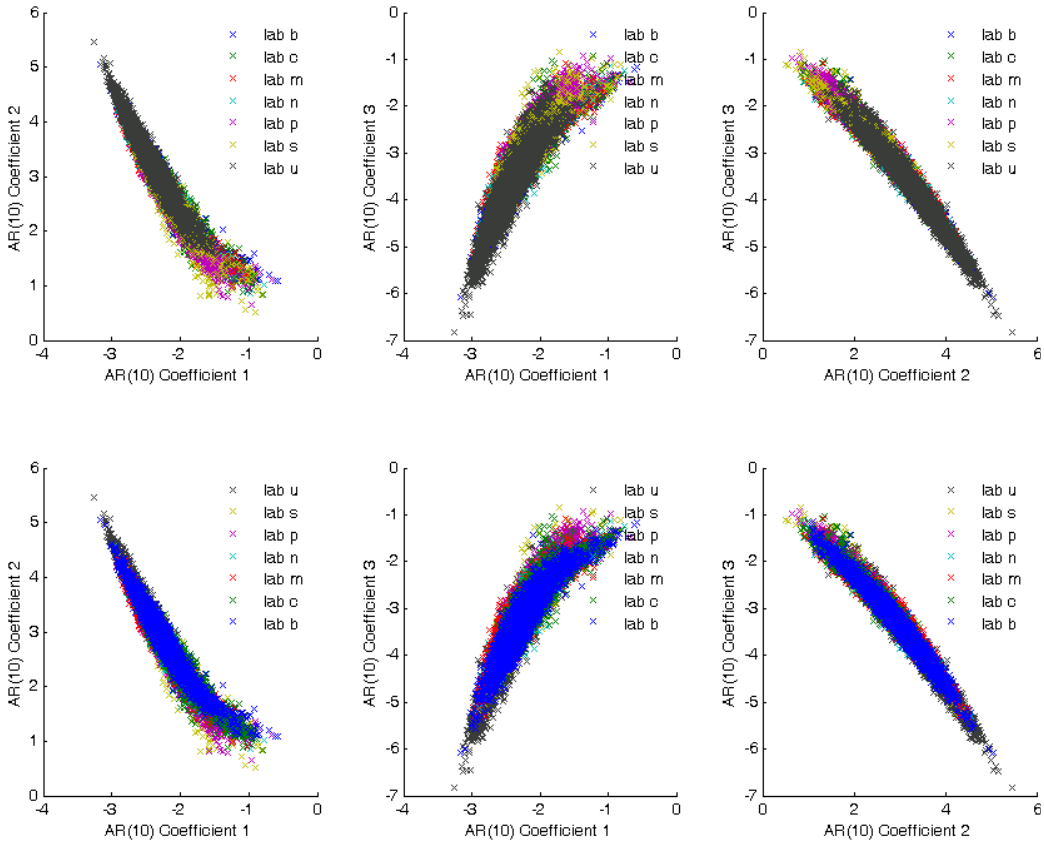

Figure 17: xy-Plot of pairs of the first three AR(10) coefficients for segments with R&K label s2. In the bottom row, the data is plotted in reversed ordering of labs to get a better impression of the overlapping regions.

The AR vectors are visualized in Figure 17 by plots of pairs of the first three AR(10) coefficients, separated by the lab where the recording was taken. The distributions appear as homogeneous clouds overlapping heavily for different labs. There are some differences in the variance of the distributions, e.g. for lab "u" the variance is slightly lower than for lab "b". The deviations are not seen to be crucial, and systematical investigation of between-lab differences is beyond the scope of this work.

To gain further insight into the structure of the data, dimensional reduction using principal component analysis (PCA) is done, see e.g. (Bishop 1996) for a review of PCA. When using all training data, more than 95 % of the variance in the data is explained by the first principal component (see Figure 18), which is astonishing. Consequently, the distribution of the data projected onto the first principal component (PC) is further explored. The probability density estimates are computed using a kernel smoothing method based on a normal kernel function.

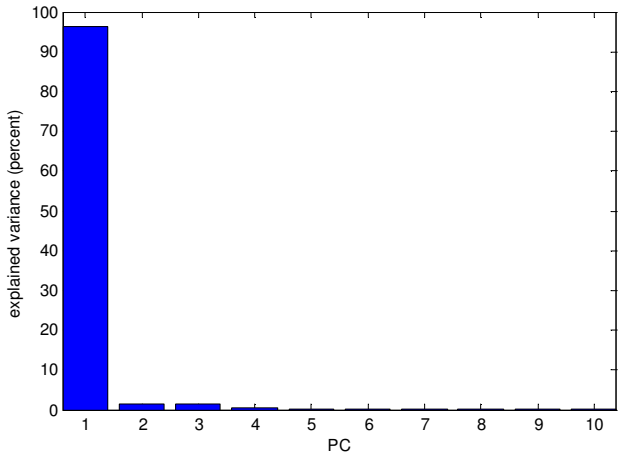

Figure 18: Part of variance of the AR(10) coefficients explained by individual principal components. All recordings from the training set have been used.

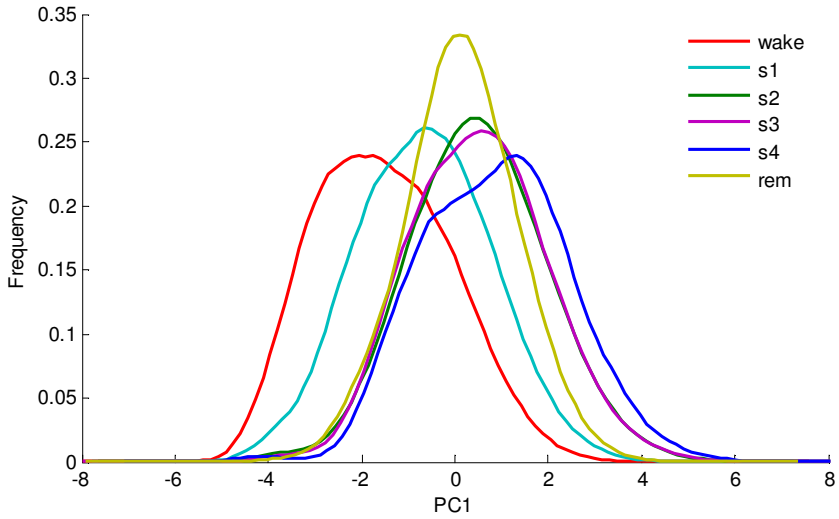

Figure 19: Distribution of the first principal component of the AR(10) coefficients for all recordings from the training set.

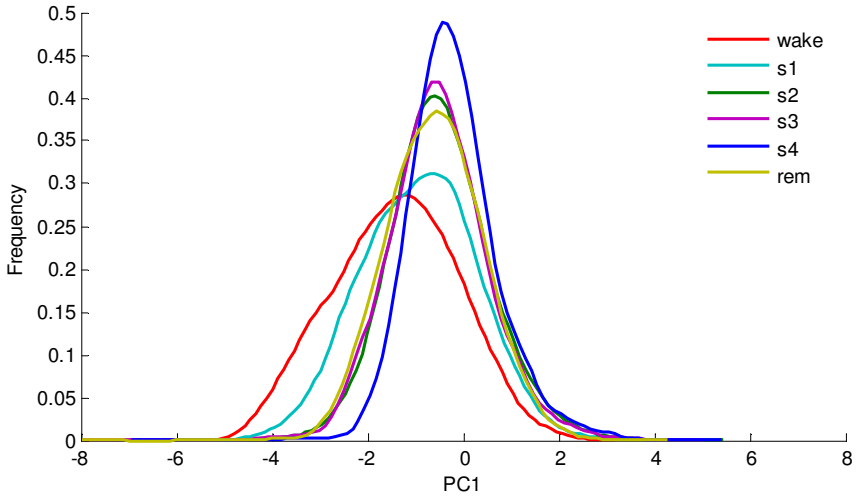

Figure 20: Distribution of the first principal component of the AR(*10*) coefficients for all recordings of lab "n" assigned to the training set.

From Figure 19 it is evident that there is some correlation between consciousness level and depth of sleep. Notably, segments from s2 and s3 are indistinguishable on the first PC, and REM resides in the midst of s1 and s2/s3.

It seems to be a contradiction to have most of the variation in the AR(10) coefficients captured by the first principal component, but to be inable to discriminate between several sleep stages. There are several possible explanations:

- Principal components explaining just a small part of the overall variance can still be the ones with high discriminative power.
- Discrimination between sleep stages might not be possible at all.
- Discrimination between sleep stages might not be possible for individual segments.

The last possibility requires further explanation. It could be possible, that variance is high for consecutive AR vectors, because of the high noise in the EEG signal. When considering a sequence of AR vectors, though, it might be seen that their distributions differ for sequences taken from different sleep stages. This idea is inherent in HMMs, but not in our hierarchical mixture. We hypothesize that the first explanation is true to a considerable extent and the discriminative power is not mainly encoded in the first principal component.

When regarding only recordings from the lab with code "n", see Figure 20, the distributions for different sleep stages are even more overlapping. Furthermore, for s2, s3, and s4 the range is altered with maximum values around 4 compared to 6 for the whole training data. That does imply that it is important to use data from more than one lab to cover a broad range of the possible data space and thus allow for good generalization properties.

It should be noted that although discrimination is an interesting task, this is not the ultimate goal of the models to follow. Therefore, the AR(10) coefficients might still be a useful representation of the data, even when lacking the ability to distinguish reliably between sleep stages.

In Figure 21, the distribution of the estimated white noise variance for all wake segments according to R&K labels for one recording is shown. It is observed that segments with extremely low $\tilde{\sigma}$ estimates for the AR(10) model are likely to result from artifacts. An example is shown in Figure 22. Exceptionally high $\tilde{\sigma}$ estimates, on the other hand, originate from perfectly valid EEG signals, at least for the randomly selected examples inspected (not shown). That indicates that low estimates could be used for artifact recognition to some extent, but this is not further addressed in this thesis as resources are limited, and furthermore the R&K labels for the data used are available and can be used to remove epochs labeled "undefined" or "movement".

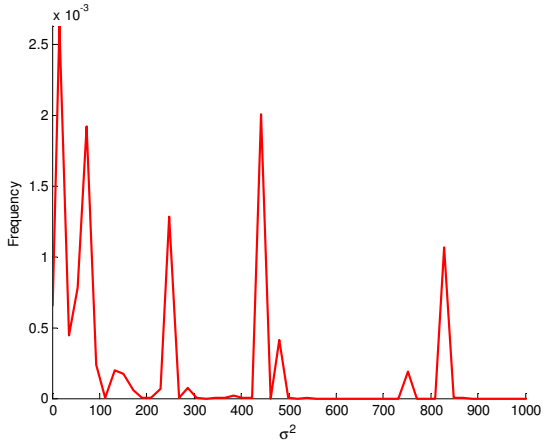

Figure 21: Distribution of the estimated white noise variance for AR(10) models considering all wake periods from recording b0001, night 2.

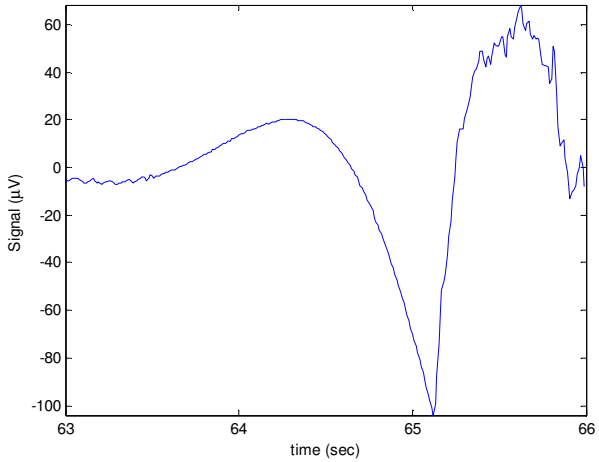

Figure 22: Artifact at the beginning of night 2 from recording b0001. The segment has R&K label wake.

5.2 Model training

For the three continuous models of sleep, the number of groups is fixed in advance and the group priors are set according to the distribution in the training data. Inference about the following parameters has to be done in the model training:

- Model family of the GMMs,
- Number of components for each class-conditional GMM,
- Parameters of the class-conditional GMMs for a fixed number of components, i.e. mixing proportions, means, and covariance matrices.

In the following section, we will show and discuss the results regarding selection of the model family and the number of components. Estimation of the parameters inside the class-conditional GMMs with known number of components is not addressed separately as this step is not crucial. Results from the reshuffling steps are shown in the last subsection.

5.2.1 Selection of the GMM model family

Regarding selection of the model family, comparison of BIC for the 28 models (see chapter 2.5) is given in Figure 23, based on two recordings, with the number of Gaussian components varying from one to six. All epochs of the recordings regardless of their R&K labels were used, in favor of simpler models, as this will increase the homogeneity of the distribution of the AR vectors. The groups with fairly big differences of BIC correspond to the three model families. As the number of components increases, the order of the models is mostly conserved. In Figure 24, results are shown with model labels for five Gaussian components. According to BIC, the models with free orientation and shape (containing "Lk" in the label) are favorable, and the selected model is the most general one with all parameters allowed to vary (pk-Lk-Ck). This model type is consequently used for all class-component GMMs.

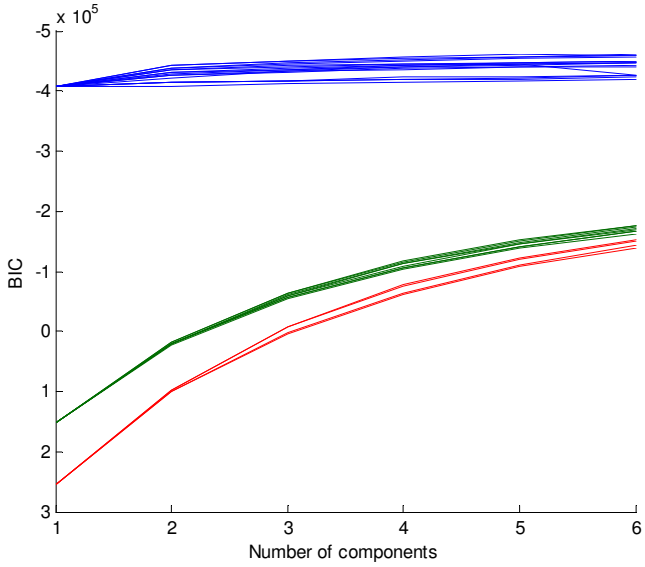

Figure 23: BIC for 28 parameterizations of GMMs with one to five components, trained on data from two whole night recordings. Blue (top): full covariance matrices; green (middle): diagonal covariance matrices; red (bottom): spherical covariance matrices.

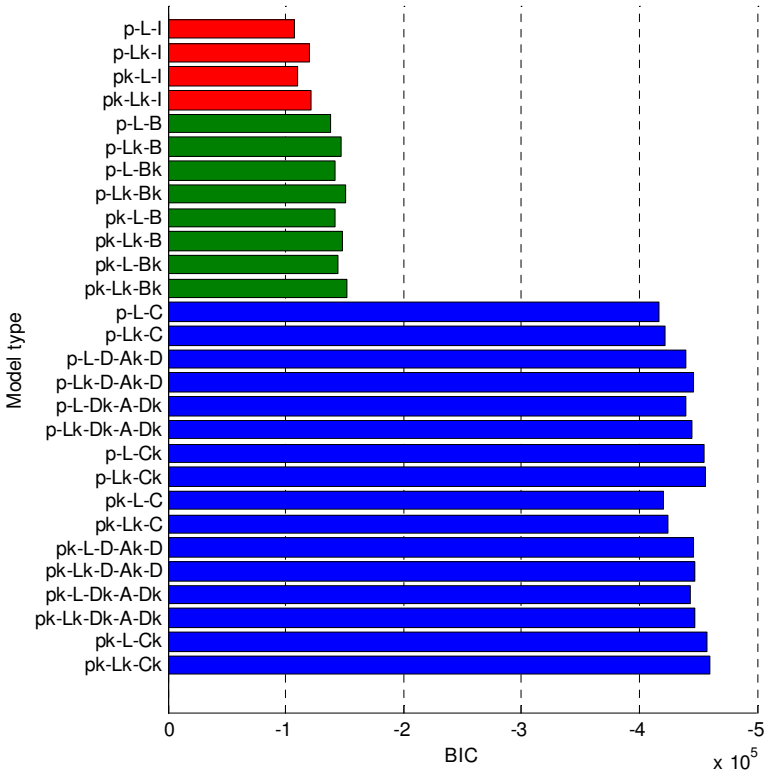

Figure 24: BIC for 28 parameterizations of GMMs with five components trained on data from two whole night recordings. Blue: full covariance matrices; green: diagonal covariance matrices; red: spherical covariance matrices.

5.2.2 Number of components for the class-conditional GMMs

The numbers of components for the class-conditional GMMs are determined by BIC in the first step of the model training. In the figures following, the value of BIC and the part of BIC without considering the penalty term, i.e. $2 \times$ log-likelihood, are shown for an increasing number of components. When the BIC value is decreasing or not does not exceed the preceding BIC by more than 10, the model search is accomplished.

In Figure 25, the BIC values for training on 79 recordings for the model of sleep cornerstones are shown. For wake, all GMMs having less than four components result in log-likelihood of minus infinity on the training data. This may happen as in the training step such outliers are excluded, but to ensure comparability between different numbers of components, they are still included in calculation of the log-likelihood. The numbers of components as determined by BIC model selection are 7 for wake, 14 for NREM sleep, and 4 for REM sleep. The influence of the BIC penalty term is small for all groups, which is highlighted in the lower part of Figure 25. Interestingly, for the NREM group the log-likelihood of the data is lower for 15 components than for 14. The likelihood theoretically increases with the number of components, but due to local maxima, this may happen. In our training regime, occurrence of such effects could be reduced by increasing the number of initializations, the number of iterations before selecting one of those settings, and finally the number of EM steps applied to the selected model. Increasing the number of iterations is computationally expensive, especially for big numbers of components. When plotting the log-likelihood after each EM step (Figure 26), we observe very slow increase in the log-likelihood after 150 iterations; thus 400 iterations does not seem to be the limiting factor.

As the BIC value for the GMM model of NREM is obviously flattening for 10 and more components, this decrease in the likelihood is not crucial in that case, and 14 components for NREM is seen as good choice.

For the NREM stages of the continuous model of the R&K sleep stages, we show the BIC values in Figure 27. The selected numbers of components are 6 for s1, 21 for s2, 9 for s3, and 12 for s4 sleep. Compared to 14 components for the NREM group in the model of sleep cornerstones, this strongly indicates that the distribution of the AR vectors is more complex when considering single NREM sleep stages, especially for s2 where a lot of training data is available, but also for s4 with considerably less training samples. For wake, the number of components is 11, and for NREM sleep we have 10 centers.

Without showing details, we give the number of components for the model of sleep substructure: 12 for wake, 11 for s1, 20 for s2, and 14 for deep sleep.

Table 10 gives an overview of the number of selected components for all three models. The biggest differences are observed for the groups REM and s1, indicating relatively high variance in the model training and model selection process. This variability could be reduced by tuning of the parameter settings, e.g. the number of initializations, or by utilizing different initialization strategies.

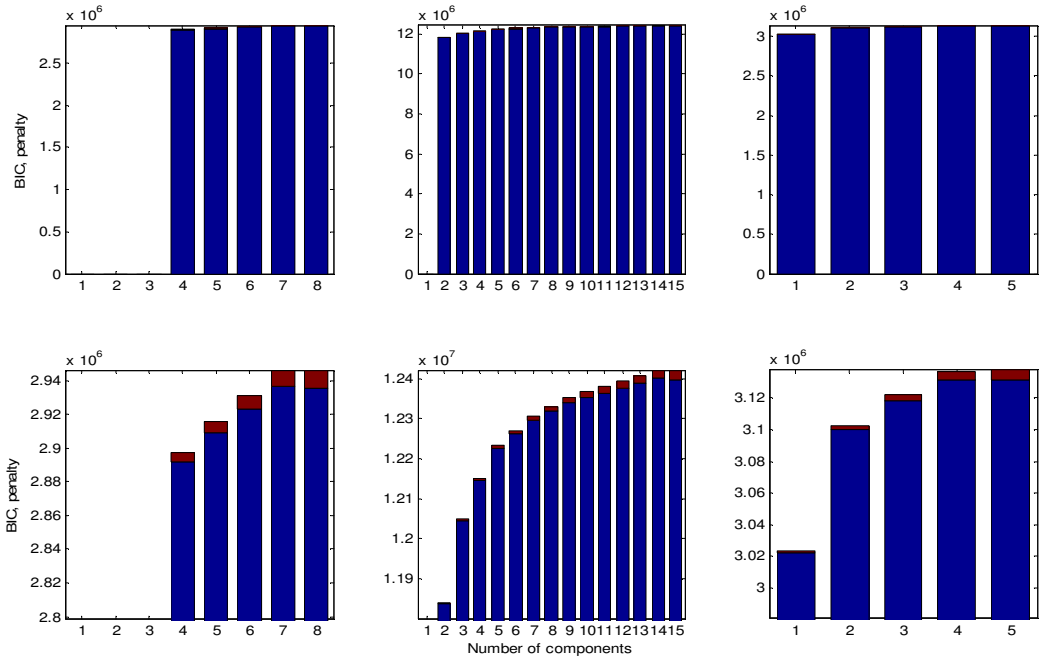

Figure 25: BIC (blue) and 2 × log-likelihood (sum of blue and red) for the model of sleep cornerstones, 79 recordings. From left to right: groups wake, NREM, REM. Bottom: zoom on the upper part of BIC.

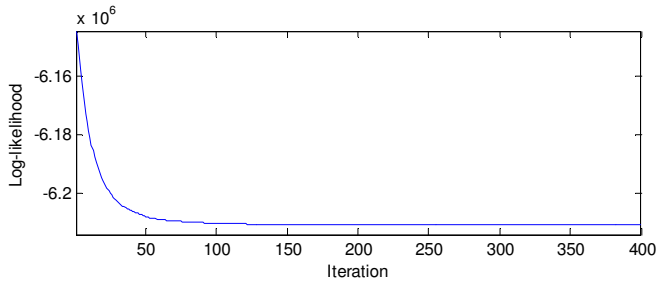

Figure 26: Course of negative log-likelihood for group wake, for 400 iterations of EM in training step 1.

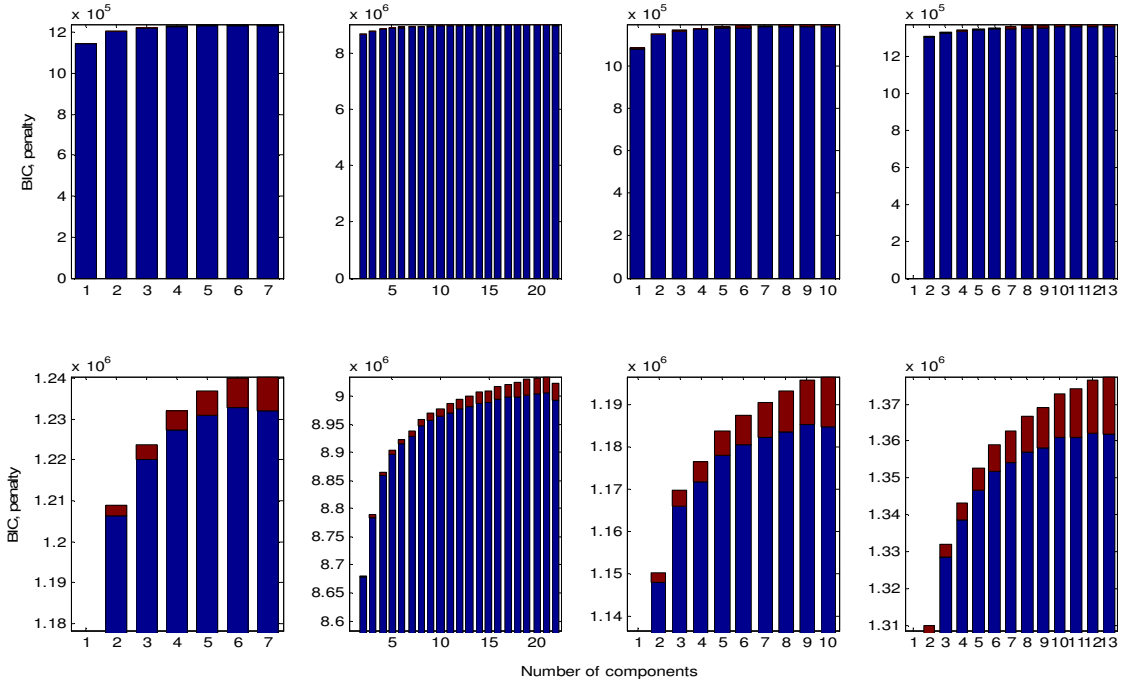

Figure 27: BIC (blue) and 2 × log-likelihood (sum of blue and red) for the NREM groups of the continuous model of R&K sleep stages, 79 recordings. From left to right: groups s1, s2, s3, s4 sleep.

Table 10: Number of components determined by BIC model selection for all continuous models

Sleep stage	Sleep cornerstones	Continuous R&K	Sleep substructure
wake	7	11	12
s1	14	6	11
s2		21	20
s3		9	14
s4		12	
REM	4	10	-

5.2.3 Reshuffling

In the reshuffling step, the structure of the model is kept, but parameters within the class-conditional GMMs are allowed to adapt to the data. In Figure 28, we show the distributions of the Gaussian components and the class-conditional densities derived thereof on the first principal component for the model of sleep cornerstones. The blue lines are assigned to class wake, where components are more clearly separated after the reshuffling process. The NREM and REM groups (red and blue respectively), on the other hand, overlap more after the reshuffling. When taking a closer look at the mixing proportions within the REM group, Figure 29, we observe that most dynamics are within the first 200 iterations. In the first few iterations of the EM algorithm, large changes are indicated by the rapid drop of the negative log-likelihood. Problems with singular covariance matrices were not encountered, thus making obsolete the alternative GMM model type with fixed mixing proportions mentioned in the model description.

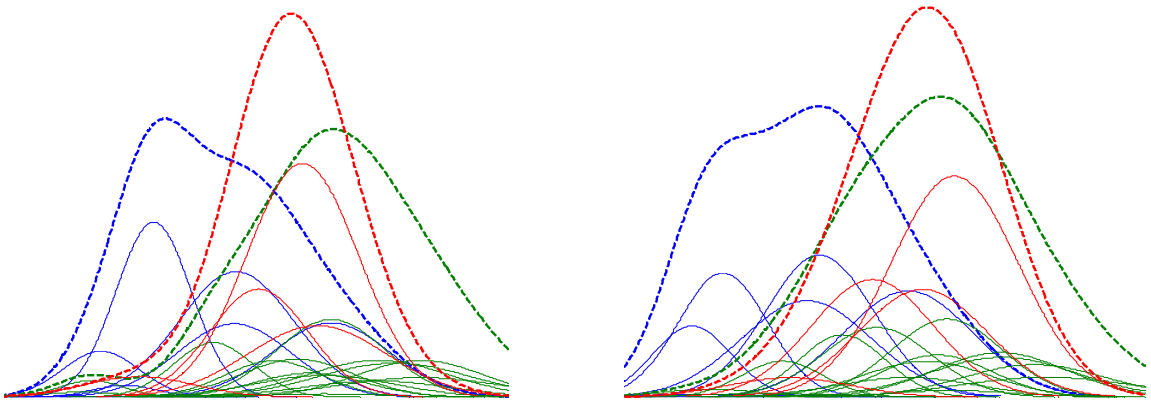

Figure 28: Impact of the reshuffling procedure on the model parameters, for the model of sleep cornerstones. The distribution of the first principal component is shown. Blue: wake, green: NREM, red: REM.

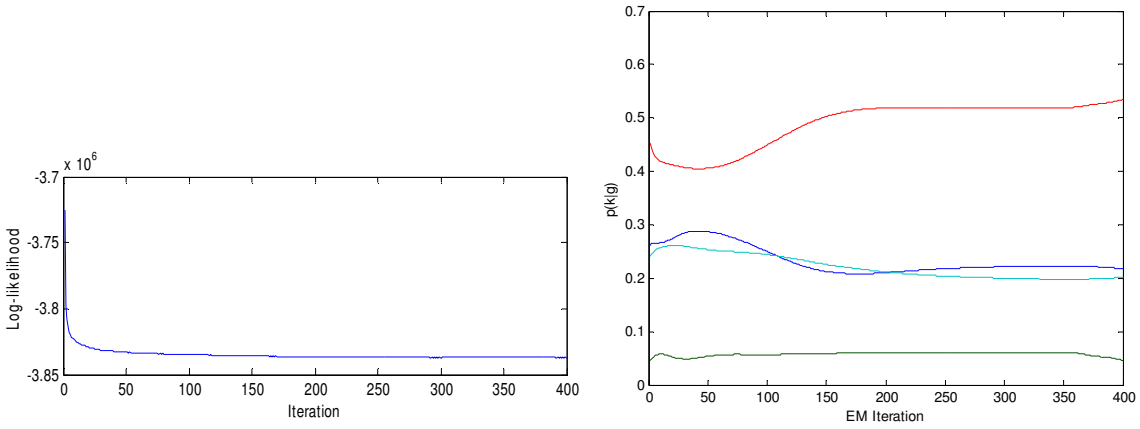

Figure 29: Development of negative log-likelihood (left) and mixing proportions (right) within the group REM, consisting of four Gaussian components, in the model of sleep cornerstone in the reshuffling step.

The impact of the reshuffling process on the model output was checked by inspection of the continuous sleep profiles. We show an example in Figure 30 using the model of sleep substructure. Some differences can be observed in the periods scored as s3 by R&K rules, and also in the awakening followed by light sleep (after min 140). Besides those differences, the characteristics are quite similar, with continuous transitions between sleep stages, and neither of the two models misses major state changes.

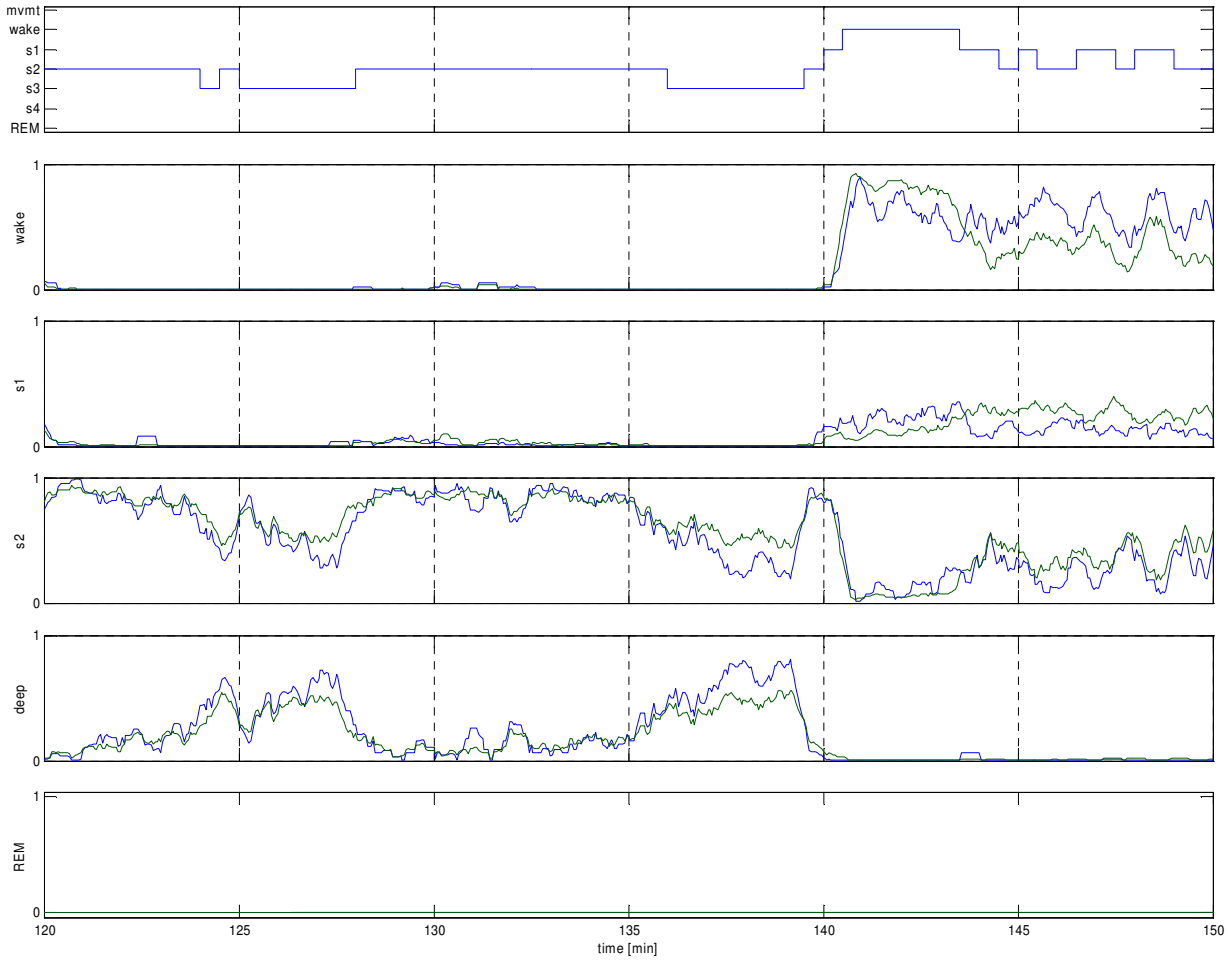

Figure 30: Sleep profile for the model of sleep substructure, recording b0002, night 1. Green traces: after training step 1. Blue trace: after reshuffling. The deep sleep before min 140 is more pronounced in the reshuffled model, and minor differences are observed in the wake/light sleep after min 140.

5.3 Visual validation and exploration

The visualization of the continuous sleep profiles combined with comparison to the R&K scoring was used as a major point of reference throughout this thesis. Although thorough interpretation of the continuous curves is not a trivial task, it is relatively easy to check whether scorings from R&K are reflected in the plots, e.g. as continuous transition from light to deep sleep in the model of sleep substructure when R&K scorings change from s2 to s4.

The continuous plots can be viewed at different levels, starting from the whole night at a glance down to parts of a second. There are inherent high-frequency changes in the probability traces, originating from the fact that the EEG has low signal to noise ratio, which is greatly but not completely reduced by using 3

second windows and the AR representation. For concise visualization of longer periods, smoothing of the traces is beneficial. We use causal moving average filters with manually adjustable filter length. To display parts of 60 minutes or more, filtering of 30 sec to 5 min is useful to pronounce trends. When using non-overlapping windows, that corresponds to an averaging of 10 to 100 consecutive samples. When using sliding windows, the highest time resolution occurring naturally is 10 ms, originating from the unified EEG frequency of 100 Hz. As reconstruction of the EEG signal at arbitrary points in time is possible, i.e. also between two sampling points, even higher time resolution could be achieved, but we limit ourselves to 10 ms. With overlapping windows, the same part of the signal is used more than once for AR parameter estimation, i.e. the AR coefficients are not independent any more. This fact is reflected in smooth transitions at this level even without filtering.

We show a few probability traces, where in the first row the R&K scoring is shown as a point of reference i.e. that is done for validation purposes only. Below the R&K scoring, there is one axis per group as defined by the sleep model, e.g. wake, NREM and REM for the model of sleep cornerstones. Those values add up to one at every point in time. When using smoothed probability traces, that is not true for the transient time of the filter at the beginning of the recording.
We start with a whole night plot for the model of sleep cornerstones in Figure 31. At the beginning of the recording, the R&K progresses from wake to s1 and further to s2. This is reflected in the continuous profile by a continuous transition from wake (close to 1) to NREM. The periods scored as wake appear as sharp peaks in the wake trace. There is also agreement on deep sleep, e.g. with high probability of NREM at min 90 and min 130. The epochs scored as REM by R&K rules are indicated by a heightened level of REM probability, but there is still a high level of NREM at the same time. This is not surprising, as information from EEG alone is problematic for distinction between those states.

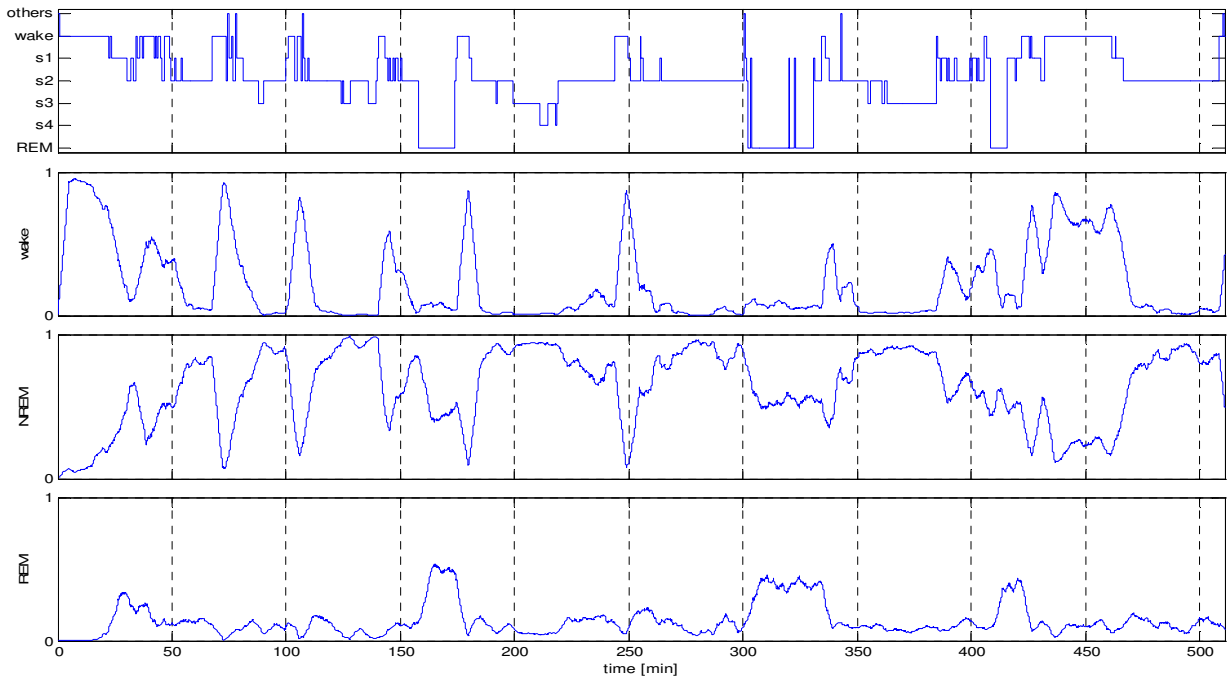

Figure 31: Continuous sleep profile for subject b0002, night 1, model of sleep cornerstones. Posterior values are smoothed by a moving average filter of length 30 sec.

For the continuous model of R&K sleep stages, we show one hour starting 2 hours after "lights off" in Figure 32. As an overall impression, the high number of continuous traces makes is hard to get an overview at a glance. From min 120 to 140, probability levels of s3 and s4 are elevated for periods scored as s3 by R&K, but probabilities for s2 is still highest. Moreover, for s3 and s4 according to R&K, the continuous s2 is at the highest level. Furthermore, changes in s3 and s4 emerge synchronously, indicating that a model merging those two states into "deep sleep" could be beneficial. The arousals at min 140 and 175 appear as transitions from s2 and s1 to wake, but the role of s1 is not that clearly visible in the continuous plot. As in the model of sleep cornerstones, the REM period is reflected by a higher level of REM probability.

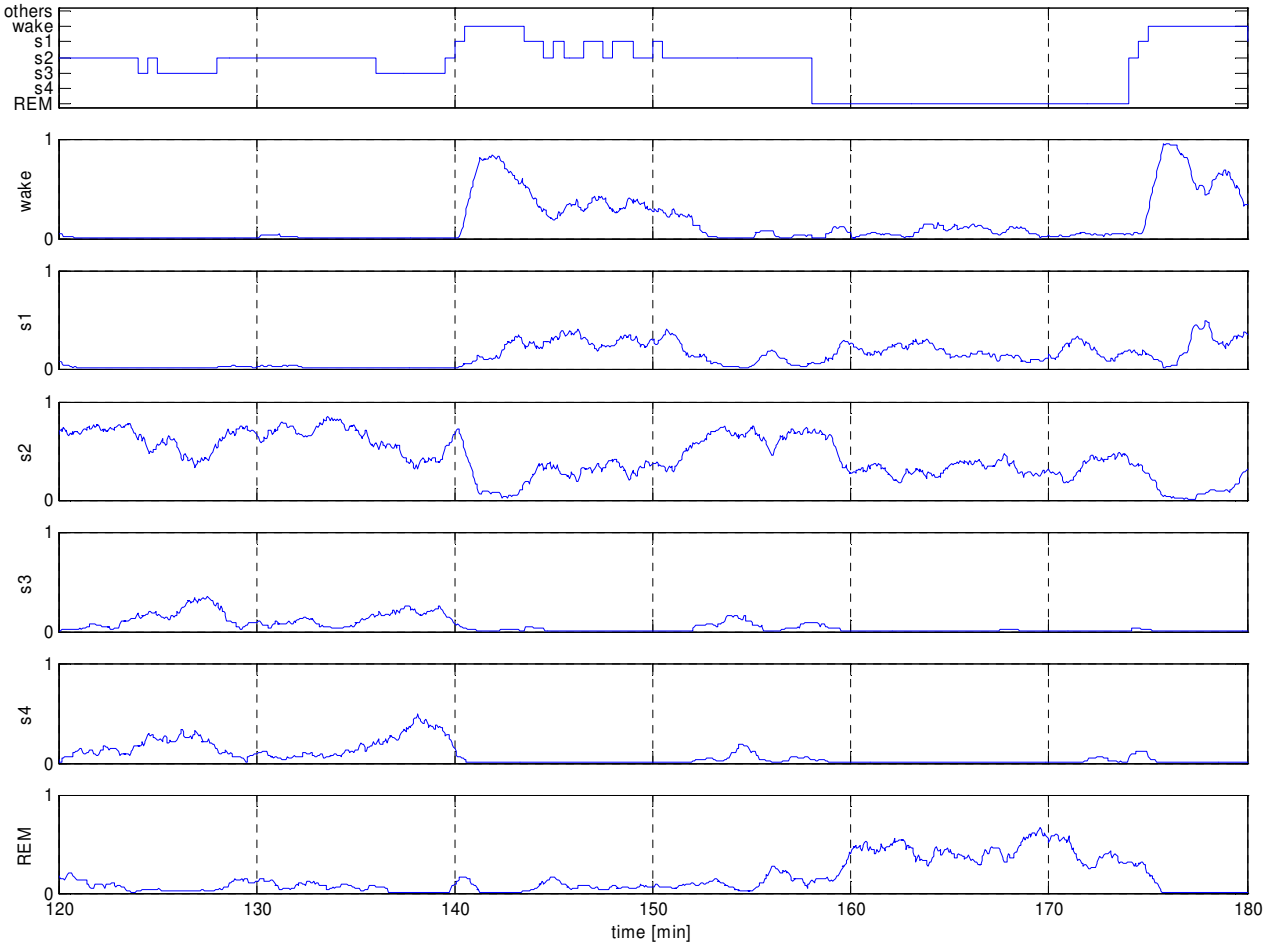

Figure 32: Continuous sleep profile for subject b0001, night 2, min 200 – 260, continuous model of R&K sleep stages. Posterior values are smoothed by a moving average filter of length 30 sec.

For the model of sleep substructure, we start by showing a continuous profile of a one hour section that will be followed by plots on a finer time scale. In Figure 33, the discrete REM state is eye-catching. As we use the R&K labels as the external classifier, the REM trace equals one when the epoch is labeled as REM in the R&K sleep profile. The transition from light sleep (s2), to deep sleep is more clearly visible than in the continuous model of R&K sleep stages. The awakenings at min 140 and 175 are nicely picked up by probabilities close to one for state wake.

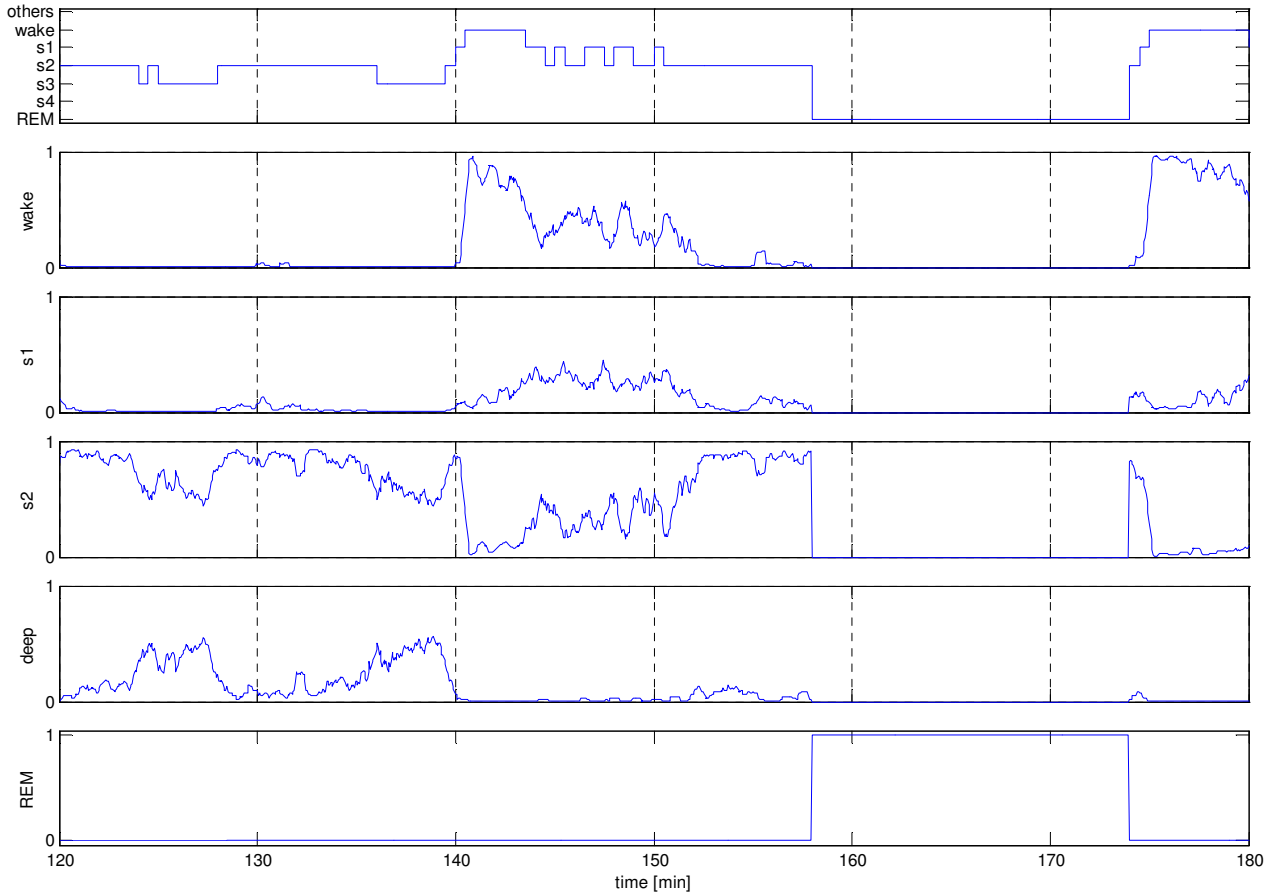

Figure 33: Continuous sleep profile for subject b0001, night 2, min 200 – 260, continuous model of sleep substructure. Posterior values are smoothed by a moving average filter of length 60 sec.

Unfiltered probability traces for the model of sleep substructure are shown in Figure 34 for a period where R&K labels pass through from s3 to wake. Compared to the same period shown in Figure 33, the continuous transition between the stages can not be seen as clearly because of the highly frequent changes. When zooming further in, as done in Figure 35, short term state changes can be traced. According to the model of sleep substructure, there is a short period of deep sleep embedded in s2 sleep, and the transition from s2 (probability close to 1) to deep sleep (probability above 0.5) is accomplished in 0.2 seconds in this case. Similar exploration of transitions can be made for other groups, e.g. to discover short periods with high probability of wake, and of course for the other two models.

The clinical relevance of such observations or events is unknown at this time, but indeed it would be very promising to derive features based on hypothesis regarding events and transitions on very high time resolution.

Based on the observations made above, it is assumedly beneficial to use the probability traces on a very high time resolution for feature calculation, e.g. also for the features described in section 4.2, that were derived from non-overlapping windows in this work.

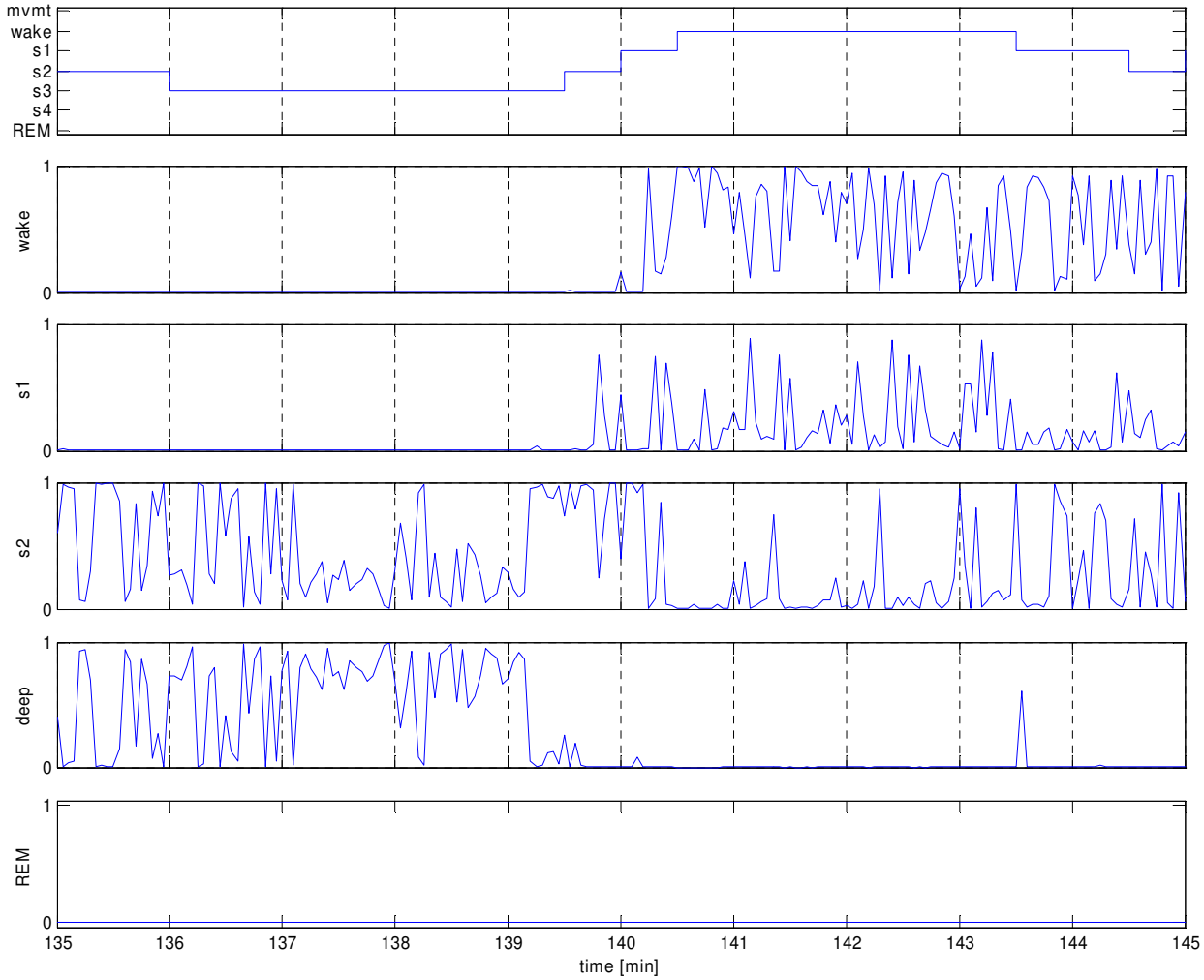

Figure 34: Continuous sleep profile for subject b0001, night 2, min 135-145, model of sleep substructure. Posterior values are not smoothed.

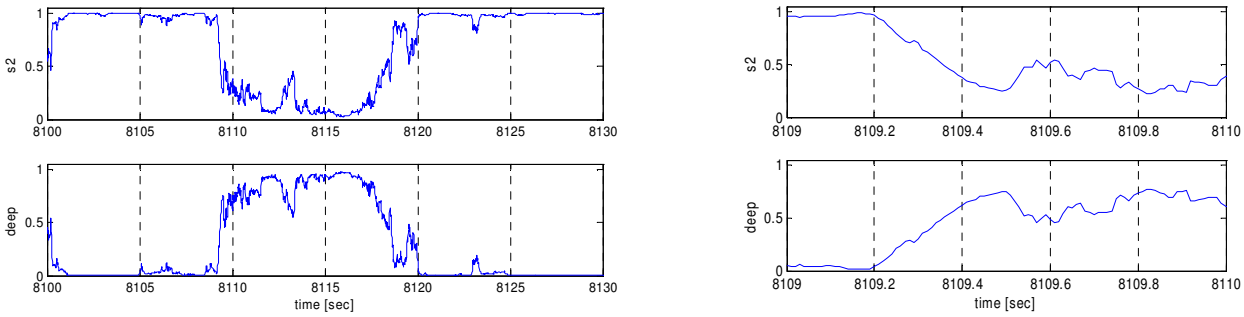

Figure 35: Traces of s2 and deep sleep for the model of sleep substructure with high time resolution, for subject b0001, night 2. Left: 30 seconds starting at min 135. Right: one second, starting at min 135:09. The epoch is scored as s2 according to R&K. In the continuous profile, rapid transition to deep sleep is observed, and deep sleep is maintained for a few seconds at a high probability level.

5.4 Classification results compared to R&K scoring

We have compared the classification of 3-second epochs from the continuous models with the R&K labels from Somnolyzer 24 × 7 (Anderer et al. 2005) for all 85 subjects of the test set. The classification was done without prior smoothing of the posterior values. As the R&K scoring is regarded as the golden standard, we will use the word *misclassified* for periods where the classification differs, even though it is possible that the classification from the continuous traces reflect the state of sleep more adequately.

As we have seen in the visual inspection of the continuous profiles (section 5.3), probability levels might not always be high enough for correct classification, which is especially true for the group REM in the model of sleep cornerstones and the continuous model of R&K sleep stages. The level of agreement is not thought to be crucial, as the discriminative power of our model is not the main goal.

For each of the three continuous models, we show the classification results in the form of a confusion matrix and as mosaic plots based on the percentages of classification for the group of R&K stages. In the form of a confusion matrix, both the absolute numbers and the percentages are given.

For the model of sleep cornerstones (Table 11), agreement is highest for epochs scored as s1, s2, s3, or s4 according to R&K rules (84 %), and lowest for REM (31 %). For wake, agreement is around two thirds (64 %). Misclassified REM periods are assigned to NREM to a large extent. NREM and REM periods are seldom classified as wake, but one quarter of the wake periods are classified as NREM. Epochs scored as movement and undefined epochs were not explicitly modeled and are mainly classified as wake.

Table 11: Confusion matrix for the model of sleep cornerstones

		Classification		
		wake	NREM	REM
R&K	wake	**0.68**	0.25	0.07
	NREM	0.06	**0.84**	0.10
	REM	0.06	0.63	**0.31**
	movement/undef	0.57	0.26	0.17

As expected, the agreement of wake and NREM periods for the continuous model of R&K sleep stages (Table 12) is about the same as for the model of sleep cornerstones. It can be seen that most of the misclassified wake epochs are assigned to s1 or s2 by the continuous model. The non-agreeing REM periods are to a large extent attracted by s2. The non-agreeing wake periods are classified to the same extent as s1 or s2 and less often as REM. In the model of sleep cornerstones we have observed high agreement on NREM stages; now we can have a look at the sleep stages s1 to s4 inside NREM.
A majority of the epochs scored as s1, s2, or s3 according to R&K rules are covered by s2 in the continuous model. The s4 epochs have agreement of 45 %, and more of the misclassified s4 epochs are assigned to s2 rather than to s3. Epochs scored as movement and undefined epochs were not explicitly modeled and are mainly classified as wake. To summarize, for the continuous model of R&K sleep

stages: wake is to some extent mixed with s1 and s2, within the NREM periods s2 is dominant where just the class s4 can be separated reasonably well. REM is to a large extent mixed up with s2 and to a lower degree with s1.

Table 12: Confusion matrix for the continuous model of R&K sleep stages

		Classification					
		wake	s1	s2	s3	s4	REM
R&K	wake	**0.69**	0.11	0.12	0.01	0.01	0.06
	s1	0.29	**0.14**	0.39	0.01	0.00	0.18
	s2	0.04	0.03	**0.77**	0.03	0.05	0.08
	s3	0.02	0.01	0.52	**0.17**	0.23	0.05
	s4	0.00	0.00	0.29	0.14	**0.53**	0.04
	REM	0.06	0.06	0.61	0.01	0.00	**0.25**
	movement/undef	0.48	0.11	0.14	0.13	0.04	0.10

Classification results from the model of sleep substructure (Table 13) show high agreement on wake and s2 periods. Inside the NREM stages, about half of s1 epochs are classified as s2 and one third as wake. Deep sleep epochs (R&K labels s3 and s4) are not confused with wake or s1, but one third is classified as s2. The agreement on REM is 100 % as the external classifier is identical to the R&K labels in our case; the same is true for periods labeled as movement and for undefined epochs.

Table 13: Confusion matrix for the model of sleep substructure

		Classification			
		wake	s1	s2	deep
R&K	wake	**0.71**	0.09	0.19	0.01
	s1	0.33	**0.17**	0.49	0.01
	s2	0.05	0.07	**0.78**	0.10
	deep (s3+s4)	0.02	0.01	0.38	**0.59**

5.5 Correlation results

We perform correlation analysis between the features derived from the sleep profiles (from R&K scoring and the continuous models, as described in chapter 4) and external criteria of sleep, see Appendix A2. We use Spearman's rank correlation test, which is a distribution-free test of independence between two variables, with H0: "there is no monotonic relation between the variables". The split of the data into a training and a testing set was kept, and correlation levels are only shown when significant correlations ($\alpha = 0.01$) are observed on both data sets. That way, false correlations induced by multiple testing are prevented and the generalization properties of the models are validated.

The analysis is done once for both nights independently and once for the differences between the two nights. The differences are especially appropriate for subjective measures like self-rating questionnaires, as that way the inter-subject variability is reduced. When both nights are tested independently, only correlations that are significant for both nights are shown.

For the results following, correlations from the features selected in the procedure described above were computed on the complete data set. For ordering by correlation level, we use the absolute value of the mean of the correlation coefficients for both-night correlations and the absolute value of the correlation coefficients for the night differences. Complete lists of observed significant correlations including coefficients for training and testing data sets are given in Appendix A3.

For the features calculated on different smoothing levels of the continuous sleep profiles, usually the unfiltered versions have higher correlations. In order to keep the tables clear, we show only the features without prior smoothing of the probability traces.

5.5.1 Correlation with Age

It is well known that sleep patterns change with age. For example, amount of deep sleep and sleep efficiency decrease with age (Kryger et al. 2005). For that reason, age can serve as a good reference variable for comparison of the features derived from the different models.

Numerous parameters correlate significantly with age (see Appendix A3, Table A3.1 for the complete list). We show correlation levels for selected features in Table 14, where features derived from the R&K scorings are compared to analogous features computed from the output of the continuous models. The highest negative correlations are observed for the number of two consecutive s4-epochs and for state changes between s2 and s4. The highest positive correlation levels are between age and the time of wakefulness within the total sleep period (wtsp). For the wake trace of the model of sleep substructure, the AUC and the AUC of 1st and 2nd derivatives also exhibit a high level of correlation (mean correlation coefficients 0.526, 0.555, 0.544).

In general, correlations with continuous features are higher than those derived from the R&K scoring. For the model of sleep cornerstones, fewer features are correlated with age, e.g. there is no corresponding correlation to tst_s1 and tst_s4. The negative correlation of state changes within the NREM stages (sc_s2_s4, sc_s4_s4), on the other hand, is nicely reflected by negative correlation to sc_nrem_nrem in the model of sleep cornerstones.

Table 14: Comparison of Spearman's rank correlation coefficient between age and features based on R&K and continuous sleep profiles

	R&K	Sleep cornerstones	Continuous R&K	Sleep substructure
eff	-0.460	-0.555	-0.559	-0.487
fw	0.460	0.533	0.542	0.526
tst_s1	0.349		0.455	0.448
tst_s4	-0.496		-0.576	-0.389
wtsp	0.519	0.561	0.562	0.525
sc_wake_wake	0.440	0.531	0.519	0.440
sc_s1_wake	0.392		0.467	0.507
sc_nrem_wake		0.438		
sc_s2_s4	-0.359		-0.614	
sc_s2_deep				-0.337
sc_s4_s4	-0.554		-0.617	
sc_deep_deep				-0.465
sc_nrem_nrem		-0.451		

5.5.2 Correlation with Subjective Sleep Quality

The Subjective Sleep Quality (SSA1, see chapter 2.2) is one subscore of the one Night differences from the Subjective Sleep Quality score have been compared with night differences from the parameters. The SSA1 reflects the sum of answers to questions on an ordinal scale, where lower values indicate better sleep quality. As we compute values of the second night (or the morning after the second night really) minus the values of the first night, negative differences indicate an improvement in subjective sleep quality and therefore negative correlation is associated with better subjective sleep quality.

All significant correlations are listed in Appendix A3, Table A3.2. The comparison of features from the different models is given in Table 15. Parameters associated with wakefulness show positive correlation coefficients, i.e. more wakefulness is correlated with decreased subjective sleep quality, as expected. This is also observed for the number of awakenings for two continuous models, but not for the corresponding R&K feature. The negative correlation for sleep efficiency is highest for the conventional R&K-based value. The number of consecutive s2 periods is correlated with better subjective sleep quality, where the significance level is highest for the model of sleep cornerstones. The correlations to state changes between wake and s1, and between wake and NREM for the model of sleep cornerstones, are not matched by the corresponding R&K features. The percentage of s1 of the total sleep time (tst_s1) is also correlated only for the continuous models; thus information from s1 taken from the continuous model reflects a higher relationship with SSA1. The total sleep time according to R&K is negatively correlated and not matched by similar continuous features.

Table 15: Comparison of Spearman rank correlation coefficient for difference between Subjective Sleep Quality (SSA1) and features based on R&K and continuous sleep profiles

	R&K	Sleep cornerstones	Continuous R&K	Sleep substructure
eff	-0.486	-0.459	-0.432	-0.458
wtsp	0.415	0.463	0.441	0.387
fw		0.358	0.356	
tst	-0.380			
tst_s1			0.399	0.416
sc_s2_s2	-0.381		-0.418	-0.319
sc_nrem_nrem		-0.425		
sc_wake_wake	0.476	0.439		0.446
sc_wake_s1			0.381	0.425
sc_wake_nrem		0.332		
sc_s1_wake			0.383	0.430
sc_nrem_wake		0.338		

For the sum of several subjective measures (variable SSA, see chapter 2.2), significant correlations are observed for a subset of features that show significant correlations to SSA. Correlation levels are below those for SSA. That indicates that those correlations are mainly due to the impact of SSA1 on the total score. For that reason, detailed inspection is omitted at this point, and the interested reader is referred to Table A3.2 in Appendix A3.

Besides the correlations stated above for the night difference, the total score of sleep and awakening quality (SSA) is positively correlated with wake within the total sleep period (wtsp) derived from the R&K sleep profile for both nights. That is remarkable because this is the only correlation found for subjective measures not computed on night differences. The concrete correlation coefficients are 0.452, 0.303 for the training set, 0.307 and 0.327 for the test set and 0.376, 0.320 for the merged data set (the value pairs refer to night 1 and night 2 respectively).

5.5.3 Correlation with Alphabetical Cross-out Test

For the alphabetical cross-out test (Grunberger 1977), several variables are available, see Appendix A2. Significant correlations between features from the sleep profiles and two of those variables are observed, namely to the total score (AD-TS) and to the total score with false responses corrected (AD-TS-ER). Correlations to the uncorrected total score are observed for a subset of features that are also significantly correlated to the corrected score, and correlation levels are lower. For that reason we will only discuss the correlations to the error corrected variable. The correlation coefficients for both variables can be found in Table A3.5.

From Table 16 it can be seen that by far the most correlations are observed for the continuous R&K model and the model of sleep cornerstones. For the significantly correlated R&K features, the corresponding continuous features, except from the model of sleep cornerstones, also show up. The AUC of the wake trace and the first two derivatives thereof show significant correlation levels only for the

model of sleep substructure, whereas the state changes between s2/s4 and wake/s2 turn up for the continuous model of the R&K sleep stages. Sleep stage s1 is represented twice, once as the time classified as s1, once as percentage within the total sleep time (tst_s1). Put together, it can be said that there is evidence that in all models, features derived from wake are correlated with values of the alphabetical cross-out test, but correlations are not consistent across the different continuous models except wake within the total sleep period, which is just "missing" for the model of sleep cornerstones.

Table 16: Comparison of Spearman rank correlation coefficient for the Alphabetical cross-out test (false responses corrected) and the features based on R&K and continuous sleep profiles

	R&K	Sleep cornerstones	Continuous R&K	Sleep substructure
wtsp	-0.371		-0.375	-0.370
fw		-0.391		-0.404
auc_tsp_wake				-0.372
auc1_tsp_wake				-0.422
auc2_tsp_wake				-0.413
tst_s1			-0.329	
s1				-0.316
sc_s2_s4			0.411	
sc_s4_s2			0.414	
sc_wake_s2			-0.390	
sc_s2_wake			-0.389	

5.5.4 Correlation with Fine Motor activity

For the test of Fine Motor activity (Grunberger 1977), scores for the left hand, the right hand, and the total score are available. Correlations to the total score are seen for all models including features derived from the R&K sleep profile, while for the separate scores only correlations to the features based on the continuous traces are significant. Detailed results are given in Table A3.5 in Appendix A3.

In Table 17 we give an overview of the significant correlations for the different models for all three scores. There is no significant correlation for features derived from the model of sleep cornerstones and just one significant correlation each for the features calculated from the R&K hypnogram and from the continuous model of R&K sleep stages. By far the biggest number of features showing significant correlations is derived from the model of sleep substructure and is associated with wake, i.e. the number of awakenings (fw), wake within the total sleep period (wtsp), and the AUC of the wake trace and the first two derivatives thereof. Interestingly, the number of REM epochs followed by another REM epoch shows up for the R&K features and the feature from the model of sleep substructure, giving light evidence that stable REM periods have positive influence on fine motor activity.
Remarkably, the level of correlation is higher for the score of the right hand in all cases while at the same time more features correlate significantly with the total score.

Table 17: Comparison of Spearman rank correlation coefficient for Fine Motor activity test (left "l:", right "r:", total "t:") and the features based on R&K and continuous sleep profiles

	R&K	Sleep cornerstones	Continuous R&K	Sleep substructure	
sc_rem_rem	t:0.312				t: 0.309
auc_tsp_wake					t:-0.332
auc1_tsp_wake				r:-0.392	t:-0.378
auc2_tsp_wake				r:-0.383	t:-0.369
fw				l:-0.311 r:-0.370	t:-0.361
wtsp				r:-0.343	t:-0.329
sc_s2_s4			t:0.330		
sc_s4_s2			r:0.357 t:0.338		
sc_wake_s2				r:-0.352	t:-0.339
sc_s2_wake				r:-0.346	t:-0.331

5.5.5 Summary of the Correlation results

An overview of the observed significant correlations is given in Table 18, where psychometric variables can also be identified where no significant correlations were observed. For all combinations of variable and sleep-model, we give the correlation coefficients for the feature with the highest level of correlation.

For subjective sleep quality, correlations are highest for features derived from the R&K sleep profiles. For the Alphabetical Cross-out Test and the Fine Motor Activity Test, features from the continuous model of R&K sleep stages and the model of sleep substructure outperform the features based on the R&K scoring. That is especially evident for the Fine Motor Activity Test, where correlations not only for the total score but also for the left hand and the right hand are significant. Most features derived from the R&K profiles with significant correlations are strongly related to wake epochs, with the exception of REM for Fine Motor activity and features correlated to age. For the features calculated from the continuous profiles, wake is also heavily involved, but sleep stages s1, s2, and deep sleep play a more important role. From the features calculated directly from the continuous traces rather than from the classification results, especially the AUC of the 1st derivative is seen with high levels of correlations.

Table 18: Overview of significant correlations between features derived from the sleep profiles and psychometric variables. The sleep stages/groups are given in order of importance together with the highest correlation coefficients observed.

	Variable	R&K	Cornerstones	Continuous R&K	Sleep Substructure
	Age	s4, wake, eff, s1 -0.554	wake, s4, eff, NREM 0.562	s4, eff, wake, s1 -0.617	wake, eff, deep, s1 0.555
For relative values (night differences)	Pittsburgh Sleep Quality Index				
	Subjective Sleep Quality (SSA1)	eff, wake, s2 -0.486	eff, wake, NREM, s1 0.466	wake, eff, s2, s1 0.445	wake, eff, s1, s2 0.477
	Subjective Awakening Quality (SSA2)				
	Somatic Complaints (SSA3)				
	Total subj. sleep and awak. quality (SSA)	eff, wake -0.410	wake, eff, NREM 0.368	wake, s1 0.341	wake, eff 0.385
	Mood				
	Drive				
	Affectivity				
	Drowsiness				
	Well Being (morning)				
	Well Being (evening)				
	SSA, night 1 + night 2	wake 0.348			
Alphabetical Cross-out	Total score	wake -0.359		s2/s4, s1 0.414	wake -0.413
	Total score, false responses corrected	wake -0.371	wake -0.391	s2/s4, wake, s1 0.414	wake, s1 -0.422
	Errors				
	Percentage of errors				
	Variability				
	Numerical memory				
Fine Motor activity	FM-left				wake -0.321
	FM-right			s2/s4 0.357	wake, wake/s2 -0.392
	FM-total	REM 0.312		s2/s4 0.338	wake, wake/s2, REM -0.378

5.6 Classification of Patient groups

In this section we will investigate the ability to discriminate between healthy controls (HC) and patients based on the features derived from the R&K and the continuous sleep profiles. The numbers of subjects are: 16 patients with Generalized Anxiety Disorder (GAD), 51 for the group suffering from Sleep Apnea Syndrome (SAS), and 15 for the Parkinson's Disease (PD) patients. Especially for the GAD and PD groups, the results will have to be verified on a larger set of patients.

In the first step, results from binary classifiers based on all features from a model are given. In the second step, a subset of features with highest discriminative power is selected. For those subsets, discriminative power is compared and selected features might serve as clinical markers e.g. for therapy control.

The classifier used is a generalized linear model with logit as link-function from the binomial-family (Venables and Ripley 2002), as results from a non-linear classifier did not give significantly better results. The data was randomly split into five training and testing partitions (5-fold cross-validation), and results shown are the mean values of the five classification results. By varying the cutoff level, the Receiving Operating Characteristic (ROC) curve is obtained for each of the five runs. As measures of the quality of classification we use the maximum performance, i.e. the maximum of the mean of specificity and sensitivity for the different cutoff levels, and the area under the curve of ROC (AUC ROC). To ease interpretation, we will also show the mean percentages in the form of confusion matrices.

The selection of the subset of the features was done by forward model selection, where in every step the feature resulting in the highest increase of model performance was added. This process was stopped when no significant improvement was achieved by adding another feature.

Results from the correlation analysis have been used to exclude continuous features based on filtered versions of the continuous sleep profiles, thus reducing the number of features without remarkable loss of information. The feature Time in bed (tib) was excluded from the features as this parameter is highly sensitive to the recording protocol and may differ significantly between sleep laboratories. This is seen to be crucial due to the fact that the patients are not distributed uniformly across labs. Furthermore the feature Total sleep period (tsp) was excluded due to the high correlation to the Time in bed.

The number of all features considered for the different models are: 72 for the traditional R&K sleep scoring, 51 for the model of sleep cornerstones, 85 for the continuous R&K, and 68 for the model of sleep substructure.

The classification is done for the patient groups separately and is embodied in the substructure of this section. Finally the results are summarized to highlight differences between the models of sleep.

5.6.1 Generalized Anxiety Disorder

Classification results using all features are best for the continuous model of R&K sleep stages (Table 19). The traditional R&K-based features perform poorest. To give an impression of the variability of the classifiers, we show the ROC curves resulting from the five-fold cross-validation in Figure 36. The relatively small number of patients is responsible for the high variability. For that reason, one can not expect to observe statistically significant differences, and results have to be interpreted cautiously.

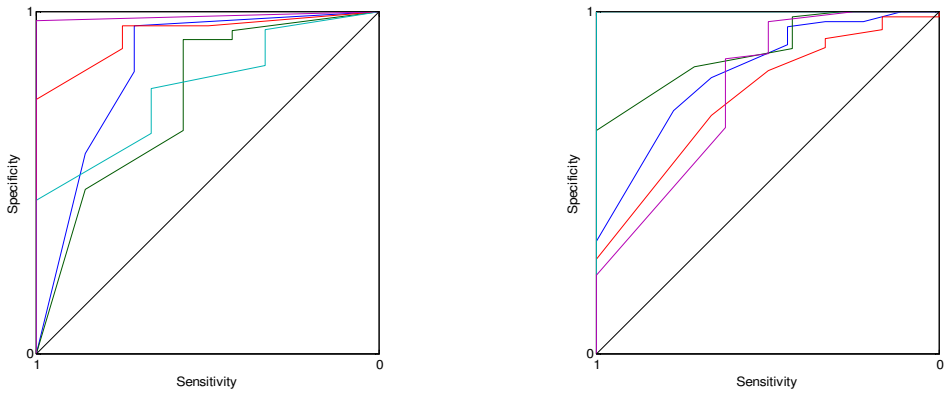

Figure 36: ROC for the 5-fold cross-validation from binary classification Healthy control vs. Generalized anxiety disorder patients for features based on the continuous R&K sleep profile. Left: all features. Right: subset of features.

Table 19: Classification results for HC vs. GAD patients using all features

	R&K		Sleep cornerstones		Continuous R&K		Sleep substructure	
	Healthy	GAD	Healthy	GAD	Healthy	GAD	Healthy	GAD
Control	0.73	0.27	0.72	0.28	0.76	0.24	0.63	0.37
GAD	0.31	0.69	0.19	0.81	0.11	0.89	0.16	0.84
Performance	71 %		76 %		83 %		73 %	
AUC ROC	0.70		0.79		0.86		0.76	

For all but the continuous model of R&K sleep stages, just one feature concerned with sleep stage 2 gives classification results at the same level (AUC ROC around 0.75) as with all features, see Table 20 for classification results:

R&K	rk_s2
Sleep cornerstones	cs_nrem
Continuous R&K	s6_s2
Sleep substructure	su_auc_s2

For the continuous model of R&K sleep stages, performance can be elevated to match that of the classifier using all features by adding the variable s6_sc_rem_wake, i.e. the number of state changes from REM to wake.

Table 20: Classification results for HC vs. GAD patients based on a subset of features

	R&K		Sleep cornerstones		Continuous R&K		Sleep substructure	
	Healthy	GAD	Healthy	GAD	Healthy	GAD	Healthy	GAD
Control	0.85	0.15	0.76	0.24	0.80	0.20	0.84	0.16
GAD	0.43	0.57	0.26	0.74	0.32	0.68	0.31	0.69
Performance	71 %		75 %		74 %		76 %	
AUC ROC	0.74		0.77		0.73		0.77	

As only one variable is involved, we compare the distributions of time in stage for the R&K and the continuous pendant in Figure 37. Wilcoxon rank sum tests verify significant differences for both features between Healthy controls and GAD patients ($p < 0.001$).

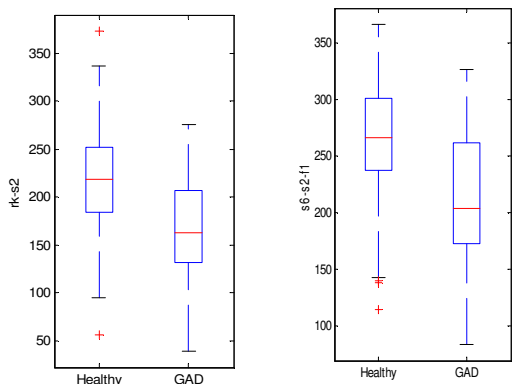

Figure 37: Comparison of time spent in s2 derived from traditional and continuous R&K sleep profiles for Healthy control vs. Generalized Anxiety Disease patients. Significant differences ($p < 0.001$, Wilcoxon rank sum test) are observed for both features.

5.6.2 Sleep Apnea Syndrome

When using all features, performance of the features derived from the continuous models is slightly higher than those from the R&K-based features, see Table 21 for the classification results.

Table 21: Classification results for HC vs. SAS patients using all features

	R&K		Sleep cornerstones		Continuous R&K		Sleep substructure	
	Healthy	Apnea	Healthy	Apnea	Healthy	Apnea	Healthy	Apnea
Control	0.80	0.20	0.76	0.24	0.78	0.22	0.83	0.17
Apnea	0.24	0.76	0.14	0.86	0.14	0.86	0.20	0.80
Performance	78 %		81 %		82 %		81 %	
AUC ROC	0.81		0.87		0.87		0.85	

The variable selection gives inhomogeneous results. While one R&K feature gives good results, two or more variables are selected for the continuous models:

R&K	rk_sc_21
Sleep cornerstones	cs_auc1_tsp_wake, cs_auc2_tsp_wake, cs_nrem
Continuous R&K	s6_auc1_tsp_wake, s6_auc2_tsp_wake, s6_tst_s2, s6_auc_tsp_wake
Sleep substructure	su_auc1_tsp_wake, su_auc2_tsp_wake

For all continuous models, the AUC of the derivatives from wake play the most important role. For all models except the model of sleep substructure, features concerned with s2 (NREM for the model of sleep cornerstones) are selected. The classification results using the listed subset of features are given in Table 22. For the traditional R&K features, AUC ROC is below those for the continuous models. Adding more R&K features does not improve classification results, thus light evidence is given that the features derived from the continuous models contain at least supplementary information.

Table 22: Classification results for HC vs. SAS patients using a subset of features

	R&K		Sleep cornerstones		Continuous R&K		Sleep substructure	
	Healthy	Apnea	Healthy	Apnea	Healthy	Apnea	Healthy	Apnea
Control	0.76	0.24	0.81	0.19	0.82	0.18	0.76	0.24
Apnea	0.25	0.75	0.23	0.77	0.22	0.78	0.15	0.85
Performance	75 %		79 %		80 %		80 %	
AUC ROC	0.78		0.84		0.84		0.84	

5.6.3 Parkinson's Disease

When using all features, the quality of classification is again higher for the features derived from the continuous models than for those from the R&K-based features, see Table 23 for the detailed results.

Table 23: Classification results for HC vs. PD patients using all features

	R&K		Sleep cornerstones		Continuous R&K		Sleep substructure	
	Healthy	PD	Healthy	PD	Healthy	PD	Healthy	PD
Control	0.79	0.21	0.75	0.25	0.86	0.14	0.82	0.18
Parkinson	0.43	0.57	0.29	0.71	0.34	0.66	0.24	0.76
Performance	68 %		73 %		76 %		79 %	
AUC ROC	0.68		0.73		0.73		0.80	

The best subsets of variables contain two variables for all models and for R&K features:

R&K	rk_sc_rem_rem, rk_tst_s1
Sleep cornerstones	cs_auc_nrem_q4, cs_nrem_q4
Continuous R&K	s6_auc1_tsp_wake, s6_tst_s1
Sleep substructure	su_auc_tsp_wake, su_entropy

From the features selected, no consensus is to be found regarding states being discriminative between Healthy controls and Parkinson's disease patients. Wake is listed for two continuous models, s1 for the traditional R&K features and for the continuous R&K model. Interestingly, the model of sleep cornerstones achieves good discrimination by taking information from the last quarter of the NREM trace. Association with s1 could be due to the fact that the amount of light sleep is increased towards the end of sleep. Classification results are given in Table 24.

The R&K features stands out in the fact that the number of consecutive REM periods is the most discriminative feature. As the same feature is available for the model of sleep substructure (as the REM state is discrete there, and identical to the R&K sleep profile), it will be interesting to compare the distributions of the most discriminative variables, as done in Figure 38. It can be seen that the medians are better separated for the continuous variable, but variance is higher compared to the R&K feature. When doing the classification using the number of REM changes and su_tst_s1, (i.e. the continuous pendant to the R&K feature from the model of sleep substructure), classification results are below those for the selected features (AUC ROC 0.76 vs. AUC ROC 0.86).

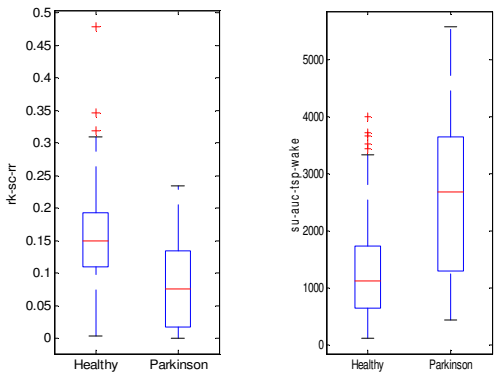

Figure 38: Comparison of the most discriminative features derived from the R&K sleep profile and the continuous profile of sleep substructure for Healthy control vs. Parkinson's Disease patients. For both features, significant differences at a high level are observed ($p < 0.001$, Wilcoxon rank sum test). Discriminative power is slightly higher for the continuous feature.

Table 24: Classification results for HC vs. PD patients using a subset of features

	R&K		Sleep cornerstones		Continuous R&K		Sleep substructure	
	Healthy	PD	Healthy	PD	Healthy	PD	Healthy	PD
Control	0.79	0.21	0.84	0.16	0.87	0.13	0.72	0.28
Parkinson	0.23	0.77	0.20	0.80	0.31	0.69	0.09	0.91
Performance	80 %		82 %		78 %		82 %	
AUC ROC	0.81		0.82		0.80		0.86	

5.6.4 Summary of the Classification results

We provide summary tables to get an overview of the entirety of the classification results. The results using all features and the best feature subsets are given in Table 25. While for all features, results are better for all continuous models, that is not true for the classification using the best subsets of features. For those, remarkably, the results for the model of sleep substructure are superior for all patient groups. When using the selected feature subsets, discrimination between Healthy controls and GAD patients is below the level for the other two patient groups.

Table 25: Classification results (AUC ROC) for all features and best subsets of the features

	All features			**Feature subsets**		
Model	**GAD**	**Apnea**	**PD**	**GAD**	**Apnea**	**PD**
Traditional R&K	0.66	0.81	0.68	0.74	0.78	0.81
Sleep cornerstones	0.79	0.87	0.73	0.77	0.84	0.82
Continuous R&K	0.86	0.87	0.73	0.73	0.84	0.80
Sleep substructure	0.76	0.85	0.80	0.77	0.84	0.86

An overview of the association of sleep stages with the selected features is given in Table 26. For GAD patients, features derived from s2 / NREM sleep are consistently selected. For Sleep Apnea Syndrome patients, features derived from the continuous wake profiles are predominant, while the changes within light sleep are discriminative for the traditional R&K features. For the group of Parkinson's Disease patients, uninterrupted REM sleep according to R&K scoring has high discriminative power whereas features regarding wake and light sleep are mainly selected from the continuous models.

Table 26: Association to sleep stages for the best features subsets for classification

Model	**GAD**	**Apnea**	**PD**
Traditional R&K	s2	s2/s1	REM
Sleep cornerstones	NREM	wake, NREM	NREM-q4
Continuous R&K	s2, REM/wake	wake, s2	wake, s1
Sleep substructure	s2	wake, entropy	wake, entropy

6 Summary

The model framework proved to be a solid basis for the continuous models of sleep/wakefulness, in accordance with the aims of this thesis. The ability to generalize was shown in the visual inspection and the correlation analysis. The signal preprocessing and feature extraction methods were carefully chosen to reduce differences between laboratories, but differences across laboratories were not analyzed systematically in this work.

In the model and parameter estimation procedures, relatively high variability was observed for the number of Gaussian components in the REM group. In the analysis done, the features associated with REM sleep turned out to play a minor role – therefore this did not have a considerable effect on the results. For further models based on the model framework, a significant increase in the number of initializations of the Gaussian Mixture Models, or the application of alternative strategies regarding model selection is nevertheless recommended.

In the most challenging part, the visualization and interpretation of the continuous sleep profiles, the usefulness of this representation of the sleep process was shown. Especially the model of sleep cornerstones and the model of sleep substructure have given a clear representation of the sleep process. The advantage of the high temporal resolution was demonstrated, and it is strongly believed that insights into the process of sleep can be gained by further investigation of the continuous sleep profiles. Ideally those insights would be implemented as features that reflect clinically relevant information.

In the correlation analysis, features derived from the continuous sleep profiles were found to be superior to those derived from the R&K scorings for the Alphabetical Cross-out test and the test of Fine motor activity. The R&K-based features, on the other hand, outperformed their continuous counterparts for the subjective measures of sleep quality. While most of the significantly correlated R&K features are concerned with wakefulness, NREM stages and deep sleep take in a more weighty position for the continuous features. Within the three continuous models, more significant correlations were observed for the finer-grained models, i.e. the continuous model of R&K sleep stages and the model of sleep substructure.

The model of sleep substructure, where external information is incorporated, appears at the top in the classification results for all patient groups. The performance is highest for discrimination between Healthy controls and Sleep Apnea patients, where a measure of complexity of the wake trace is selected as the main feature for all continuous models. For the group of Generalized Anxiety Disorder patients, one feature regarding stage s2 is selected for each model for discrimination from Healthy controls, and classification is accomplished at the same performance level for all models. For Parkinson's Disease patients, a single feature from the model of sleep substructure was observed to be superior to the most discriminative R&K-based feature. As the results were derived from small numbers of patients, further evaluation on a wider set of patients is needed.

When comparing the three continuous models presented, the model of sleep cornerstones seems to be useful for visualization, but has the disadvantage that deep sleep is not explicitly distinguished from light sleep. The continuous model of R&K sleep stages performs well in the correlation analysis but is too complex for visual exploration. The model of sleep substructure gives good results in all aspects and is favorable from an overall point of view.

One important aspect of further exploration of the continuous sleep profiles is the extraction of more clinically relevant features. We suggest the application of pattern recognition techniques for the detection of micro-events. Furthermore, the temporal information should be taken into account to a greater extent, for example by investigating cyclic alternating patterns is suggested in several studies for the assessment of sleep quality.

Artifact processing was beyond the scope in this work, but will be included in the future. The model will possibly be extended to take into account more biosignals, e.g. from electrooculogram and electromyogram to improve the level of discrimination between REM and light NREM sleep. Considering more than one EEG channel could provide information about local characteristics of sleep, that might be useful for patients with narcolepsy or parasomnias (Mahowald and Schenck 2005).

In context of the SENSATION project, the model was recently adapted to be used on daytime data, where the level of vigilance is of main interest. Here the flexibility of the model framework proved to be valuable, as emphasis can be put on the processes of particular interest and an adapted model can be derived easily.

To combine results of this work with related work, a sleep model based on Hidden Markov Models (HMM) as described in (Flexer et al. 2005) could be extended to model the observation distributions by Mixtures of Gaussians rather than by single Gaussians, or a hierarchical HMM with similar structure to the model of sleep substructure could be considered.

Appendix

A1 EM for GMM with constraints on mixing proportions

A short introduction to the EM algorithm was given in section 2.6, including the results for GMM without constraints on the variances or the mixing proportions. In the step called *reshuffling*, equation (21) of section 3.2, a modified update equation is used, emerging from the constraints

$$\forall g \in G : \sum_{k' \in K_g} \alpha_{k'} = p(g)$$

placed on the mixing proportions, where we dropped the identifier [(b)] from the α_k's. In this section, the derivation of this adapted update equation is given.

The incomplete-data log-likelihood is given by

$$\log L(\Theta \mid X) = \log \prod_{n=1}^{N} p(x_n \mid \Theta) = \sum_{n=1}^{N} \log \left(\sum_{k \in K} \alpha_k \, p(x_n \mid \theta_k) \right)$$

With introduction of unobserved data Z, with $z_n \in K$, and the assumption of the joint density function

$$p(x, z \mid \Theta) = p(x \mid z, \Theta) \, p(z \mid \Theta)$$

we arrive at the complete-data log-likelihood:

$$\log L(\Theta \mid X, Z) = \log p(X, Z \mid \Theta) = \sum_{n=1}^{N} \log \{ p(z_n \mid \Theta) \, p(x_n \mid z_n, \Theta) \} = \sum_{n=1}^{N} \log \{ \alpha_{z_n} \, p(x_n \mid \theta_{z_n}) \}$$

In the E-step, the expected value of the complete-data log-likelihood is computed, with respect to observed data X and current parameter estimates Θ^{old}. To achieve that, we need an expression for the distribution of the unobserved data. By using the current parameter estimates and applying Bayes' theorem, we have

$$p(z \mid x, \Theta^{old}) = \frac{\alpha_z \, p(x \mid \theta_z)}{\sum_{k \in K} \alpha_k \, p(x \mid \theta_k)}$$

and for a vector $\mathbf{z} = (z_1, ..., z_N)$:

$$p(\mathbf{z} \mid X, \Theta^{old}) = \prod_{m=1}^{N} p(z_m \mid x_m, \Theta^{old})$$

Note that the computation of these posterior probabilities is often called the E-step in the literature. That is not exactly correct, but the estimation of the distribution of the unobserved data is used for the evaluation of the expected value as done in the E-step:

$$\begin{aligned} Q(\Theta, \Theta^{old}) &= E\left[\log p(X, Z \mid \Theta) \mid X, \Theta^{old} \right] \\ &= \sum_{\mathbf{z}} \log \{ p(X, \mathbf{z} \mid \Theta) \} \, p(\mathbf{z} \mid X, \Theta^{old}) \\ &= \sum_{\mathbf{z}} \sum_{n=1}^{N} \log \{ \alpha_{z_n} \, p(x_n \mid \theta_{z_n}) \} \prod_{m=1}^{N} p(z_m \mid x_m, \Theta^{old}) \\ &= \sum_{z_1 \in K} \cdots \sum_{z_N \in K} \sum_{n=1}^{N} \log \{ \alpha_{z_n} \, p(x_n \mid \theta_{z_n}) \} \prod_{m=1}^{N} p(z_m \mid x_m, \Theta^{old}) \end{aligned} \tag{24}$$

By rearrangement and using the identity

$$\sum_{k\in K} p(k \mid x) = 1$$

the equation (24) can be greatly simplified:

$$
\begin{aligned}
& Q(\Theta, \Theta^{old}) \\
&= \sum_{z_1 \in K} \cdots \sum_{z_N \in K} \sum_{n=1}^{N} \log\left\{\alpha_{z_n} p(x_n \mid \theta_{z_n})\right\} \prod_{m=1}^{N} p(z_m \mid x_m, \Theta^{old}) \\
&= \sum_{z_1 \in K} \cdots \sum_{z_N \in K} \sum_{n=1}^{N} \sum_{k\in K} \delta_{k,z_n} \log\left\{\alpha_{z_n} p(x_n \mid \theta_{z_n})\right\} \prod_{m=1}^{N} p(z_m \mid x_m, \Theta^{old}) \\
&= \sum_{n=1}^{N} \sum_{k\in K} \log\left\{\alpha_k p(x_n \mid \theta_k)\right\} \sum_{z_1 \in K} \cdots \sum_{z_N \in K} \delta_{k,z_i} \prod_{m=1}^{N} p(z_m \mid x_m, \Theta^{old}) \\
&= \sum_{n=1}^{N} \sum_{k\in K} \log\left\{\alpha_k p(x_n \mid \theta_k)\right\} \left(\sum_{z_1 \in K} \cdots \sum_{z_{n-1} \in K} \sum_{z_{n+1} \in K} \cdots \sum_{z_N \in K} \prod_{\substack{m=1 \\ m\neq n}}^{N} p(z_m \mid x_m, \Theta^{old}) \right) p(k \mid x_n, \Theta^{old}) \\
&= \sum_{n=1}^{N} \sum_{k\in K} \log\left\{\alpha_{z_n} p(x_n \mid \theta_k)\right\} \left(\sum_{z_1 \in K} \cdots \sum_{z_{n-1} \in K} \sum_{z_{n+1} \in K} \cdots \sum_{z_{N-1} \in K} \prod_{\substack{m=1 \\ m\neq n}}^{N-1} p(z_m \mid x_m, \Theta^{old}) \right) \left(\sum_{k'\in K} p(k' \mid x_N, \Theta^{old}) \right) p(k \mid x_n, \Theta^{old}) \\
&= \sum_{n=1}^{N} \sum_{k\in K} \log\left\{\alpha_{z_n} p(x_n \mid \theta_k)\right\} \left(\sum_{z_1 \in K} \cdots \sum_{z_{n-1} \in K} \sum_{z_{n+1} \in K} \cdots \sum_{z_{N-1} \in K} \prod_{\substack{m=1 \\ m\neq n}}^{N-1} p(z_m \mid x_m, \Theta^{old}) \right) p(k \mid x_n, \Theta^{old}) \\
&= \sum_{n=1}^{N} \sum_{k\in K} \log\left\{\alpha_{z_n} p(x_n \mid \theta_k)\right\} \left(\sum_{z_1 \in K} \cdots \sum_{z_{n-1} \in K} \sum_{z_{n+1} \in K} \cdots \sum_{z_{N-2} \in K} \prod_{\substack{m=1 \\ m\neq n}}^{N-2} p(z_m \mid x_m, \Theta^{old}) \right) p(k \mid x_n, \Theta^{old}) \\
&\vdots \\
&= \sum_{n=1}^{N} \sum_{k\in K} \log\left\{\alpha_k p(x_n \mid \theta_k)\right\} p(k \mid x_n, \Theta^{old}) \\
&= \sum_{k\in K} \sum_{n=1}^{N} p(k \mid x_n, \Theta^{old}) \log \alpha_k + \sum_{k\in K} \sum_{n=1}^{N} p(k \mid x_n, \Theta^{old}) \log p(x_n \mid \theta_k)
\end{aligned}
\tag{25}
$$

We see that the two terms can be maximized independently, whereby the first term is determined by the mixing proportions α_k and the second term is governed by the parameters θ_k. As a consequence, the update equations for the parameters θ_k are not affected by our constraints.

Maximization of the first term in equation (25) is equivalent to minimization of

$$\tilde{Q} = -\sum_{k\in K} \sum_{n=1}^{N} \log(\alpha_k)\, p^{old}(k \mid x_n)$$

To facilitate minimization under the constraints

$$\forall g \in G : c_g \equiv p(g) - \sum_{k'\in K_g} \alpha_{k'} = 0$$

Lagrange multipliers λ_g are introduced:

$$0 = \frac{\partial}{\partial \alpha_k} \left\{ \sum_{k' \in K} \sum_{n=1}^{N} \log(\alpha_{k'}) p^{old}(k' \mid x_n) + \sum_{g \in G} \lambda_g \left(\sum_{k' \in K_g} \alpha_{k'} - p(g) \right) \right\}$$

The derivative of the first summand vanishes for $k \neq k'$, the second term for all k' that are components not belonging to group(k):

$$\begin{aligned} 0 &= \frac{\partial \tilde{Q}}{\partial \alpha_k^{new}} + \sum_{g \in G} \lambda_g \frac{\partial c_g}{\partial \alpha_k^{new}} \\ 0 &= \sum_{n=1}^{N} \frac{p^{old}(k \mid x_n)}{\alpha_k^{new}} + \sum_{g \in G} \lambda_g \mathrm{elem}(k, K_g) \end{aligned} \tag{26}$$

where elem(k, K_g) is defined as

$$\mathrm{elem}(k, K_g) = \begin{cases} 1 & k \in K_g \\ 0 & k \notin K_g \end{cases}$$

We proceed with multiplication by α_k^{new}

$$0 = \sum_{n=1}^{N} p^{old}(k \mid x_n) + \alpha_k^{new} \sum_{g \in G} \lambda_g \mathrm{elem}(k, K_g)$$

and summation over all components of g, $k \in K_g$:

$$0 = \sum_{n=1}^{N} \sum_{k \in K_g} p^{old}(k \mid x_n) + \sum_{k \in K_g} \alpha_k^{new} \sum_{g' \in G} \lambda_{g'} \mathrm{elem}(k, K_{g'})$$

$$0 = \sum_{n=1}^{N} \sum_{k \in K_g} p^{old}(k \mid x_n) + \lambda_g \sum_{k \in K_g} \alpha_k^{new}$$

By transformation, the values of the Lagrange multipliers are obtained:

$$\lambda_g = -\frac{\sum_{n=1}^{N} \sum_{k \in K_g} p^{old}(k \mid x_n)}{\sum_{k \in K_g} \alpha_k^{new}}$$

From the constraints, we know the α_k add up to $p(g)$ for $k \in K_g$, which gives us an expression for the Lagrange multipliers:

$$\lambda_g = -\frac{1}{p(g)} \sum_{n=1}^{N} \sum_{k \in K_g} p^{old}(k \mid x_n) \tag{27}$$

The terms for the Lagrange multipliers λ_k are inserted into (26):

$$\frac{1}{\alpha_k^{new}} \sum_{n=1}^{N} p^{old}(k \mid x_n) = \sum_{g \in G} \frac{\mathrm{elem}(k, K_g) \sum_{n=1}^{N} \sum_{k' \in K_g} p^{old}(k' \mid x_n)}{p(g)}$$

We see that summation over the groups G and components within the several groups is not necessary and reformulate as summation only over the components belonging to the same group as k:

$$\frac{1}{\alpha_k^{new}} \sum_{n=1}^{N} p^{old}(k \mid x_n) = \frac{\sum_{n=1}^{N} \sum_{k' \in K_{group(k)}} p^{old}(k' \mid x_n)}{p(g)}$$

The update equation for the mixing proportion is therewith

$$\alpha_k^{new} = p(g)\frac{\sum_{n=1}^{N} p^{old}(k \mid x_n)}{\sum_{k' \in K_{group(k)}} \sum_{n=1}^{N} p^{old}(k' \mid x_n)}$$

which is intuitively obtained by application of the unconstrained update equation (14) followed by normalization of the mixing proportions for the group components to sum up to $p(g)$:

$$\begin{aligned}
\alpha_k^{new} &= \frac{\alpha_k^{new,unconstrained}}{\sum_{k' \in K_{group(k)}} \alpha_{k'}^{new,unconstrained}} p(g) \\
&= \frac{\frac{1}{N}\sum_{n=1}^{N} p^{old}(k \mid x_n)}{\sum_{k' \in K_{group(k)}} \frac{1}{N}\sum_{n=1}^{N} p^{old}(k' \mid x_n)} p(g) \\
&= p(g)\frac{\sum_{n=1}^{N} p^{old}(k \mid x_n)}{\sum_{k' \in K_{group(k)}} \sum_{n=1}^{N} p^{old}(k' \mid x_n)}
\end{aligned}$$

A2 Psychometric variables

The measures listed in Table A2.1 are described in section 2.2.

Table A2.1: Psychometric variables

Variable	Description
PSQI	Pittsburgh Sleep Quality Index (Buysse et al. 1989) Retrospective questionnaire (19 questions) over 1 month. Total score from 0 (best) to 21 (worst).
SSA scores	Self-rating scale for sleep and awakening quality (Saletu et al. 1987).
SSA	Total score of the sleep and awakening quality Sum of SSA1, SSA2 and SSA3 (see below).
SSA1	Subjective Sleep Quality Seven questions regarding subjective sleep quality.
SSA2	Subjective Awakening Quality Eight questions regarding subjective sleep quality.
SSA3	Somatic Complaints Five questions regarding awareness of physical symptoms.
ASES	100 mm visual analogue scales (Folstein and Luria 1973)
Mood	Visual analogue scale for mood
Drive	Visual analogue scale for drive
AFF	Visual analogue scale for affectivity
Drows	Visual analogue scale for drowsiness
Bf-scale	Zerssen Bf –scale (Osgood et al. 1975)
WB-M	Well Being (morning) – self assessment scale
WB-E	Well Being (evening) – self assessment scale
AD-Test	Alphabetical Cross-out Test (Grunberger 1977)
AD-TS	Total score
AD-TS-ERR	Total score, false responses corrected
AD-ER	Number of errors
AD-ERP	Percentage of errors
AD-SV	Variability of the scores
NUM-M	Digit span test (Wechsler 1955)
FM-Test	Fine Motor activity test (Grunberger 1977)
FM-T	Fine Motor activity test – total score
FM-L	Fine Motor activity test – left hand
FM-R	Fine Motor activity test – right hand

A3 Correlation results

Table A3.1: Spearman's rank correlation coefficients between age and the features computed from the R&K and continuous sleep profiles

feature	train n1	train n2	test n1	test n2	n1	n2	mean
rk_sc_44	-0.592	-0.586	-0.518	-0.489	-0.565	-0.542	-0.554
rk_s4	-0.582	-0.575	-0.485	-0.464	-0.537	-0.525	-0.531
rk_wtsp	0.493	0.589	0.454	0.524	0.474	0.563	0.519
rk_tst_s4	-0.538	-0.521	-0.463	-0.440	-0.505	-0.487	-0.496
rk_sc_w1	0.402	0.459	0.427	0.539	0.417	0.508	0.463
rk_eff	-0.457	-0.586	-0.380	-0.452	-0.406	-0.514	-0.460
rk_fw	0.463	0.412	0.445	0.503	0.453	0.466	0.460
rk_sc_ww	0.444	0.581	0.354	0.403	0.388	0.492	0.440
rk_sc_12	0.295	0.320	0.464	0.464	0.389	0.416	0.403
rk_sc_1w	0.359	0.339	0.381	0.462	0.379	0.405	0.392
rk_sc_24	-0.405	-0.361	-0.315	-0.331	-0.365	-0.353	-0.359
rk_tst_s1	0.354	0.325	0.351	0.372	0.347	0.351	0.349
cs_wake_f1	0.631	0.658	0.456	0.515	0.532	0.593	0.562
cs_wtsp_f1	0.629	0.657	0.455	0.519	0.530	0.592	0.561
cs_eff_f1	-0.638	-0.649	-0.437	-0.501	-0.524	-0.586	-0.555
cs_fw_f1	0.513	0.641	0.477	0.521	0.492	0.574	0.533
cs_sc_wake_wake_f1	0.627	0.642	0.407	0.462	0.502	0.559	0.531
cs_sc_nrem_wake_f1	0.448	0.589	0.441	0.553	0.441	0.573	0.507
cs_sc_wake_nrem_f1	0.436	0.569	0.421	0.542	0.429	0.554	0.491
cs_sc_nrem_nrem_f1	-0.520	-0.529	-0.380	-0.382	-0.447	-0.455	-0.451
cs_nrem_f1	-0.509	-0.504	-0.290	-0.300	-0.389	-0.402	-0.396
cs_auc_nrem_q3	-0.366	-0.501	-0.345	-0.307	-0.364	-0.407	-0.386
s6_sc_s4_s4_f1	-0.680	-0.698	-0.534	-0.517	-0.617	-0.617	-0.617
s6_sc_s4_s2_f1	-0.627	-0.644	-0.585	-0.584	-0.609	-0.622	-0.616
s6_sc_s2_s4_f1	-0.625	-0.649	-0.577	-0.574	-0.607	-0.620	-0.614
s6_s4_f1	-0.665	-0.701	-0.526	-0.527	-0.606	-0.621	-0.613
s6_tst_s4_f1	-0.617	-0.652	-0.500	-0.495	-0.570	-0.583	-0.576
s6_wake_f1	0.589	0.657	0.491	0.533	0.538	0.595	0.566
s6_wtsp_f1	0.589	0.651	0.486	0.519	0.536	0.587	0.562
s6_eff_f1	-0.598	-0.661	-0.463	-0.518	-0.525	-0.594	-0.559
s6_fw_f1	0.547	0.672	0.469	0.485	0.505	0.579	0.542
s6_auc_s4_q2	-0.492	-0.624	-0.499	-0.500	-0.507	-0.570	-0.538
s6_sc_wake_wake_f1	0.558	0.621	0.444	0.486	0.484	0.554	0.519
s6_sc_s1_wake_f1	0.448	0.574	0.382	0.457	0.414	0.520	0.467
s6_auc_s4_q3	-0.482	-0.533	-0.394	-0.442	-0.434	-0.496	-0.465
s6_sc_wake_s1_f1	0.439	0.577	0.362	0.436	0.404	0.510	0.457
s6_tst_s1_f1	0.374	0.513	0.439	0.469	0.416	0.494	0.455
s6_sc_s2_wake_f1	0.457	0.506	0.379	0.429	0.419	0.468	0.443
s6_auc_s4_q4	-0.552	-0.462	-0.368	-0.374	-0.459	-0.419	-0.439
s6_tst_f1	-0.485	-0.545	-0.287	-0.354	-0.382	-0.464	-0.423
s6_sc_wake_s2_f1	0.437	0.497	0.347	0.377	0.389	0.439	0.414
s6_sc_wake_rem_f1	0.359	0.535	0.357	0.403	0.354	0.466	0.410
s6_s1_f1	0.291	0.453	0.416	0.456	0.362	0.449	0.405
s6_sc_rem_wake_f1	0.367	0.534	0.361	0.345	0.360	0.437	0.399
s6_sc_s2_s2_f1	-0.312	-0.348	-0.355	-0.351	-0.330	-0.359	-0.344
su_auc1_tsp_wake_f1	0.506	0.624	0.509	0.569	0.514	0.597	0.555
su_auc2_tsp_wake_f1	0.505	0.616	0.492	0.553	0.500	0.588	0.544
su_fw_f1	0.485	0.612	0.490	0.523	0.485	0.567	0.526
su_auc_tsp_wake	0.498	0.603	0.449	0.547	0.473	0.579	0.526

feature	train n1	train n2	test n1	test n2	n1	n2	mean
su_wtsp_f1	0.499	0.605	0.445	0.546	0.470	0.579	0.525
su_sc_s2_wake_f1	0.445	0.595	0.494	0.506	0.474	0.556	0.515
su_sc_wake_s2_f1	0.428	0.590	0.495	0.518	0.464	0.560	0.512
su_wake_f1	0.470	0.609	0.415	0.516	0.436	0.568	0.502
su_eff_f1	-0.451	-0.608	-0.384	-0.501	-0.420	-0.555	-0.487
su_sc_deep_deep_f1	-0.559	-0.561	-0.365	-0.373	-0.467	-0.464	-0.465
su_deep_q1_f1	-0.521	-0.521	-0.342	-0.404	-0.434	-0.466	-0.450
su_tst_s1_f1	0.413	0.510	0.412	0.457	0.411	0.485	0.448
su_deep_f1	-0.557	-0.528	-0.342	-0.375	-0.441	-0.453	-0.447
su_sc_wake_s1_f1	0.451	0.563	0.360	0.399	0.404	0.484	0.444
su_sc_wake_wake_f1	0.392	0.551	0.333	0.475	0.363	0.517	0.440
su_sc_s1_wake_f1	0.443	0.555	0.344	0.415	0.391	0.485	0.438
su_s1_f1	0.358	0.455	0.419	0.424	0.390	0.442	0.416
su_tst_deep_f1	-0.471	-0.447	-0.311	-0.320	-0.397	-0.380	-0.389
su_deep_q2_f1	-0.308	-0.453	-0.294	-0.355	-0.320	-0.406	-0.363
su_auc_deep_q2_f1	-0.309	-0.455	-0.295	-0.346	-0.319	-0.406	-0.362
su_tst_s2_f1	0.318	0.465	0.303	0.343	0.308	0.405	0.357
su_sc_deep_s2_f1	-0.368	-0.360	-0.322	-0.321	-0.345	-0.348	-0.346
su_fs_f1	0.307	0.318	0.418	0.362	0.346	0.338	0.342
su_sc_s2_deep_f1	-0.355	-0.357	-0.313	-0.312	-0.332	-0.342	-0.337

Table A3.2: Spearman's rank correlation coefficients between night differences of the Subjective Sleep Quality (SSA1) and the features computed from the R&K and continuous sleep profiles

feature	train	test	all
rk_eff	-0.405	-0.555	-0.486
rk_sc_ww	0.410	0.536	0.476
rk_wtsp	0.350	0.487	0.415
rk_sc_22	-0.298	-0.445	-0.381
rk_tst	-0.316	-0.434	-0.380
cs_wake_f1	0.331	0.582	0.466
cs_wtsp_f1	0.326	0.582	0.463
cs_eff_f1	-0.324	-0.573	-0.459
cs_sc_wake_wake_f1	0.289	0.550	0.439
cs_sc_nrem_nrem_f1	-0.334	-0.506	-0.425
cs_fw_f1	0.334	0.391	0.358
cs_sc_nrem_wake_f1	0.317	0.387	0.338
cs_sc_wake_nrem_f1	0.301	0.384	0.332
s6_wake_f1	0.309	0.571	0.445
s6_wtsp_f1	0.302	0.567	0.441
s6_eff_f1	-0.305	-0.552	-0.432
s6_sc_s2_s2_f1	-0.296	-0.510	-0.418
s6_tst_s1_f1	0.348	0.442	0.399
s6_sc_s1_wake_f1	0.333	0.439	0.383
s6_sc_wake_s1_f1	0.332	0.424	0.381
s6_fw_f1	0.331	0.394	0.356
su_wake_f1	0.356	0.590	0.477
su_eff_f1	-0.335	-0.566	-0.458
su_sc_wake_wake_f1	0.329	0.559	0.446
su_sc_s1_wake_f1	0.308	0.548	0.430
su_sc_wake_s1_f1	0.310	0.538	0.425
su_tst_s1_f1	0.291	0.518	0.416
su_wtsp_f1	0.291	0.493	0.387
su_sc_s2_s2_f1	-0.289	-0.346	-0.319

Table A3.3: Spearman's rank correlation coefficients between night differences of the Total score of the sleep and awakening quality (SSA) and the features computed from the R&K and continuous sleep profiles

feature	train	test	all
rk_eff	-0.378	-0.438	-0.410
rk_sc_ww	0.371	0.438	0.401
rk_wtsp	0.336	0.361	0.339
cs_wake_f1	0.311	0.411	0.368
cs_wtsp_f1	0.301	0.412	0.365
cs_eff_f1	-0.297	-0.397	-0.356
cs_sc_nrem_nrem_f1	-0.335	-0.340	-0.333
s6_wake_f1	0.298	0.386	0.341
s6_wtsp_f1	0.291	0.381	0.337
s6_sc_s1_wake_f1	0.365	0.294	0.323
s6_sc_wake_s1_f1	0.349	0.293	0.322
s6_tst_s1_f1	0.335	0.293	0.319
su_wake_f1	0.346	0.416	0.385
su_eff_f1	-0.322	-0.392	-0.360
su_sc_wake_wake_f1	0.315	0.389	0.351

Table A3.4: Spearman's rank correlation coefficients between the Alphabetical cross-out scores (ad-ts: total score. ad-ts-er: total score. false responses corrected) and the features computed from the R&K and continuous sleep profiles

variable	feature	train n1	train n2	test n1	test n2	n1	n2	mean
ad-ts	rk_wtsp	-0.424	-0.379	-0.289	-0.324	-0.356	-0.362	-0.359
ad-ts	s6_sc_s4_s2_f1	0.450	0.437	0.411	0.358	0.424	0.404	0.414
ad-ts	s6_sc_s2_s4_f1	0.442	0.441	0.395	0.348	0.420	0.401	0.410
ad-ts	s6_tst_s1_f1	-0.378	-0.359	-0.320	-0.311	-0.354	-0.325	-0.339
ad-ts	s6_s1_f1	-0.316	-0.315	-0.303	-0.306	-0.320	-0.295	-0.308
ad-ts	su_auc1_tsp_wake_f1	-0.469	-0.448	-0.357	-0.385	-0.410	-0.416	-0.413
ad-ts	su_auc2_tsp_wake_f1	-0.475	-0.450	-0.335	-0.371	-0.398	-0.407	-0.403
ad-ts	su_fw_f1	-0.482	-0.452	-0.317	-0.347	-0.392	-0.396	-0.394
ad-ts	su_sc_wake_s2_f1	-0.444	-0.441	-0.300	-0.362	-0.367	-0.394	-0.381
ad-ts	su_sc_s2_wake_f1	-0.452	-0.446	-0.300	-0.346	-0.372	-0.389	-0.380
ad-ts	su_auc_tsp_wake	-0.404	-0.400	-0.306	-0.346	-0.352	-0.374	-0.363
ad-ts	su_wtsp_f1	-0.391	-0.406	-0.302	-0.342	-0.349	-0.373	-0.361
ad-ts-er	rk_wtsp	-0.398	-0.391	-0.309	-0.355	-0.356	-0.386	-0.371
ad-ts-er	cs_fw_f1	-0.533	-0.486	-0.282	-0.319	-0.396	-0.386	-0.391
ad-ts-er	s6_sc_s4_s2_f1	0.427	0.442	0.412	0.381	0.414	0.414	0.414
ad-ts-er	s6_sc_s2_s4_f1	0.421	0.447	0.398	0.371	0.411	0.411	0.411
ad-ts-er	s6_wake_f1	-0.492	-0.460	-0.290	-0.316	-0.381	-0.376	-0.379
ad-ts-er	s6_wtsp_f1	-0.491	-0.456	-0.286	-0.306	-0.379	-0.371	-0.375
ad-ts-er	s6_tst_s1_f1	-0.337	-0.372	-0.312	-0.312	-0.327	-0.331	-0.329
ad-ts-er	su_auc1_tsp_wake_f1	-0.442	-0.462	-0.367	-0.420	-0.404	-0.440	-0.422
ad-ts-er	su_auc2_tsp_wake_f1	-0.445	-0.464	-0.347	-0.408	-0.393	-0.434	-0.413
ad-ts-er	su_fw_f1	-0.454	-0.462	-0.330	-0.384	-0.386	-0.421	-0.404
ad-ts-er	su_sc_s2_wake_f1	-0.429	-0.453	-0.321	-0.372	-0.373	-0.407	-0.390
ad-ts-er	su_sc_wake_s2_f1	-0.418	-0.447	-0.320	-0.386	-0.367	-0.412	-0.389
ad-ts-er	su_auc_tsp_wake	-0.376	-0.411	-0.320	-0.379	-0.346	-0.398	-0.372
ad-ts-er	su_wtsp_f1	-0.363	-0.415	-0.315	-0.375	-0.343	-0.397	-0.370
ad-ts-er	su_s1_f1	-0.339	-0.355	-0.282	-0.292	-0.311	-0.322	-0.316

Table A3.5: Spearman's rank correlation coefficients between the Fine motor activity test (fm-l: left side. fm-r: right side. fm-t: total score) and the features computed from the R&K and continuous sleep profiles

var	feature	train n1	train n2	test n1	test n2	n1	n2	mean
fm-l	su_auc1_tsp_wake_f1	-0.411	-0.296	-0.291	-0.304	-0.341	-0.302	-0.321
fm-l	su_fw_f1	-0.404	-0.291	-0.292	-0.284	-0.336	-0.287	-0.311
fm-r	s6_sc_s4_s2_f1	0.417	0.370	0.358	0.289	0.379	0.335	0.357
fm-r	su_auc1_tsp_wake_f1	-0.488	-0.364	-0.370	-0.353	-0.423	-0.362	-0.392
fm-r	su_auc2_tsp_wake_f1	-0.483	-0.359	-0.356	-0.338	-0.411	-0.354	-0.383
fm-r	su_fw_f1	-0.467	-0.354	-0.358	-0.317	-0.404	-0.335	-0.370
fm-r	su_sc_wake_s2_f1	-0.476	-0.325	-0.325	-0.292	-0.395	-0.310	-0.352
fm-r	su_auc_tsp_wake	-0.379	-0.309	-0.348	-0.343	-0.361	-0.333	-0.347
fm-r	su_sc_s2_wake_f1	-0.485	-0.318	-0.315	-0.279	-0.391	-0.300	-0.346
fm-r	su_wtsp_f1	-0.365	-0.311	-0.345	-0.346	-0.353	-0.334	-0.343
fm-t	rk_sc_rr	0.322	0.322	0.280	0.355	0.288	0.336	0.312
fm-t	s6_sc_s4_s2_f1	0.378	0.355	0.350	0.299	0.356	0.321	0.338
fm-t	s6_sc_s2_s4_f1	0.366	0.359	0.334	0.286	0.345	0.314	0.330
fm-t	su_auc1_tsp_wake_f1	-0.466	-0.344	-0.350	-0.367	-0.402	-0.353	-0.378
fm-t	su_auc2_tsp_wake_f1	-0.462	-0.339	-0.334	-0.355	-0.391	-0.348	-0.369
fm-t	su_fw_f1	-0.451	-0.335	-0.342	-0.341	-0.389	-0.333	-0.361
fm-t	su_sc_wake_s2_f1	-0.434	-0.306	-0.317	-0.304	-0.374	-0.305	-0.339
fm-t	su_auc_tsp_wake	-0.363	-0.291	-0.317	-0.360	-0.339	-0.326	-0.332
fm-t	su_sc_s2_wake_f1	-0.445	-0.294	-0.302	-0.289	-0.368	-0.293	-0.331
fm-t	su_wtsp_f1	-0.351	-0.294	-0.314	-0.363	-0.331	-0.327	-0.329
fm-t	su_entropy_rem	0.327	0.325	0.280	0.354	0.282	0.336	0.309
fm-t	su_sc_rem_rem_f1	0.326	0.322	0.278	0.353	0.282	0.336	0.309

List of Abbreviations

AR	Autoregressive
AUC	Area under the curve
AUC ROC	Area under the ROC curve
BIC	Bayesian information criterion
EEG	Electroencephalogram
EM	Expectation Maximization
EMG	Electromyogram
EOG	Electrooculogram
GAD	Generalized Anxiety Disorder
GLM	Generalized linear model
GMM	Gaussian Mixture Model
HC	Healthy control
HMM	Hidden Markov Model
ICD	International Statistical Classification of Diseases and Related Health Problems
NREM	Non Rapid Eye Movement
PC	Principal Component
PCA	Principal Component Analysis
PD	Parkinson's Disease
QQ plot	Quantile-Quantile plot
R&K	Rechtschaffen and Kales
REM	Rapid Eye Movement
ROC	Receiver Operator Characteristic
SAS	Sleep Apnea Syndrome

List of Figures

List of Tables

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory, Akademia Kiado, Budapest.

Anderer, P., G. Gruber, S. Parapatics, M. Woertz, T. Miazhynskaia, G. Klosch, B. Saletu, J. Zeitlhofer, M. J. Barbanoj, H. Danker-Hopfe, S. L. Himanen, B. Kemp, T. Penzel, M. Grozinger, D. Kunz, P. Rappelsberger, A. Schlogl and G. Dorffner (2005). "An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 x 7 utilizing the Siesta database." Neuropsychobiology 51(3): 115-33.

Benfield, J. D. and A. E. Raftery (1993). "Model-Based Gaussian and Non-Gaussian Clustering." Biometrics 49: 803-821.

Biernacki, C., G. Celeux, G. Govaert and F. Langrognet (2005). Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. Preprint. 2005.

Bishop, C. M. (1996). Neural Networks for Pattern Recognition, Oxford University Press.

Blinowska, K. J. and M. Malinowski (1991). "Non-Linear and Linear Forecasting of the EEG Time Series." Biol Cybern 66(2): 159-165.

Bohning, D. and W. Seidel (2003). "Editorial: recent developments in mixture models." Computational Statistics & Data Analysis 41(3-4): 349-357.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse Fisher information matrix. Information and Classification. O. Opitz, B. Lausen and R. Klar, Springer-Verlag: 40-54.

Broersen, P. M. T. (1997). The ABC of autoregressive order selection criteria. Sysid Conf., Kitakyushu, Japan.

Burg, J. P. (1972). "The relationship between maximum entropy spectra and maximum likelihood spectra." Geophysics 38: 375-376.

Burg, J. P. (1975). Maximum entropy spectral analysis. Dep. Geophys. Stanford, Stanford Univ.

Buysse, D. J., C. F. Reynolds, 3rd, T. H. Monk, S. R. Berman and D. J. Kupfer (1989). "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research." Psychiatry Res 28(2): 193-213.

Crochiere, R. E. (1979). A General Program to Perform Sampling Rate Conversion of Data by Rational Ratios. Programs for Digital Signal Processing. New York, IEEE Press: 8.2-1 to 8.2-7.

Danker-Hopfe, H., D. Kunz, G. Gruber, G. Klosch, J. L. Lorenzo, S. L. Himanen, B. Kemp, T. Penzel, J. Roschke, H. Dorn, A. Schlogl, E. Trenker and G. Dorffner (2004). "Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders." J Sleep Res 13(1): 63-9.

Feinberg, L. and T. C. Floyd (1979). "Systematic Trends Across the Night in Human Sleep Cycles." Psychophysiology 16(3): 283-291.

Flexer, A., G. Dorffner, P. Sykacek and I. Rezek (2002). "An automatic, continuous and probabilistic sleep stager based on a Hidden Markov Model." Applied Artificial Intelligence 16(3): 199-207.

Flexer, A., G. Gruber and G. Dorffner (2005). "A reliable probabilistic sleep stager based on a single EEG signal." Artif Intell Med 33(3): 199-207.

Folstein, M. F. and R. Luria (1973). "Reliability, validity, and clinical application of the Visual Analogue Mood Scale." Psychol Med 3(4): 479-86.

Fraley, C. and A. E. Raftery (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation." Journal of the American Statistical Association Jun 2002(97): 611-631.

Grunberger, J. (1977). Psychodiagnostik des Alkoholkranken. Ein methodischer Beitrag zur Bestimmung der Organizität in der Psychiatrie. Wien, Maudrich.

Hasan, J. (1996). "Past and future of computer-assisted sleep analysis and drowsiness assessment." J Clin Neurophysiol 13(4): 295-313.

Hastie, T. and R. Tibshirani (1996). "Discriminant analysis by gaussian mixtures." Journal of the Royal Statistical Society(58, Ser. B): 155-176.

Himanen, S. L. and J. Hasan (2000). "Limitations of Rechtschaffen and Kales." Sleep Med Rev 4(2): 149-167.

Hori, T., Y. Sugita, E. Koga, S. Shirakawa, K. Inoue, S. Uchida, H. Kuwahara, M. Kousaka, T. Kobayashi, Y. Tsuji, M. Terashima, K. Fukuda and N. Fukuda (2001). "Proposed supplements and amendments to 'A Manual of Standardized Terminology,Techniques and Scoring System for Sleep Stages of Human Subjects', the Rechtschaffen & Kales (1968) standard." Psychiatry and Clinical Neurosciences 55: 305-310.

Iber, C. (2004). "Development of a new manual for characterizing sleep." Sleep 27(2): 190-2.

Kaipio, J. P. and P. A. Karjalainen (1997). "Simulation of nonstationary EEG." Biol Cybern 76(5): 349-56.

Karlis, D. and E. Xekalaki (2003). "Choosing initial values for the EM algorithm for finite mixtures." Computational Statistics & Data Analysis 41(3-4): 577-590.

Kass, R. E. and A. E. Raftery (1995). "Bayes Factors." Journal of the American Statistical Association 90(430): 773-795.

Kemp, B. (1993). "A proposal for computer-based sleep/wake analysis." J Sleep Res 2(3): 179-185.

Keribin, C. (2000). "Consistent estimation of the order of Mixture Models." Sankhya: The Indian Journal of Statistics 62, Series A, Pt. 1: 49-66.

Klosch, G., B. Kemp, T. Penzel, A. Schlogl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeitlhofer, B. Saletu, W. M. Herrmann, S. L. Himanen, D. Kunz, M. J. Barbanoj, J. Roschke, A. Varri and G. Dorffner (2001). "The SIESTA project polygraphic and clinical database." IEEE Eng Med Biol Mag 20(3): 51-7.

Kryger, M. H., T. Roth and W. C. Dement (2005). Principles and Practice of Sleep Medicine, 4th ed. Philadelphia, Elsevier Saunders.

Kubicki, S. and W. M. Herrmann (1996). "The future of computer-assisted investigation of the polysomnogram: sleep microstructure." J Clin Neurophysiol 13(4): 285-94.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Fifth Symposium on Math, Statistics and Probability, Berkeley, CA, University of California Press.

Mahowald, M. W. and C. H. Schenck (2005). "Insights from studying human sleep disorders." Nature 437(7063): 1279-85.

Malvimvuo, J. and R. Plonsey (1995). Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields. New York, Oxford University Press.

McLachlan, G. and D. Peel (2000). Finite Mixture Models, John Wiley & Sons.

Miloslavsky, M. and M. J. van der Laan (2003). "Fitting of mixtures with unspecified number of components using cross validation distance estimate." Computational Statistics & Data Analysis 41(3-4): 413-428.

Nabney, I. T. (2002). Netlab: Algorithms for Pattern Recognition. Berlin, Springer.

Osgood, C. D., G. J. Suci and P. H. Tannenbaum (1975). The measurement of meaning. Urbana, Urbana University Press.

Pardey, J., S. J. Roberts, L. Tarassenko and J. Stradling (1996). "A new approach to the analysis of the human sleep/wakefulness continuum." J Sleep Res 5(4): 201-10.

Parks, T. W. and C. S. Burrus (1987). Digital Filter Design, John Wiley & Sons.

Rechtschaffen, A. and A. Kales (1968). A Manual of Standardized Terminology, Techniques, and Scoring System for Sleep Stages of Human Subjects, Brain Information / Brain Research Institute UCLA, Los Angeles.

Saletu, B., G. Kindshofer, P. Anderer and J. Grunberger (1987). "Short-term sleep laboratory studies with cinolazepam in situational insomnia induced by traffic noise." Int J Clin Pharmacol Res 7(5): 407-18.

Schwarz, G. (1978). "Estimating the dimension of a model." Annals of Statistics 6: 461-464.

Siegel, J. M. (2005). "Clues to the functions of mammalian sleep." Nature 437(7063): 1264-71.

Stephens, M. (2000). "Bayesian Analysis of Mixture Models with an Unknown Number of Components: an alternative to reverstible jump methods." Annals of Statistics 28(1): 40-74.

Sugimoto, H., N. Ishii, A. Iwata, N. Suzumura and T. Tomita (1978). "On the stationarity and normality of the electroencephalographic data during sleep stages." Comput Programs Biomed 8(3-4): 224-34.

Sykacek, P., S. J. Roberts, I. Rezek, A. Flexer and G. Dorffner (1999). Classification in the sampling paradigm: a predictive approach towards a SIESTA sleep analyzer. Proceedings of the European Medical & Biomedical Engineering Conference: 1652-1653.

Sykacek, P., S. J. Roberts, I. Rezek, A. Flexer and G. Dorffner (2001). A Probabilistic Approach to High-Resolution Sleep Analysis. Artificial Neural Networks - ICANN 2001, International Conference, Vienna, Austria, Lecture Notes In Computer Science 2130. D. G. e. al, Springer: 617-624.

Thomson, D. J. (1990). "Time series analysis of Holocene climate data." Philosophical Transactions of the royal Societa A 330: 601-616.

Vaz, F., P. G. De Oliveira and J. C. Principe (1987). "A study on the best order for autoregressive EEG modelling." Int J Biomed Comput 20(1-2): 41-50.

Venables, W. N. and B. D. Ripley (2002). Modern Applied Statistics with S. Fourth Edition, Springer.

Verbeek, J. J., N. Vlassis and B. Krose (2003). "Efficient Greedy Learning of Gaussian Mixture Models." Neural Comp. 15(2): 469-485.

Wechsler, D. (1955). Manual for the Wechsler Adult Intelligence Scale. New York, The Psychological Corporation.

WHO (1992). International Statistical Classification of Diseases and Related Health Problems, Tenth Revision. Geneva, World Health Organization.