Bayesian Assessment of Newborn Brain Maturity from Sleep Electroencephalograms

Livija Jakaite

A thesis submitted to the University of Bedfordshire,
in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

June 2012

Abstract

In this thesis, we develop and test a technology for computer-assisted assessments of newborn brain maturity from sleep electroencephalogram (EEG). Brain maturation of newborns is reflected in rapid development of EEG patterns over a number of weeks after conception. Observing the maturational patterns, experts can assess newborn's EEG maturity with an accuracy ± 2 weeks of newborn's stated age. A mismatch between the EEG patterns and newborn's physiological age alerts clinicians about possible neurological problems. Analysis of newborn EEG requires specialised skills to recognise the maturity-related waveforms and patterns and interpret them in the context of newborns age and behavioural state. It is highly desirable to make the results of maturity assessment most accurate and reliable. However, the expert analysis is limited in capability to estimate the uncertainty in assessments. To enable experts quantitatively evaluate risks of brain dysmaturity for each case, we employ the Bayesian model averaging methodology. This methodology, in theory, provides the most accurate assessments along with the estimates of uncertainty, enabling experts to take into account the full information about the risk of decision making. Such information is particularly important when assessing the EEG signals which are highly variable and corrupted by artefacts. The use of decision tree models within the Bayesian averaging enables interpreting the results as a set of rules and finding the EEG features which make the most important contribution to assessments. The developed technology was tested on approximately 1,000 EEG recordings of newborns aged 36 to 45 weeks post conception, and the accuracy of assessments was comparable to that achieved by EEG experts. In addition, it was shown that the Bayesian assessment can be used to quantitatively evaluate the risk of brain dysmaturity for each EEG recording.

Contents

List of Figures

List of Tables

Acknowledgements

This work is part of the research project "Automated Electroencephalographic Assessment of Brain Maturation in Newborns" funded by the Leverhulme Trust. The EEG data used within this project were granted by the University of Jena, Germany.

It was a pleasure to receive an award at the University of Bedfordshire annual poster day 2010, and I am thankful to VC of the University of Bedfordshire Prof. Les Ebdon and Pro VC Research and Enterprise Prof. Carsten Maple for this encouragement.

I also wish to acknowledge the support and administrative assistance of the Research Graduate School, Prof. Angus Duncan, the Department of Computer Science and Technology, Prof. Yong Yue, and the Institute for Research Into Applicable Computing, Prof. Edmond Prakash. I am grateful to Dr Dayou Li and Dr Mehmet Aydin for their constructive comments and advices on this PhD project.

Special thanks to Dr Vitaly Schetinin for the opportunity to work on this project and explore application of Bayesian methodology to a unique set of EEG data.

For consultations on the EEG data preparation and analysis, thanks to Dr Joachim Schult, University of Hamburg, Germany.

Finally, I wish to thank Kelvin Hopkins MP for taking interest in this project at the SET for Britain 2012 poster exhibition.

Abbreviations

aEEG - Amplitude-Integrated EEG
AD-test - Anderson-Darling test
AR - Auto-Regression
BMA - Bayesian Model Averaging
DT - Decision Tree
EEG - Electroencephalogram
FFT - Fast Fourier Transform
KS-test - Kolmogorov-Smirnov test
MCMC - Markov Chain Monte Carlo
NLEO - Nonlinear Energy Operator
PCA - Post-Conceptional Age
SPS - Spectral Power Statistics

Chapter 1

Introduction

The electroencephalogram or EEG is the recording of electrical activity measured by electrodes placed on the scalp. Analysis of frequency, amplitude and form of EEG waves reveals important information about brain function for clinical diagnostics as well as for cognitive research.

The first report on human EEG was published in 1929 by Berger, who observed that different types of brain electrical activity could be described in terms of their frequency range. He measured the frequencies manually and proposed the terms Alpha and Beta processes to designate the frequency ranges of waves found in most recordings. Within 10 years, new results were published in the main areas of EEG research: sleep study (Loomis et al., 1937; Davis et al., 1938, 1939), distribution of activity over brain regions (Rubin, 1938), potentials related to muscle control (Jasper and Andrews, 1938), and localization of tumours (Walter, 1936). The latter publication also described the slow Delta and Theta waves found along with the Alpha and Beta activity, thus establishing the main four frequency bands used today in EEG interpretation.

Analysis of EEG frequency composition was first automated by the improved low frequency analyser (Walter, 1943). Brazier and Casby (1952) applied auto-correlation and cross-correlation techniques to extract features representing EEG frequencies. By 1970, the introduction of the fast Fourier transform made automated analysis of EEG frequencies practical (Dumermuth et al., 1970). With the increase in computational power over the following decades it became feasible to develop pattern recognition methods to detect EEG waves and classify types of activity (Robert et al., 2002).

The Bayesian methodology of probabilistic inference, which only recently has been made computationally practical, can be used to quantify the uncertainty in results of EEG analysis, and during the last ten years it has been widely applied

for EEG-based localisation of brain activity sources, see e.g. (Trujillo-Barreto et al., 2004; Jun et al., 2008).

Analysis of newborn EEG has been developing as a special branch of electroencephalography since the 1960s, when advances in neonatal care markedly improved survival rates of babies born pre-term. Nevertheless, the pre-term newborns remained at higher risk of brain injuries, and continuous EEG monitoring has been proposed to assess their brain function.

Brain maturation of newborns is reflected in rapid development of EEG patterns over a number of weeks since conception. Observing the maturational patterns, experts can assess newborn's EEG maturity with an accuracy of ± 1 or ± 2 weeks. Analysis of newborn EEG requires specialised skills to recognise the maturity-related waveforms and patterns and interpret them in the context of newborns age and behavioural state (Mizrahi et al., 2003).

In this thesis, we develop and test a technology for computer-assisted assessments of newborn brain maturity from EEG. To enable experts quantitatively evaluate risks of brain dismaturity for each case, we employ the Bayesian methodology of probabilistic reasoning. The use of decision tree models for assessments will enable interpreting the results as a set of rules and finding the EEG features which make the most important contribution to assessments.

1.1 Motivation

During the past 20 years, the number of pre-term newborns has increased by almost 25%, and, despite the improved survival, the risks of neurological problems are still high (Niemarkt et al., 2008). In the UK, over 80,000 babies every year are born premature or sick (Bliss, 2012), and their brain development can be affected by birth injuries, oxygen deprivation, or stress of pre-term birth. Brain injuries are mainly diagnosed by ultrasound scanning. However, a newborn may still be at risk caused by the development abnormalities which experts can only recognise in sleep EEG.

These abnormalities are expressed as altered rates of maturation of EEG patterns (Scher, 1997; Lombroso, 1985). Observing these maturational patterns, experts can assess newborn's EEG maturity with an accuracy of ± 1 or ± 2 weeks. If the EEG assessment matches a newborn's age, the brain maturity is likely normal. However, if the patterns are mismatched by more than two weeks, the maturity is likely abnormal.

Abnormal brain maturation has been found strongly associated with increased risk of sudden infant death syndrome (Scher et al., 2003b) and with neurological problems in later life (Lombroso, 1985; Tharp et al., 1989; Okumura et al., 2010; Bihannic et al., 2012). Developmental care procedures have

been shown promising to correct brain dysmaturity (Beckwith and Parmelee, 1986; Scher et al., 2009). Therefore it is crucial to timely recognise abnormal EEG maturation in order to provide the necessary care. Monitoring the developmental patterns in weekly EEG recordings is important to detect abnormality and predict outcome for a newborn (Tharp et al., 1989).

The maturity-related patterns are difficult to recognise in EEG as they widely vary during the first weeks after birth (Pressler et al., 2003). This makes the analysis of EEG laborious, and as a consequence of that, EEG analysis cannot be made available for all newborns at risk. Experts can sadly make a mistake, as there are no regular rules for interpretation of EEG maturity patterns.

To assist experts with analysis of EEG, computer-based classification methods have been proposed (Crowell et al., 1978; Holthausen et al., 2000). Within these methods, it has been attempted to classify at most three levels of brain maturity; however, automated assessment of in the range of ± 2 weeks of newborn's stated age has not been explored. Consequently, the existing approaches cannot assess brain maturity with the accuracy comparable to that provided by experts.

Another limitation of the existing approaches is that, being based on a single model, they cannot provide accurate estimates of confidence in assessments. Using the methods of probabilistic inference is of crucial importance to allow experts to quantitatively evaluate the confidence. The Bayesian theory of probabilistic inference enables the confidence to be estimated most accurately, see e.g. (Robert and Casella, 2004).

In addition to assisting experts with accurate estimates of confidence, it is important to provide an explanatory model readable as a set of rules so that the experts can understand how assessments are made in each case. The use of decision tree models which are transparent for users will allow EEG experts to make new finding in the neurological assessment of newborn brain.

1.2 Aim and scope

The aim of this research is to develop and test a computer-assisted technology for assessments of newborn EEG maturation. The maturity assessment is formulated as classification of EEG into ages expressed in weeks since conception. We expect to achieve the accuracy of the assessment comparable to that obtained by experts.

The assessments will be made within the methodology of Bayesian averaging over classification models to allow experts to obtain the exhaustive information on risk of brain dysmaturity. The use of decision trees for classification enables

making the results interpretable for experts and selecting EEG features making important contribution to the assessments.

To obtain accurate assessments of newborn EEG, first, it is important to minimise the negative influence of noise and artefacts on these weak signals. Finding of most informative features to represent newborn EEG is an open research area, and so extraction and evaluation of the features describing EEG maturation will be explored in this work. Then, aiming to classify a large range of newborns' ages, we will look into ways of improving the accuracy of multi-categorical classifiaction. Finally, we will test the computer-based assessments on ca 1,000 recordings of newborns aged between 36 and 45 weeks.

In this thesis, the post-conceptional ages of the newborns are used as labels of brain maturity. This approach, however, imposes limitations on estimating the accuracy of maturity assessments because of the following reasons.

- Maturation of EEG patterns does not necessarily progress at the same rate for all healthy newborns, and the normal patterns can vary for newborns of the same age.
- Accurate estimation of post-conceptional age is not always possible and the ages of some of the newborns could be estimated with an error of ± 1 or ± 2 weeks.
- The brain maturity of some of the newborns could be mismatched with their post-conceptional ages because of their health conditions.

For accurate evaluation of the developed technology it would be necessary to compare the results with expert assessments of brain maturity for each recording. Ideally, each recording would have to be analysed by a few experts in order to make the assessments more objective. Unfortunately, such analysis would be infeasible for a large collection of recordings.

Nevertheless, as most of the newborns in the database were neurologically normal, we assume that the maturational patterns in most of the EEG will be appropriate for the stated ages. We expect that under the above conditions the developed technology will be capable of recognising the age-related patterns in EEG, and in the cases when the assessment will considerably mismatch the stated age, brain dysmaturity may be suspected. Overall, we expect that the rates of match between the stated age and the assessed maturity for the developed technique will be comparable to those of expert assessments performed on smaller sets of newborn EEG recordings.

1.3 Thesis outline

Chapter 2 provides the background information about the expert approach to EEG maturity assessment as well as about previous work on computer-based assessments. We describe the expected accuracy of maturity assessment based on results found in literature. Next, we provide background on existing technologies of EEG maturity assessment. First, we describe the maturational patterns analysed by experts and the rules proposed for assessing maturity from EEG. Second, we review the existing approaches to computer-based assessments. Lastly, we describe the EEG data available for our experiments.

Chapter 3 introduces Bayesian methodology of model averaging and the use of decision tree models within this methodology. It discusses how the Bayesian methodology is made computationally feasible by using the Markov chain Monte Carlo method.

Chapter 4 explores how the influence of EEG artefacts can be reduced to improve the accuracy of maturity assessments. The artefacts are typically detected and marked by experts. Unfortunately, there are no standard rules for detection of artefacts, and the marks may be subjective and inconsistent. We assume that automatic removal of artefacts will provide better results within Bayesian assessments than the manual removal by experts. We test this assumption in experiments.

In **Chapter 5** we assume that the accuracy can be improved further by selecting the most informative EEG features. The standard spectral features form a multidimensional representation of EEG data. It is unrealistic to expect that all these features are equally important — some of them may be making a weak contribution to maturity assessments. We assume that the weak features unnecessarily increase the dimensionality of a model parameter space, making it difficult to be be explored in detail within Bayesian integration. The lack of detailed exploration can negatively affect the accuracy of Bayesian assessments. Fortunately, the use of DT models within Bayesian model averaging enables the importances of EEG features to be estimated in post analysis. We propose and test a technique that uses the information on estimated importances to improve the results of Bayesian assessments.

In **Chapter 6** we assume that complementing the standard spectral features with the new features which represent time-domain information will increase the accuracy of assessments. Specifically, we explore extraction of features related to EEG discontinuity, which is the most important maturational feature. We propose a new technique to estimate EEG discontinuity as non-stationarity, and show that the new feature outperforms the conventional discontinuity estimates. Used in combination with the standard spectral features, the new

feature representing EEG non-stationarity significantly improves the accuracy of assessments.

Having extracted the features which improve the assessment accuracy, in **Chapter 7** we extend the assessment technology to a larger range of ages. When the number of age groups is increased, it becomes more difficult to learn to distinguish the multiple age groups (classes). Therefore the accuracy of the standard multiclass approach becomes negatively affected. We show that the assessment accuracy can be improved by converting the multiclass problem into a set of two-class tasks.

Finally, in **Chapter 8** we explore the accuracy of Bayesian assessments in the typical intervals of ± 1 and ± 2 weeks. We compare the accuracy with that of expert assessments obtained for similar age groups. We show how the Bayesian assessments of the posterior probabilities can be used for evaluating the risk of possible errors and explore the shape of the class posterior distribution for patient cases with normal and with dysmature assessments. Additionally, for the found cases of dysmature assessments we tested two assumptions abut possible causes of the dysmaturity. We show a statistically significant relationship between the mismatch and very pre-term birth.

1.4 Main contributions

The main contributions of this work in the order of significance are as follow:

- *Automated assessment of newborn brain maturation from EEG.* The accuracy of the proposed automated assessment was comparable to that of expert assessments. We tested the assessments on EEG of newborns aged 36 to 45 weeks and counted the accuracy within the intervals of ± 1 and ± 2 weeks of newborns stated age, as typically done by EEG experts. In contrast, the automated assessments found in literature were capable of estimating up to three levels of maturation. We also showed that maturation of newborns aged 36–45 weeks can be assessed from two-channel EEG, without the conventionally used multiple channels and a polysomnogram (Chapter 8).

- *A new feature describing discontinuity of newborn EEG.* The new feature, estimated as the rate of pseudo-stationary EEG intervals, was shown strongly correlated with newborns' age in weeks since conception.The use of the new feature significantly improved the accuracy of maturity assessment. The new feature provided better accuracy than the conventional discontinuity estimates based on variability of amplitude (Chapter 6).

- *Technique of refining Bayesian ensembles of decision trees.* The idea of the proposed refining technique is to discard the models which use the features making weak contribution to assessments. The refining was shown capable of increasing the performance and decreasing the ensemble entropy or uncertainty. At the same time, it enabled selecting a subset of most important EEG features (Chapter 5).

- *Technique for binarisation of multiclass EEG assessment problem.* We showed that converting a multiclass problem into a set of binary ones improves the accuracy of maturity assessments in case of multiple classes and when the class labels, being ages in weeks, are naturally ordered. A limitation of this technique is that the assessment decision is combined from outputs of multiple binary classifier so that it becomes difficult to interpret. To simplify interpretation of maturity assessments, we proposed a meta-tree classifier which provides a performance comparable to that of the pairwise binarisation while using fewer binary classifiers, whose contributions are defined within a hierarchical structure (Chapter 7).

- *Evaluation of techniques for removal of EEG artefacts to improve the accuracy of Bayesian assessments.* EEG artefacts vary in the form and appear within different types of activity, so that experts cannot apply standard rules to artefact detection. Therefore, the detection can be inconsistent between experts, and the inconsistencies in removal of artefacts can affect the accuracy of computer-based assessments. We assumed that automated removal, providing consistent results, will enable achieving better accuracy of Bayesian assessments of brain maturity. In our experiments, automated removal of artefacts was shown improving the accuracy of maturity assessments, outperforming the manual removal (Chapter 4).

1.5 Publications

Parts of this research have been described in the following publications:

Schetinin, V., Jakaite, L. (2012). Classification of newborn EEG maturity with Bayesian averaging over decision trees. *Expert Systems with Applications*, 39(10):9340-9347. DOI: 10.1016/j.bbr.2011.03.031.

Jakaite, L., Schetinin, V. and Maple, C.(2012). Bayesian assessment of newborn brain maturity from two-channel sleep electroencephalograms. *Computer and Mathematical Methods in Medicine*. DOI: 10.1155/2012/629654.

Jakaite, L., Schetinin, V., Maple, C. and Schult, J. (2011). Feature extraction from electroencephalograms for Bayesian assessment of newborn brain maturity. *The 24th International Symposium on Computer Based Medical Systems*, CBMS 2011, Bristol, UK. DOI: 10.1109/CBMS.2011.5999109.

Schetinin, V., Jakaite, L., Maple, C. and Schult, J. (2011). Informativeness of sleep cycle features in Bayesian assessment of newborn electroencephalographic maturity. *The 24th International Symposium on Computer Based Medical Systems*, CBMS 2011, Bristol, UK. DOI: 10.1109/CBMS.2011.5999111.

Jakaite, L., Schetinin, V., Maple, C. and Schult, J. (2010). Bayesian model averaging over decision trees for assessing newborn brain maturity from electroencephalogram. *The 9th IEEE International Conference on Cybernetic Intelligent Systems*, CIS 2010, University of Reading, UK. DOI: 10.1109/UKRICIS.2010.5898148.

Jakaite, L., Schetinin, V., Maple, C. and Schult, J. (2010). Bayesian decision trees for EEG assessment of newborn brain maturity. In *the 10th Annual Workshop on Computational Intelligence*, UKCI 2010, University of Essex, UK. DOI: 10.1109/UKCI.2010.5625584.

Posters presenting research results and related techniques have participated in competitions:

Jakaite, L., Schetinin, V. (2012). Computer-Assisted Assessemnt of Newborn Brain Maturity from EEG. *SET for Britain 2012*, The House of Commons, London. Selected for the exhibition.

Jakaite, L., Schetinin, V. (2010). Feature Selection for Evaluation of Trauma Death Risk. *University of Bedfordshire Conference*, Luton. Awarded 1st prize.

Techniques employed in this research have also been used in the following publications:

Uglov, J., Jakaite, L., Schetinin, V. and Maple, C. (2008). Comparing robustness of pairwise and multiclass neural-network systems for face recognition. *EURASIP Journal of Advances in Signal Processing.* DOI: 10.1155/2008/468693.

Jakaite, L., Schetinin, V. (2008). Feature selection for Bayesian evaluation of trauma death risk. In *IFMBE Proceedings of the 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics*, 20(3):123-126, Riga, Latvia. DOI: 10.1007/978-3-540-69367-33_3.

The author's contribution to these publications was in part of the application of Bayesian model averaging over decision trees to real data, including contribution to the conception of research, design and implementation of experiments, and interpretation of results.

Chapter 2

Maturational Patterns in EEG

EEG patterns reflect rapid maturational changes in the brain from prematurity to a few weeks after birth. Analysing these patterns experts can estimate a newborns EEG age and compare it to the stated physiological age to assess brain development. The maturational patterns were being validated during the past fifty years, whereas finding of the EEG features for automated assessments is an open research question. This chapter provides the background on the expert and computer-assisted assessments of newborn EEG maturation.

Section 2.1 introduces the technique of recording newborn EEG and the main characteristics of the resulting signal. Section 2.2 explains why EEG maturation is typically assessed within the interval of ± 2 weeks of a newborn's stated age, and summarise the results of assessments described in literature. We also state the possible causes of abnormal EEG maturity. Next, in Section 2.3 we provide the background on existing technologies of EEG maturity assessment. We describe and give examples of the maturity-related patterns interpreted by experts, and discuss the assessment scales based on these patterns. Section 2.4 discusses the existing approaches to computer-based assessments. First, we describe extraction of features carrying the information on EEG maturity. These features can be correlated to ages of newborns or used to represent EEG data for automated classification. We discuss the existing attempts to classify the maturational levels. Finally, in Section 2.5 we give characteristics of the data available for our research, and in Section 2.6 we summarise the chapter.

2.1 EEG recording technique and properties

EEG enables monitoring a newborn's brain function continuously during a few hours to make real-time assessments. This makes EEG the most sensitive technique for predicting outcome in newborns with brain injury (Boylan, 2008). To take full advantage of EEG, it is important to record during at least one hour to capture the maturational patterns which vary with a patient's sate, specifically the two main states, called the quiet and active sleep (Pressler et al., 2003; Scher, 2006).

EEG is typically recorded with non-invasive electrodes placed on patient's scalp according to the standard "10–20" electrode placement system in positions shown in Fig. 2.1. Electrical activity of different brain regions is recorded from a number of EEG channels, each of which measures the difference between two electrodes. The channels may be set up according to a given "montage" system. The two most common montages are the bipolar and referential. In the bipolar montage two adjacent electrodes are linked into one channel, whereas in the referential one each of the scalp electrodes is linked to a common reference electrode.

EEG signal is composed of waves with different frequencies and amplitudes. The frequencies are typically measured in the range of 0.1 to 30 Hz. The amplitudes are measured in a range of -100 to 100 μV (microvolts) with a mean amplitude of zero and the minimal and maximal amplitudes are symmetrically distributed around the mean. The absolute amplitudes, measured from peak of a wave to its trough, usually are in the range of 50-100μV (Boylan, 2008).

The average absolute amplitude and dominant frequencies of newborn EEG change in time, reflecting short term variations in brain activity as well as changes between the quiet and active sleep stages. These variations make the EEG highly non-stationary. Being a weak signal, EEG is easily contaminated by artefacts which can be biological or caused by external sources. The biological artefacts can be caused by patient's movements, breathing, pulse, etc., whereas external artefacts — by electrical interference from other monitoring devices in neonatal intensive care units. It is important to recognise and exclude the artefacts, some of which may appear similar to EEG waves (Clarencon et al., 1996; van de Velde et al., 1999, 1998).

In general, it is desirable to record EEG from the complete set of electrodes to capture the whole range of maturity-related patterns which are observed at different brain regions (Mizrahi et al., 2003). The complete set includes more than 20 electrodes.

In practice, it may be too difficult or time-consuming to place all the electrodes on a small newborn's head, especially when other care procedures must be

carried out (Boylan, 2008). Long preparation for EEG additionally causes stress for a newborn and consequently affects the quality of sleep so that artefacts occur more frequently. Moreover, the interpretation of some of the electrodes may be difficult as they are highly affected by artefacts from eye and head movement (Chang et al., 2005).

Therefore a minimal set of electrodes often needs to be chosen so that most of the maturity-related patterns can be captured and at the same time the preparation costs and the influence of artefacts can be minimised. One way of achieving such a trade-off is to use four electrodes linked into two bipolar channels C3T3 and C4T4 (Holthausen et al., 2000). Electrodes in these positions show maturational changes most clearly (Niemarkt et al., 2011).

EEG is often recorded as part of a polysomnogram comprising measurements of heart and respiration rates, movements of eyes, limbs and chin, as well as blood oxygen saturation. The full polysomnographic information is typically taken into account for assessment of developing sleep cycles, which however after the 36 weeks are so developed that they become apparent in EEG (Mizrahi et al., 2003). Therefore, maturity assessment at this age may be done from EEG exclusively.

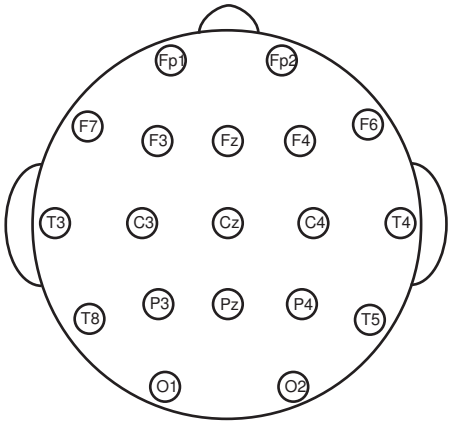

Figure 2.1: Positions of electrodes according to the standard system. Letters denote head regions: prefrontal (Fp1, Fp2), frontal (F2-F7), temporal (T3-T8), central (Cz, C3, C4), parietal (P2-P4) and occipital (O1, O2). The odd and even numbers denote the left and right hemispheres.

2.2 Accuracy of maturity assessments

This section explains why EEG maturation is typically assessed within the interval of ±2 weeks of newborn's stated age. The performance of computer-assisted method will be assessed in the same interval.

First, we give the terms referring to ages of newborns, used for discussing maturity assessment techniques. We discuss the assessment accuracy that can be theoretically achieved based on the conventional criteria. We also note that assessments can sometimes mismatch because of uncertainty in the estimate of newborn's physiological age. We then summarise the results of assessments described in literature. We also state the common factors of abnormal EEG maturity which can lead to mismatched assessments.

2.2.1 Terminology of newborn ages

Newborn's *gestational age* refers to the time in weeks from conception to birth, and is counted since mother's last menstrual period. The *Post-Conceptional Age* (PCA) is a sum of the newborns gestational age and weeks since birth. It is considered that EEG patterns of a newborn are determined by PCA rather than gestational age (Lombroso, 1985; Pressler et al., 2003), therefore EEG maturity is assessed in PCA.

Newborns' ages are conventionally classified into a number of groups. Most broadly, babies born before 37 weeks are classified as born *pre-term*. Births at 38 to 41 weeks are *full-term*, and after 41 weeks – *post-term*. To distinguish between different developmental stages from conception till birth, ages of pre-term babies can be additionally classified as *very pre-term* (before 32 weeks) and *late pre-term* (34-37 weeks).

2.2.2 Assessment intervals and uncertainty

The criteria proposed in literature for estimating EEG maturity are given in ranges of 3 to 5 weeks PCA (Lombroso, 1985; Mizrahi et al., 2003; Pressler et al., 2003; Niedermeyer, 2005). According to these criteria, EEG recorded at neighbouring PCA weeks are very similar, whereas EEG recorded a few weeks apart can be distinguished more easily. Brain development is, however, a continuous process, and the criteria corresponding to certain weeks should be viewed as a guide only. In theory, these criteria enable brain maturity to be estimated with accuracy of ±2 weeks, and the maturity is considered abnormal if an EEG estimate is out of this range (Tharp, 1990).

When evaluating the gap between the EEG estimate and stated age, we must keep in mind that PCA can be mistaken. The weeks of PCA are most often

counted on the base of information obtained from a questionnaire of the mother. Such information can be imprecise so that the PCA estimate can be misleading. Ultrasound dating has been shown more accurate than the questionnaire estimate and normally is undertaken during the first three and six months of pregnancy. The dates are typically replaced by the ultrasound estimates if the difference exceeds ± 1 week in the first three months and ± 2 weeks, in the six months (Hoffman et al., 2008).

Unfortunately, few publications have compared the EEG estimated ages with PCA in practice. In one of the first publications on EEG assessment of newborn brain development (Parmelee et al., 1968), the experts have estimated maturity of 47 EEG recordings made at 30 to 43 weeks PCA. The estimates matched the PCA in a range of ± 2 weeks for 85% of cases. The rate of match was 60% within the range of ± 1 week and 27% of recordings matched the PCA exactly. It remains unclear, however, whether the newborns for whom the assessments were mismatched by more than ± 2 weeks had any neurological problems.

Recently, a modified version of the approach by Parmelee et al. has been tested on 146 recordings of newborns aged 27 to 37 weeks PCA (Kato et al., 2011). The maturity was assessed for 129 recordings which were found of acceptable quality. In 77.5% the assessment was within the interval of ± 1 week of stated age and in 96.9% — within ± 2 weeks. However, in 17 recordings experts could not assess brain maturity because of artefact contamination. It must be also noted that some PCA, especially 32 and 34 weeks, are easy to identify by characteristic EEG waves, whereas at other ages there are no such markers.

A comparison between the EEG, ultrasound and questionnaire-based estimates of PCA has been made by Scher et al. (1994a). They assessed the EEG maturity of 13 very pre-term newborns who were subsequently reported as healthy at one to three years of age. It was concluded that average PCA estimates based on EEG, mothers questionnaire, and a number of ultrasound measurements counted over the 13 subjects were not significantly different. However, for individual patients, the various PCA estimates differed by three to 12 weeks in 10 of the cases. These results illustrate the difficulty of estimating the PCA with any of the techniques. To obtain a reliable ultrasound-based estimate, an expert has to review and compare the different ultrasound measurements, however, the final estimate was not shown in the paper. From the shown results, EEG estimate matched the questionnaire-based one in a range of ± 2 weeks in 10 of the cases (approximately 77%).

Based on the above discussion, we must keep in mind that the uncertainty in stated PCA will affect the accuracy of automated maturity assessments which we aim to obtain in our research. Under the uncertainty, the EEG data of neighbouring weeks will overlap and we cannot expect that the maturity assessments

will exactly match the stated PCA in most of the cases. However, we aim to achieve the accuracies in the ranges of ± 1, ± 2 weeks, which are at least as good as those provided by experts, as reported by Parmelee et al. (1968).

2.2.3 Causes of dysmaturity

In some of the cases when the maturity assessments mismatch the stated age, the reason is that PCA estimates were mistaken. However, the causes of mismatch may be physiological, and so may alert about development problems. Persistent maturational delays of more than two weeks are usually found in patients with brain injuries or chronic hypoxia. On the other hand, transient dysmaturity which becomes resolved within a few weeks has not been found associated with neurological problems (Lombroso, 1985; Tharp, 1990).

The pioneering studies in newborn EEG reported that maturation progresses independently of gestational age and birth weight (Parmelee et al., 1967, 1968). More recently, it has been found that rates of brain development in pre-term newborns may be altered. Conde et al. (2005) found prolonged dysmature patterns in babies born at or before 28 weeks gestational age. The dysmaturity was more evident in those of the newborns who were diagnosed with brain lesions. In (Scher et al., 1994c, 2003a, 2011) it was shown that the maturational trends are altered even in healthy pre-term newborns. In general, clinical significance of dysmature patterns in the absence of apparent brain abnormalities is yet to be established. In this direction, (Beckwith and Parmelee, 1986) showed that altered maturation in pre-term newborns is associated with lower intelligence in childhood, however such outcomes were avoided if babies were raised in attentive environments.

Some studies have attempted to find association between known risk factors and brain dysmaturity. Sterman et al. (1982) studied EEG maturation in 20 newborns who had a family history of sudden infant death, and therefore were considered at higher risk. It was found that some of the maturational patterns for these newborns appeared approximately four weeks earlier, than for a control group. Authors explained the accelerated maturation by adaptive response to a mild inborn hypoxia sustained by at-risk infants. Holthausen et al. (2000) explored association between dysmaturity and apnoea. They analysed 71 EEG recordings of which seven were found dysmature. All seven were from newborns at high risk of apnoea.

Based on these findings, we can expect that factors such as pre-term birth and risk of apnoea can alter the maturational patterns in some of the recordings included in our data, and so the portion of matching assessments can become decreased. These factors will be explored in this thesis. In the next section

we describe the maturity-related patterns which experts analyse in sleep EEG. We also review some approaches to scoring of these patterns to brain maturity assessment.

2.3 Expert assessment technology

Experts typically view newborn EEG at time scale of 1.5 cm/sec and amplitude scale of 50-100 μV/cm. Under these stetting, experts can detect characteristic individual waveforms as well as patterns of reoccurring waves and changes in sleep states (Scher, 2006).

First, we introduce continuity, which is one of the most important characteristics of newborn EEG, and describe how it varies with age. The frequency is discussed with reference to the main rhythms of EEG patterns evolving with age as well as to separate waveforms typical for certain weeks. We then describe how EEG varies over sleep states of newborns and how these variations are developing with age. Finally we review the maturity assessment scales based on visual analysis of continuity, frequency and sleep state.

2.3.1 Continuity

EEG of very pre-term and full-term are easily distinguished based on their continuity (or contrary discontinuity). By definition, an EEG pattern is discontinuous if the intervals with the normal voltage range are interchanged with periods of low voltage below $< 20\mu$V, called the *inter-burst intervals*. Otherwise, if the average amplitude is relatively constant with no periods of inactivity, the pattern is continuous.

EEG of very pre-term newborns is discontinuous during most of the recording. Long inter-burst intervals lasting up to 60 seconds, during which there may be no measurable activity, are interrupted by shorter high-amplitude bursts of mixed-frequency waves. This pattern is called *tracé discontinu* (Pressler et al., 2003; Boylan et al., 2008).

Fig. 2.2 shows segments with different continuity from an EEG recorded at 30 weeks PCA. Here, EEG is plotted at 50% scale of typical display (0.7 cm/sec, 150 μV/cm) to fit sufficient signal on page while keeping the proportions which are familiar to experts. The upper plot shows a segment with tracé discontinu pattern. An inter-burst interval is seen during the first ten seconds and then a burst with amplitude 100-150μV appears. The lower plot shows a continuous segment from the same EEG recording. Although the amplitude varies, there are no obvious periods of inactivity.

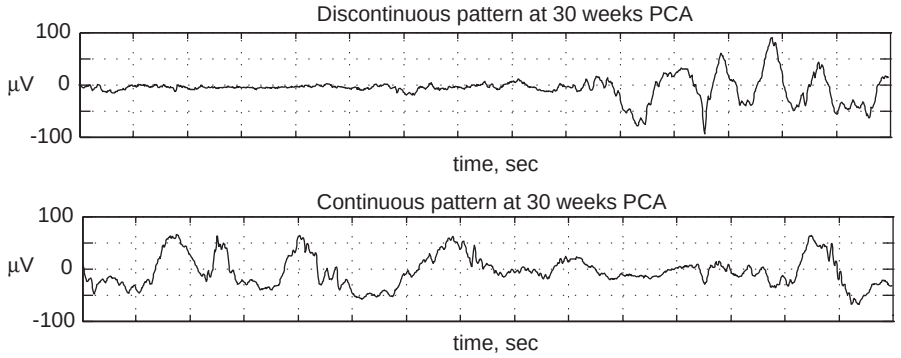

Figure 2.2: Continuous and discontinuous segments from an EEG recorded at 30 weeks PCA.

It has been observed that during brain development, the portion of discontinuous patterns is decreased while continuous patterns become longer. At the same time, the discontinuous patterns become "less discontinuous" as inter-burst intervals become progressively shorter while bursts become longer. Although this tendency for the decrease in discontinuity has been observed in several studies (Scher et al., 1994b; Niemarkt et al., 2010; Pressler et al., 2003; Boylan et al., 2008), the proportion of discontinuous patterns and the mean and maximal lengths of bursts and inter-burst intervals varied significantly (Niemarkt et al., 2008). This variation in results can be caused by the lack of standard rules for detection of inter-burst intervals. Hahn et al. (1989) found that the mean and maximal durations of inter-burst intervals were significantly affected by the choice of criteria for detection. Defining the criteria is problematic mainly because the amplitudes of bursts and inter-burst intervals change during maturation (Pressler et al., 2003), so that a single rule cannot be applied for detection of discontinuity at different PCA. Because of the lack of agreed criteria, the assessments of discontinuity can be subjective.

Nevertheless, it is generally considered that at full-term age, EEG is mostly continuous, except during one pattern named *tracé alternant*, shown in Fig. 2.3. In this pattern, inter-burst intervals are only 4-5 sec long and bursts last 2-4 sec. This pattern starts to replace tracé discontinu at 36 weeks PCA, and it may be observed till 44 weeks (Mizrahi et al., 2003; Boylan et al., 2008). In contrast to the very pre-term discontinuous patterns, in the full-term tracé alternant, the inter-burst intervals show higher amplitude and are never completely inactive. The amplitude of burst gradually decreases with maturation. As the difference in the amplitudes of the bursts and inter-burst intervals becomes less prominent,

the full-term tracé alternant pattern cannot be strictly defined as discontinuous and sometimes is said to be semi-discontinuous (Pressler et al., 2003).

It is important to note that tracé discontinu and tracé alternant are normal patterns for healthy newborns, if observed at appropriate PCA. However, these patterns must be distinguished from a specific pathological discontinuous pattern called burst-suppression, which is always associated with brain injuries. Burst suppression appears similar to the tracé discontinu with long inactive inter-burst intervals and high-amplitude bursts. However a distinct feature of this pattern is that it does not vary during the whole recording (Mizrahi et al., 2003).

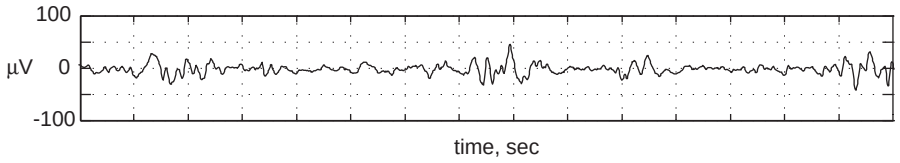

Figure 2.3: Tracé alternant pattern at 40 weeks PCA.

2.3.2 Frequency

Conventionally, the range of EEG frequencies is subdivided into a number of bands: Subdelta (0-1.5 Hz), Delta (1.5-3.5 Hz), Theta (3.5-7.5), Alpha (7.5-13.5), Beta 1 (13.5-19.5 Hz), and Beta 2 (19.5-25 Hz). These frequencies can be represented by specific maturity-related waveforms as well as by patterns consisting of mixed waves. Experts typically assess the frequencies of different EEG waves by comparing them with a time scale. The frequency of a wave in Hz can be estimated by comparing its cycle length with a one-second interval. In newborn EEG, the slow waves normally are higher in amplitude and vice versa the faster waves are lower.

Fig. 2.4 shows examples of characteristic waves from each of the frequency bands. The waves which are higher in amplitude, rhythmic and with clearly visible frequency composition are easiest to detect by visual analysis. Therefore, experts have explored such waves as maturational features. The overall frequency composition of a recording, influenced by multiple different waves, may provide new features, which cannot be assessed visually (Richards et al., 1986).

Experts have observed that, during maturation, the development of continuity is accompanied by an increase in the dominant frequency from Subdelta-Delta to mixed frequencies (Pressler et al., 2003). In the very pre-term EEG,

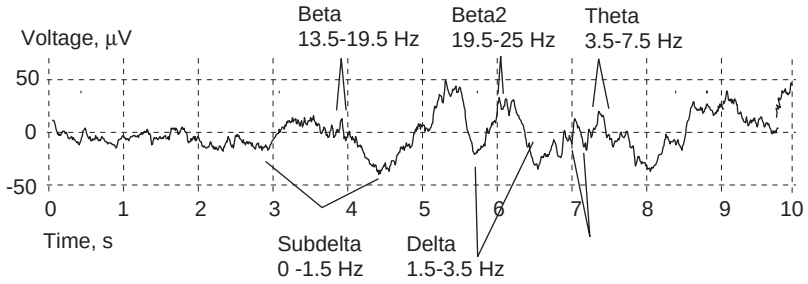

Figure 2.4: Examples of EEG waves from each of the frequency bands: Subdelta, Delta, Theta, Alpha, Beta and Beta2.

high-amplitude Subdelta and Delta waves are predominant during bursts and in the developing continuous patterns. At full-term PCA, most of the continuous patterns show a mixture of frequencies.

A specific waveform with high amplitude in the Subdelta-Delta frequency bands often seen during 26-36 weeks PCA is a *Delta brush*, a wave consisting of high-amplitude Subdelta or Delta wave with superimposed Alpha-Beta activity of lower amplitude forming a "brush" on the wave, see Fig. 2.5. The Delta brushes are seen most frequently at approximately 32 weeks PCA, and then become more rare, finally disappearing at full-term age.

The Theta burst is a rhythmic wave with Theta frequency and amplitude of 100-200μV, observed at temporal (T*) electrodes. This waveform appears at 28 weeks and is maximal at 32 weeks. After this age during weeks 34-35, the frequency of the sawtooth shifts to the Alpha band.

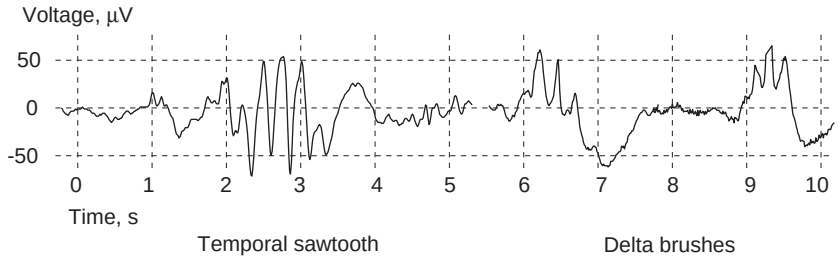

Figure 2.5: Age-related EEG waveforms.

By 36-37 weeks PCA, the largest part of recording is occupied by continuous patterns visible as a mixture of Delta and Theta waves of a moderate amplitude of 20-100μV intermixed with some low-amplitude Alpha-Beta activity. Closer to full-term age, it becomes possible to visually distinguish a number of patterns

with characteristic frequency composition, which appear during the quiet and active sleep. A pattern of high-amplitude delta waves remains during a part of quiet sleep. During the remaining part, tracé discontinu becomes replaced by tracé alternant, which contains a larger portion of low-voltage high-frequency activity during the inter-burst intervals. Tracé alternant, in turn, gradually diminishes during weeks 44 to 48 (Koszer et al., 2006). Half of a recording contains active sleep patterns of mixed Delta, Theta, Alpha and Beta waves in low or medium amplitudes. Next, we look in more detail at the development of the quiet and active sleep as a maturity-related feature.

2.3.3 Sleep states

The development of a cycle of active and quiet sleep states is an important characteristic of normal brain maturation. At full-term age, the sleep cycle variations should be clearly visible in EEG. However, some authors consider that a rudimentary sleep cycle appears in EEG as early as 27 weeks PCA (Curzi-Dascalova et al., 1993; Koszer et al., 2006).

In general, the active sleep patterns tend to be with higher frequency and lower amplitude, whereas quiet sleep patterns, contrary, tend to be with lower frequency but higher amplitude (Boylan et al., 2008). The difference in the average amplitudes during the active and quiet sleep is clearly seen in full-term EEG. This difference in amplitudes starts to become distinguishable at 36 weeks PCA. In the pre-term EEG, the sleep states differ mainly in continuity: the active sleep patterns tend to be more continuous than the quiet sleep ones. The cyclic variations in continuity are present in normal EEG at 30 weeks (Olischar et al., 2004; Koszer et al., 2006). As EEG matures from pre-term to full-term, the percentage of quiet sleep in a recording increases to approximately 50%.

At full-term age, four patterns typify the sleep stages. Two patterns can be seen during the quiet sleep: the continuous slow-wave sleep pattern of Delta and Theta waves with amplitudes of 25-100μV, and the semi-discontinuous tracé alternant. The active sleep is represented by two continuous patterns of mixed Delta, Theta, Alpha and Beta waves: the low voltage irregular pattern with amplitude of 10-15μV, and the mixed pattern with some higher-amplitude (50μV) Theta waves.

Fig. 2.6 shows the sleep cycle in an EEG recorded at 40 weeks PCA. The upper plot shows approximately 3.5 h of the recording. The quiet sleep states, lasting approximately 30 min each, can be identified by an increase in the average amplitude. The 20 sec segments shown in the lower plots exemplify the slow-wave sleep, tracé alternant, low voltage irregular and mixed pattern, corresponding to marks a, b, c and d.

It is also possible to recognise the quiet and active awake states in newborn EEG. However, awake EEG is difficult to record and prone to artefacts, and therefore is not widely used for brain maturity assessment.

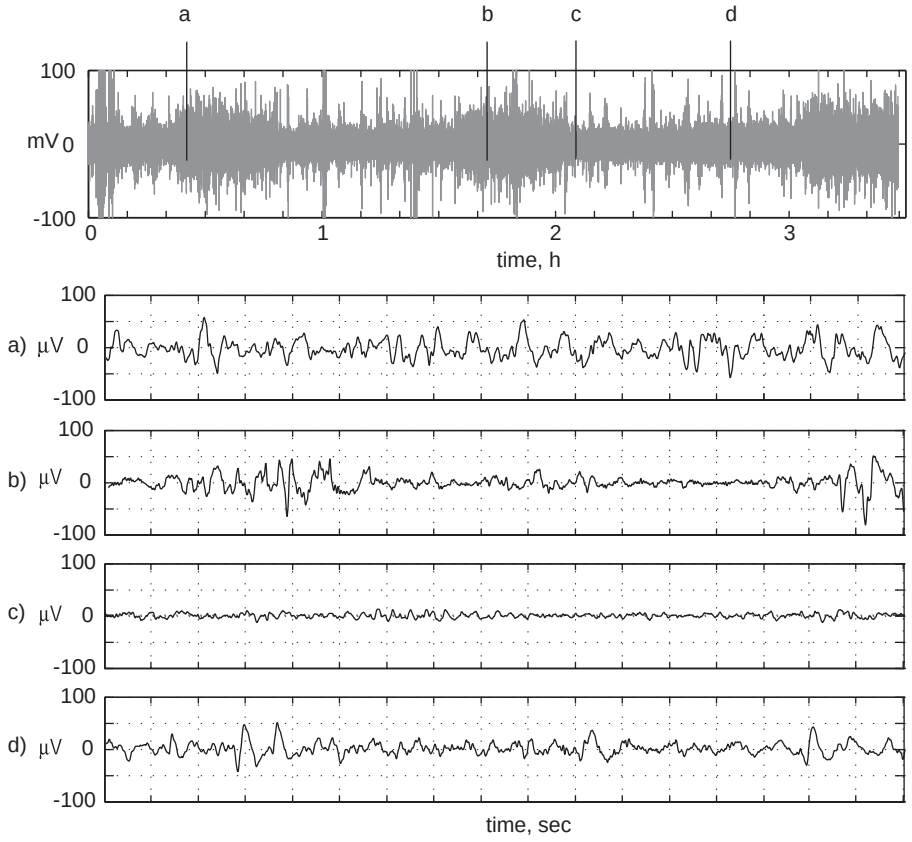

Figure 2.6: Sleep cycle and corresponding patterns: a) slow-wave sleep, b) tracé alternant, c) low voltage irregular, d) and mixed pattern.

2.3.4 Scales for assessment

In one of the first publications on newborn brain maturity assessment from sleep EEG, Parmelee et al. (1968) proposed an EEG pattern coding system for visual assessment of brain maturity. The coding system defined 5 basic patterns with characteristic continuity level, frequency composition, and amplitude range. Additionally, the system distinguished maturity-related variations of the basic patterns for weeks 28, 32, 36 and 40. In total, the system used 10 codes, ad the descriptions and illustrations were provided for the corresponding patterns. During analysis, one of the 10 codes was assigned to each 20 sec EEG segment.

The assessment of maturity was made based on distribution of the codes in the recording. The resultant assessments were found matching the PCA within the range of ± 2 weeks in 85% of cases.

To use the coding system, it is necessary to learn to identify the patterns based on descriptions and illustrations. Each 20 sec segment of a recording has to be coded to obtain the distribution of patterns for maturity assessment. In practice, the patterns widely vary, and considerable experience is needed to recognise them. It was found that two experts scoring a set of EEG segments assigned the same patterns in 65% of cases.

Tharp et al. (1989) employed a simplified technique to quickly estimate maturation in late pre-term and full-term newborns. The technique has been intended to detect significant dysmaturity of EEG after 37 weeks PCA rather than to study ongoing maturation. The following features were considered appropriate for the age group: infrequent Delta brushes, moderate amplitude ($\geq 25\mu$V) of inter-burst intervals during the tracé alternant, synchronous start and duration of bursts observed at both hemispheres, and rare Theta bursts. In the absence of these features, a dysmaturity of at least 2 weeks was defined. The system may be unsuitable to assess ongoing maturation after 37 weeks.

Recently, (Kato et al., 2011) published results of assessments in a range of ± 2 weeks made for newborns aged 26-38 weeks. The assessment was mainly based on the following parameters: amplitude and frequency of Delta waves, rate of Delta brushes, characteristic waveforms, and percentage of continuous activity. The latter was counted as the percentage of uninterrupted continuous segments which lasted at least 20 sec. The rate of Delta brushes was given qualitatively. To detect some of the waveforms, 8-channel recordings were used. The technique was tested on 129 recordings of 37 patients aged 27-34 weeks, and the assessment was matched in the range of ± 2 weeks for 97% of recordings.

Kato et al. also noted that the technique required a high level of EEG interpretation skills. Specifically, the observation of the maturity-related waveforms over the 8 channels was necessary for achieving the high accuracy of assessments, whereas confident assessment from 2 channels was found difficult. In general, the analysis of waveforms such as Theta and Alpha bursts enables accurate estimation of maturity for 32-35 weeks PCA, whereas for later ages no such developmental markers are known. Therefore, the accuracy is expected lower at these ages.

2.4 Computer-assisted maturity assessments

For automated assessments, the continuity, frequency and sleep state variations need to be estimated automatically. This estimation is termed EEG feature

extraction. The correlation between the extracted features and ages of newborns can be explored to establish their relevance to brain maturation. It has also been attempted to use the features to represent EEG data for automated classification.

This section first reviews the approaches to EEG feature extraction and the associations found between the features and newborn brain maturation. Spectral powers, describing EEG frequency composition, are the most widely used features for automated assessments, and we first describe the main approaches to computing the spectral powers, as well as the correlations between spectral powers and maturation reported in previous studies. Then we introduce amplitude integration, which is becoming an established technique to assist experts with visual analysis of newborn EEG. The existing approaches to extracting continuity features are reviewed next. We also review some of the more rarely used features including spectral edge frequency, entropy and complexity. Some of these features were employed to differentiate between the sleep states of newborns. Finally, we turn to automated classification of EEG data represented by the maturity-related features and review the previous attempts to classify different levels of maturity.

2.4.1 Spectral powers

Typically, computer-aided analysis of sleep EEG is carried out with the spectral powers computed in the standard frequency bands. To compute the spectral powers, EEG is transformed into the frequency domain using the discrete Fourier transform. The idea of the transform is to multiply the signal with a pair of sine and cosine waves for each frequency. These products provide the total amplitude of activity within the corresponding frequencies. The computation has been made efficient within the Fast Fourier Transform (FFT) algorithm.

The number of frequencies, or spectral resolution, in the FFT of a signal is equal to the number of samples divided by two. Obviously, the lowest frequency that can be represented, in Hz, depends on the duration of signal in seconds, and the highest frequency depends on the sampling rate. That is, given a signal with duration of T seconds, sampling rate S, and length $N = T \times S$, the computable frequency range will be from $\frac{1}{T}$ Hz to $\frac{S}{2}$ Hz (Nyquist frequency), and there will be $\frac{N}{2}$ frequency components.

The plot of the amplitudes over frequencies is the frequency spectrum of a signal. To obtain the powers within the standard frequency bands, the spectral components falling within each band are averaged and squared. These are called the absolute powers. The relative powers are obtained by dividing the power in each band by the summed power of all bands.

One problem with spectral analysis of EEG is, however, that Fourier transform assumes stationary signals. That is, signals whose mean amplitude and frequency do not vary over time. This means that the Fourier transform gives a measurement of the total power of waves with different frequencies, but contains no information about the time distribution of EEG activity. For example, within a given frequency band, the power of continuous low voltage activity may be the same as that of rarely occurring waves with high amplitude. Spectra may also become corrupted by high-amplitude EEG artefacts which introduce non-stationarities in data (Clarencon et al., 1996).

To cope with the non-stationarity of EEG, it is typical to apply the transform to short epochs which are assumed to be pseudo-stationary, that is, their amplitude and frequency vary in a small range. The durations of segments are usually 2-30 sec (Cooper et al., 2003; Victor et al., 2005; Estevez et al., 2002). Generally it is desirable to transform longer segments in order to obtain a higher resolution and a better representation of low frequencies. However, the longer the segments, the more likely they are to be non-stationary. Given that even relatively short (8 sec) segments are often non-stationary (McEwen and Anderson, 1975), the choice of segment length is *ad hoc*. To avoid presetting the segment lengths, adaptive segmentation can be used to automatically split EEG into pseudo-stationary intervals (Barlow, 1985; Aufrichtig et al., 1991; Appel and Brandt, 1983; Bodenstein et al., 1985; Agarwal et al., 1998).

In a typical few hours long EEG recording, there will be thousands of pseudo-stationary segments, in which the spectral bands have been computed. Classification of such large amount of data would be infeasible, and sometimes individual segments have to be selected for anlysis (Holthausen et al., 1999; Paul et al., 2003). However, the selected segments may not be representative of the whole recording. One way to obtain a compact representation of the whole EEG recording is to combine the spectral estimates computed in all the segments. A standard approach is to average spectra over segments. An advantage of such averaging of the spectra is that transient variations and artefacts, that affect the individual segments, become suppressed, so that a more robust representation of a patient's state can be obtained (Cooper et al., 2003; Victor et al., 2005; Kropotov, 2009).

Alternatively to using the FFT, spectral characteristics of EEG can be represented by coefficients of autoregression (Crowell et al., 1978, 1977), which are can be more robust to non-stationarity according to (Blinowska et al., 1981). However, the coefficients are not as clearly interpretable as the powers in spectral bands, and there is no guarantee that all important bands will be represented.

Another alternative is the discrete wavelet transform which provides a time-frequency representation of EEG. The discrete wavelet transform produces a

result similar to that of applying a series of band-pass filters. This transform enables analysing changes of frequency over time and can be useful for detecting transient pattens with known frequencies, such as the repetitive bursts in tracé alternant (Turnbull et al., 2001). However, to obtain a compact representation of an EEG recording, the results of discrete wavelet transform need to be processed further, as the data size of the transform is equal to that of the original signal.

Maturational changes of the powers in frequency bands have been explored in several studies (Bell et al., 1991b; Scher et al., 1995; Holthausen et al., 2000; Niemarkt et al., 2011). The absolute powers in Delta and Theta bands as well as the relative Delta, Alpha and Beta powers have been found correlated with maturation from pre-maturity to full-term. Surprisingly, there was no agreement between these studies about the direction of the correlation for some of the bands. For example, the groups of Scher and Holthausen found that the absolute Theta power increased with maturation. Contrary, Niemarkt et al. found that the Theta power decreased. Likewise, the absolute Delta power was found decreasing in two studies (Bell et al., 1991b; Niemarkt et al., 2011), whereas Holthausen et al. found that it was increasing.

The causes of this variability between studies remain unclear. It is possible that the powers could be affected by variations, especially when the number of subjects was small as in (Niemarkt et al., 2011). It is also possible that the relationship of powers and PCA is non-linear and the direction of correlation depends on the age range analysed. To obtain reliable results, the correlations need to be studied in various PCA groups and on a large set of recordings.

2.4.2 aEEG features

The amplitude-integrated EEG (aEEG) assists experts in assessing some of the maturity-related features, such as continuity, maximal amplitude and sleep cycle. The idea of aEEG is to present experts with a compact display of the EEG envelope, or the line connecting peaks in the signal. The detection of envelope is based on the principle of a smoothing capacitor, which is charged on the peaks and gradually discharged when no peaks are encountered. Consequently, the minimal and maximal amplitudes of the envelope reflect EEG amplitude as well as continuity. During continuous patterns, the envelope will stay "charged", whereas during discontinuous patterns the envelope will be "discharged" on inter-burst intervals. The longer the inter-burst intervals, the more will the envelope decrease in amplitude.

The minimal and maximal amplitudes of the compressed envelope make up the lower and upper borders of aEEG. The difference between the lower and

upper borders, analogous to the ripple of the smoothing capacitor, is called the aEEG bandwidth.

Within the aEEG technique, the envelope is presented in a time-compressed way to enable experts view the whole recording and assess cyclic variations in continuity and amplitude from the borders and bandwidth. Guides for aEEG interpretation describe the values which are characteristic to patterns with different levels of amplitude and continuity, including the normal maturational as well as pathological patterns (Thornberg and Thiringer, 1990; Olischar et al., 2004; Hellstrom-Westas et al., 2006, 2008). These values are typically measured manually by experts as there is no established technique for automated assessment.

Experts have observed that with advancing PCA the amplitude of the lower border of aEEG becomes elevated and the bandwidth decreases (Viniker et al., 1984; Thornberg and Thiringer, 1990). Burdjalov et al. (2003) proposed an aEEG-based scoring system for brain maturity assessment (Burdjalov score). The scoring system includes four components: continuity, sleep cycle, lower border amplitude and bandwidth. For the amplitude and bandwidth, thresholds were proposed, whereas the assessments of continuity and sleep cycle were qualitative. The total score was shown correlated with brain maturity between 24 and 39 weeks PCA.

Recently, Kato et al. (2011) evaluated results of aEEG-based maturity assessments on 129 recordings from newborns aged 27-37 weeks PCA. They compared the accuracies of the Burdjalov aEEG score and Parmelee coding system (Parmelee et al., 1968). The visual assessment based on Parmelee coding was found more accurate; the assessments were in the range of ± 2 weeks for 96.7% of recordings, whereas for the Burdjalov score 79.8% of the recordings were assessed within this range. Authors argued that the parameters of aEEG scores were limited in describing brain maturation, compared to visual assessment of patterns.

2.4.3 Continuity features

Most techniques for estimation of continuity are based on measuring the variability of maximal EEG amplitudes. The most widely known approach is to assess continuity from aEEG by manually measuring the lower border and bandwidth, as well qualitatively assessing the density of tracing.

To measure continuity automatically, it has been suggested to apply amplitude thresholds to segment the bursts and inter-burst intervals (Jennekens et al., 2011; West et al., 2011). A disadvantage of this approach is that a threshold needs to be adjusted for every recording.

Another approach is to evaluate the distribution of envelope amplitudes, as proposed in (Wong and Abdulla, 2008). Unlike in the aEEG, within this technique the envelope was detected as a vector of the mean amplitudes of pseudo-stationary segments. The distribution of the envelope amplitudes was then approximated with a lognormal distribution, and the mean and variance of the distribution were proposed as quantitative continuity features. In (Wong, 2008) it was shown that these features were correlated with PCA between 25 and 35 weeks. However the variability between subjects was high and the features could not be employed for classification of maturity.

Results presented in (Paul et al., 2003) reveal a promising new approach to estimating continuity, although this work was mainly focused on comparing frequency and variability features of newborn EEG during the quiet and active sleep. To extract the features, EEG are first segmented into pseudo-stationary intervals. An interesting observation was that durations of these intervals were informative for differentiation of the sleep states. Specifically, shorter segments were detected in the quiet sleep which is more discontinuous, whereas longer segments were found in the more continuous active sleep. This finding suggests that the lengths of pseudo-stationary intervals may be a promising feature to estimate continuity.

2.4.4 Other features

In addition to the standard frequency bands, other spectral characteristics are sometimes used for analysis of newborn EEG. Spectral edge frequency (SEF) is the frequency below which 95% of the total power of a signal is located. Bell et al. (1991a) found that SEF increases with maturation. Recently, West (2006) studied SEF in newborns aged 28–38 weeks and found the increase during weeks 28–33, after that SEF stabilised. The increase of SEF during the very pre-term PCA was explained by the development of low amplitude beta activity in the delta brushes.

Shannon entropy applied to the powers in spectral bands, or the spectral entropy, measures the peakedness or conversely uniformity of the spectrum. Spectral entropy tends to be lower for peaked and higher for uniform spectra. As the high powers, and consequently peaks, in EEG spectra usually correspond to low frequencies, decreased spectral entropy tends to be correlated with slow-wave patterns (Inouye et al., 1991). This feature has been used to estimate discontinuity within burst-suppression patterns during anaesthesia in adult patients (Vakkuri et al., 2004), however has not yet been applied to assess neonatal discontinuous patterns.

Korotchikova et al. (2009) compared the spectral entropy and SEF during quiet and active sleep of full-term newborns and found that both values are significantly higher during active sleep, which tends to be with high frequency and low amplitude. Based on this observation, spectral entropy is potentially useful feature to assess the development of the slow wave sleep pattern in full-term EEG.

Alternatively to using the conventional spectral representation, it has been recently attempted to apply the chaos theory to analysis of newborn EEG maturation. The idea of the approach is to consider the brain as a non-linear dynamical system with an attractor whose complexity increases with maturation (Scher et al., 2005). Janjarasjitt et al. (2008) estimated the complexity for EEG recorded from 50 healthy newborns aged 28–42 weeks PCA. The median complexity, counted over each 1 min EEG segment, was shown to increase with maturation approximately during weeks 31–36.

2.4.5 Detection of sleep states

Barlow et al. (1981) proposed the temporal profiles technique for detecting and encoding main patterns in EEG. The first step of this technique is to segment the EEG into pseudo-stationary intervals, usually a few seconds long. The intervals are then clustered based on their mean frequency and amplitude. To create a temporal profile, EEG segments are coded with the number of the closest cluster. The different states could then be assessed from their temporal profile.

More recently, temporal profiles have been employed to classify the active and quiet sleep states in newborn EEG (Paul et al., 2003; Krajca et al., 2009; Djordjevic et al., 2009). A potential weakness of the temporal profiles technique is that clustering of segments may produce different results depending on input data. Barlow et al. (1981) noted that the number of clusters had to be set by EEG experts after reviewing the results of clustering.

Alternatively to composing the temporal profiles from short segments, Piryatinska et al. (2009) proposed that each EEG sleep state can be viewed as a long pseudo-stationary segment. The quiet and active sleep states were distinguished based on the standard spectral bands as well the less widely used spectral characteristics including spectral entropy, SEF, and complexity estimates.

A similar extended set of features, aimed to represent diverse characteristics of the signal, has been employed to classify EEG segments into four labelled sleep and awake states as well as the pathological burst-suppression pattern (Lofhede et al., 2010). In total, 22 features, including statistics of power spectrum, amplitude distribution, entropy and cepstral coefficients, were counted in 1 sec sliding windows. Each of the resulting 22 "feature signals" was summarised

with four statistical parameters, so that each segment was finally represented by 88 "meta features". Selecting the most important of these features was the main problem of the approach, and the set of features for classification with linear discriminant analysis was optimised using a genetic algorithm. The quiet sleep and burst-suppression patterns were classified with accuracy of 93 and 100%. For the other patterns the accuracy was around 50%. One limitation of the approach is that the employed meta features, describing shapes of distributions of spectrum, cepstrum and entropy, are difficult to interpret for EEG experts.

2.4.6 Classification of brain maturity

Although various quantitative maturity-related features have been proposed, few attempts have been made to employ these features in computer-assisted assessments of newborn brain maturity. In one of the few such works, Crowell et al. (1978) used autoregression-based estimates of EEG spectrum as features for classification of brain maturity levels. They trained a logistic discriminant classifier to distinguish three PCA groups: 35 weeks or less, 40 weeks, and 46 weeks to 3 months. Note that the groups were taken with gaps of 4–6 weeks so that not all PCA weeks were used in training. The classifier was tested on two sets of EEG recorded at 40 weeks PCA, each set included approximately 50 recordings. From each recording, they analysed four-second segments taken from two different electrode channels. The average performances, counted as the portion of patients classified into the 40 weeks group, ranged between 72% and 96% for the two datasets. No relationship between the performance and the electrode channels has been found. The authors stressed that the high performance could be obtained because the spectral features in the three age groups were significantly different. Classification of a broader range of PCA has been proposed for future work, however no subsequent publications on this topic were found.

Holthausen et al. (1999) employed an Artifical Neural Network (ANN) to classify EEG recordings of 71 newborns. This study also used three age groups: 28 to 35, 36 to 40, and 41 to 100 weeks PCA. The spectral features were obtained with the FFT, and the average absolute powers within the frequency bands along with their variance and entropy were computed within 10 sec epochs. The performance of classification was counted within the leave-out cross validation, so that the ANN was trained on 69 patients and tested on the remaining two. The training and testing was run around 2000 times for each patient. As a result, the average performance over these runs was 96%. For seven of the newborns, the classification performance was lower than the average. Three of

four of these newborns were at high risk of apnoea, one was born extremely premature, and one recording was with abnormally low-voltage.

In a subsequent publication (Holthausen et al., 2000), the same authors analysed the importance of spectral powers for classification of EEG maturity in the PCA groups 28–35 and 36–112 weeks. The importance was estimated from the synaptic weights of ANN classifying the age groups. The most important features were the absolute Delta and Theta powers as well as the the ratios of Beta/Theta and Beta/Delta powers.

The above studies were aimed to distinguish a number of levels of brain maturity, each level included a range of weeks PCA. Therefore these methods may be inapplicable to detect a mismatch of two weeks, or to make weekly EEG assessments for a newborn. Classification of exact weeks PCA has been attempted by Schetinin and Schult (2005). An ANN technique has been employed to classify EEG segments from 65 newborns in 16 PCA groups from 35 to 51 weeks. As a result, 80% of segments have been classified correctly. However, segments from the same recording were used for training an for testing of the classifier, possibly enabling it to adapt to patient-specific variations. Training and testing a classifier on disjoint sets of patients is yet to be explored.

In general, problems with a large number of classes, such as maturity assessment, are more difficult to solve than two-class problems, because a boundary separating all classes needs to be learned, see e.g. (Bishop, 2007). In case of EEG classification, the boundary between classes becomes affected by overlapping of data samples, which creates additional difficulties for training. Some of the factors causing EEG data to overlap are artefact corruption, between-patient variations, shifts of sleep states and uncertainty in PCA estimates. To better handle the difficulties of multiclass EEG assessment, Schetinin and Schult (2005) employed a pairwise classification system, converting a multiclass problem into a set of two-class tasks. The pairwise approach has been shown promising to deal with multiple classes (Friedman, 1996; Hastie and Tibshirani, 1998), particularly when the class labels are naturally ordered (Frank and Hall, 2001; Fürnkranz, 2002) as in the maturity assessment problem. We will explore the pairwise approach further in this thesis.

Another important limitation of conventional methods of EEG maturity classification is that they are based on learning a single model from a given set of data and cannot provide accurate estimates of the uncertainty in assessments. Such estimates are required to count risk of brain dysmaturity for each patient. Bayesian classification enables the uncertainty to be accurately estimated by averaging over areas of high densities of the likelihood (Chipman et al., 1998; Denison et al., 2002). This motivates us to employ the Bayesian methodol-

ogy for classification of brain maturity to allow experts to obtain exhaustive information on uncertainty or risks in EEG assessment.

2.5 EEG data

The newborn EEG dataset used in this research includes 1993 recordings made at different ages. The EEG have been recorded at the main German university hospitals, and the dataset has been collected by the University of Jena, Germany.

The distribution of recordings over weeks PCA is shown in Fig. 2.7. Weeks 36-46 take up the largest part of this set (64%) and include at least 100 recordings each, so that it becomes possible to learn classification models from this age group. On the contrary, from 26 to 35 weeks, few recordings are available, and so we cannot expect to learn to reliably classify EEG in this age group.

The recordings were made during sleep hours of newborns with 2 channels C3-T3 and C4-T4. The durations of recordings varied between 2 and 10 hours, and the average was 4 hours. The recordings were available unprocessed. The sampling rate was 100 Hz and the range of amplitudes was -128 to 127. Experts have viewed the recordings and provided artefact markings as separate files. Additionally, for each 10-sec segment, 72 spectral features have been computed. The first 36 features comprise the average powers in the six frequency bands for the two channels and their sum, the other half of the features are the variances of the same measurements.

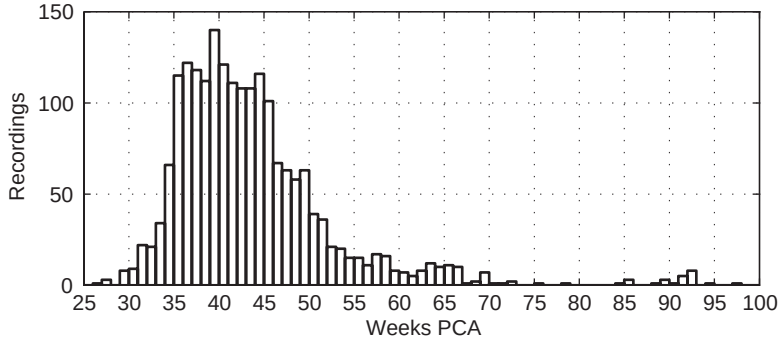

Figure 2.7: Numbers of recordings in each week.

2.6 Summary

This chapter reviewed the existing approaches to newborn EEG maturity assessment. We described the maturational features typically analysed by experts, namely, continuity, frequency, and sleep cycle. By analysing these patterns experts can estimate a newborn's EEG age and compare it to the physiological age to assess brain development. The accuracy of EEG-estimated age normally is within an interval of ± 2 weeks of a newborn's stated age. If the mismatch in ages exceeds this interval, the brain maturity may be abnormal. In cases of severe mismatch, abnormal maturity may be caused by brain injuries. However, experts have also found that abnormal EEG maturation is often associated with pre-term birth, family history of sudden infant death and high incidence of sleep apnoea.

Assessment of brain maturity requires a high level of skills and experience in recognising and scoring the EEG patterns, which vary between patients and change over the weeks PCA, so that the analysis becomes laborious and costly. In the absence of standard rules for interpretation of maturational patterns the assessment results may be subjective.

Computer-based assessment technologies have been proposed to assist experts with assessments. Within these approaches, the EEG data are first represented by spectral features, and then classified into up to three predefined levels of maturation. These approaches may be unsuitable for making assessments in the typical range of ± 2 weeks. Development and testing of a technology capable of making such assessments will be explored in this thesis.

Another area to be explored is the use of features representing EEG continuity in automated classification of brain maturity. Current approaches are limited to employing the spectral powers in the standard EEG frequency bands. These features cannot adequately represent the information on continuity which is one of the most important maturational characteristics. The extraction and use of continuity features will be explored in Chapter 6.

An important limitation of the existing methods of automated assessment of EEG maturity is that they are based on learning a single classification model, and so cannot provide accurate estimates of the uncertainty in assessments. Such estimates are required to count the risk of brain dysmaturity for each patient.

To allow experts to obtain exhaustive information on uncertainty or risks in EEG assessment, we propose to employ Bayesian methodology for classification of brain maturity. Within this methodology the uncertainty in assessments is accurately estimated by averaging over areas of high densities of the likelihood.

Chapter 3

Bayesian Model Averaging

In this chapter, first we introduce the methodology of Bayesian inference and then discuss how this methodology can be used for averaging over decision tree models. Second, we introduce Markov chain Monte Carlo method which makes the Bayesian methodology computationally efficient. Lastly, we discuss problems with the implementation of this method for Bayesian averaging over decision tree models.

3.1 Introduction

Let us consider a problem of inference from data that are represented by a set of data points which were assigned to one of C categories or classes. Each data point is represented by an m-element *feature vector* or *input* $\mathbf{x} = (x_1, \ldots, x_m)$ and has a *class label* $y \in \{1, C\}$. Therefore, we can consider a data set $\mathcal{D}$ that consists of pairs $(\mathbf{x}, y)$. Our problem is now to learn an *inference rule* or *model* that allows us to predict the class y for a given input $\mathbf{x}$. Such an inference model can be learnt from data $\mathcal{D}$ within the probabilistic framework (MacKay, 1998; Hoeting et al., 1999; Duda et al., 2001; Bishop, 2007; Kruschke, 2011).

Using this framework, we can calculate the *prior* probabilities of classes, $P(y = 1), \ldots, P(y = C)$, such that $\sum_{i=1}^{C} P(y = i) = 1$. Value of $P(y = i)$ is a probability that a given input $\mathbf{x}$ belongs to the ith class. When the numbers of data points, n_i, included in class y are known, then the probability $P(y)$ are

$$P(y = i) = \frac{n_i}{\sum_{i=1}^{C} n_i}. \tag{3.1}$$

Observing values of $\mathbf{x}$ for different classes, we can see that the distribution of $\mathbf{x}$ depends on the class y. We therefore can consider a *class-conditional* probability distribution or density function $p(\mathbf{x}|y)$ for a given class y.

When we observe that an input vector has value $\mathbf{x}$ and that it is assigned to the class y, we can consider a *joint* probability density $p(\mathbf{x}, y)$. Taking into account that the prior probabilities $P(y)$ and the class-conditional densities $p(\mathbf{x}|y)$ are known, the densities $p(\mathbf{x}, y)$ are written as follows:

$$p(\mathbf{x}, y) = p(\mathbf{x}|y)P(y). \tag{3.2}$$

At the same time, we can consider the probability of the class y given that an input vector has value $p(\mathbf{x})$. This is the *class posterior* probability $P(y|\mathbf{x})$ which is defined as

$$P(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})}. \tag{3.3}$$

The distribution $p(\mathbf{x})$ in Eq. 3.3 is the *evidence* factor which is the sum of joint density $p(\mathbf{x}, y)$ over all classes $y = 1, \ldots, C$:

$$p(\mathbf{x}) = \sum_{i=1}^{C} p(\mathbf{x}, y = i). \tag{3.4}$$

This summation is called *marginalization* over y, and and the resultant distribution $p(\mathbf{x})$ is called the *marginal* probability density.

It is important to note that both equations Eq. 3.2 and Eq. 3.3 include the joint probability density $p(\mathbf{x}, y)$. Therefore we can rewrite these equations in the following form:

$$p(\mathbf{x}|y)P(y) = P(y|\mathbf{x})p(\mathbf{x}). \tag{3.5}$$

From this equality, we observe that when the densities $p(\mathbf{x}|y)$ and $p(\mathbf{x})$ and the probability $P(y)$ are known, the class posterior probability $P(y|\mathbf{x})$ can be written as

$$P(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)P(y)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y)P(y)}{\sum_{i=1}^{C} p(\mathbf{x}|y = i)P(y = i)}. \tag{3.6}$$

This is *Bayes' formula* which allows us to calculate the class posterior probability $P(y|\mathbf{x})$ when we observe an input vector x for the given densities $p(\mathbf{x})$ and $p(\mathbf{x}|y)$ and probabilities $P(y)$.

In Bayes' context, density $p(\mathbf{x}|y)$ is called the *likelihood* of y with respect to $\mathbf{x}$. This term is used to indicate that the larger the likelihood $p(\mathbf{x}|y)$, the more likely that the point $\mathbf{x}$ belongs to the true class (Duda et al., 2001).

The evidence factor $p(\mathbf{x})$ given by Eq. 3.4 is the denominator in Bayes' formula 3.6. It is typically interpreted as a scale factor that allows the class posterior probabilities $P(y|\mathbf{x})$ to be normalised such that $\sum_{i=1}^{C} P(y = i|\mathbf{x}) = 1$.

Given an input $\mathbf{x}$, the class posterior probabilities $P(y|\mathbf{x})$ are calculated for each class $y = 1, \ldots, C$. The given $\mathbf{x}$ is assigned to the class c^* which has the largest probability $P(y|\mathbf{x})$. This decision is made according to the *Bayesian decision rule*:

$$c^* = \max_{1 \leq i \leq C}(P(y = i|\mathbf{x})). \tag{3.7}$$

This rule assigns a given input $\mathbf{x}$ to the true class $y = c^*$ with the largest probability. The input is assigned to a false class $y = i : i \neq c^*$ with a smaller probability; such a decision is called *misclassification*.

The consequences of misclassification can be serious and when this is the issue, we consider a *cost function* λ. Such a function can assign a unit loss to any error:

$$\lambda(y = i|y = j) = \begin{cases} 0, & \text{if } i = j\,, \\ 1, & \text{if } i \neq j\,. \end{cases}$$

Given a cost function λ, the conditional *risk* R is written as follows

$$\begin{aligned} R(y = i|\mathbf{x}) &= \sum_{j=1}^{C} \lambda(y = i|y = j) P(y = j|\mathbf{x}) \\ &= \sum_{j \neq i} P(y = j|\mathbf{x}) = 1 - P(y = i|\mathbf{x}), \end{aligned} \tag{3.8}$$

where $P(y|\mathbf{x})$ are the probabilities calculated by Bayes' formula 3.6 for the given input $\mathbf{x}$.

The notations introduced above allow us to analyse models which can be learnt from data in the Bayesian framework. Next we consider the problem of model comparison.

3.2 Bayesian model comparison

It is often that we can learn a number of suitable models from given data. In such cases, we can define a set of such models, $\mathcal{M}_i, i = 1, \ldots, L$, and then compare them in terms of fitness to the given data $\mathcal{D}$. To make such comparison, we can consider the *prior probability* density $p(\mathcal{M}_i)$ that $\mathcal{M}_i$ is the true model which is capable of generating the data $\mathcal{D}$, see e.g. (Hoeting et al., 1999).

Observing these models in this light, we can define the *posterior densities* $p(\mathcal{M}_i|\mathcal{D})$ as:

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)P(\mathcal{M}_i)}{p(\mathcal{D})}. \tag{3.9}$$

Here $p(\mathcal{D}|\mathcal{M}_i)$ is the probability density of observing the data $\mathcal{D}$ given that a model $\mathcal{M}_i$ is true. This density is called the *evidence* for model $\mathcal{M}_i$, see e.g. (Duda et al., 2001), or the *marginal likelihood*, see e.g. (Denison et al., 2002).

The above evidence $p(\mathcal{D}|\mathcal{M}_i)$ can include all available information about a model $\mathcal{M}_i$ including its parameter Θ. Therefore we can consider the posterior distribution of the model parameter Θ for a given model $\mathcal{M}_i$, $p(\Theta|\mathcal{D}, \mathcal{M}_i)$. In this case the evidence is determined by integrating of $p(\mathcal{D}, |\Theta, \mathcal{M}_i)$ over Θ:

$$p(\mathcal{D}|\mathcal{M}_i) = \int_{\Theta} p(\mathcal{D}|\Theta, \mathcal{M}_i) p(\Theta|\mathcal{D}, \mathcal{M}_i) d\Theta. \tag{3.10}$$

For comparison of competitive models $\mathcal{M}_i$ and $\mathcal{M}_j$, we can use the posterior densities $p(\mathcal{M}_i|\mathcal{D})$ written as a *posterior odds ratio*:

$$\frac{p(\mathcal{M}_i|\mathcal{D})}{p(\mathcal{M}_j|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)} \times \frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)}. \tag{3.11}$$

There exist $\binom{L}{2}$ pairs of models $\mathcal{M}_i$ and $\mathcal{M}_j, i \neq j, j = 1, \ldots, L$. The best model has the highest odds ratio, larger than one.

For a given model $\mathcal{M}(\Theta)$, we can find the parameter Θ' that maximises the odds ratio respect to other model parameters Θ:

$$\frac{p(\mathcal{M}|\mathcal{D}, \Theta')}{p(\mathcal{M}|\mathcal{D}, \Theta)} \geq 1. \tag{3.12}$$

The model $\mathcal{M}(\Theta')$ is called the maximum *a posteriori* model (Duda et al., 2001).

3.3 Bayesian learning

The desired class posterior probability $P(y|\mathbf{x})$ is calculated by the Bayes' formula 3.6 for the given prior probabilities $P(y)$ and the class-conditional densities $p(\mathbf{x}|y)$, $y \in \{1, C\}$. The prior probabilities $P(y)$ can be given by a domain expert or calculated by Eq. 3.1. However, the class-conditional densities $p(\mathbf{x}|y)$ are often unknown and their estimation can be the main problem, see e.g. (Duda et al., 2001).

One way to estimate the densities $p(\mathbf{x}|y)$ is to use all available data $\mathcal{D}$ that include the pairs $(\mathbf{x}, y)$. Using the class labels $y = 1, \ldots, C$, we can consider the given data $\mathcal{D}$ as a set of independent subsets $\mathcal{D}_1, \ldots, \mathcal{D}_C$. This will allow us to simplify the notation.

Observing the data $\mathcal{D}$, the desired class posterior probability $P(y|\mathbf{x})$ becomes conditional on $\mathcal{D}$, that is denoted as $P(y|\mathbf{x}, \mathcal{D})$. Then Bayes' formula 3.6 can be

rewritten as follows:

$$P(y = i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|y = i, \mathcal{D}_i)P(y = i)}{\sum_{j=1}^{C} p(\mathbf{x}|y = j, \mathcal{D}_j)P(y = j)}. \tag{3.13}$$

As the formula is calculated separately for each class $y = i$ and subset $\mathcal{D}_i$, we can simplify the notation of $p(\mathbf{x}|y, \mathcal{D}_i)$ by omitting the term y and use $p(\mathbf{x}|\mathcal{D})$ instead. Further we can consider parameters of density function $p(\mathbf{x}|\mathcal{D})$ and denote them as an unknown parameter vector Θ which has to be fitted to the data.

In the Bayesian context, we consider the parameter Θ as a *random variable* with a prior distribution $p(\Theta)$. When we observe the available data $\mathcal{D}$, this distribution is considered as a posterior density $p(\Theta|\mathcal{D})$.

Observing values of $\mathbf{x}$ and Θ in the data, we can consider a joint posterior density $p(\mathbf{x}, \Theta|\mathcal{D})$. Then the integration of this density over Θ allows us to define the desired density $p(\mathbf{x}|\mathcal{D})$ as:

$$p(\mathbf{x}|\mathcal{D}) = \int_{\Theta} p(\mathbf{x}, \Theta|\mathcal{D}) d\Theta. \tag{3.14}$$

According to the definition of the joint probability, the density $p(\mathbf{x}, \Theta|\mathcal{D}) = p(\mathbf{x}|\Theta, \mathcal{D})p(\Theta|\mathcal{D})$. Here both functions $p(\mathbf{x}|\Theta, \mathcal{D})$ and $p(\Theta|\mathcal{D})$ are conditioned on the same data $\mathcal{D}$, and so we can omit one of these conditions. This allows us to rewrite the function $p(\mathbf{x}|\Theta, \mathcal{D}) = p(\mathbf{x}|\Theta)$, and finally the density $p(\mathbf{x}|\mathcal{D})$ can be written as:

$$p(\mathbf{x}|\mathcal{D}) = \int_{\Theta} p(\mathbf{x}|\Theta)p(\Theta|\mathcal{D}) d\Theta. \tag{3.15}$$

The integral in Eq. 3.15, and therefore Bayes' formula 3.13, are analytically tractable for cases when prior and likelihood distributions are given as the *conjugate* functions, see e.g. (Denison et al., 2002; Bishop, 2007). The numerical integration over parameter space Θ with the common techniques is limited and becomes computationally infeasible in a high dimensional space.

However, there is a family of widely known *Monte Carlo* methods which have been developed for such calculations. In our case, the posterior distribution $p(\Theta|\mathcal{D})$ cannot be directly simulated. However we can use the Monte Carlo method to generate random samples in order to approximate the desired distribution. The classical Monte Carlo method has been extended to *Markov chain Monte Carlo* (MCMC) in order to avoid limitations and expand areas of applications, see e.g. (MacKay, 1998; Robert and Casella, 2004; Robert, 2007; Webb et al., 2011).

3.4 Markov chain Monte Carlo method

The desired class (or predictive) posterior distribution $p(y|\mathbf{x}, \mathcal{D})$ is calculated by Bayes' formula 3.13 for given input $\mathbf{x}$, densities $p(\mathbf{x}|y, \mathcal{D}_i)$, $p(y)$, and $p(\mathbf{x}|\mathcal{D})$. The class-conditional densities $p(\mathbf{x}|y, \mathcal{D}_i)$ are determined by Eq. 3.15, where the term y has been omitted for simplicity as described in the above section. In cases when the available data are representative, the class posterior density $p(y|\mathbf{x}, \mathcal{D})$ becomes independent from $\mathcal{D}$, and so can be replaced by $p(y|\mathbf{x})$. However, we keep the full notation $p(y|\mathbf{x}, \mathcal{D})$.

These notations allows us to rewrite the desired distribution $p(y|\mathbf{x}, \mathcal{D})$ in the following form:

$$\begin{aligned} p(y|\mathbf{x}, \mathcal{D}) &= \frac{p(\mathbf{x}|y, \mathcal{D})p(y)}{p(\mathbf{x}|\mathcal{D})} \\ &= \frac{\left(\int_\Theta p(\mathbf{x}|y, \Theta)p(\Theta|\mathcal{D})d(\Theta)\right)p(y)}{p(\mathbf{x}|\mathcal{D})} \\ &= \int_\Theta p(y|\mathbf{x}, \Theta, \mathcal{D})p(\Theta|\mathcal{D})d(\Theta). \end{aligned} \tag{3.16}$$

The posterior distribution $p(\Theta|\mathcal{D})$ in this equation cannot be evaluated as noted in the above section. However, we can sample parameters Θ from the posterior distribution $p(\Theta|\mathcal{D})$ in order to approximate the integral 3.16 that determines the desired class posterior distribution $p(y|\mathbf{x}, \mathcal{D})$. Having drawn the samples $\Theta^{(1)}, \ldots, \Theta^{(N)}$, we can rewrite the class posterior $p(y|\mathbf{x}, \mathcal{D})$ as follows:

$$\begin{aligned} p(y|\mathbf{x}, \mathcal{D}) &= \int_\Theta p(y|\mathbf{x}, \Theta, \mathcal{D})p(\Theta|\mathcal{D})d(\Theta) \\ &\approx \sum_{i=1}^{N} p(y|\mathbf{x}, \Theta^{(i)}, \mathcal{D})p(\Theta^{(i)}|\mathcal{D}) \\ &= \frac{1}{N}\sum_{i=1}^{N} p(y|\mathbf{x}, \Theta^{(i)}, \mathcal{D}). \end{aligned} \tag{3.17}$$

Such approximation of integrals is known as the Monte Carlo method, see e.g. (MacKay, 1998; Denison et al., 2002; Robert and Casella, 2004). The accuracy of this method is increased with the number of samples N.

The Monte Carlo method has been extended to the MCMC simulation method in order to draw $\Theta^{(i)}$ from the posterior distribution $p(\Theta|\mathcal{D})$ when the information about this distribution is limited. In our case, the main idea of MCMC method can be defined as follows.

A sequence of random samples $\Theta^{(0)}, \Theta^{(1)}, \ldots, \Theta^{(k)}$ is a *Markov chain* such that the conditional density of $\Theta^{(k)}$ given the past samples $\Theta^{(k-1)}, \ldots, \Theta^{(1)}, \Theta^{(0)}$

depends only on the state $\Theta^{(k-1)}$. This transition density is defined as q:

$$q(\Theta^{(k)}|\Theta^{(k-1)}, \ldots, \Theta^{(1)}, \Theta^{(0)}) = q(\Theta^{(k)}|\Theta^{(k-1)}). \tag{3.18}$$

The Markov chain posses the following properties.

- It is said a Markov chain is *stationary* if there exists a density $f(\Theta)$ such that if $\Theta^{(k)} \sim f(\Theta)$, then $\Theta^{(k+1)} \sim f(\Theta)$.
- A transition density q is designed so that to make moves over all possible states of Θ.
- A transition density q is called *irreducible* if for any given $\Theta^{(0)}$ a Markov chain reaches any area of the parameter space Θ with a non-zero probability, that is $q(\Theta^{(k)}, \cdot) > 0$.
- Having a stationary density, a Markov chain can return to any arbitrary state an infinite number of times. Such chains are called *recurrent*.
- If a transition density q generates a Markov chain $\Theta^{(1)}, \ldots, \Theta^{(n)}$ with a stationary density, then according to the Law of Large Numbers the average of $p(y|\mathbf{x}, \Theta^{(i)}, \mathcal{D})$ over $\Theta^{(i)}$ converges to the integral $\int_\Theta p(y|\mathbf{x}, \Theta, \mathcal{D})p(\Theta|\mathcal{D})d(\Theta)$.

The use of these properties allows us to generate random samples $\Theta^{(i)}$ from a posterior distribution $p(\Theta|\mathcal{D})$ by running a Markov chain which has achieved a stationary distribution $f(\Theta)$. The samples generated by the Markov chain should be omitted during so-called *burn-in* phase when its distribution is non-stationary.

3.5 Bayesian decision tree models

3.5.1 Decision tree models

Classification or Decision Tree (DT) models are multilevel hierarchical structures consisting of *splitting* and *terminal* nodes, see e.g. (Breiman et al., 1984; Buntine, 1998). The *root* node at the first hierarchical level of a DT model allocates a given input $\mathbf{x}$ to one of the nodes at the next level. The allocation is made until the input $\mathbf{x}$ falls into a terminal node which finally assigns the input to one of the given classes C.

The *size* of a DT model is defined by the number of terminal nodes k, and the number of splitting nodes is equal to $k - 1$. The number of possible configurations, S_k, is defined by the number of nodes, k, in a DT accordingly to the *Catalan number*:

$$S_k = \frac{1}{k+1}\binom{2k}{k}. \tag{3.19}$$

This number exponentially grows with the size k; for example, $S_{k=5} = 42, S_{k=10} = 16,796$, etc.

Each splitting node has the attributes (p, v, r), where $p \in \{1, k-1\}$ is the position of the node in the DT, $v \in \{x_1, \ldots, x_m\}$ is the predictor variable or feature employed by the pth node for splitting, and $r \in \{v^{min}, v^{max}\}$ is the rule or threshold used for splitting at this node, min and max denote the minimum and maximum of variable v, respectively. The structure of a DT model is described by a sequence of the splitting nodes $\{p_i, v_i, r_i\}_{i=1}^{k-1}$, which determines the model parameter Θ.

Figure 3.1 shows an example of a DT model consisting of two splitting nodes s_1 and s_2, and three terminal nodes t_1, t_2, and t_3. The first node s_1 splits the entire data into two disjoint subsets so that data samples from one subset fall into node s_2 via the left branch, and samples from the other subset fall into the terminal node t_2 via the right branch. The node s_2 further partitions the data samples which fall into the terminals t_2 or t_3 via the left and right branches. Finally, one of the terminal nodes assigns the given input to one of the given classes.

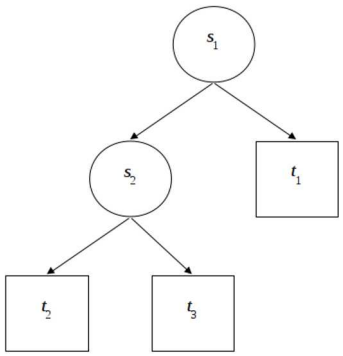

Figure 3.1: An example of DT model with two splitting nodes s_1, s_2 and three terminal nodes t_1, t_2, and t_3.

The parameter Θ of a DT model is learnt from a given set of labelled data samples, $\mathcal{D}$. The size of a DT model learned from the data depends on a minimal number of data samples, p_{min}, which are allowed to be in the terminal nodes. Setting a smaller number p_{min} increases the DT size, and *vice versa* setting a larger number decreases the size.

It is often that the structure of a DT is unknown, and we need to grow DT models of a reasonable size to achieve the maximal accuracy of predicting unseen data that have not been included in the set of labelled data. Such an

ability of models is called *generalisation*. To grow such DT models we can grow DT models with various numbers p_{min} and then select the most suitable DT model.

When DT models learn from data, the attributes of nodes, (p, v, r), are changed within the given *priors*, which are defined as available information on a specified attribute. In the Bayesian context, we define these priors as follows.

In the absence of information on the importance of the variables $x_1, \ldots, x_m$ used for prediction, a variable v which is assigned to the ith splitting node is randomly drawn from a set of variables $x_1, \ldots, x_m$; it is said a variable v is drawn from the uniform discrete distribution U: $v \sim U(1, m)$. Similarly, a rule r can be drawn from the uniform discrete distribution of the vth variable: $r \sim U(v^{min}, v^{max})$.

Having defined the parameter Θ of a DT model, we can calculate the class posterior probability with which a given input $\mathbf{x}$ is assigned by a terminal node to a class y. The samples of Θ are generated by the MCMC simulation method to calculate the desired class posterior density $p(y|\mathbf{x}, \mathcal{D})$ by using Eq. 3.17.

3.5.2 MCMC integration

When the Markov chain becomes stationary, we collect samples $\Theta^{(1)}, \ldots, \Theta^{(N)}$ to approximate the class posterior distribution $p(y|\mathbf{x}, \mathcal{D})$ determined by Eq. 3.17. In cases when data $\mathcal{D}$ are representative, the class posterior density becomes independent from the data and we can rewrite this equation as follows:

$$\begin{aligned} p(y|\mathbf{x}) &\approx \sum_{i=1}^{N} p(y|\mathbf{x}, \Theta^{(i)}) p(\Theta^{(i)}) \\ &= \frac{1}{N} \sum_{i=1}^{N} p(y|\mathbf{x}, \Theta^{(i)}). \end{aligned} \tag{3.20}$$

The desired approximation of the posterior distribution $p(y|\mathbf{x})$ is achieved when the samples $\Theta^{(i)}$ are drawn from a stable Markov chain, and the number of samples, N, is sufficiently large.

Having specified priors on DT models, we can consider algorithms for MCMC simulation. The most general form of MCMC simulation is known as *Metropolis-Hastings* (MH) algorithm or sampler, see e.g. (Robert and Casella, 2004; Denison et al., 2002).

The MH algorithm requires to define the *proposal density* q for updating model parameter Θ. Proposals Θ^p are made dependent on the current state Θ_i. Proposal Θ^p drawn from a given proposal density q is always accepted if the model likelihood $p(\mathcal{D}|\Theta^p)$ is greater than that of the current model, $p(\mathcal{D}|\Theta)$.

Otherwise the proposal is accepted with an *acceptance probability* α which is determined by the ratio:

$$\alpha = min\left(1, \frac{p(\mathcal{D}|\Theta^p)p(\Theta^p)q(\Theta, \Theta^p)}{p(\mathcal{D}|\Theta)p(\Theta)q(\Theta^p, \Theta)}\right). \tag{3.21}$$

Here the marginal likelihood $p(\mathcal{D}|\Theta)$ is defined as the probability density of observing the data $\mathcal{D}$ given the model parameter Θ, that can be written as:

$$p(\mathcal{D}|\Theta) = \frac{p(\Theta|\mathcal{D})p(\mathcal{D})}{p(\Theta)}. \tag{3.22}$$

Particularly, when the costs of misclassification, α, are equal for each class, the above marginal likelihood for DT models with the k splitting nodes is written as follows (Denison et al., 2002):

$$p(\mathcal{D}|\Theta) = \left[\frac{\Gamma(\alpha C)}{\Gamma(\alpha)^C}\right]^k \prod_{i=1}^{k} \frac{\prod_{j=1}^{C}\Gamma(m_{ij}+\alpha_j)}{\Gamma(n_i + \sum_{j=1}^{C}\alpha_j)}, \tag{3.23}$$

where n_i is the number of data points fallen in the ith terminal node, m_{ij} is the number of data points of the jth class in this node, and Γ denotes the Gamma function.

The number of splitting nodes is typically unknown and the desired DT models must be grown to a proper size in order to provide the best generalisation as discussed above. However, when the size of DT models varies, the MCMC integration must be done over a parameter space Θ of variable dimensionality. It is expected to explore as many as possible of DT configurations given by Eq. 3.19. For such cases, MCMC is extended by Reversible Jump (RJ) proposed in (Green, 1995).

3.5.3 Reversible jump MCMC

It is important to note that the configurations of DT models with different numbers of splitting nodes, k, have to be explored in the same proportions – that is, the samples from the posterior $p(\Theta|\mathcal{D})$ have to be collected in the proportions to the numbers of S_k.

The integration over DT models of variable size is achieved by using the *birth*, *death*, *change-split*, and *change-rule* moves (Chipman et al., 1998; Denison et al., 2002). The first two moves, birth and death, reversibly change the dimensionality of a parameter space Θ, whilst the third and fourth moves change the model parameters within a current dimensionality. These moves are as follow:

Birth move randomly splits the data points falling in one of the DT terminal nodes by inserting a new splitting node with a variable and rule drawn from the given priors.

Death move randomly picks a DT splitting node with two terminal splits and then assigns it to be one terminal node with the united data points.

Change-split move randomly picks a splitting node and assigns it to be with a new splitting variable and rule drawn from the given priors.

Change-rule move randomly picks a splitting node and assign it to be with a new rule drawn from the given prior.

We can see that the birth move adds a new splitting node and the number of k increases by one. On the contrary, the death move unites the two terminal nodes that decreases the k by one.

The change-split move, assigning a new splitting variable, can make a potentially large change in the parameters that increases the chance to properly sample areas of interest with high posterior. In contrast, the change-rule move makes a small change which is required for detailed exploration of a surrounding area.

When there is no prior information on DT models, MCMC algorithm starts to explore a DT model consisting of one splitting node. Making the above moves, a DT model is grown in the size and its parameters Θ are changed so that to increase the likelihood of the model.

The likelihood is gradually increased and then becomes oscillating around a stable value. This phase is named *burn-in* and must be preset sufficiently long in order to achieve a stationary distribution $p(\Theta|\mathcal{D})$. During the second phase called *post burn-in*, the samples of a random variable with this distribution, $\Theta^{(1)}, \ldots, \Theta^{(N)}$, are collected to approximate the desired class posterior distribution $p(y|\mathbf{x}, \Theta)$.

3.5.4 Implementation of RJ MCMC

The above moves are proposed randomly with the proposal probabilities given for the birth, death, change-split, and change-rule moves. The values of these probabilities are dependent on the complexity of the problem – more complex problems require larger DT models, and so the MCMC algorithm has to change the dimensionality more frequently. However, there is no guidance for setting the proposal probabilities, and their proper values have to be found empirically (Chipman et al., 1998; Denison et al., 2002).

When one of the moves change the dimensionality of a DT model, the Markov chain has to remain reversible in order to ensure the integration over all areas

of high posterior density $p(\Theta|\mathcal{D})$ in a parameter space. The effect of variable dimensionality can be accounted for with a proposal ratio R inserted in the acceptance probability α as follows:

$$\alpha = min\left(1, \frac{p(\mathcal{D}|\Theta^p)}{p(\mathcal{D}|\Theta)} \times R\right), \tag{3.24}$$

When the birth or death move changes the dimensionality, the corresponding ratio R_b or R_d are calculated as follows. First, we define a conditional probability distribution, $q(\Theta^p|\Theta)$, that a DT model with the current vector Θ is moved to a proposed vector Θ^p. Similarly, we can define a density of a reverse move, $q(\Theta|\Theta^p)$. Then the desired reversibility of a Markov chain is kept if these densities are equal:

$$q(\Theta^p|\Theta) = q(\Theta|\Theta^p). \tag{3.25}$$

For the birth moves, the Θ^p is the $(k+1)$-dimensional vector, and the number of DT configurations is $S_{k+1} : S_{k+1} > S_k$. Therefore the Markov chain is kept reversible when the ratio R_b is written as follows:

$$R_b = \frac{q(\Theta|\Theta^p)p(\Theta^p)}{q(\Theta^p|\Theta)p(\Theta)}, \tag{3.26}$$

where $p(\Theta)$ and $p(\Theta^p)$ are the prior densities of parameters Θ and Θ^p, respectively.

When we assume that all the configurations of a DT with k terminal nodes are equally likely, the prior density $p(\Theta)$ is written as follows:

$$p(\Theta) = \left(\prod_{i=1}^{k-1} p(s_i^{rule}|s_i^{var})p(s_i^{var})\right) p(\{s_i^{pos}\}_1^{k-1}). \tag{3.27}$$

Here s_i^{var} is the predictor of the ith splitting node which is drawn from the uniform distribution, $s_i^{var} \sim U(1, m)$. The $p(s_i^{rule}|s_i^{var})$ denotes the conditional density of the rule s_i^{rule} given the predictor s_i^{var}, so that $s_i^{rule} \sim U(X_1, X_n)$, where X_i are the values of the variable s_i^{var}, and n is the number of data samples that represent the variable. The $\{s_i^{pos}\}_1^{k-1}$ denotes the set of $(k-1)$ splitting nodes a DT model consists of. For a DT model with k splitting nodes there are S_k combinations given by Eq. 3.19, and therefore the probability of such a DT model is $p(\{s_i^{pos}\}_1^{k-1}) = 1/S_k$, when there are no preferences on the number of splitting nodes.

Taking the above notations, we can rewrite Eq. 3.27 as follows:

$$p(\Theta) = \left(\prod_{i=1}^{k-1} \frac{1}{N(s_i^{var})}\frac{1}{m}\right) \frac{k!}{S_k}\frac{1}{K}, \tag{3.28}$$

where $N(s_i^{var})$ is the number of possible rules for the variable s_i^{var}, the factorial $k!$ denotes the number of all possible configurations of DT with k terminal nodes, and $K \geq k$ is the maximal number of terminal nodes.

Using these notations, the proposal distribution $q(\Theta^p|\Theta)$ can be ultimately written as follows:

$$q(\Theta^p|\Theta) = \frac{b_k}{k} \frac{1}{N(s_i^{var})} \frac{1}{m}, \tag{3.29}$$

where b_k is the given proposal probability of the birth move.

According to this equation, a new splitting node is made from one of k terminal nodes, chosen with a probability $1/k$. This node gets the variable s_i^{var} drawn randomly from the m variables, $s_i^{var} \sim U(1, m)$. It gets also the rule s_i^{rule} drawn randomly from the $N(s_i^{var})$ possible values of this variable, $s_i^{rule} \sim U(X_1, X_n)$.

Likewise, we can write the proposal distribution for the reverse (death) move:

$$q(\Theta|\Theta^p) = \frac{d_{k+1}}{D_Q}, \tag{3.30}$$

where d_{k+1} is the given proposal probability of the death move, and D_Q is the number of splitting nodes both branches of which are are the two terminal nodes.

Using the above notations, finally we can rewrite Eq. 3.26 for the desired ratio R_b as follows:

$$R_b = \frac{d_{k+1}}{b_k} \frac{k}{D_Q} \frac{S_k}{S_{k+1}}. \tag{3.31}$$

Here $D_Q \leq k$ and $S_k \leq S_{k+1}$, the probabilities d_{k+1} and b_k can be set equal. Then the ratio $R_b : 0 < R_b < 1$.

Likewise, the proposal ratio for the death moves, R_d, is written as follows:

$$R_d = \frac{b_k}{d_{k-1}} \frac{D_Q}{k-1} \frac{S_k}{S_{k-1}}, \tag{3.32}$$

and $R_d > 1$ under the above conditions.

We see that the acceptance probability defined by Eq. 3.24 depends on the two factors: first on the type of moves (birth or death) and second on the size of a DT. For the birth move the number of configurations of a DT with $(k+1)$ terminal nodes increases, and so the ratio becomes smaller than 1, $R < 1$. On the contrary, the death move decreases the number of combinations, and $R > 1$. Such variations in R allow the reversibility of a Markov chain to be kept during MCMC integration over a model parameter space of variable dimensionality (Denison et al., 2002).

3.5.5 Problems with RJ MCMC implementation

As DT models are hierarchical structures, changes at a node located at the upper levels close to the root node can cause drastic changes in the distribution of data points over nodes at the lower levels. For this reason there is a very small chance to accept the change in a node near the root. This means that the MCMC algorithm tends to explore the DT models in which the nodes selected to be changed are far from the root. Most of these nodes are close to the terminal nodes, and so contain small numbers of data points that makes a little change in the likelihood values. As a result such moves are mostly accepted. This affects the accuracy of integration because the RJ MCMC algorithm cannot explore all possible areas of interest in a model parameter space (Chipman et al., 1998; Denison et al., 2002; Schetinin et al., 2006; Jakaite and Schetinin, 2008).

During MCMC integration the moves are made to change distributions of data samples falling into DT terminal nodes. A move can change the number of samples in a terminal node so that their number becomes less than the given number p_{min}. When it happens such a move is assigned unavailable and a new move has to be proposed.

In the Bayesian context, this action is determined by a prior on p_{min}:

$$p(p_{min}) = \begin{cases} 0, & \text{if } min(n_1, n_2) < p_{min}\,, \\ 1, & \text{otherwise}\,, \end{cases} \tag{3.33}$$

where n_1 and n_2 are the numbers of data samples falling into the left and right branches of the terminal node.

The choice of p_{min} is dependent on such factors as the class boundaries and noise level in data. The complex class boundaries typically require large DTs for which p_{min} has to be small enough. In practice, the prior knowledge on favourite shape or size of DT models is absent, and then the appropriate p_{min} has to be found experimentally. Setting an inappropriate small p_{min} can lead to excessive growth of DT models which is correlated with an accelerated growth of likelihood values during MCMC averaging. When it happens, the DT models of a smaller size will not be explored in detail, and the results of integration will be most likely biased (Buntine, 1998; Domingos, 2000).

The other reason of the excessive growth is that the birth move is favoured to be accepted when the MCMC algorithm starts to grow a DT model (Chipman et al., 1998; Denison et al., 2002). The growth of DT size is typically monitored and the excessive growth can be restricted by setting a larger number p_{min} as well as by setting a smaller value of the proposal probability for the birth move.

The negative effect of sampling oversized DT models has been mitigated by using RJ MCMC with the restarting strategy (Chipman et al., 1998). The main idea of this strategy is to grow a DT within a limited period and then average the multiple runs of DTs randomly initialized. It has been shown that this strategy provides a better accuracy when a length of the period and a number of the runs are properly defined.

A similar idea of restricting the growth of DTs has been proposed in (Denison et al., 2002). The growth is banned within a given interval in order to let MCMC explore a model parameter space in more detail.

One general drawback of the restriction strategies is that they require additional settings for MCMC which must be tuned experimentally during the burn-in phase. Another side effect appears when a DT becomes grown up and most of the birth and change moves are made unavailable. This deteriorates the given proposal probabilities of the moves and forces MCMC to replicate samples that finally affects the accuracy of integration (Schetinin et al., 2004).

3.5.6 Sweeping strategy of RJ MCMC integration

The sweeping strategy proposed in (Schetinin et al., 2004) exploits a new prior on moves that assign or update a rule of a DT splitting node so that to decrease the probability of making unavailable moves – the problem with such moves was discussed in the previous section. The idea behind this prior is to use a proposal variable uniformly distributed within a *min-max* range of data points assigned to the chosen node. It has been found that this prior is able to prevent DTs from an excessive growth which affects the ability to generalise unseen data.

For each birth or change move, the proposal parameters are drawn from the given priors to be assigned to a chosen node. The proposed change can be made so that one or more terminal nodes in the DT will contain fewer data points than that allowed by p_{min}. If such a change is accepted, within the sweeping strategy a node with the fewer samples is removed from the DT being counted as the death move. If, however, there are more than one such nodes, the MCMC algorithm assigns the proposal unavailable in order to keep the balance between the death and birth moves.

When the birth move adds a new splitting node with the parameters drawn from the given priors, the MCMC assigns a new splitting variable s_i^{var} as well as a new rule s_i^{rule} taken from a uniform distribution over values of variable s_i^{var} at the partition l:

$$s_i^{rule} \sim U(min(X^{(l)}), max(X^{(l)}), \tag{3.34}$$

where $X^{(l)}$ are the values of the variable s_i^{var} at the lth partition, $l = 1, \ldots, k$.

According to this prior, the first partition is made over all data points $X^{(1)}$ that come at the root node. The second partition is made over a subset of data points, $X^{(2)}$, coming from one of two branches of the root node. The data are further split while the terminal nodes contain, at least, $2p_{min}$ data points. Figure 3.2 illustrates the changes in the boundaries x_{min} and x_{max}, that are determined by the above Eq. 3.34, for the first two partitions.

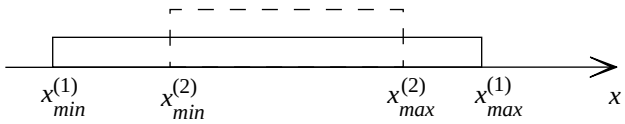

Figure 3.2: Illustration of changes in the boundaries x_{min} and x_{max} for the first and second partitions.

During MCMC integration, the birth or change moves can produce a splitting node in which one of two branches contains fewer data points than p_{min}. If this condition is met for one splitting node, this node is removed. More rarely, this condition is met for a branch with two or more nodes. When this happens, the proposal is assigned unavailable and the MCMC algorithm makes a new proposal.

The above condition can be met for the change move, when one terminal node is removed. In this case, the likelihood of the new DT model can be slightly smaller than that of the previous model, and so the proposed change will be most likely accepted. The sweeping strategy removes a node in which after the change move the number of data samples becomes fewer than p_{min} from the DT model. This strategy is applied during the both, burn-in and post burn-in, phases.

3.6 Summary

We introduced the Bayesian methodology of averaging over decision tree models and showed that the Markov chain Monte Carlo simulation method allows us to implement this methodology. The analysis of this method has revealed a number of problems which are mainly related to variable dimensionality of decision tree models, their hierarchical structure and large number of possible configurations. The main approaches to the problems were described and reviewed in the light of both accuracy of approximation and usage for solving real-world applications.

Chapter 4

Influence of EEG Artefacts

Newborn EEG are often contaminated by artefacts which can affect the accuracy of maturity assessments. To improve the accuracy, it is important to detect the artefacts and mark the affected segments to be removed. Experts can spend hours to recognise the various types of artefacts within the context of changing EEG patterns. The wide variations in artefacts and EEG patterns make it difficult to apply standard rules to artefact detection. In the absence of rules, the manual marking of artefacts may be inconsistent between experts and recordings. The inconsistencies in artefact removal may affect the accuracy of Bayesian assessments.

In this chapter we hypothesise that automated techniques, removing artefacts consistently in all recordings, will provide better results within Bayesian assessments than the manual removal. To test the hypothesis, we explore how the removal of marked artefacts and automatic artefact detection with various techniques improve the accuracy of Bayesian assessments of brain maturity.

The manual and automated artefact removal techniques are discussed in Section 4.1. In Section 4.2, we describe experiments with the artefact removal techniques. The first experiments test whether the removal of marked artefacts improves the assessments; we compare the accuracies on EEG data including artefacts and on clean data after removal of artefacts marked by experts. Next, we test a standard technique of averaging over EEG segments to suppress the influence of artefacts. We then describe and test two techniques for automatic removal of artefacts with abnormally high amplitudes. We summarise the results of the techniques and conclude the chapter.

4.1 Manual and automated artefact removal

EEG artefacts need to be recognised and removed to reduce the chance of mistaken assessments. In case of visual assessments, the artefacts can be mistaken for an EEG pattern. For example, the electrode movement artefacts may be confused with high-amplitude delta waves characteristic of very pre-term patterns. In case of automated assessment, the features extracted from contaminated EEG will be affected by artefacts. In particular, the spectral features computed within the FFT, which assumes a stationary signal, may provide biased results, because the artefacts make EEG data highly non-stationary (Clarencon et al., 1996).

To remove the artefacts, EEG experts analyse the recordings and mark the affected segments. The detection of artefacts is time consuming and difficult, as the artefacts widely vary in appearance and can occur within various EEG patterns, so that developing and applying standard rules for detection becomes infeasible. Under the lack of rules, the marking of artefacts becomes subjective.

According to van de Velde et al. (1999) the agreement between two experts marking artefacts in the same recording is on average 76%, whereas one expert analysing the same recording repeatedly marks only around 80% of the artefacts that were detected the first time. The inconsistencies in marking of artefacts may affect the accuracy of automated maturity assessment.

Computer-based techniques of EEG artefact removal provide consistent results, and thus we hypothesise that the use of such techniques will improve the accuracy of Bayesian assessments of brain maturity. The artefact removal techniques are typically based on deleting EEG samples with abnormal features. In general, artefacts can be considered as abnormal events whose characteristics are different from normal EEG. For example, the artefacts caused by patient's movements have much higher amplitudes than those of normal EEG (Nolan et al., 2010). Therefore, movement artefacts appear as outliers in the distribution of EEG amplitudes. A simple technique for removing the artefacts is to delete the samples whose amplitudes exceed a threshold given as the mean plus standard deviation of the EEG amplitude distribution.

A weakness of this technique is that a single threshold is used for the whole recording, and variations of EEG amplitudes over the patterns are not taken into account. This means that in EEG patterns with low dominant amplitudes the artefacts can be missed, whereas in patterns with high amplitudes EEG data can be lost. Therefore, it is desirable to adapt the threshold to EEG variations.

In cases when the frequency the artefact is well defined, the artefacts can be removed by band-pass filtering without significant loss of EEG information. For example, a notch filter set to 50 Hz can be used to remove the electrical

mains interference (Sanei and Chambers, 2007). Such filtering can be useful if high-frequency EEG waves need to be analysed. In assessment of newborn EEG however, frequencies above 30 Hz are not typically considered.

When EEG has been recorded from multiple channels, the Independent Component Analysis (ICA) can be applied to minimise the artefacts. This technique attempts to separate the EEG signal into statistically independent sources. The sources that are found most strongly affected by the artefact are then eliminated and the remaining sources are mixed to obtain a cleaned EEG signal. This technique has been shown successfully reducing the influence of artefacts, however, EEG experts have raised concerns that ICA can distort the power spectrum of EEG (Castellanos and Makarov, 2006). The ICA-based artefact removal requires the number of EEG channels to be at least the same as the number of sources, and this technique cannot be applied to recordings with only two-channels.

Alternatively to removing the artefacts, another standard approach is to suppress the influence of artefacts by averaging over EEG features computed in multiple short segments. The averaging suppresses the transient variations and artefacts occurring in the individual segments, and therefore the averaged features are more reliable for EEG analysis (Cooper et al., 2003; Kropotov, 2009). Importantly, the short segments can often be considered as pseudo-stationary, unlike the whole EEG which is highly variable. This means that the FFT applied to the segments can provide more reliable results. The choice of segment length is a trade-off between frequency resolution and stationarity. The lengths from 2 to 20 sec are typically chosen (Cooper et al., 2003).

4.2 Experiments

We compare the performance of Bayesian assessments of brain maturity on EEG data after manual and automated removal of artefacts. In the experiments, we use 210 recordings from newborns in 2 age groups, pre-term (36 weeks PCA) and full-term (41 week PCA). Each group contains 105 recordings.

First, to test how the manual removal of artefacts improves the assessment accuracy, we compare the accuracies on EEG data including artefacts and on clean data after removal of artefacts marked by experts. Next, we compare the automated artefact removal techniques, starting with the technique of averaging over segments to obtain more reliable EEG features which are less affected by artefacts. We hypothesise that the averaging will improve the accuracy of assessments. Next, we describe and test two techniques for automatic removal of artefacts with abnormally high amplitudes. The first technique removes the samples which exceed a threshold found as the mean plus standard deviation

of amplitudes in the recording. As described above, such a single threshold is unlikely to provide the best artefact detection within EEG patterns with different dominant amplitudes. The second technique is more complex and aims to adapt the threshold to the variable patterns.

4.2.1 Removing the marked artefacts

Experts have analysed the EEG and the corresponding polysomnogram in 10 sec segments, and marked each segment as artefact or non-artefact. The percentage of artefact segments in all recordings was around 7%.

To compute the spectral EEG features, the FFT was applied to the whole recordings to obtain the power spectrum, and then the spectral powers in the six frequency bands were calculated. Each recording was represented with the six absolute and relative spectral powers for the two channels C3T3-C4T4 and their sum, in total 36 features.

The Bayesian classification was run with the following settings: during the burn-in and post burn-in phases we collected 100,000 and 10,000 DTs. The pruning factor was set to 4. The variance of change-rule proposals was 1.0, and the probabilities of birth, death, change-split and change rule-moves were set to 0.15, 0.15, 0.1 and 0.6, respectively. Given these settings, the acceptance rate was on average 0.26 and the average DT had 5 nodes.

Table 4.1 compares the performances (P) and entropies (E) of Bayesian classification on the raw EEG data including artefacts and on the data from which the marked artefact segments were removed.

Here and in the subsequent chapters, the performance of a technique is defined as the percentage of test data samples that were classified correctly. The entropy of the ensemble, expressing the classification uncertainty (Kuncheva, 2004), is defined as follows.

$$E = \frac{1}{t}(-\sum_{i=1}^{t}\sum_{j=1}^{c} P(j|x_i)log_2(P(j|x_i))), \tag{4.1}$$

where t is the number of test data samples x_i, c is the number of classes, and P are the class posterior probabilities. The entropy is measured in the number of bits per test data sample.

In table 4.1, the performances and entropies are listed along with 2σ confidence intervals counted within the five-fold cross-validation.

We can see that after the removal of artefacts the performance becomes improved by ca 4%, on average. However, according to the Mann-Whitney U test this improvement is not statistically significant ($p \approx 0.47$). Next, we test

whether the accuracy can be improved further by averaging over EEG segments to suppress the influence of artefacts.

Table 4.1: Performance (P) and entropy (E) before and after the removal of marked artefacts.

Data	P, %	E, bits
With artefacts	78.10±13.21	0.126±0.037
Without marked artefacts	82.38±13.72	0.125±0.025

4.2.2 Averaging spectral features over segments

In the experiments, the EEG were split into 10 sec segments, and the 36 spectral features were computed in the segments. These features were averaged over all the segments in a recording. From Table 4.2 we can observe that the averaging over segments has improved the mean performance to 85.7% ($p \approx 0.06$). The entropy has been significantly reduced to 0.111 ($p < 0.01$). The performance has been improved by approximately 8%, compared to that achieved on EEG data with the artefacts.

For comparison, the artefact segments were removed, and only normal segments were averaged. As shown in Table 4.2, the averaging over the normal segments provided the performance in a similar range. The uncertainty in terms of entropy was slightly decreased to 0.096.

Overall, the averaging of spectral features over the segments has improved the accuracy of maturity assessment by 7%, in comparison to the accuracy achieved with the features extracted on the whole recording. As the averaging suppresses the influence of artefacts, the removal of artefact segments prior to the averaging had insignificant impact on the results. In the next subsection we will explore how the accuracy can be improved further by using a simple thresholding technique for automatic removal of artefacts.

Table 4.2: Performance and entropy on EEG data represented by averaged features.

Data	P, %	E, bits
All segments	85.71±5.83	0.111±0.022
Without marked artefacts	85.00±10.61	0.096±0.030

4.2.3 Removing artefacts with statistical thresholding

Artefacts with high amplitudes, such as those caused by eye and head movement, can be considered as outliers in the distribution of EEG amplitudes. A simple

technique to detect such outliers is to estimate the distribution of amplitudes in an EEG and then to remove the samples whose amplitudes exceed a threshold given as the mean plus standard deviation of the distribution.

Fig. 4.1 shows 1 min of EEG with samples considered as outliers given thresholds of 1, 2 and 3 standard deviations of data. The outlier samples are marked in grey. An artefact can be seen during approximately the first 20 sec.

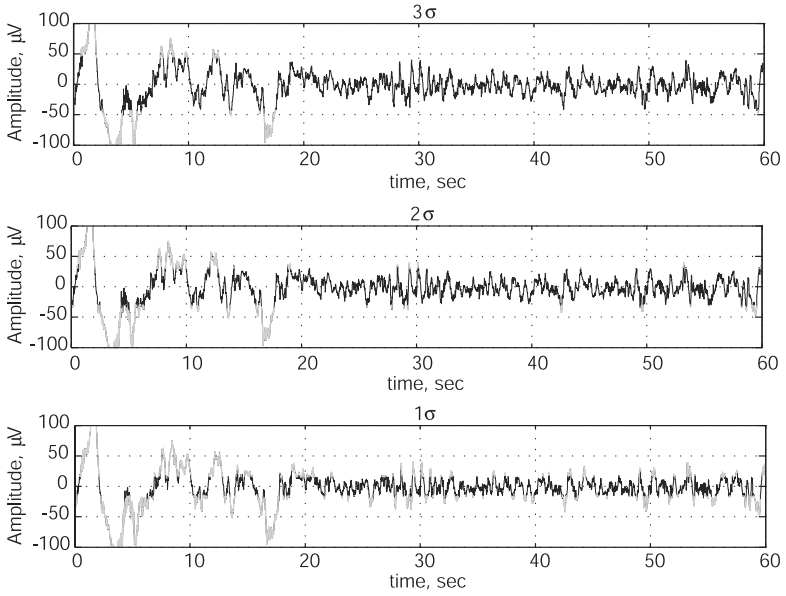

Figure 4.1: Artefacts (grey) and normal EEG (black) given different thresholds

With the threshold of 3σ only the highest peaks within the artefact are detected as outliers and the EEG after the first 20 sec remains intact. With the threshold of 1σ most samples within the artefact are removed. However, some peaks within EEG signal are also cut off, making the EEG more difficult to analyse visually.

Table 4.3 shows the accuracy of maturity assessment on EEG data from which the outlier samples were removed given confidence intervals ranging from 2.5 to 0.75σ. The percentage of samples considered as artefacts (A) ranged from 2.5% to 31%. The highest performance, 87.6%, and the lowest entropy, 0.092, are obtained given 1σ confidence intervals, when approximately 20% of EEG samples are removed as artefacts. Thus, the removal of artefacts by the statistical thresholding technique improved the average performance of maturity assessment by 9.5% and decreased the entropy by 26% ($p \approx 0.02$ and $p < 0.01$).

An obvious weakness of the described technique is that a single threshold is used for the whole recording, and variations of EEG amplitudes during different

types of brain activity are not taken into account. Specifically, this may lead to false detection or missing of artefacts within the quiet and active sleep stages, during which the average EEG amplitudes are significantly different. In the next subsection, we explore a more complex artefact removal technique capable of adapting the threshold to variable EEG activity.

Table 4.3: Performance, entropy and percent of artefacts (A) after removal of outliers.

Threshold, σ	P, %	E, bits	A, %
2.5	83.8±11.4	0.113±0.028	2.5±1.3
2.0	83.8±11.9	0.097±0.030	4.6±2.1
1.5	83.3±12.9	0.094±0.006	9.3±4.3
1.0	87.6±9.2	0.092±0.011	20.4±9.9
0.75	81.0±12.1	0.102±0.019	31.1±13.4

4.2.4 Removing artefacts based on local amplitude statistics

Instead of finding a single threshold amplitude on the whole recording, the detection can be made in the context of local EEG variations. We assume that artefacts are events with abnormally high amplitudes which significantly deviate from normal EEG.

Based on this assumption, in order to detect artefacts with abnormal amplitudes, we estimate the standard deviation of amplitudes in a sliding window. The technique is summarised in Algorithm 1.

Algorithm 1 Artefact Detection

Inputs: X, $wlen_1$, $wlen_2$, q
Initialise: sliding windows W_1 and W_2 with given $wlen_1$ and $wlen_2$
$X \leftarrow |X|$
while $W_1 \in X$ **do**
    $D(W_2) \leftarrow Deviation(X(W_1))$
    Increment positions of W_1 and W_2
end while
$f(d) \leftarrow ProbabilityDensityEstimation(D)$
$d_0 \leftarrow argmax(f(d))$
$d_{thresh} \leftarrow d_0 + (max(d) - d_0) * q$
$A \leftarrow (D > d_{thresh})$
return A

According to the algorithm, a window W_1 of length $wlen_1$ is moving along a rectified EEG signal X. For each position of W_1, the standard deviation of samples of X is estimated. When the deviation has been counted for all window

positions in X, the probability density function of the deviations is estimated in order to find the most frequent value d and the maximal value d_{max}.

We expect that normal EEG samples appear most frequently and so their deviation will be less than or equal to d. On the other hand, artefacts will appear with a higher deviation. We can set a threshold d_{thresh} for artefact removal to be proportional to the difference between d and d_{max}: $d_{thresh} = d + (d_{max} - d)q$. In practice, when setting the constant q, we need to find a trade-off between the accuracy of artefact detection and the amount of normal EEG samples being removed. In our experiments, such a trade-off has been achieved with $q = 0.15$, and a sliding window was set 10 sec in duration. With these settings, the average percentage of data considered as artefacts was 16.14±22.35% for all recordings.

Fig.4.2 shows the detection and removal of artefacts in an EEG. The fist plot shows the EEG with artefacts visible as peaks with high amplitude mostly during the minutes 40 to 110 and 160 to 210. The second plot shows the deviations d_i with the threshold $d_{thresh} = 13.2$ shown as a dashed line. The third plot shows the marks of detected artefacts whose rate was 30%, and the fourth plot shows the EEG cleaned from the artefacts. We can see that most of the artefacts are removed, and at the same time, the amplitude variations of normal EEG samples are preserved.

For EEG data, from which the artefacts were removed using the described technique, the performance and entropy of Bayesian classification were 86.36 ± 6.43 and 0.096±0.016. The performance has improved by approximately 8.2% and entropy by 19.4% in comparison to those obtained on the EEG data with artefacts. The results are comparable to those achieved after removing artefacts with the simple statistical thresholding technique. A benefit of using the local amplitude statistics for artefact removal is that the sleep cycle variations are preserved in EEG.

4.3 Chapter discussion and conclusions

EEG artefacts have a negative influence on the accuracy of maturity assessments, which can be improved by removing the artefacts. We hypothesised that the automatic detection of artefacts and averaging over segments provides more consistent results than the removal of artefacts marked by experts, and so enables achieving a better accuracy of Bayesian assessments.

We tested two techniques for automatic removal of artefacts with abnormally high amplitudes. The first technique removes the EEG samples whose amplitudes exceed a threshold found as the mean plus standard deviation of amplitudes in the whole recording. This technique, however, does not take into account variations of EEG patterns, and therefore the detection of artefacts

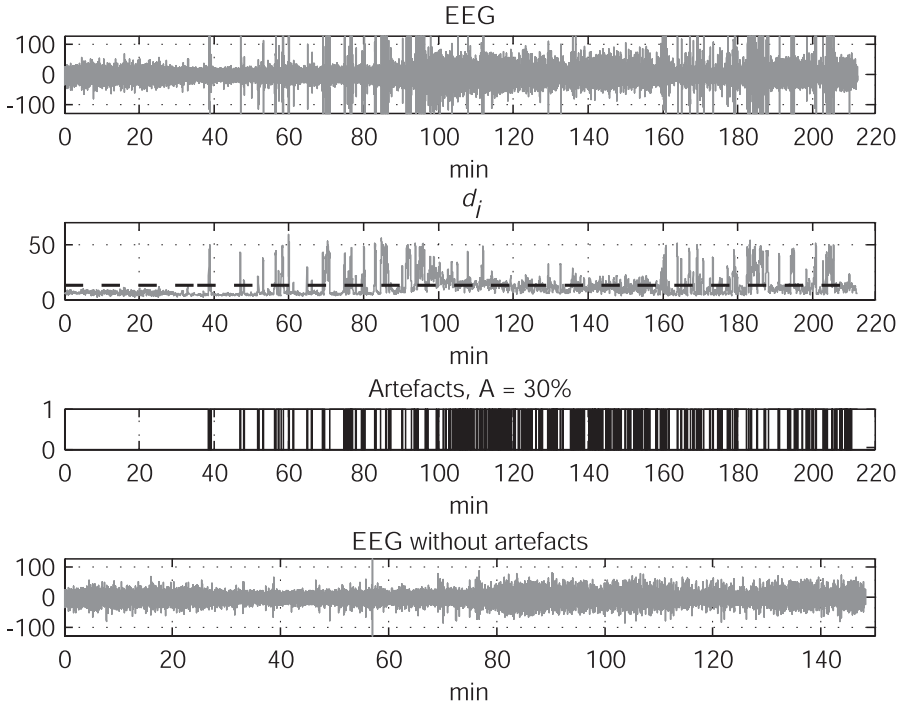

Figure 4.2: Detection and removal of artefacts

could be misleading. The second technique calculates the deviation of amplitudes in windows sliding over EEG in order to detect artefacts in the context of EEG variations. We also expected that averaging over multiple EEG segments can suppress the influence of artefacts.

Fig. 4.3 compares the performances and entropies of Bayesian assessments of brain maturity obtained on raw EEG (a), after expert removal of artefacts (b), after the averaging over segments (c), and automated removal of artefacts (d, e). The boxplots summarise the results obtained within the five-fold cross-validation. As expected, the removal of artefacts and the averaging over segments improved the accuracy of maturity assessments. In our experiments, the averaging over segments as well as the automatic techniques of artefact detection provided better results than the removal of expert marked artefacts, which supports our hypothesis.

The statistical thresholding technique enabled obtaining a slightly better performance and lower entropy, than the local amplitude deviation technique. A weakness of thresholding technique, however, is that it may often delete normal EEG samples with high amplitudes, possibly leading to loss of normal data. Although in our experiments this did not create problems for Bayesian assess-

ments, it may make the visual analysis difficult. The detection of artefacts based on local amplitude deviation was shown to remove most of the artefacts while preserving the normal EEG samples, so that the sleep cycle variations could still be analysed. A possible negative effect was, however, that the preserved variations in EEG slightly increased in the assessment uncertainty in terms of the entropy.

In the next chapter, we can employ the artefact removal techniques to prepare EEG data for experiments with classification of six age groups. We expect that the accuracy of Bayesian assessments can be further improved by selecting the most informative EEG features.

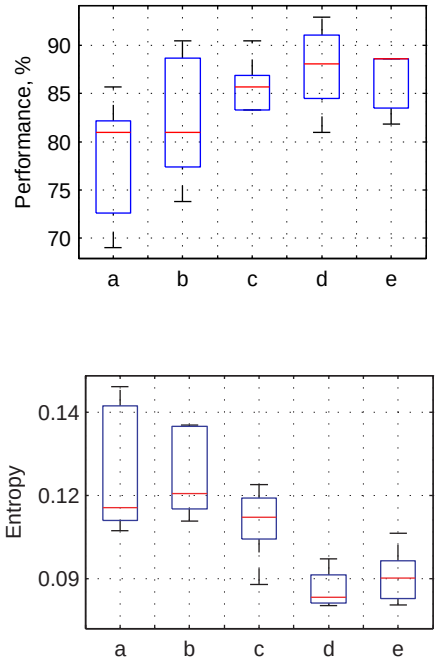

Figure 4.3: The performance and entropy of Bayesian classification: on raw EEG (a), after removal of expert marked artefacts (b), averaging over multiple segments (c), artefact removal by statistical thresholding (d), and artefact removal based on local amplitude deviation (e).

Chapter 5

Importance of spectral features

The absolute and relative spectral powers in the standard frequency bands are considered to be informative EEG features. However, the importance of these features for the assessment of newborn brain maturity is likely to be unequal — some of them may be weakly associated with brain maturation. The presence of such weak features can negatively affect the accuracy of maturity assessments obtained within the methodology of Bayesian averaging over DTs.

First, the use of weak features obstructs interpretation of DTs. Second, weak attributes increase dimensionality of a model parameter space, making it more difficult to be explored in detail within a reasonable time. The lack of detailed exploration can negatively affect the results of Bayesian averaging. Prior information about EEG features can be used to reduce the dimensionality, however, in our case no such prior information is available.

The use of DTs for Bayesian assessment enables obtaining posterior information on the importance of EEG features. In this chapter, we hypothesise that the posterior information about feature importance can be used to refine the DT ensemble from models using features found making weak contribution.

In Section 5.1 we discuss in more detail the reasons for selecting the important EEG features to be used with Bayesian averaging over DTs and outlines the principle of refining the DT ensemble for feature selection. Section 5.2 describes the refining technique, which can be used to find a subset of the most important EEG features. Section 5.3 describes experiments with the proposed technique for six PCA groups or classes, 36 to 41 weeks. This multiclass problem is expected to be difficult as the EEG from the neighbouring age groups are hard to differentiate and so the classes can overlap. Section 5.4 describes experiments

on two age groups, 36 and 41 weeks. We expect that these groups, including EEG of pre-term and full-term newborns, will be easier to distinguish, and we find that these groups can be classified by DTs using only a small portion of the EEG features. We hypothesise that the subset of features can be further reduced by setting a larger pruning factor to encourage growing shorter DTs which use fewer features, and test this hypothesis in experiments. Section 5.5 concludes the chapter.

5.1 Feature selection within Bayesian averaging over DTs

The spectral features along with their statistical characteristics form a multi-dimensional representation of the EEG data. As described in Section 2.5 EEG Data, the spectral features comprise the absolute and relative spectral powers in the six standard frequency bands, calculated for the two channels C3T3, C4T4, and their sum. Additionally the statistical variances are calculated for these features, so that the total number of EEG attributes becomes 72.

Previous research has shown that the importances of the spectral bands for maturity assessment may be unequal and only some of the bands may make a significant contribution, however, in different studies different bands have been identified as most important (Holthausen et al., 2000; Scher et al., 1995). Thus, we have no reliable prior information on the importance of the 72 features, but we can expect that some of them do not make a significant contribution.

Under the lack of prior information on feature importance, the MCMC technique, described in Chapter 3, can sometimes accept a DT model with one or more of the weak features, even with a slight decrease in performance. In presence of few weak features, the portion of such DTs included in an ensemble will likely be insignificant and results will not be affected. Contrary, in presence of many weak features, the portion of such DTs can become substantial.

The negative impact of this is twofold. First, the use of weak EEG features obstructs the interpretation of maturity assessments. Employing an excessive number of features is likely to result in growing oversized DTs. Second, weak features increase the dimensionality of a model parameter space that needs to be explored within the MCMC technique. The success in implementation of Bayesian model averaging is critically dependent on the diversity and proportion of models sampled for averaging. The models should be diverse in parameters and structure, and the portion of models whose likelihood is high should be largest to ensure unbiased estimates.

To ensure the diversity, the model parameter space must be explored in detail in order to sample models from diverse areas of interest with highest likelihoods. The larger the model parameter space, the more difficult it is to be explored. Thus the results of Bayesian averaging will likely suffer from disproportionally sampling the posterior distribution, as we cannot expect that a multidimensional model space will be explored in detail, and the areas of interest will be properly explored within a reasonable time.

Information about feature importance could be used to specify areas of a model parameter space to be explored. In our case, this information is unavailable and we are forced to make an unrealistic assumption that all the EEG attributes make an equal contribution to the maturity assessment. However, the use of DT models provides the feature selection, and therefore the Bayesian averaging over such models will give us the posterior information about EEG feature importance, which can be used to improve the results of averaging. This information is estimated as the frequencies of using each feature by the DTs in the ensemble. The importance of feature k is estimated as follows:

$$\gamma_k = \frac{\sum_{i=1}^{np} \sum_{j=1}^{Si} (v_j == k)}{\sum_{i=1}^{np} S_i}, \tag{5.1}$$

where np is the number of DTs in the ensemble, S_i is the size of ith DT, and v_j is the index of feature used by the jth split of the DT.

If a feature is rarely used in the ensemble, then it likely makes a weak contribution. We hypothesise that the DTs which use the weak attribute could be discarded from the ensemble without a decrease in performance.

In the next section we propose a technique for refining the DT ensembles from features found making weak contribution. A subset of the most important features can be found within a sequential-forward strategy of eliminating the weak ones.

5.2 Refining ensembles from DTs using weak features

Having obtained a range of the posterior probabilities of EEG features, we can define a threshold value to cut off the ones with the probabilities below this threshold; we define such features as weak. A trivial way of using the posterior information on weak features is to rerun the Bayesian averaging on data from which such features were deleted. This reduces a model parameter space so that it can be explored in more detail. However this techniques requires multiple reruns to find the best threshold.

The other way is to refine the DT ensemble by discarding DTs which use weak features. For each threshold, we can find the DT models which use these weak features and discard these DT models from the ensemble. We expect that such a refining strategy will reduce the original set of features without rerunning the Bayesian averaging, keeping its performance high. We can also expect that there is an optimal threshold probability at which the largest number of weak features can be discarded. It is interesting to explore whether the discarding of weak features will improve the results of Bayesian model averaging. In a series of experiments, we could increase the threshold probability in steps and evaluate the performance of the refined ensemble on the test data.

Alternatively to such try-and-see approach, we can search for the smallest set of important EEG features by discarding the models using weak features and monitoring the accuracy of the refined ensemble on the training data. We use a sequential forward strategy of finding DT models using a weak feature in order to eliminate these models from the ensemble. The search continues while the training accuracies of the refined and original ensembles are comparable within a given p-value of a statistical hypothesis test, such as the two-sample Kolmogorov-Smirnov test (KS-test). The accuracies are said comparable as long as the test cannot reject the null hypothesis. The null-hypothesis assumes that samples of the accuracies are drawn from the same distribution. The test rejects the null-hypothesis if the modifications made for kth attribute decrease the accuracy, and then the procedure stops.

To compare the training accuracies with a hypothesis test, the distribution of the accuracies given each feature subset need to be estimated. To estimate the distributions it is required to collect sufficient independent samples representing the accuracies of each of the ensembles. Such samples could be obtained by calculating the accuracies on multiple independent data sets. When the training data are limited, the independent data sets can be simulated by resampling the available data. One of the techniques enabling the multiple independent datasets to be generated is to randomly subsample two-thirds of data without replacement. In cases when the simulated data sets are required to be with the same number of samples as the original data set, bootstrapping with replacement is typically used. In our case, however, there is no such requirement.

The proposed technique of finding a subset of the most important features can be summarized by Algorithm 2.

The algorithm returns the number of features which were found weak within a given p-value. Thus, the indexes of weak features are in positions from 1 to k of the list F. Obviously, the greater the number of attributes found weak, the larger is the portion of DT models discarded from the ensemble. As a result,

Algorithm 2 Refining a DT Ensemble

Inputs: training data D represented by m features, ensemble of DTs, number of subsamples n, p-value, number of attempts v_{max},
Initialise: counter of attempts $v = 0$, number of weak features $k = 1$
Estimate the posterior feature importance
Sort the list of features, F, in the order of their importance
for $i = (1, n)$ **do**
    Subsample D and calculate the ensemble accuracy A_i
end for
while $v \leq v_{max}$ and $k < m$ **do**
    Find the DT models using feature F_k and delete them from the ensemble
    for $i = (1, n)$ **do**
        Subsample D and calculate the ensemble accuracy AR_i
    end for
    Run the KS-test to compare the samples $\{AR_i\}_1^n$ and $\{A_i\}_1^n$
    if null-hypothesis rejected **then**
        $v \leftarrow v + 1$
    else
        Reset the counter of attempts $v, v \leftarrow 0$
    end if
    $k : k \leftarrow k + 1$
end while
$k \leftarrow k - v$
return k

we expect to find the smallest set of attributes making the most important contribution and keep the performance of the refined ensemble high.

A potential criticism of the refining technique is that the sequential forward strategy of eliminating the weak features does not take into account the possible interactions between the features. However, the technique assumes that the feature interactions have been considered by the collected DT models. Our hypothesis is that the combinations of the features which make valuable contributions to the classification have been used by the largest portion of the ensemble's DT models. On the contrary, the weak features, which are sometimes added to the DT even with a slight decrease in the likelihood, are used by a much smaller portion of the models. When the MCMC technique adds a weak feature to a DT by making a birth move, a new "version" of the model with the weak feature is included in the ensemble. The fact that a weak feature is rarely used by ensemble's DTs means that proposals to add this feature tend to decrease the model's likelihood and are rarely accepted by the sampler. The refining technique is aimed to remove those DT versions which include the weak features while keeping those DTs which have employed the successful feature combinations. The efficiency of this technique is evaluated in experiments in terms of performance and accuracy of uncertainty assessment.

5.3 Experiments with six age groups

The first experiments were run on classification of EEG of newborns in six age groups or classes, 36 to 41 weeks PCA. This multiclass problem is expected to be difficult as the EEG from the neighbouring age groups are hard to differentiate and so the classes can overlap. For this problem, we first run the Bayesian DT technique described in Chapter 3, and then obtain the posterior probabilities of EEG features being used in the DT ensemble. Having found the ranges of the posterior probabilities, we assign thresholds probabilities to define features as weak, and test our hypothesis that an ensemble can be refined from DTs using weak features without a decrease in performance. We then test the proposed technique of searching for the minimal subset of important features. For the comparison, we rerun the Bayesian averaging on the data of reduced dimensionality, having eliminated the weak attributes.

5.3.1 Bayesian classification

The experiments were run with the set of EEG recordings of 686 newborns aged between 40 and 45 weeks so that the number of age groups was six. Each of these groups (classes) included around 100 recordings. The EEGs have been segmented in 10-sec intervals, and the 72 spectral features, namely the spectral powers and their variances within the standard frequency bands, were computed within these segments. We averaged the segments of each patient to suppress the artefacts and transient variations in EEG as described in Chapter 4.

The Bayesian technique was run with the following settings. In a burn-in phase we collected 200,000 DTs, and in a post burn-in phase 10,000 DTs. During the post burn-in phase each 7th model was collected to reduce the correlation between DT models. The minimal number of data samples allowed to be in DT nodes, p_{min}, was set to six. Proposal variance was 1.0, and probabilities of making moves of birth, death, change variable, and change threshold were set to 0.15, 0.15, 0.1, and 0.6, respectively. The performance and entropy of the DT ensemble collected in the post burn-in phase were evaluated using a five-fold cross-validation.

The rate of acceptance of DT models was around 0.13 in both phases. In the burn-in phase, the log-likelihood as well as the size of DT were stabilized after 10,000 samples, as seen from Fig. 5.1, so that the remaining samples were drawn from an approximately stationary Markov Chain. The average performance of the Bayesian technique (exact match of weeks) was 27.41±3.9% and the entropy was 0.414.

In comparison, a single DT, trained with the same p_{min} setting, provided a performance of 24.6±3.9%. The Bayesian averaging over DT models provided

an almost 3% better performance than the single DT, and this result shows that the sampler has acceptably explored the parameter space.

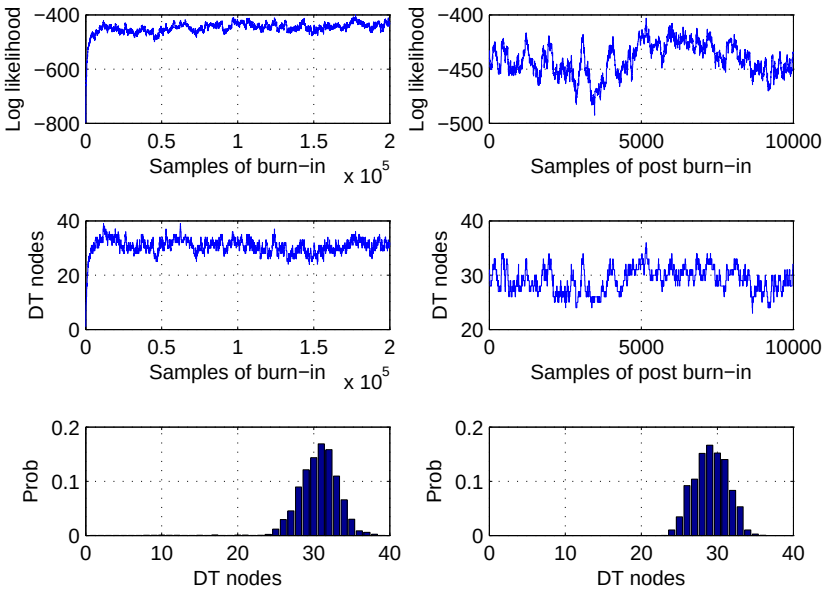

Figure 5.1: Log-likelihood, number of DT nodes and distribution of DT sizes during the burn-in and post burn-in phases.

5.3.2 Feature importance

According to the proposed technique, we estimated the importance of all the 72 attributes in terms of the posterior probabilities of using these attributes by the DT models collected in the post burn-in phase. The posterior probabilities (frequencies) of using the attributes ranged between 0.0 (exactly zero) and 0.048 as shown in Fig. 5.2. Here, the probabilities were averaged over the 10 folds. We can observe that the three most important features with probabilities near 0.048 are the mean relative and absolute powers in the Delta range. The probabilities of all of the mean spectral powers are generally higher than those of their variances; 12 mean powers are with probabilities above 0.02, but only seven variance features have probabilities above this threshold. The probabilities of the absolute power variances generally are the lowest, all below 0.02.

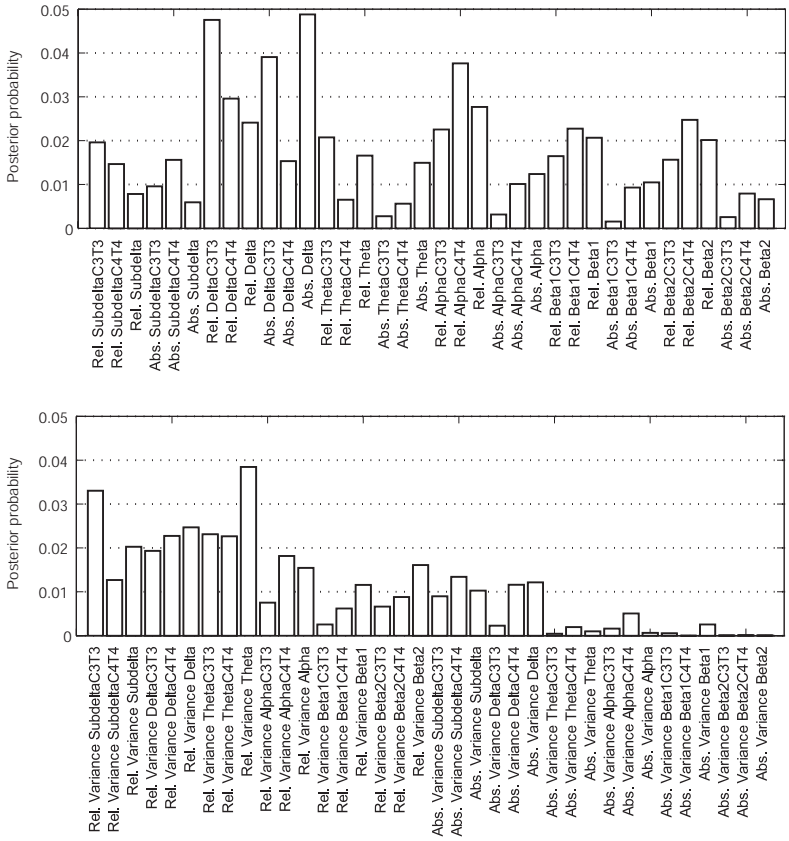

Figure 5.2: Posterior probabilities of 72 EEG attributes characterising the relative and absolute spectral powers (the upper plot) and their variances (the lower plot).

5.3.3 Refining the ensemble

Having found the range of feature importances, we applied the proposed technique to refine the DT ensemble. Table 5.1 shows the number of weak features, k, versus the threshold values within a 5-fold cross-validation. At threshold value 0.001 the average number of weak features, k, was 15, whilst at level 0.005 their number has increased to 30. We found that around 30 weak attributes could be discarded without a significant decrease in performance, P. At the same time, when the threshold was gradually increased from 0.0 to 0.005, the uncertainty in decisions insignificantly decreased from 0.414 to 0.403 in terms of entropy E of the ensemble.

Having confirmed that the DT models using weak EEG features can be discarded from the ensemble without a decrease in performance, we test the

proposed sequential-forward strategy of finding the minimal subset of important EEG features. Fig. 5.3 shows the training accuracy, performance, ensemble size and p-value of the KS-test calculated within the proposed technique for one of the five folds. We can observe that for ($k = 29$) weak features, p-value becomes lower than 0.5, the given confidence interval. Further discarding of weak features did not increase the accuracy. Thus we define 28 weak features and select the remaining 44 as most informative ones.

Table 5.2 compares the performance and entropy of the original ensemble with that of the refined ensemble excluding the 26 weak features. The performance, entropy and the number of weak features are counted within the five-fold cross-validation. We can see that after refining the performance has slightly increased by 1.8% and the entropy has slightly decreased.

Fig. 5.4 shows the distributions of performances provided by the original and refined DT ensembles on the test data. We can see that the size of the refined ensemble becomes significantly smaller. Most of the DTs with performance above 32.0% have been kept, whilst most of the DTs with performance below 24.0% have been discarded from the refined ensemble.

Table 5.1: Performance (P) entropy (E) and the number of weak features (k) for the thresholds

Threshold	P, %	E, bits	k
0.001	27.8±4.5	0.414±0.014	15
0.002	26.8±3.6	0.414±0.014	20
0.003	27.6±3.4	0.413±0.009	23
0.004	27.8±6.2	0.409±0.001	28
0.005	27.6±5.0	0.403±0.011	30

Table 5.2: Performance and entropy of the DT ensembles

Original ensemble		Refined ensemble		Rerunning	
P, %	E, bits	P, %	E, bits	P, %	E, bits
27.4±3.9	0.414±0.015	29.2±6.9	0.410±0.014	29.3±6.5	0.416±0.024

5.3.4 Rerunning the Bayesian classification with a reduced set of features

Having found a minimal subset of important EEG features, we can rerun the Bayesian classification on a dataset of reduced dimensionality. Table 5.2 shows the performance and entropy of DT ensemble rerun on the EEG data represented

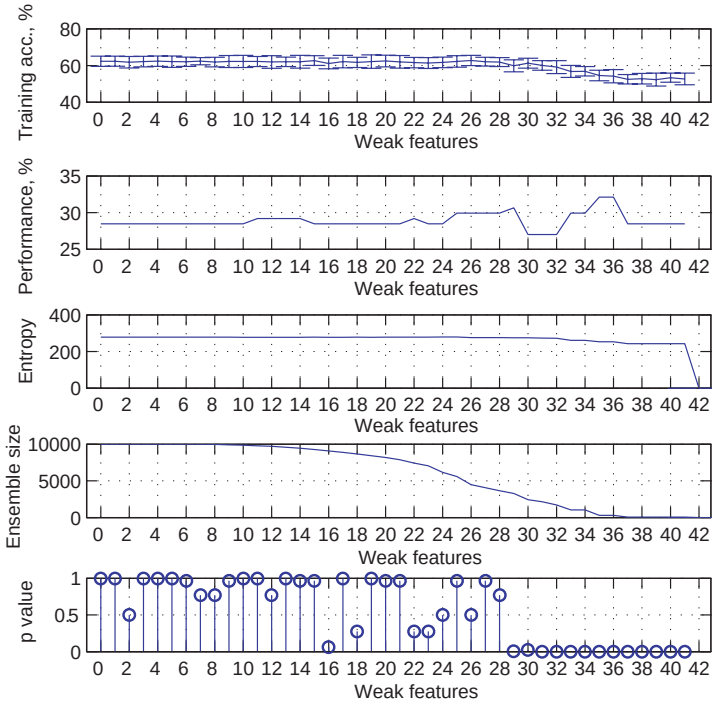

Figure 5.3: Finding a minimal feature subset. From top: training accuracy, performance, ensemble size, and p value of KS-test.

by the features found most important. We can see that the performance is similar to that of the refined ensemble. Compared to the original ensemble, the increase in performance is 1.9%. This result supports our hypothesis that dimensionality reduction provides better conditions for proportional sampling.

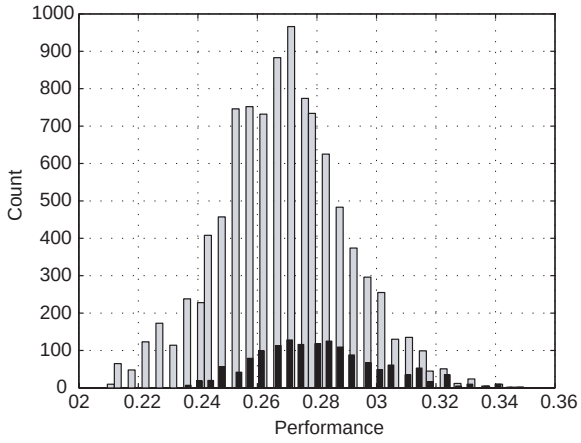

Figure 5.4: Distributions of performances of DTs included in the original (grey) and refined (black) ensembles

5.4 Experiments with two age groups

In this section we describe experiments with classification of pre-term and full-term EEG recorded at 36 and 41 weeks respectively. The EEG of pre-term and full-term newborns are very different, so the classification performance is expected to be higher than that for the six-class problem. Having run the Bayesian classification, we use the proposed technique to find the minimal subset of important features. We then explore whether the set of EEG features could be further reduced by increasing the minimal number of data points allowed in DT nodes, or the so-called pruning factor.

5.4.1 Bayesian classification

We used the EEG recorded from 200 newborns in two age groups, 36 and 41 weeks PCA. Each of the groups contained 100 patients. The EEG have been segmented in 10-s intervals and represented by 36 features, the mean absolute and relative spectral powers in the six frequency bands.

We ran the BMA with the same settings as for the six-class problem. The pruning factor, that is the minimal number of data points allowed in DT nodes, was set to 2 or 1% of data samples. Under these settings, the acceptance rate was around 0.3 in the burn-in and post burn-in phases. The DT size became stationary after growing to 5 nodes. The average performance counted within the five-fold cross-validation was 86.5±7.6% and the entropy was 0.107±0.023.

5.4.2 Feature importance for EEG recorded at 36 and 41 weeks

Fig. 5.5 shows the posterior probabilities of the 36 features representing the two-class EEG problem. These probabilities ranged between 0.004 and 0.22. For this problem, the most important features were the relative powers in the Theta and Alpha bands with the probabilities around 0.2. The importance of more than two thirds of the features was approximately 10 times lower, around 0.02.

The relatively low probabilities of most of the features may mean that the EEG of pre-term and full-term newborns can be distinguished based on a small set of features. Thus, we can expect that many weak features making insignificant contribution can be excluded from the maturity assessment without a decrease in performance.

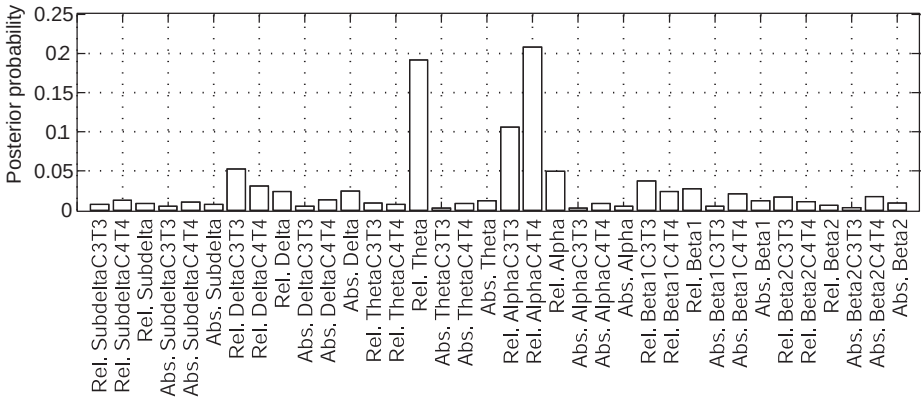

Figure 5.5: Posterior probabilities of 36 EEG attributes characterising the relative and absolute spectral powers.

5.4.3 Refining the ensemble

Table 5.3 lists the performance and ensemble entropy obtained in the experiments for the original and refined ensembles without the 27 weak features. The results show that the performances of the original and refined ensembles are comparable.

Fig. 5.6 shows the likelihood, performance, ensemble size and p-value of KS-test calculated within the proposed refining technique for one of the five folds. We can observe that for ($k = 25$) weak attributes, p-value becomes below 0.05, the given confidence interval, and does not improve significantly when the next

Table 5.3: Performance and entropy of the DT ensembles

Original ensemble		Refined ensemble	
P, %	E, bits	P, %	E, bits
86.5±7.6	0.107±0.038	86.5±7.6	0.106±0.021

weak features are removed. Thus we define 24 weak attributes and select the remaining 12 as most informative features.

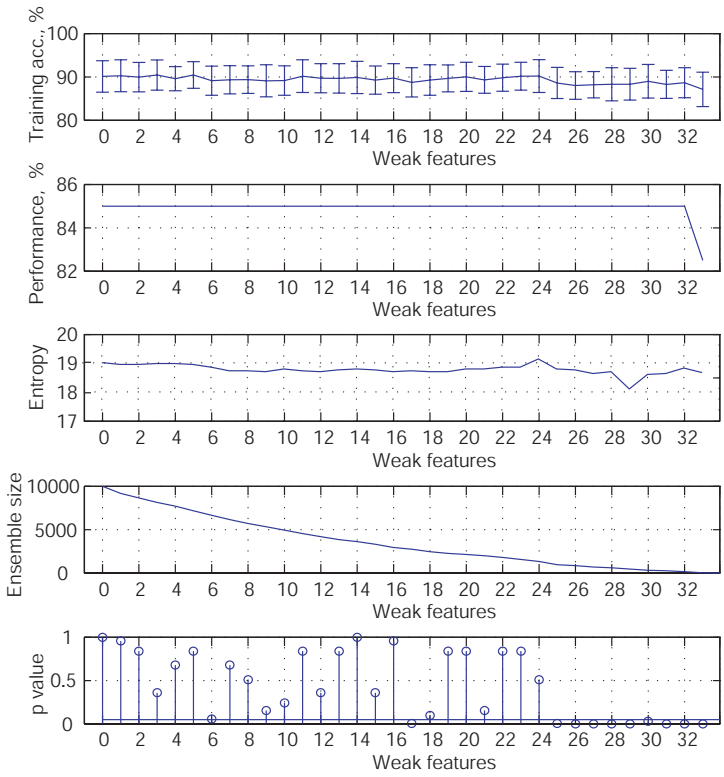

Figure 5.6: Finding a minimal feature subset for the classification of EEG in the two age groups, $p_{min} = 2$. From top: training accuracy, performance, ensemble size, and p-value of KS-test.

Fig. 5.7 compares the probabilities of using the 36 features in the original and refined ensembles. Here, the probabilities were counted within the five folds. From the upper plot we can see that the average probabilities in the original ensemble range between 0.0 and 0.2. We expect that the features with probabilities close to zero make insignificant contribution to the results. The

lower plot shows that refining has eliminated some of these features in the Subdelta, Beta and Alpha bands.

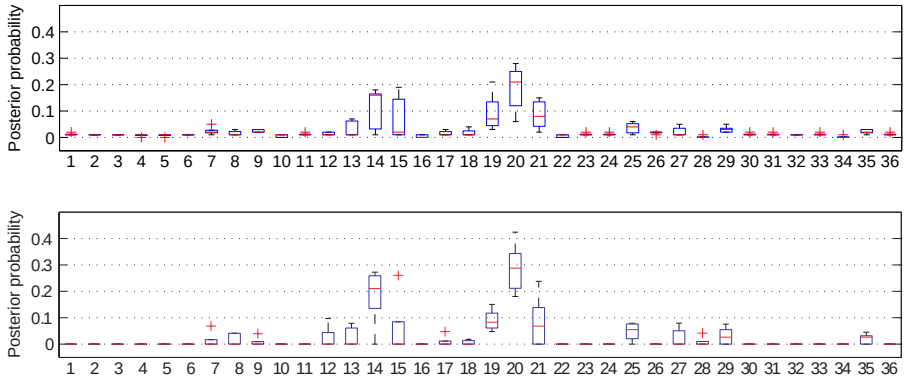

Figure 5.7: Probabilities of using features in the original and the refined ensemble, $p_{min} = 2$.

5.4.4 Pruning of Decision Trees

We investigated how the minimal number of data points in DT nodes, p_{min}, affects the results. In our experiments we ran the Bayesian averaging over DTs with p_{min} ranging from 2 to 10. Table 5.4 shows the average performance, entropy, and size of the DT ensembles within the five-fold cross-validation. We can see that DT size, on average, decreases while p_{min} increases from 2 to 10. For larger p_{min}, the ensemble performance becomes slightly lower, remaining within a given confidence interval, while the entropy is slightly growing.

Table 5.4: Performance, entropy and average DT size of the ensemble given different pruning factor, (p_{min})

p_{min}	P, %	E, BITS	DT SIZE
2	86.5±7.6	0.107±0.023	5.6±1.0
4	86.5±10.9	0.115±0.018	4.7±0.9
6	85.0±8.7	0.112±0.014	4.1±0.5
8	84.5±14.3	0.112±0.017	3.7±0.2
10	85.0±10.6	0.113±0.026	3.5±0.3

Fig. 5.8 shows the likelihood, performance, ensemble size, and p-value of KS-test calculated for DT ensembles refined with the proposed technique for $p_{min} = 10$. Comparing the results shown in Fig. 5.6 and Fig. 5.8, we can see that a larger number of the features can be discarded without significant decrease in

the training accuracy. We can also see that the ensemble size decreases less steeply for $p_{min} = 10$. This is likely because fewer DT models use the weak features.

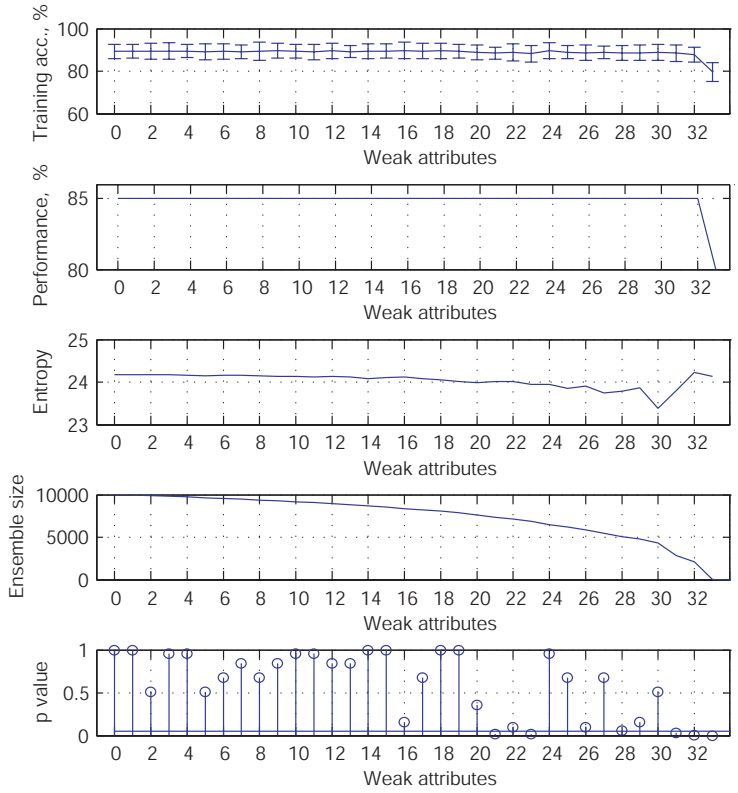

Figure 5.8: Finding a minimal feature subset for the classification of EEG in the two age groups, $p_{min} = 10$. From top: training accuracy, performance, ensemble size, and p-value of KS-test.

The p-value finally becomes below 0.05 for ($k = 31$) weak fatures, and subsequently 30 features are assigned as weak, and the remaining 6 features were most informative. For comparison, we reran the Bayesian technique on the data represented by these 6 features. The results presented in Table 5.5 show that the performances of the DT ensembles are comparable within a given confidence interval.

Fig. 5.9 shows the boxplots of the posterior probabilities of the 36 EEG features within five-fold cross-validation for the original and refined ensembles. Features which are most frequently used in the original ensemble represent the Theta and Alpha bands. The refining technique discarded the least frequently used (and therefore weakest) features without a decrease in ensemble's perfor-

Table 5.5: Performance and entropy of the DT ensembles

Original ensemble		Refined ensemble		Rerunning	
P, %	E, bits	P, %	E, bits	P, %	E, bits
85.0±10.6	0.113±0.026	84.5±14.3	0.106±0.025	85.5±9.6	0.107±0.018

mance. Overall, we can see that setting a larger number of data points in DT nodes enables selecting a smaller subset of features.

Examining the median importances of the features in Fig. 5.7 and Fig. 5.9, we can see that two of the most important features are found among the attributes representing the relative powers in both the Theta and Alpha bands. We can therefore hypothesise that the most accurate separation of the age groups is achieved by using the features from these bands. The electrode channels (right side, left side or their sum) within each of the bands are interchangeable in terms of the contribution to classification.

Fig. 5.10 shows the scatter plot of the EEG recorded at 36 and 41 weeks of PCA in the space of two of the most important features. We can see that most EEGs in the two groups can be separated based on the two features in the Alpha and Theta bands.

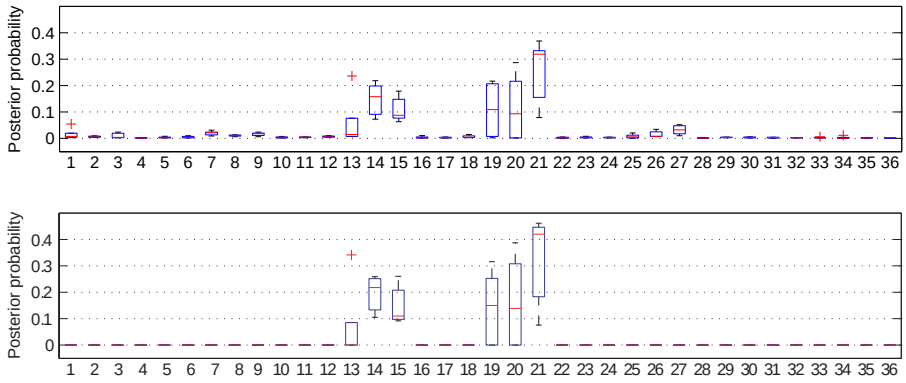

Figure 5.9: Probabilities of using features in the original and the refined ensemble, $p_{min} = 10$.

5.5 Chapter discussion and conclusions

We explored how the posterior information can be used within the methodology of Bayesian averaging over DTs in order to select a subset of EEG features mak-

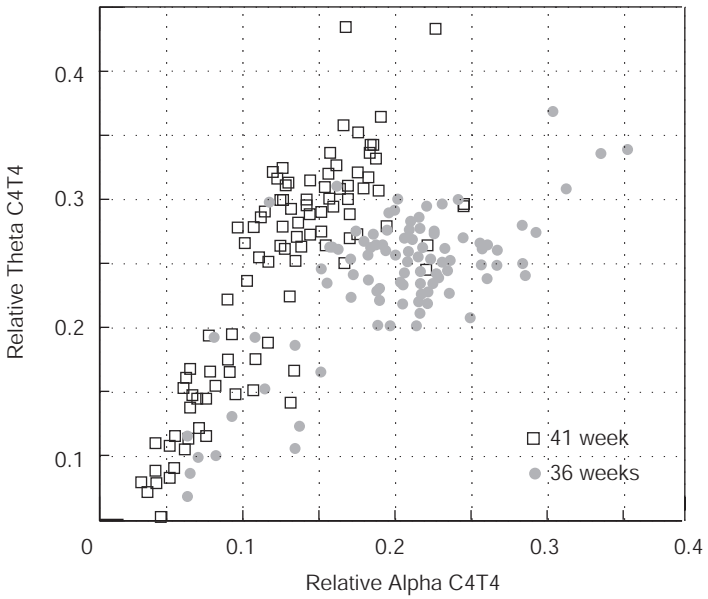

Figure 5.10: EEG recorded at 36 and 41 weeks of PCA in the space of two of the most important features

ing important contribution in the brain maturity assessment. We hypothesised that the posterior information about feature importance can be used to refine the DT ensemble from models using features found making weak contribution.

We assumed that the use of weak features within Bayesian averaging over DTs unnecessarily increases the model parameter space, which needs to be explored in detail to achieve proportional sampling from areas of interests. Besides, the use of weak features obstructs the interpretation of the ensemble. In general, the larger the number of weak attributes, the greater is the number of DT models using these features, and the greater is their negative impact on performance. We expect that the discarding of models using weak attributes will reduce the negative influence on the classification.

A technique we proposed for refining DT ensemble has been tested on two EEG assessment problems. The first problem was classification of EEG in six age groups from 36 to 41 weeks of PCA, represented by 72 spectral features. The results showed that the set of features could be reduced, on average, to 46. At the same time, the mean performance was increased by 1.8% to 29.2±6.9% and the uncertainty was slightly decreased. When we rerun the Bayesian averaging on EEG data of reduced dimensionality represented by the most important features, a similar increase in performance was observed. This result supports

our assumption that dimensionality reduction provides better conditions for proportional sampling.

The second problem was classification of EEG of pre-term and full-term newborns in two age groups, 36 and 41 weeks of PCA. The EEG were represented by 36 spectral features. The results of experiments run on this problem showed that the set of EEG features can be reduced from 36 to 19 on average, keeping the performance 86.5±7.6 on average.

We expected that the set of EEG features could be further reduced by increasing the minimal number of data points allowed in DT nodes. We found that a larger pruning factor encourages growing shorter DTs, which use a smaller set of EEG features. As a result, an ensemble of DTs using 6 features was selected without a significant decrease in the accuracy of assessment. These features represent mainly Theta and Alpha bands.

The results of the proposed technique were comparable with the results obtained by rerunning the Bayesian averaging on the EEG data without the weak features. To find the minimal set of most important features, the rerunning technique requires multiple runs of the Bayesian averaging each of which takes hours to complete. The proposed technique has been shown to provide the comparable performance without the need of reruns.

Chapter 6

Extraction of EEG features

In this chapter, we extract EEG features to obtain more information for Bayesian assessment of brain maturity. We hypothesise that the new features will complement the standard spectral powers, and their use will increase the accuracy of assessments. To test the hypothesis, we run Bayesian classification on EEG data represented by the new feature sets.

First, we explore extraction of time-domain EEG characteristics, assuming that discontinuity is the most important maturational feature. The techniques explored in Sections 6.1 and 6.2 assume that discontinuity is conventionally defined in terms of durations of inter-burst intervals and variability of amplitude. Specifically, in Section 6.1 we extract information on durations of inter-burst intervals, bursts, and continuous intervals, and show that this information is relevant for maturity assessment. Section 6.2 describes application of an aEEG technique to extract features related to continuity. The amplitudes of the aEEG borders, reflecting the variability of EEG amplitudes, are quantified to be used as new features. These features are shown to slightly improve the accuracy of assessments, when used with the spectral powers.

In Section 6.3 we hypothesise that discontinuity can be estimated as EEG non-stationarity, and propose a new feature extraction technique. The new features representing EEG non-stationarity significantly improve the assessment accuracy, and are shown to outperform the conventional discontinuity estimates.

In Section 6.4 we extract new spectral features. We hypothesise that the ratios of spectral powers are more informative than the individual powers, as shown in (Holthausen et al., 2000; Lippe et al., 2007). The ratio of absolute powers in the Alpha and Theta bands is found most informative, and its use increases the assessment accuracy. Overall, the highest accuracy is achieved by supplementing the standard spectral features with both the non-stationarity estimate and the Alpha/Theta ratio.

6.1 Detection of bursts, inter-burst intervals and continuous activity

EEG experts have observed that durations of bursts, inter-burst intervals and periods of continuous activity reflect EEG maturation levels. Based on these observations, we hypothesise that EEG features describing these durations will be informative for assessment of brain maturity.

To extract the features, first, we detect the bursts, inter-burst intervals and continuous intervals in EEG. In contrast to conventional approaches to event detection requiring the amplitude thresholds to be set (West et al., 2011; Jennekens et al., 2011), we use a detector based on an Artificial Neural Network (ANN) trained on examples of the EEG events from the three classes. ANNs have been shown promising to detect different types of EEG waves (Cooper et al., 2003).

We estimate the durations of the events of each class in an EEG recording to represent the continuity information. The durations, summarised by histograms, are then used as new features for the assessment of EEG maturity. In our experiments, we us a DT classifier to differentiate two age groups, each including 65 recordings. To test our hypothesis, we compare the classification accuracies obtained with the new features and with the standard spectral powers.

6.1.1 Codebook of events

To train an ANN detecting the EEG events, a codebook of 90 segments with 1-sec durations representing bursts, inter-burst intervals and continuous activity was collected. The segments were picked manually from real recordings according to descriptions of EEG events in the literature (Cooper et al., 2003; Boylan et al., 2008; Mizrahi et al., 2003). Bursts and inter-burst intervals were selected from discontinuous EEG recorded from pre-term newborns. Burst segments contained Subdelta, Delta and Theta waves with maximum amplitude of 50-80μV. Inter-burst intervals were with very low amplitude close to 0μV. Segments containing Theta, Alpha and Beta waves with maximum amplitudes of 0-20μV were marked as continuous. Fig. 6.1 shows examples of segments chosen to represent the three types of EEG events.

6.1.2 Detection of events

Having collected the codebook, we trained an ANN to classify the events. The 1-sec segments in the codebook were preprocessed using the fast Fourier transform into 64 spectral powers. The spectrum represents the information on the frequency and amplitude of EEG activity, which enables the types of events to

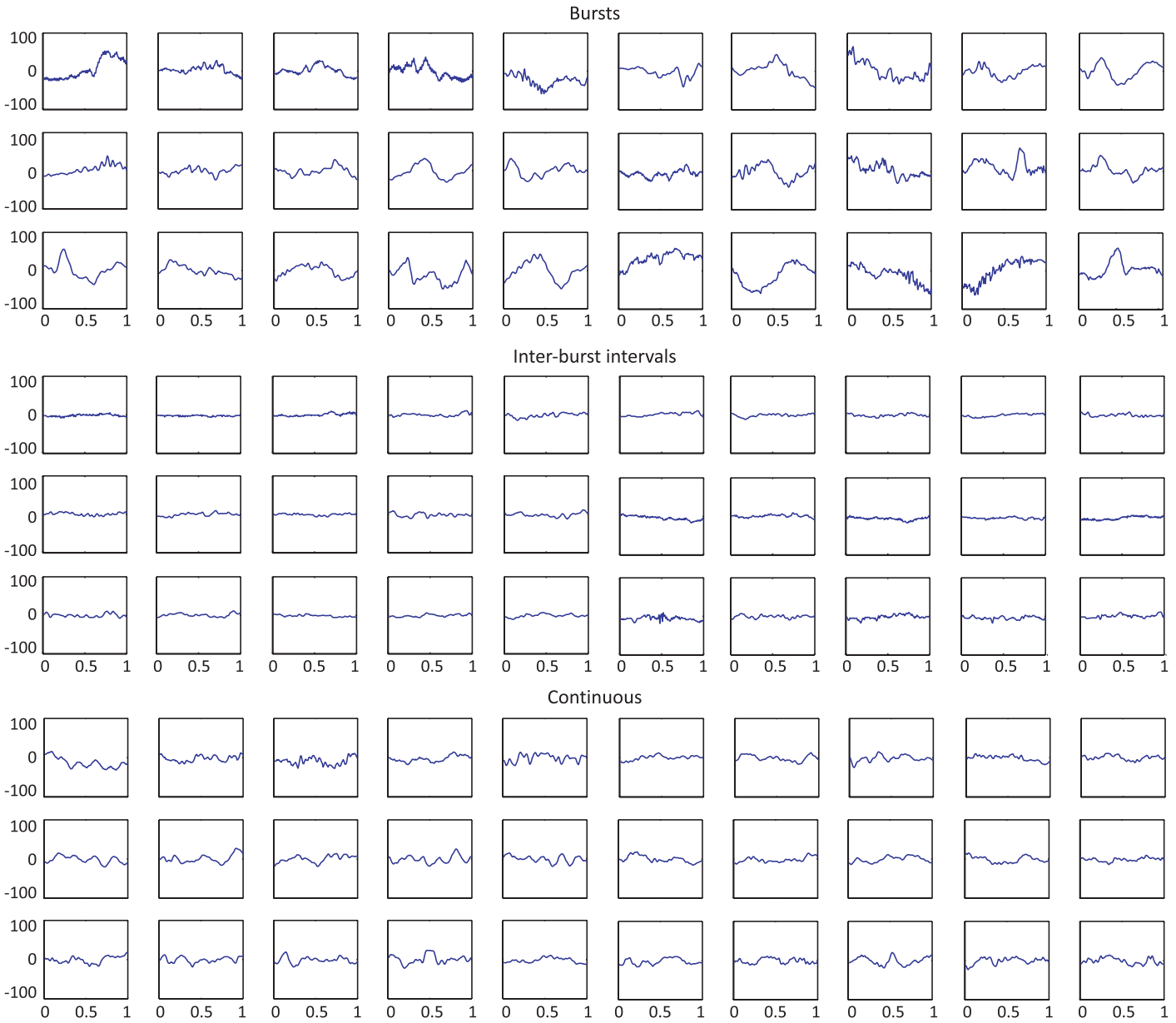

Figure 6.1: Examples of EEG segments representing the bursts, inter-burst intervals and continuous activity. The horizontal axes show seconds, the vertical axes show μV.

be distinguished. To reduce the dimensionality of data, Principal Component Analysis was applied to the 64 spectral powers. The principal components that contributed less than 0.1% to the total variation in the data set were eliminated and the remaining 14 components were used to represent the data. Then the data were classified with an ANN and DT using the Matlab Neural Network and Statistics Toolboxes. The performances were tested within the 10-fold cross-validation so that 81 of the 90 segments were used for training and 9 for testing.

The best performance of ANN was achieved with five hidden neurons with sigmoid activation. The network was set up with three output neurons, and the classification outcomes were determined by the unit with the highest activation value. It was found that either a linear or sigmoid activation function could be used for the output units without significantly affecting the performance. The resilient backpropagation method with the gradient descent momentum, and the learning rate 0.01 were found to be the best settings for training. Each fifth sample of the training data was chosen for validation. For the DT, the minimal number of data points in the terminal nodes was set to one, and the splitting criterion was maximum deviance.

The ANN provided the performance of 85.7±29.4% and the DT 76.9±28.3. For the ANN, on most of the folds, the performance was close to 90 to 100%. However on one fold it was 56%. The large variation of performances may be caused by subjective or erroneous labelling of some of the codebook segments. Table 6.1 shows the confusion matrix for the classification of EEG events. We see that

Table 6.1: Confusion matrix for classification of the EEG events

	Iinter-burst	Continuous	Burst
Iinter-burst	28	3	0
Continuous	2	21	2
Burst	0	6	28
Total	30	30	30

We applied the trained ANN providing the best performance to classify segments in EEG recordings. EEG signals were processed using a 1-sec window sliding with a step of 25 points. In each window, the fast Fourier transform was applied and the trained ANN was used to classify the 1-sec segments as burst, inter-burst intervals or continuous activity. Fig. 6.2 shows the classification outputs for 20 sec of two EEG recordings. In the absence of reliable labels for all portions of these recordings, the classification performance cannot be determined. However, we can see that the first EEG recorded at 28 weeks PCA

is discontinuous and there are obvious low-voltage inter-burst intervals during approximately the seconds 0–2, 12–16 and 18–20. Bursts can be seen during the seconds 8-10 and 16-18. Observing the maximal outputs of the burst and inter-burst interval classes, we can see that the classification of the EEG events is plausible during these periods.

The second EEG recorded at 41 weeks has no discontinuities and we can see that the outputs for the inter-burst class are small. Most of the EEG has been classified as burst or continuous. We see that no bursts were confused with the inter-burst intervals, and all the errors were made in discrimination of continuous activity from the other types.

6.1.3 Segmentation of events

Having obtained the classification outcomes for the bursts, inter-burst intervals and continuous activity on the windowed signal, we can segment the events in the signal according to the maximal outcome for one of the three classes. Because of the variations in data, some of the windows were assigned to a different class than most of the surrounding activity. Such random variations caused interruptions in the segmented events.

To suppress the random variations and enhance the dominant type of activity, the classification outcomes were smoothed with a moving average over 100 samples. Furthermore, the classification was uncertain for some of the windows, especially on transition between the different events, so that the outcomes for the classes were in a similar range. Therefore, to avoid misleading segmentation of events, the windows for which the difference between the maximal classification outcome and the next highest outcome was less than 5% were marked as "transitional". An example of the final segmentation is shown below in Fig. 6.3. Such segmentation enables the durations of events to be estimated.

6.1.4 Maturity assessment

In the experiments we tested a hypothesis that the durations of the segmented EEG events can be informative for brain maturity assessment. To test the hypothesis, the durations of bursts, inter-burst intervals and continuous periods were computed for 130 recordings in two age groups, 36 and 41 week PCA, each group included 65 recordings.

From each recording we extracted 75 features representing distribution of the durations of the events. The number of the event features was made similar to that of the spectral features in order to compare the informativeness the features, keeping the dimensionality of the model parameter space similar.

To extract these features, first the durations of the segmented events lasting from 0.5 to 50 sec were counted and then the counts for each of the three event types were summarised with a 25-bin histogram. Thus, the first 25 of the features represented the durations of bursts, next 25 – durations of continuous activity, and the last 25 – durations of inter-burst intervals.

A single DT used to classify this data provided the performance of 70.0±20.17% over five data folds. For comparison, the same recordings were represented by the conventional 36 EEG attributes and classified by a DT. The performance achieved on this data was 77.2±20.13%. Combining the event durations and spectral features did not improve the performance.

6.1.5 Conclusions on section

The average performance of age classification based on the durations of segmented EEG events was 70.0%, that is, 20% better than random guess. We can conclude that the durations of events can be promising features for age assessment. However, in the current experiments, the use of standard spectral powers enabled achieving a significantly higher performance, 77.2%.

To provide more information for the age classification, the technique of EEG event segmentation needs to be improved. One drawback of the current implementation is that the categorisation of EEG events into bursts, inter-burst intervals and continuous activity may be too general, and more types of events need to be distinguished to improve the discrimination of the neighbouring age groups. For example, the continuous activity may be presented as slow wave sleep or low-voltage irregular pattern, and the proportions of these patterns vary at different PCA. Detection of such events will be explored in future work.

It will be interesting to explore how hidden Markov models can be applied to segmentation of EEG events. This technique enables the prior information about durations of EEG events to be employed in segmentation, and it has been successfully applied to recognition of patterns related to mental and motor tasks in EEG see e.g. (Lederman and Tabrikian, 2012).

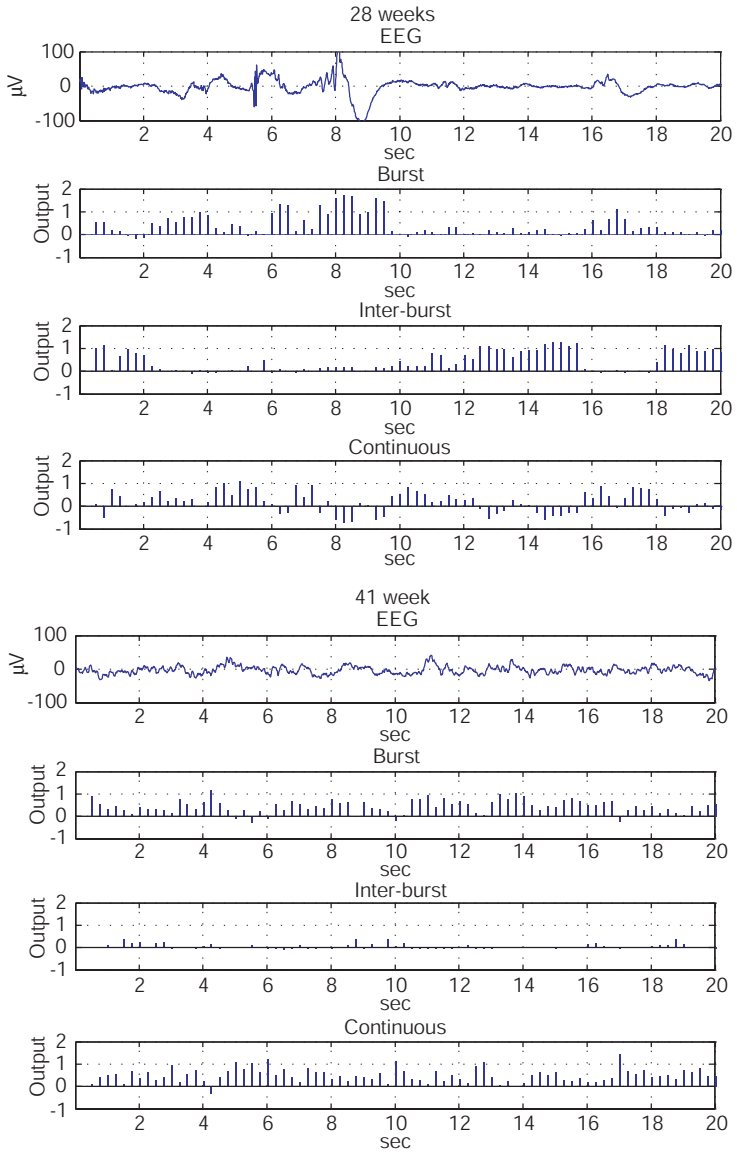

Figure 6.2: Classification outputs for EEG recorded at 28 weeks (upper plot) and 41 week (lower plot).

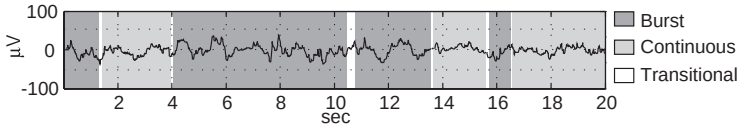

Figure 6.3: Segmentation of EEG recorded at 40 weeks PCA.

6.2 Envelope and aEEG

The amplitude-integrated EEG (aEEG) shows the peak-to-peak amplitudes, or envelope (Cooper et al., 2003). Looking at aEEG, clinicians can recognise different levels of continuity. To assess the continuity, the amplitudes of the lower and upper borders of the aEEG are measured, and according to (Burdjalov et al., 2003),the EEG at different PCA can be distinguished based on these amplitudes.

Typically, the aEEG amplitudes are measured manually by experts. In this section we attempt to measure the amplitudes automatically to extract important features for brain maturity assessment. We implement an envelope detection technique similar to the aEEG technique described in (Hellstrom-Westas et al., 2008). Then we use the statistics on the lower and upper borders as features for maturity assessment. We hypothesise that these features, characterising EEG changes in time domain, supplement the standard spectral features, and so their use will improve the performance of EEG age classification. To test this hypothesis, we use the envelope and spectral features with Bayesian classification of 210 EEG in two PCA groups.

6.2.1 Envelope detection

Similarly as in the aEEG technique (Hellstrom-Westas et al., 2008), in our experiments the raw EEG signal was filtered by a band-pass filter with cut-off frequencies of 2 and 15 Hz to suppress the artefacts. Then the signal was rectified, and the peak-to-peak amplitudes were detected using a simulated smoothing capacitor.

The implementation of the smoothing capacitor that was used to detect the peak-to-peak amplitudes is summarised in Algorithm 3. According to this algorithm, the peak-to-peak amplitude follows the voltage of a capacitor (peak rectifier), which is charged when the amplitude of EEG signal, X, exceeds the capacitor's current voltage, V. As long as the EEG amplitude is lower than V, the capacitor discharges exponentially until a new peak in the signal is encountered. An example of the resultant peak-to-peak amplitude is shown in Fig. 6.4.

As described in (Hellstrom-Westas et al., 2008) the peak-to-peak amplitudes counted for the signal were finally smoothed with a window of 50 samples. After that the upper and lower borders of the envelope were found as the average maximum and minimum amplitudes of the peak-to-peak output. To obtain the upper border, the maximum amplitudes were found in non-overlapping windows of 10 sec in duration, and then the maximums in each ten subsequent windows were averaged. Similarly, for the lower border, the the minimal amplitudes were computed in the 10-sec windows and averaged.

Algorithm 3 Detection of peak-to-peak amplitudes

Inputs: X, τ
Initialise: current amplitude $V = 0$, initial amplitude $V_0 = 0$, $t = 0$
$X \leftarrow |X|$
$n \leftarrow \text{NumberOfSamples}(X)$
for $i \leq n$ **do**
    if $X(i) > V$ **then**
        $V \leftarrow S(i)$
        $t \leftarrow 0$
        $V_0 \leftarrow V$
    else
        $t \leftarrow t + 1$
        $V \leftarrow V_0 e^{(-t/\tau)}$
    end if
    $A(i) \leftarrow V$
end for
return A

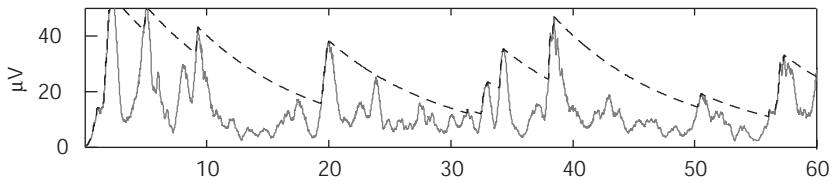

Figure 6.4: Peak-to-peak amplitudes (dashed) in a rectified EEG (solid).

Fig. 6.5 shows an example of the obtained aEEG-like signal with the upper and lower borders along with the corresponding original EEG. We can see that the amplitude of the lower border decreases, whereas the amplitude of the upper border becomes higher during the quiet sleep stages approximately at the minutes 0–30 and 70–100. Overall, the bandwidth, or the difference between the lower and the upper border, is greater during the quiet sleep, similarly as described in (Hellstrom-Westas et al., 2008).

The amplitude of the lower border is dependant on the rate of capacitor's discharge. The slower the capacitor discharges, the higher will be the amplitude during EEG periods with frequent bursts. The rate of discharge can be controlled by tuning the constant τ in the Algorithm 3. This rate has to be tuned so that the resultant envelope would enhance the different levels of EEG continuity.

The left column in Fig. 6.6 shows the envelope along with the lower and upper borders given different values of τ for an EEG recorded at 36 weeks PCA. The quiet sleep periods can be seen approximately during min 0-30, 90-120 and

190- 220. We can observe that setting a greater value of τ enhances the difference in the amplitude of the lower border during the quiet and active sleep stages. In our experiments with EEG age classification, we will extract the features from the envelopes constructed with $\tau = 100$ and $\tau = 200$, as these settings enhance the sleep cycle variations.

We can also observe high-amplitude points in the envelope. These points correspond to the peaks in the EEG signal, and some of these peaks with the highest amplitudes are likely caused by artefacts. We can expect that the characteristics of the upper border of the envelope will be negatively affected by the artefacts. Attempting to reduce the negative influence we can detect the envelope of EEG signal from which the artefacts were removed. The right column in Fig. 6.6 shows the envelope along with the borders detected in the same EEG recording after removal of high-amplitude artefacts. We can see that the sleep cycle variations of the lower border remain similar to those of the border detected in the raw EEG. However, for the upper border we see that during the quiet sleep stages the amplitude became lower than during the active sleep, which is not typical for aEEG. In our experiments, we will explore the informativeness of aEEG features extracted from raw signal as well as after removal of artefacts.

6.2.2 The envelope features

Having found the lower and upper borders of the EEG envelope, we can calculate the statistics of the amplitudes of the borders to be used as features for maturity assessment. It was observed that the distributions of the amplitudes were slightly skewed, and so they could be better described by statistics of an asymmetrical distribution rather than by those of the normal one. Assuming that the data have to be modelled by a family of distributions other than the normal one, the statistics of the data have to be estimated as parameters of the model distribution.

Following (Wong and Abdulla, 2008) we chose the log-normal distribution to model the distribution of EEG envelope amplitudes. The μ and σ of the amplitudes were then estimated as parameters of a log-normal distribution fitting the data. Fig. 6.7 shows the histograms of the amplitudes of the upper and lower borders along with the fitted distributions.

Additionally, to the μ and σ the minimum and maximum amplitudes of the borders were counted. In total, the envelope borders provided eight new features: the four statistics of the upper border: μ, σ, minimum and maximum, and the same four statistics of the lower border.

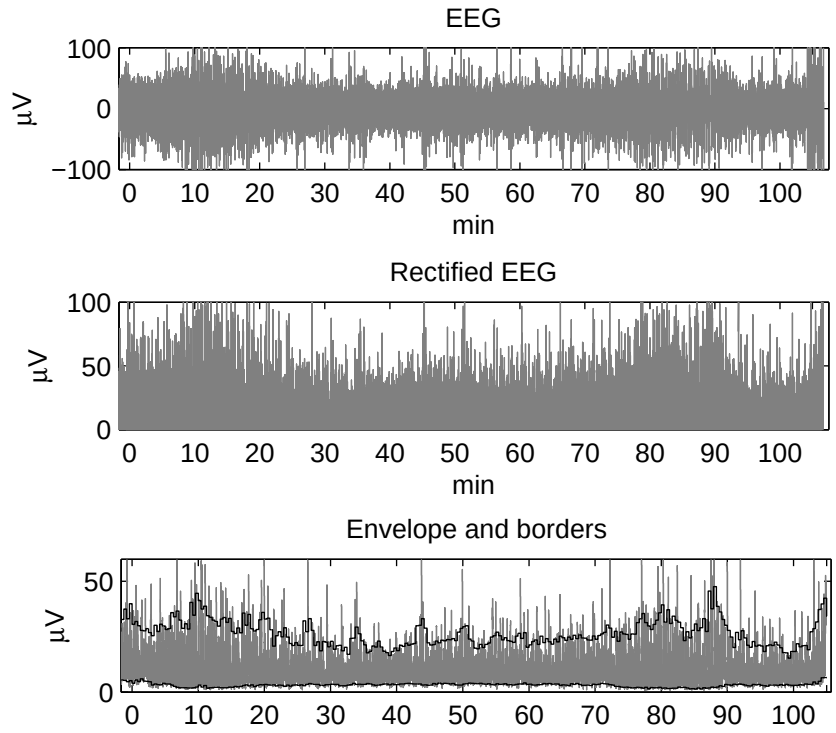

Figure 6.5: Detection of the lower and upper borders of EEG envelope. From top: original EEG, EEG after filtering and rectification, the envelope (in grey) with the lower and upper borders (in black).

6.2.3 Maturity assessment

To compare the new envelope features with the standard spectral features, we used the Bayesian averaging over DTs with the extracted features for classification of EEG maturity of newborns at ages 36 and 41 (36/41), and 37 and 39 (37/39) weeks PCA. Each age group was represented by 105 sleep EEG recordings. As the EEG maturity of pre-term (36 weeks) and full-term (41 weeks) newborns is different, the accuracy of classification of these EEG is expected to be high. In contrast, we expect that EEG maturity patterns of newborns at ages of 37 and 39 weeks are more close, and the classification accuracy is expected lower.

The Bayesian technique was run with the following settings. In a burn-in phase we collected 200,000 DTs, and in a post burn-in phase 10,000 DTs. During the post burn-in phase each 7th model was collected to reduce the correlation between DT models. The minimal number of data samples allowed to be in DT nodes (pruning factor) was set to six. Proposal variance was 1.0, and proba-

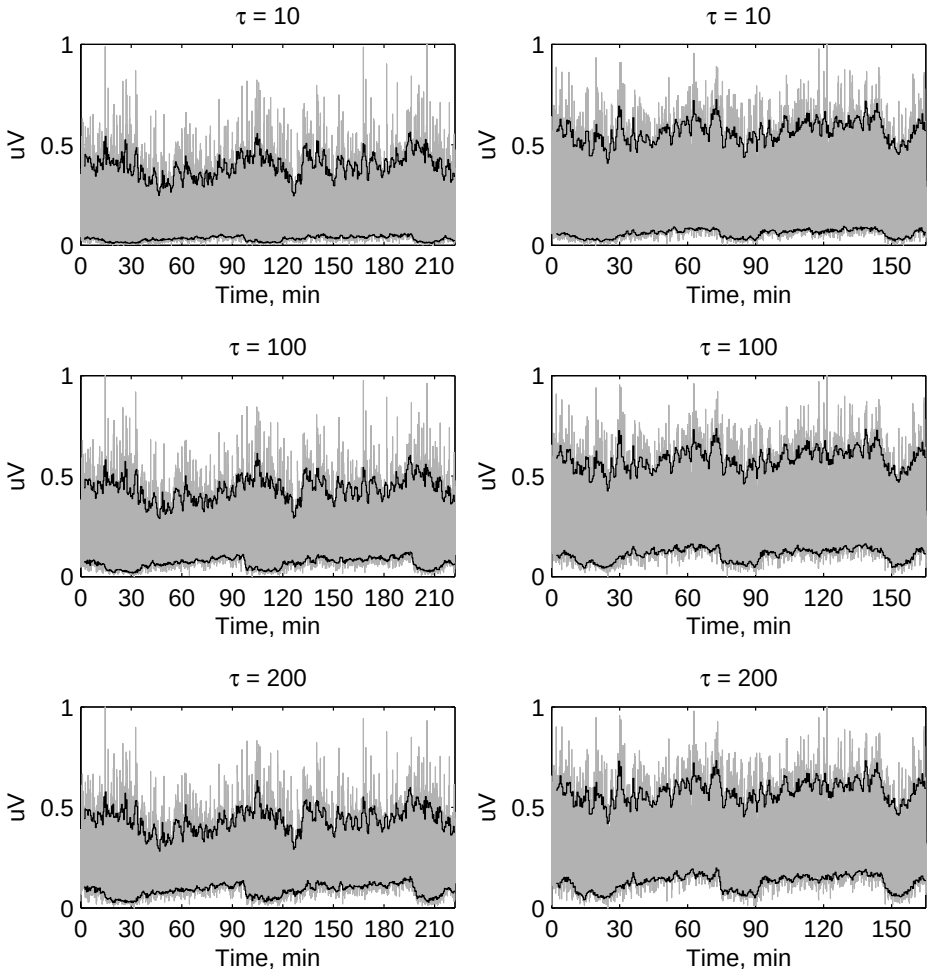

Figure 6.6: EEG envelope (grey) along with the lower and upper and borders (black) given different constant τ. In the left column, no artefact removal was applied to the input EEG. In the right column, the artefacts with high amplitude were removed using statistical thresholding. The envelope has been scaled in the range of 0 to 1.

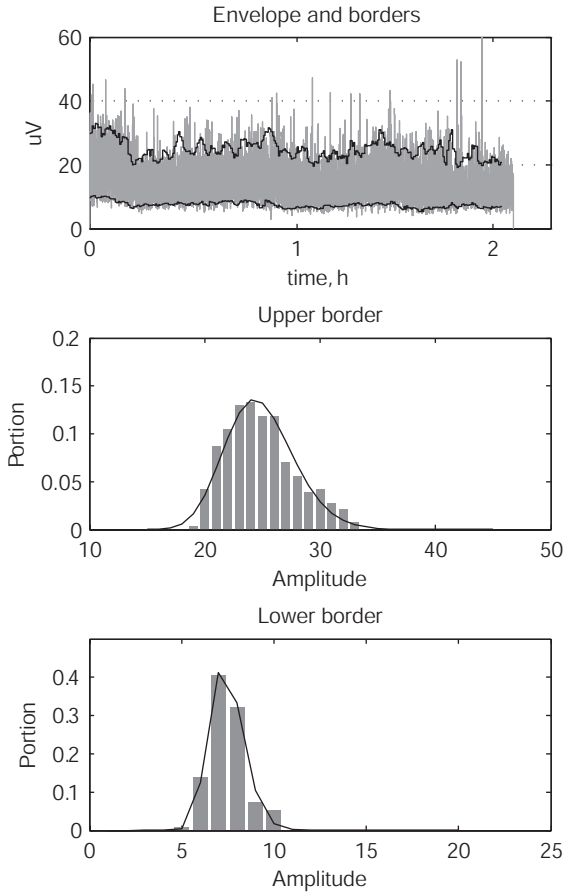

Figure 6.7: Distributions of the upper and lower borders of EEG envelope. The top plot shows EEG envelope (in grey) along with borders (in black). The middle and bottom plots show the histograms of the border amplitudes (in grey) approximated with log-normal distribution (in black).

bilities of making moves of birth, death, change variable, and change threshold were set to 0.15, 0.15, 0.1, and 0.6, respectively. For the two-class problem, running the MCMC technique for 200,000 DTs in Matlab on a 64-bit Linux PC took approximately 3 minutes.

First, in our experiments we compared the informativeness of the envelope features extracted from raw EEG and after the removal of artefacts. In both cases, the eight envelope features were appended to the standard 36 spectral features calculated on EEG without the artefacts. The artefacts were detected as high-amplitude outliers, as shown in Chapter 4. The best results were obtained when the envelope was constructed wit the setting $\tau = 100$.

Table 6.2 shows the performance and entropy of classifying EEG represented by the above features. The EEG were recorded at 36 and 41 weeks. The difference in the performances with the envelope features extracted before and after the removal of artefacts was not significant ($p > 0.48$).

Table 6.2: Performance and entropy for classification of EEG recorded at 36 and 41 weeks, represented by spectral features as well as by envelope features extracted before and after the removal of artefacts

With artefacts		Without artefacts	
P, %	E, bits	P, %	E, bits
87.1±8.0	0.07±0.03	84.3±8.0	0.08±0.03

In the experiments, we compared three different sets of EEG features. The first set (Spectral) comprised the 36 standard spectral features. The features were computed from the EEG after the removal of high-amplitude outliers, as shown in Chapter 4. The second set (Envelope) included the eight envelope features, and the third set (Combined) included the spectral as well as the new envelope features, in total 44.

Table 6.3 compares the performance and entropy of maturity assessment on the EEG data represented by the three feature sets. The performance and entropy are calculated within the five-fold cross-validation. For classification of 36 and 41 weeks, the use of the Combined set improved the average performance by approximately 2%. However, the improvement was not statistically significant, when tested with the Mann-Whitney U test ($p > 0.7$). For 37 and 39 weeks no improvement was observed.

Fig. 6.8 shows the boxplots of the performance and entropy counted within the five folds. For 36 and 41 weeks with the Combined set the minimal and maximal performances became higher whereas the entropies became lower.

Table 6.3: Performance and entropy for the EEG classification with different sets of features

	Spectral (36)		Envelope (8)		Combined (44)	
Weeks	P, %	E, bits	P, %	E, bits	P, %	E, bits
36/41	85.2±9.2	0.09±0.01	80.5±7.8	0.10±0.03	87.1±8.0	0.07±0.03
37/39	66.2±14.4	0.17±0.02	62.9±19.2	0.19±0.01	65.2±7.2	0.17±0.01

Fig. 6.9 shows the average frequencies of the 44 features being used by the DT models classifying the 36 and 41 weeks. We can see that the feature 14, Relative Theta power, and feature 43, the minimal amplitude of the lower bor-

der, were used with the highest probabilities, and so made the most important contributions to the maturity assessment.

It is interesting to explore the performance of Bayesian assessments using the two most important features. We found that the performance was 88.1 ± 10.6 and entropy was 0.08 ± 0.02. Thus, for classifying 36 and 41 weeks, the performance with the two most important spectral and envelope features was comparable to that obtained with the full set of spectral features ($p > 0.77$).

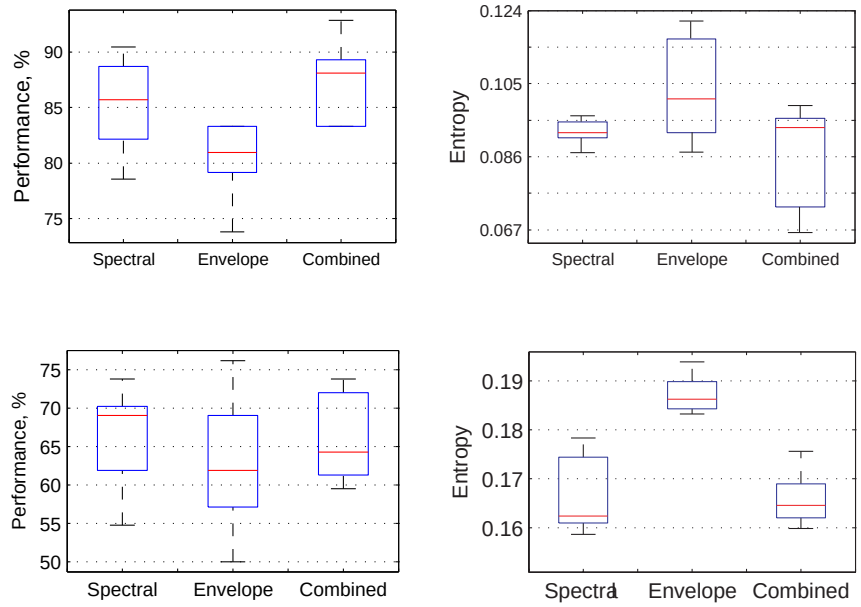

Figure 6.8: Boxplots of the performance and entropy with different sets of features for 36/41 week (top row) and 37/39 weeks (bottom row).

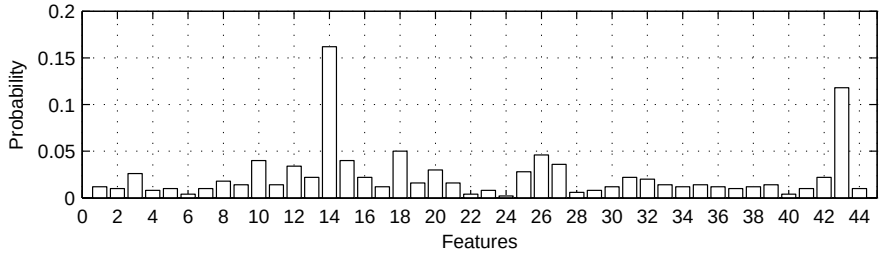

Figure 6.9: Posterior probabilities of features in the combined set.

6.2.4 Conclusion on section

We hypothesised that important features for the assessment can be automatically extracted from the EEG envelope which reflects the continuity. We estimated the envelope with a technique similar to that used for aEEG. We also hypothesised that the envelope features, carrying time-domain information, can complement the spectral features, and thus their use will increase the performance of maturity assessment.

To test the hypothesis, we used the BMA over DTs with the extracted features for classification of EEG maturity of newborns at ages 36 and 41 (36/41), and 37 and 39 (37/39) weeks PCA. In comparison to the standard spectral features, the use of the extracted envelope features improved the classification performance by 2% for weeks 36/41, for which the continuity is expected to be different. For weeks 37/39, the use of the new features provided no improvement in performance.

We explored the importances of the standard and new EEG features for classifying weeks 36/41, and found that the most important features were the relative Theta power and the minimal amplitude of the lower border of the envelope. The amplitude of the lower border was previously described as the most important envelope feature in (Viniker et al., 1984). This amplitude is most strongly influenced by the duration and amplitude of inter-burst intervals, and therefore it reflects the development of EEG continuity.

The remaining envelope features were making a much smaller contribution. One explanation of this may be that these features were affected by high-amplitude EEG artefacts.

Attempting to reduce the influence of artefacts, we extracted the envelope features after the removal high-amplitude artefacts from EEG. However, in our experiments, the removal of artefacts did not significantly affect the informativeness of the envelope features. A negative side effect of removing the artefacts before the detection of envelope was that the shape of the upper border of the resultant aEEG became distorted.

The performance of Bayesian assessment rerun with only the two most important spectral and envelope features was comparable to that obtained with the full set of spectral features.

We have confirmed our hypothesis that important time domain features, can be automatically extracted from the aEEG signal. However, the use of the new features did not improve the performance significantly. Future work will explore how the informativeness of the envelope features can be improved by using more complex techniques for removing EEG artefacts before detecting the envelope.

The informativeness of the envelope features will also be tested for classification of a larger range of ages.

6.3 EEG segmentation for measuring discontinuity

In this section we propose a new technique to evaluate the discontinuity of EEG as the rate of "non-stationarity". To evaluate the non-stationarity, we employ adaptive segmentation techniques for splitting the EEG into pseudo-stationary intervals. We assume that longer pseudo-stationary intervals are detected in signals which are more stationary, hence the rate of intervals will reflect non-stationarity of the signal. The rate of of pseudo-stationary intervals (or segments) refers to the number of intervals detected in an EEG relative to the length of the recording.

We show that the rate of the intervals is highly correlated with brain maturity. The statistics of the intervals are used as new features for the age classification, and the use of the proposed EEG features is shown significantly increasing the accuracy of age differentiation.

First, we discuss the rationale for using adaptive segmentation to assess the EEG discontinuity. We then briefly describe the main conventional techniques of adaptive segmentation of EEG and propose a new technique based on spectral power statistics. We use the techniques to extract the non-stationarity feature from synthetic and real EEG. We compare the non-stationarity feature with the conventional amplitude-based continuity estimates (Wong, 2008) in terms of their correlation with brain maturity. Having extracted the new EEG features, we run the experiments with Bayesian classification of newborn ages.

6.3.1 Adaptive segmentation in assessment of discontinuity

Discontinuous EEG is characterised by sharp changes in the amplitude and frequency during the bursts and inter-burst intervals. Because of these changes the EEG becomes highly non-stationary and complicated for analysis with conventional methods assuming a stationary signal. Adaptive segmentation aims to split EEG into pseudo-stationary intervals in which the EEG amplitudes and frequencies vary within an acceptably small range, see e.g. (Barlow et al., 1981). These intervals are then used for extracting EEG features. However within this technique few attempts have been made to measure the discontinuity which is an important maturity-related feature.

Recently, adaptive segmentation was employed in a technique for extracting the discontinuity features from newborn EEG (Wong and Abdulla, 2008). It was assumed that the discontinuity of EEG can be defined as the variability of its amplitude, and it was proposed to estimate the variability from statistics of the EEG envelope, which was composed from the mean amplitudes of the pseudo-stationary intervals. Within this technique, the mean absolute voltage of each interval was computed and repeated for the duration of the intervals. The discontinuity was then quantified by the parameters of the distribution of the envelope amplitudes. The parameters of the distribution were shown to be correlated with brain maturation for infants aged between 25 and 35 weeks PCA (Wong, 2008). However, the obtained values varied between patients and could only be used for EEG age classification.

The decrease of envelope variability during maturation of pre-term newborns may reflect the fact that the portion of the tracé discontinu, a pattern of long inter-burst intervals and high-amplitude bursts, becomes less pronounced between the 25 and 35 weeks (Pressler et al., 2003; Niemarkt et al., 2008). However we expect that this feature will not be as informative for assessing the maturation in late pre-term or term newborns, for whom the tracé discontinu normally disappears and the amplitude variability in the discontinuous patterns becomes lower. At the same time, the development of low-voltage and slow wave sleep stages may affect the EEG amplitudes.

A potential way of obtaining a more informative measure of EEG discontinuity for newborns aged older than 35 weeks is to take into account the durations of the pseudo-stationary intervals, which were not considered in the above approach. It is reasonable to expect that longer pseudo-stationary intervals will be detected in more continuous patterns with lower variations in frequency and amplitude, whereas shorter intervals will be detected in the more variable discontinuous patterns. A paper by Paul et al. (2003) reported interesting results obtained with such an approach. In this work, adaptive segmentation of EEG was employed in analysis of the quiet and active sleep stages of newborns. It was found that longer pseudo-stationary intervals were detected during the active sleep EEG which is mostly continuous, whereas shorter intervals were detected during the more discontinuous quiet sleep. This observation suggests that the rate of pseudo-stationary intervals is dependent on the discontinuity of EEG.

In this section, we hypothesise that the adaptive segmentation can be used to evaluate EEG discontinuity as non-stationarity or the rate of pseudo-stationary segments. We also hypothesise that this rate will decrease with brain maturation. Our hypothesis is based on the clinical observation that during brain development the continuous EEG patterns become longer, while the discontinuous patterns become shorter and more continuous, that is, their amplitude and

frequency variation decreases, see e.g. (Pressler et al., 2003). Hence, a lower rate of pseudo-stationary segments will be detected in EEGs of more mature newborns.

Next, we briefly describe the concepts of adaptive segmentation techniques used for the proposed assessment of non-stationarity. The rates of segments found within the techniques will be compared on model and real EEG data.

6.3.2 Conventional segmentation techniques

The aim of adaptive segmentation is to automatically detect boundaries of pseudo-stationary intervals in EEG. The segment boundaries are typically found by evaluating the dissimilarity of EEG in one or more successive intervals. Typically, the dissimilarity is evaluated between two intervals called the reference and test windows. If the dissimilarity is small, samples of the reference and test windows are considered taken from a single stationary process. On the contrary, if the dissimilarity exceeds a given threshold, the samples are assumed taken from the different processes, and a segment boundary is assigned between the reference and test windows. The reference and test windows are typically made adjoined and sliding along the EEG (Paul et al., 2003; Krajca et al., 2009). Alternatively, the reference window can be fixed at the beginning of the segment while the test window is moved until the dissimilarity exceeds a given threshold (Bodenstein et al., 1985; Aufrichtig et al., 1991).

A more complex approach proposed by Appel and Brandt (1983) uses a two-stage segmentation technique to find the boundaries most precisely. At the first stage, the reference window with a fixed starting position is being grown as the adjoined test window is sliding forward until the dissimilarity threshold is exceeded. Then, the test window starts shrinking while the reference window keeps growing. The test window's starting position at which the dissimilarity becomes highest is assigned to be the segment boundary. The same segmentation technique has been employed in (Wong and Abdulla, 2008) for estimation of EEG envelope.

Within these techniques, the dissimilarity is often evaluated as the difference between the coefficients of auto-regression (AR) models fitted to the reference and test windows. This technique requires finding the proper threshold as well as the proper number of the coefficients.

An adaptive EEG segmentation technique proposed by Agarwal et al. (1998) aimed to estimate the dissimilarity between the adjoined reference and test windows by using the Nonlinear Energy Operator (NLEO). The NLO, Ψ, enabling to estimate "frequency-weighted energy" is computed as follows:

$$\Psi(x(n)) = x^2(n) - x(n-1)x(n+1) \tag{6.1}$$

The summed frequency-weighted energy computed in the test window is subtracted from that computed in the reference window to produce the dissimilarity criterion $G_{nleo}(n)$.

$$G_{nleo}(n) = \sum_{m=n-N+1}^{n} \Psi(m) - \sum_{m=n+1}^{n+N} \Psi(m), \tag{6.2}$$

where N is the window size.

To find the segment boundaries, a threshold is applied to $G_{nleo}(n)$. The threshold is adaptive and is automatically found in a sliding threshold adaptation window of a predefined size. For each position of this window, the threshold is found as the local maximum of $G_{nleo}(n)$. The proper length of the threshold adaptation window has to be found from experiments.

6.3.3 Segmentation using Spectral Power Statistics

Alternatively to the above techniques, we propose to segment the EEG by using estimates of the spectral dissimilarity of intervals. In our implementation, the dissimilarity is assessed over spectral power bands computed in the reference and test windows. The spectral estimates can be compared within a two-sample statistical hypothesis test such as the standard KS-test. The main idea behind our technique is to employ statistical hypothesis tests for comparing distributions of spectral powers for making decisions on the dissimilarity of EEG intervals. We refer to this technique as Spectral Power Statistics (SPS). A similar approach has been used for evaluating EEG non-stationarity (McEwen and Anderson, 1975); however, the motivation has been to test stationarity before applying the Fourier transform, rather than to extract a new feature.

Algorithm 4 summarises the main steps of the SPS segmentation technique. According to the Algorithm, two adjoined sliding windows W_1 (reference) and W_2 (test), both of length L, are moving along EEG signal X. At each position of the windows, the FFT is applied to segments of the signal within W_1 and W_2 to compute frequency spectra, S_1 and S_2. The components of the spectra falling within each frequency band defined in $Bands$ are summed to form the estimates of spectral powers within the bands, B_1 and B_2. Next, B_1 and B_2 are compared with a two-sample statistical test. Each of the band powers represents a value in the two data vectors B_1 and B_2 of size $nofbands$. The p-value of the test is compared to the threshold p_0. If the value is below the threshold, the signal portions within W_1 and W_2 are assumed to have different characteristics, and

thus a boundary of a pseudo-stationary segment is assigned at the point between the two windows. The algorithm returns the locations of all the boundaries, T.

Algorithm 4 Adaptive segmentation using Spectral Power Statistics

Inputs: X, L, $Bands$, $nofbands$, p_0
Initialise:
$T \leftarrow 0$, $tind \leftarrow 0$
$i_1 \leftarrow 1$, $i_2 \leftarrow i_1 + L$,
while $i_2 + L <$ length(X) **do**
    $W_1 \leftarrow (i_1 : i_1 + L)$
    $W_2 \leftarrow (i_2 : i_2 + L)$
    $S_1 \leftarrow \mathrm{FFT}(X(W_1))$
    $S_2 \leftarrow \mathrm{FFT}(X(W_2))$
    $B_1 \leftarrow \mathrm{Sum}(S_1, Bands)$
    $B_2 \leftarrow \mathrm{Sum}(S_2, Bands)$
    $p \leftarrow \mathrm{Test}(B_1, B_2)$
    if $p < p_0$ **then**
        $tind \leftarrow tind + 1$
        $T(tind) \leftarrow i_2$
    end if
    $i_1 \leftarrow i_1 + L$
    $i_2 \leftarrow i_2 + L$
end while
return T

The non-stationarity of the signal can then be estimated as the segment rate. The segment rate is calculated as the portion of segments in which a boundary has been detected: $sr = \frac{\|X\|/L}{\|T\|}$, where $\|X\|$ and $\|T\|$ denote the lengths of the vectors X and T. The larger the segment rate, the higher is the level of EEG non-stationarity.

In our implementation, the best results were obtained with the window length L set to two seconds, and the threshold $p_0 = 0.95$. Although the spectral powers are typically calculated for the six frequency bands, within our method, this set is extended in order to meet the requirements of a statistical test. In particular, the number of bands was increased to nine to provide enough samples for the standard KS-test. Because the Theta (3.5-7.5 Hz), Alpha (7.5-23.5 Hz) and Beta (13.5-19.5 Hz) bands are wider than the Subdelta (0-1.5 Hz) and Delta (1.5-3.5 Hz) bands, it was decided to split the Theta, Alpha and Beta bands into sub-bands and make the widths more uniform.

In the comparison with the conventional segmentation techniques, the proposed technique needs to define a smaller number of parameters. These parameters mainly include the following: the number and widths of the spectral bands, durations of the reference and test windows, a threshold level of rejecting the alternative hypothesis, and a form of distribution of the tested samples.

In the next section, we test the concept of using adaptive segmentation for assessing the non-stationarity in model EEG data. We compare the NLEO, AR and SPS-based segmentation techniques.

6.3.4 Measuring non-stationarity on model data

To test how the segment rate reflects the non-stationarity, we run the NLEO, AR and SPS segmentation techniques on model EEG with various levels of non-stationarity. To model a stationary signal, we first generated 10,000 samples (100 seconds) of white noise and then smoothed the noise with a moving average filter. To model different levels of non-stationarity we generated on this signal 3 to 15 random "bursts" by increasing the amplitude of randomly picked segments by 3 to 5 times and smoothing them with a moving average filter with adjustable window of size 5 to 20 samples.

The segment rates (SR) were calculated as the ratio of EEG segments, in which a boundary has been detected. We generated 1,000 signals for each level of non-stationarity to obtain statistics of the SR. We evaluated the relationship between the rates and the number of bursts with the Spearman rank correlation coefficient ρ.

The first segmentation technique employed the NLEO applied to the reference and test windows, each of one second duration. The duration of the threshold adaptation window sliding over the EEG was two seconds.

The correlation between the segment rate and the number of bursts was low ($\rho = 0.2$) as the numbers of pseudo-stationarity intervals were similar for all five non-stationarity levels. This likely happens because the NLEO technique assigns the maximal energy in the threshold adaptation window to be a threshold, so that pseudo-stationary intervals are most likely broken within the 2-sec threshold adaptation window.

The second EEG segmentation technique employs the auto-regression (AR) models computed in the reference and test windows of 2-sec duration. This technique starts a new interval when the cross-validation errors of the AR models exceeded a given threshold. For this technique, the correlation between the rate of intervals and the number of bursts was much higher ($\rho = 0.96$).

Within the proposed SPS technique we used the KS-test to compare the reference and test windows. The windows were adjoined, and their durations were set to two seconds. The spectral powers required by the SPS were computed with the fast Fourier transform and then summed within the standard frequency bands, whose number has been increased from six to nine, as described in the previous section. A segment boundary was assigned when the p-value of the KS-

test dropped below 0.95. For the SPS technique, the correlation was slightly lower ($\rho = 0.87$).

Fig. 6.10 shows examples of segmented signals with 3 and 15 bursts. For three bursts, we can see that the segment boundaries were assigned at the beginnings and ends of the bursts (seconds 12, 18, 72, 76, 84 and 88). Because of random variations in the signal, some unexpected boundaries were detected in the intervals in which no bursts were generated. Nevertheless, are relatively few segment boundaries within the more stationary parts of the signal. When the number of bursts is increased, the segment boundaries are assigned more frequently.

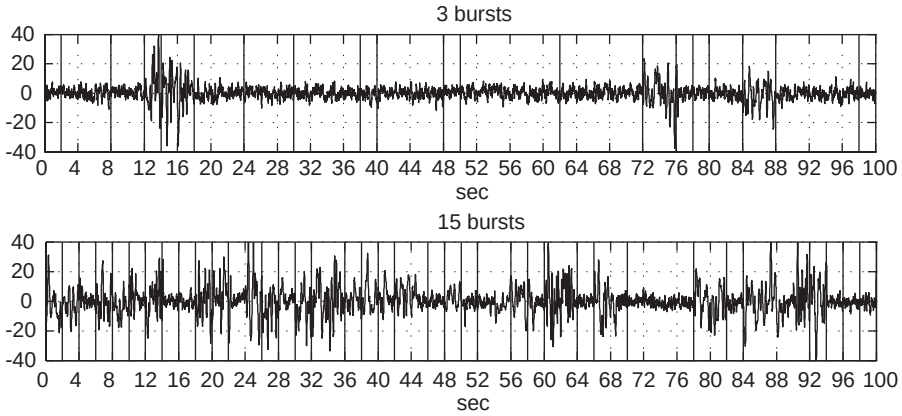

Figure 6.10: Segmentation of model signals with three and 15 bursts using the SPS technique.

From Fig. 6.11 we can see that for both the AR and SPS techniques the average segment rate (SR) tends to increase linearly as the number of bursts is increased from three to 15. Fig. 6.12 shows the distributions of the segment rates in the 1,000 generated signals with 3 and 15 bursts. The distributions of segment rates for three and 15 bursts are significantly different ($p = 0$).

In the next section we apply adaptive segmentation to estimate the non-stationarity of EEG recordings made at different PCA. We apply the described AR technique as well as the SPS technique. We compare the results of the SPS technique used with a number of hypothesis tests.

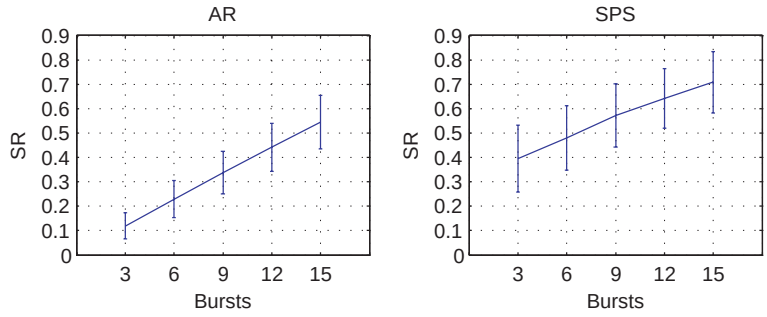

Figure 6.11: Correlation between the number of bursts and segment rate (SR) found with the AR and SPS-based segmentation techniques.

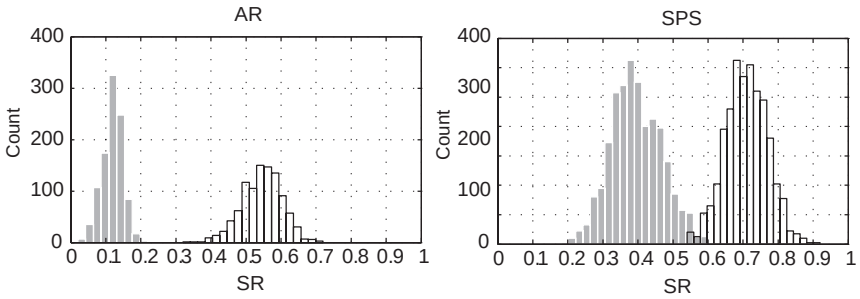

Figure 6.12: Distribution of segment rates (SR) for signals with three bursts (dashed) and 15 bursts (solid) for the AR and SPS-based segmentation techniques.

6.3.5 Correletion of non-sationarity features

The experiments were run on 260 EEGs recorded from newborns in 13 age groups between 32 and 44 weeks PCA, with 20 recordings in each group. The non-stationarity of an EEG recording has been evaluated by the segment rate, as described in Subsection 6.3.3 as well as by ten-bin histogram summarising the durations of pseudo-stationary intervals detected by a segmentation technique. The ten bins of the histogram represented the portions of segments the durations of which were from two to 20 sec.

In our experiments, we used the AR-based segmentation technique as well as the SPS technique with a number of statistical hypothesis tests used to compare the distributions of spectral powers counted in the adjacent sliding windows, according to Algorithm 4. Namely, we used the standard two-sample X^2 (chi-squared test), t-test, KS, and AD tests.

We found that the rates of segments found using the tests correlated with the PCA differently. Table 6.4 shows the relationships between the rates and ages, represented by the Spearman rank correlation coefficient ρ. All correlations were significant ($p < 0.05$).

The two sample X^2-test was the first test that we tried to employ within the SPS technique. This test is commonly used for comparing two distributions represented by histograms with the same bin locations. Therefore the X^2-test could be applied to compare the samples represented by spectral powers in the given frequency bands. We found that the highest value of correlation for the segment rates computed using this test was $\rho = -0.67$.

The X^2-test, whose critical values are based on the squared differences between the bin heights of the two histograms, could be excessively sensitive to EEG variations in the reference and test windows. Specifically, the variations of the Subdelta and Delta powers, that are typically the highest, can affect the test more strongly than the variations in the other bands. Furthermore, the waves in these bands are slowest and therefore their powers will vary within the short epochs. Increasing the window length would enable the Subdelta and Delta powers to be estimated more reliably. However, this could lead to poor detection of short discontinuities. Another possible problem is that the X^2-test is can be highly affected by the choice of the histogram bins, and thus it can be sensitive to slight variations of EEG power in the frequencies at which the edges of the bands are defined.

The KS-test, which evaluates the difference between the cumulative distribution functions of two samples, is expected to be more robust to such variations. This test also differs from the X^2 one in that it will treat each band power as an independent observation in a sample. This means that the test will not use the information about the band corresponding to each of the power values, and the distributions of powers will be assessed irrespectively of the order of bands. In the experiments we explore whether the KS-test with the above limitation is suitable for assessing EEG non-stationary. We also compare the results with those obtained using the more widely used t-test, which has the same limitation in analysis of EEG spectral powers.

The rates of segments found with the SPS technique employing both the t-test and KS test have shown similar correlation with PCA, ρ = -0.75, although t-test is typically applied to samples from a normal distribution, and KS test is suitable for arbitrary distributions.

A modification of KS-test known as the A-D test tends to assign larger weight to the tails of distributions (Scholz and Stephens, 1987; Trujillo-Ortiz et al., 2007). In our experiments, this test provided a better correlation (ρ = -0.81).

Table 6.4: Correlation between the PCA and segment rate.

Technique	ρ
AR	-0.57
X^2-test SPS	-0.67
t-test SPS	-0.75
KS-test SPS	-0.75
AD-test SPS	-0.81

Fig. 6.13 shows examples of segmentation of continuous and discontinuous EEG patterns. We can see that the segment rates are higher for a discontinuous pattern (a) as well as a pattern which can be considered as semi-discontinuous because of large variation in amplitudes (b). The rates are much lower for full-term newborns' EEGs with smaller variation in amplitudes (c and d). We can also see that during the quiet sleep (c) the segment rate is slightly higher than that during the active sleep (d).

In our experiments we used sleep EEGs which are typically contaminated by artefacts, which can affect the informativeness of the extracted features. For this reason, we are interested in investigating how artefact removal can improve the results of the proposed SPS-based segmentation technique.

For these experiments, we used the two techniques of removal of artefacts. The first technique (AR1) removes EEG samples with amplitudes exceeding a threshold assigned as the sum of the mean plus standard deviation calculated in a 10-min sliding window. The second technique (AR2) employs the NLEO to remove EEG samples with abnormal energy detected in a 3-min sliding window, as suggested in (Agarwal et al., 1998). The proportions of data removed as artefacts by these techniques were similar.

Table 6.5 shows the correlation coefficients after the artefact removal, and we can observe that the first technique (AR1) has increased the correlation for the KS and AD tests by 3%. The second technique has not improved the result.

Fig. 6.14 shows the rates of segments estimated by the KS-test after the removal of artefacts by the technique AR1 for newborns aged between 32 and 44 weeks PCA.

Table 6.5: Correlation between the PCA and rate of segments detected by the KS, AD and t-test after removal of EEG artefacts

Technique	Proportion of artefacts,	t-test	KS-test	AD-test
AR1	0.28±0.08	-0.76	-0.85	-0.85
AR2	0.29±0.20	-0.70	-0.71	-0.79

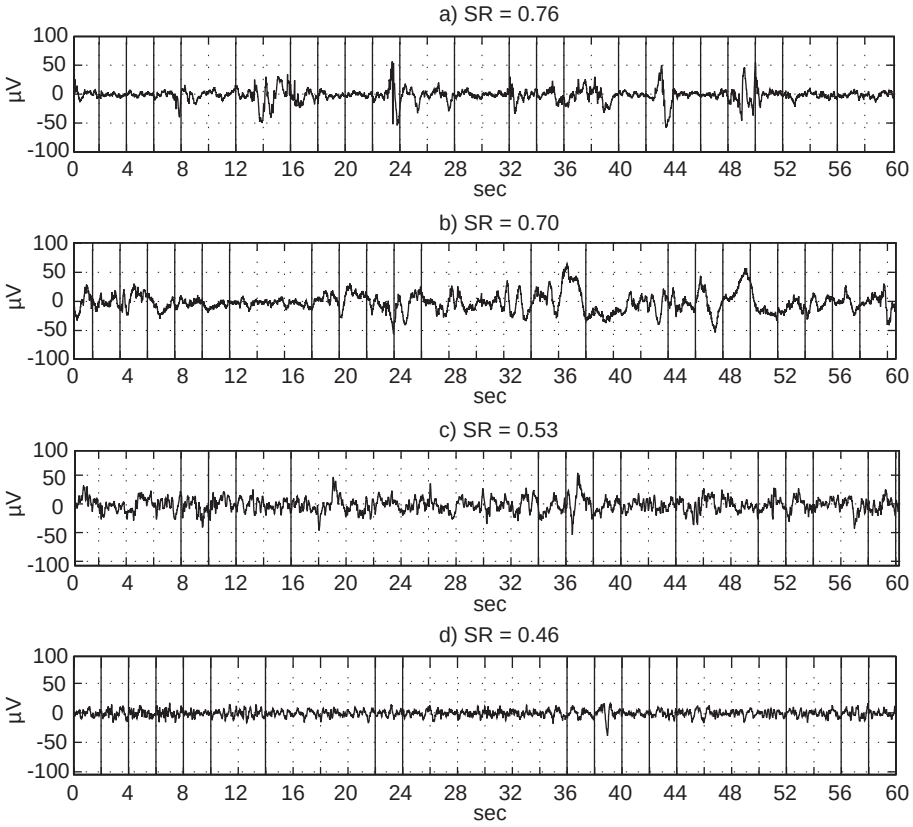

Figure 6.13: Segment rates (SR) for different EEG patterns: a) discontinuous pattern at 34 weeks, b) semi-discontinuous quiet sleep at 36 weeks, c) continuous quiet sleep at 41 week, d) continuous active sleep at 41 week.

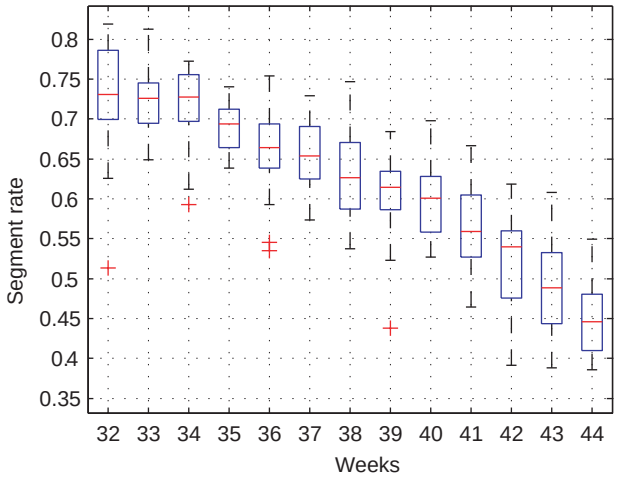

Figure 6.14: Correlation between the PCA and segment rate after the removal of artefacts.

6.3.6 Comparison with amplitude statistics

For comparison, we evaluated the correlation between the PCA and statistics of the amplitude vector, as proposed in (Wong, 2008). To obtain the amplitude vector, each pseudo-stationary segment was represented by its mean absolute amplitude. The value of the mean absolute amplitude was repeated for the duration of the segment, so that the dimensionality of the resultant amplitude vector was the same as that of the EEG. The distribution of the amplitude vector was modelled with a log-normal distribution. To assess the continuity of a recording, the mean (μ) and standard deviation (σ) of the log-normal distribution were counted.

Fig. 6.15 shows examples of two EEG along with the μ and σ counted in 10-min sliding windows with a step of 1 min. For the EEG recorded at 34 weeks PCA, there are two quiet sleep stages during the minutes 0–40 and 100–140, which are distinguished from the active sleep by the slightly higher average amplitude, and are typically more discontinuous. Similarly, as reported by (Wong, 2008), during these more discontinuous patterns, the μ of the amplitude vector's distribution becomes lower and the 2σ interval becomes wider. On the contrary, for the EEG recorded at 41 weeks, the μ increases during the quiet sleep. We can explain this change in the properties of the mean absolute amplitude by the following: closer to the term age, the inter-burst intervals become so short that their influence on the amplitude distribution is negligible. Although in the pre-term EEG the mean amplitude is highly influenced by inter-burst intervals, at term age, it reflects the differences in the dominant amplitudes of the continuous quiet and active sleep patterns, as seen in the lower plot.

Fig. 6.16 shows the average μ and σ values for the 260 recordings in the 13 PCA groups. We can see that between 32 and 34 weeks the both values are decreasing with maturation ($\rho = -0.42$ and -0.36) as shown in (Wong, 2008; Wong and Abdulla, 2008). However, after 34 weeks, the μ is increasing with maturation ($\rho = -0.48, p < 0.05$). This may be in part due to the brain growing larger and emitting stronger signals, in part due to the increase in the quiet sleep ratio. For the σ value, no correlation was found with PCA between 35 and 44 weeks.

Overall, the rates of segments found with the proposed SPS technique have shown stronger correlation with PCA during the weeks 32-44. In the next section, we use the segment rates as new features for Bayesian assessment of brain maturity.

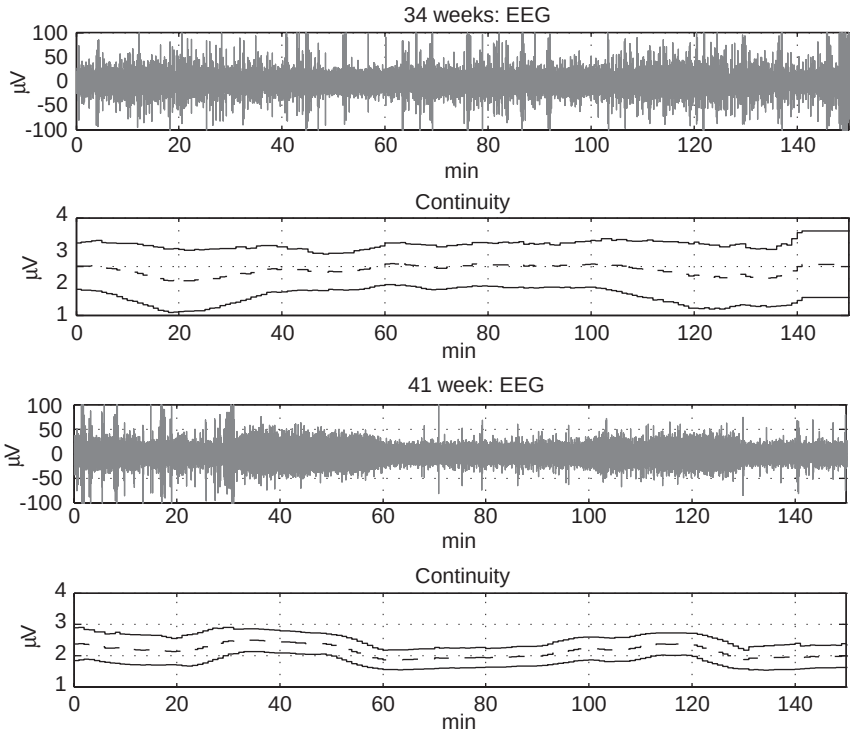

Figure 6.15: Continuity feature represented by μ (dashed) and 2σ (solid) of the amplitude vector's distribution for a pre-term (34 weeks) and full-term (41 week) EEG.

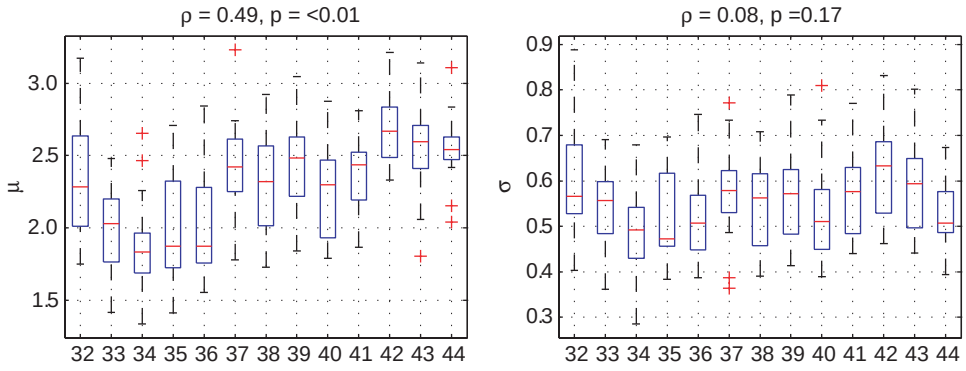

Figure 6.16: Correlation of the μ and σ of the amplitude distribution.

6.3.7 Classification of EEG maturity

In our experiments, we used the BMA over DTs with the features extracted from EEG for classification of EEG maturity of newborns at ages 36 and 41 (36/41), and 37 and 39 (37/39) weeks PCA. Each age group was represented by 110 sleep EEG recordings.

To clean the EEG from artefacts, the samples whose amplitudes exceeded the sum of the mean plus 1σ of amplitude were deleted. Then the EEG were segmented into pseudo-stationary intervals by using the SPS technique. In the experiments, we compared three different sets of EEG features.

The Set 1 comprised the conventional 36 spectral features which are the absolute and relative powers computed in the standard six frequency bands for the C3 and C4 electrodes and their sum. The Sets 2 and 3 comprised the features included in Set 1 and the new features representing the non-stationarity of an EEG recording estimated with the proposed SPS-based techniques which employ the KS and AD tests, respectively. The new features include the segment rate and 10-bin histograms of the counts of pseudo-stationary intervals ranging from 2 sec to 20 sec.

Table 6.6 shows the average performances (P) and entropies (E) of DT ensembles along with 2σ intervals obtained for the Bayesian classification of newborns aged 36/41 and 37/39. The performances of the BMA over DTs employing the Sets 1, 2, and 3 were compared within the five-fold cross-validation.

From Table 6.6 we can observe that for age classification at 36 and 41 weeks, the features of Set 3 improve the performance, on average, by 6.4%, and Set 2 only by 3.7%. For classification of ages at 37 and 39 weeks, the gains are 10.0%, and 8.7%, respectively. Overall, the use of the features from Set 3 increased the accuracy of age classification by 6.3% and 10.0%, respectively, for the two age groups. For the first group the increase in performance did not reach statistical signifficance when tested with the Mann-Whitney U test ($p < 0.06$), however for the second group the improvement was signifficant ($p < 0.02$).

Besides, the new features included in Set 3 decrease the uncertainty of DT ensembles in terms of entropy. The mean entropy of the DT ensembles has been reduced from 24.2 to 15.0 for classification of newborns at 36 and 41 weeks, and from 38.4 to 34.0 for classification of newborns at 37 and 39 weeks.

6.3.8 Conclusion on section

We proposed to evaluate EEG discontinuity as non-stationarity, or the rate of pseudo-stationary intervals found with adaptive segmentation. We hypothesised that the rate of intervals would be lower for more continuous EEG patterns and higher for the discontinuous ones. For continuous patterns the short-term vari-

Table 6.6: Performance (P) and entropy (E) for the EEG age classification with different sets of features.

	Set 1		Set 2		Set 3	
Weeks	P, %	E, bits	P, %	E, bits	P, %	E, bits
36/41	83.6±8.7	0.115±0.031	87.3±12.3	0.080±0.029	90.0±9.4	0.071±0.017
37/39	64.5±6.1	0.183±0.013	73.2±12.6	0.171±0.014	74.5±10.8	0.162±0.023

ability of amplitudes and frequencies is lower, and thus longer pseudo-stationary intervals can be segmented.

We also hypothesised that the rate of intervals would be correlated with brain maturation. Our hypothesis was based on the clinical observation that during brain development the continuous EEG patterns become longer, while the discontinuous patterns become shorter. Hence, we expected that the rate of intervals is an important feature for assessment of EEG maturity.

To segment the EEG into pseudo-stationary intervals, we proposed the new Spectral Power Statistics (SPS) technique testing the hypothesis of similarity between the distributions of spectral powers in the adjoined EEG intervals. We found that the SPS technique provides better results, in terms of correlation between the rates and PCA, than the conventional adaptive segmentation techniques based on auto-regression coefficients and the nonlinear energy operator. The performance of the proposed SPS-based segmentation technique has been explored with a number of standard hypothesis tests, namely X^2-test, t-test, Anderson-Darling (AD) and Kolmogorov-Smirnov (KS) tests. We found that the AD test and KS test provide a better performance in terms of the correlation coefficient value. We also compared the rates of intervals with the amplitude-based continuity estimates (Wong, 2008) and found that the rates provide a stronger correlation with PCA between 32 and 44 weeks.

The histograms of the pseudo-stationary intervals were used to extend the conventional set of spectral features for the Bayesian assessment of EEG maturity of newborns at ages 36 and 41 weeks as well as 37 and 39 weeks. The use of the new features has been shown to increase the accuracy of age classification by 6.3% and 10.0%, respectively, for these age groups. Overall, the classification accuracies have been increased to 90.9% and 74.5%, respectively.

6.4 Ratios of spectral powers

In this section we explore the interactions between the spectral powers in the six frequency bands with the aim to extract new information for maturity assess-

ment. The individual absolute and relative powers have been shown correlated with maturation, see e.g. (Bell et al., 1991b). Holthausen et al. (2000) showed that the correlations were stronger for the ratios Beta/Delta and Beta/Theta than for the individual absolute spectral powers. They explored the ratios in 94 EEG recorded between 28 and 112 weeks and found a significant decrease in the ratios between approximately 30 and 50 weeks.

A later study (Lippe et al., 2007) explored spectral power ratios in EEG recoded after visual stimulation from subjects in different age groups. The Alpha/Theta ratio has been shown associated with brain maturation in one month old newborns, children and adults. However the trend of this ratio has not yet been explored in pre-term and full-term newborns.

The above findings motivate us to explore the correlation between the spectral power ratios and PCA during 32-44 weeks. We expect that the correlations of the ratios will be stronger than those of the absolute powers.

6.4.1 Correlations of features with PCA

We explored the correlations on a dataset of 260 EEG recorded from newborns aged between 32 and 44 weeks PCA. Each age was represented by 20 recordings. Fig. 6.17 shows the correlations of the absolute amplitude spectra in the six bands over the weeks PCA. The powers were computed in 10 sec segments and then averaged for each recording. The strongest correlations ($\rho = 0.53$) were observed for the Delta and Theta bands.

Fig. 6.18 shows the correlations of the ratios Beta/Delta, Beta/Theta and Alpha/Theta. The ratios were calculated as quotients of the absolute powers. We can observe that all the ratios tend to decrease with advancing PCA between the weeks 34 to 44, whereas between the weeks 32 to 34 no correlation with PCA can be observed.

The Alpha/Theta ratio showed the strongest overall correlation ($\rho = -0.84$), although the decrease is not consistent over the whole range of weeks. On the contrary, the ratio seems growing between weeks 32-34, however, to confirm this trend, more data from this group have to be analysed.

6.4.2 Classification of EEG maturity

We run the Bayesian classification to evaluate the importance of the the Alpha/Theta ratio for classification of EEG maturity. It is interesting to compare the importance of the ratio with that of the non-stationarity features described in Section 6.3. The Alpha/Theta ratio and the non-stationarity features provide a comparably strong correlation with PCA, however we hypothesise that these

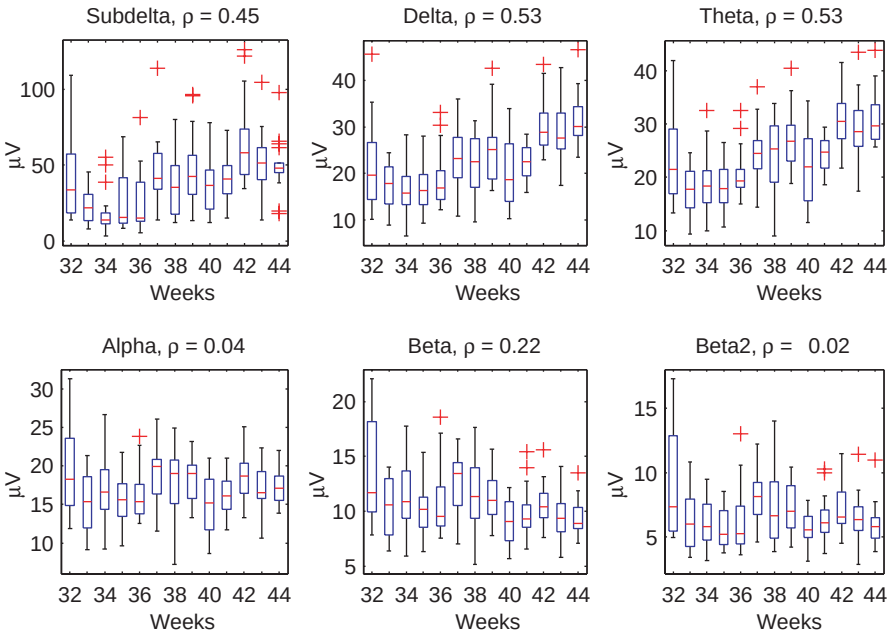

Figure 6.17: Correlations between the absolute amplitudes in the frequency bands and PCA.

features, describing spectral and time-domain characteristics, will be complimentary for the maturity assessment.

To test the hypothesis, we count the ratio on the same EEG recordings as those used in the experiments with features representing EEG non-stationarity. Namely, we use 220 recordings of newborns at ages 37 and 39 (37/39) weeks. The performances obtained on this data represented with the standard 36 spectral features (Set 1) and with the non-stationarity features (Sets 2 and 3) were shown in Table 6.6. For comparison, we represented these EEG with two new feature sets: Set 4 comprised the conventional features together with the Alpha/Theta ratio, and in Set 5 the previous features were combined with the non-stationarity estimates.

Table 6.7 shows the average performances and entropies of DT ensembles along with 2σ intervals obtained for the Bayesian classification. The performances were calculated within the five-fold cross-validation. From the results obtained with Set 4, we can see that supplementing the standard spectral features with the the Alpha/Theta ratio has increased the average performance by approximately 10% from 64.5% to 74.2%. The gain in performance is comparable to that provided by the non-stationarity features.

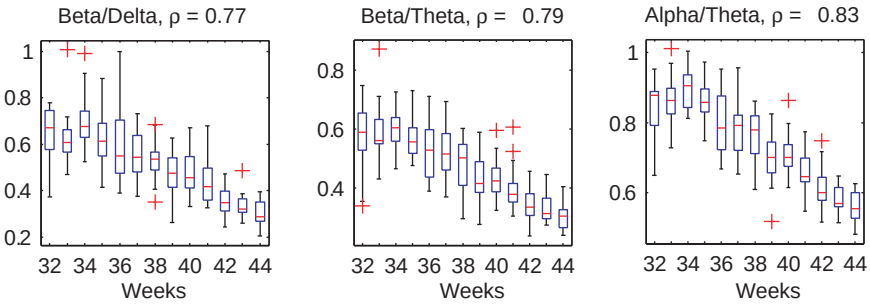

Figure 6.18: Correlations between the ratios of absolute powers and PCA.

However, the use of Set 5 increased the performance further to 80.7%. Altough the difference in performances for Set 4 and Set 5 is not statistically significant according to the Mann-Whitney U test ($p \approx 0.1508$), the Alpha/Theta ratio and the non-stationarity features seem to be complementary for improving the performance of maturity assessment. We can also observe that the average entropy decreased from 0.183 for Set 1 to 0.146 for Set 5.

Table 6.7: Performance and entropy of age classification with the different sets of features

Set 4		Set 5	
P, %	E, bits	P, %	E, bits
74.2±6.5	0.159±0.014	80.7±10.2	0.146±0.017

6.4.3 Conclusion on section

We hypothesised that the ratios of spectral powers Beta/Delta, Beta/Theta and Alpha/Theta are more strongly correlated with brain maturation than the absolute powers. We explored the correlations on a set of 260 EEG recorded at 32-44 weeks. The ratios were shown significantly correlated with advancing weeks PCA, and the correlations were stronger than those of the absolute powers. The trends of Beta/Delta and Beta/Theta ratios after 34 weeks were consistent with those reported by Holthausen et al. (2000).

The strongest correlation was, however, found for the Alpha/Theta ratio ($\rho = -0.84$), which has been previously shown to increase with maturation in EEG of infants, children and adults (Lippe et al., 2007). In contrast to these findings, we found that the ratio was decreasing. This difference in findings can be caused by the states of subjects: we used sleep EEG while Lippe et al.

recorded from awake subjects. Overall, the correlation of the Alpha/Theta ratio was similar to that of the EEG non-stationarity estimate counted as the rate of pseudo-stationary segments.

In the experiments with Bayesian classification of EEG recorded at 37 and 39 weeks compared the importance of the Alpha/Theta ratio with that of the non-stationarity features. We hypothesised that the features, describing frequency and time-domain characteristics, will provide complementary information for maturity assessment. We found that the use either of these feature types increases the assessment accuracy on average by approximately 10% making it 74%. However, in combination, the features increased the performance further to 81%. In subsequent chapters the combined features will be used to classify a larger range of PCA.

6.5 Chapter conclusion

We hypothesised that the new features extracted from EEG can complement the standard spectral powers and thus increase the accuracy of assessments. To test the hypothesis, we run Bayesian classification on EEG data represented by the new feature sets.

We explored extraction of features describing the level of EEG discontinuity, based on clinical observations that discontinuity is one of the most important maturational characteristics. We proposed to evaluate EEG discontinuity as its non-stationarity, or the rate of pseudo-stationary intervals found with adaptive segmentation. The new features were used to extend the conventional set of spectral powers for the Bayesian assessment of EEG maturity of newborns at ages 36 and 41 weeks as well as 37 and 39 weeks. The use of the new features has been shown increasing the accuracy of age classification by 6.3% and 10.0%, respectively.

We also explored conventional approaches to estimating discontinuity. The first was to extract features describing the durations of bursts, inter-burst intervals and continuous intervals, which we detected automatically. The features were found relevant for maturity assessments, but unfortunately, they were less informative than the standard spectral powers. The second approach was to quantify the amplitudes of aEEG borders. Used in combination with the standard spectral features, the aEEG amplitudes increased the accuracy of Bayesian assessment by 2%. The lower border amplitude was found most informative, similarly as reported by Viniker et al. (1984).

Finally, we looked into extraction of new spectral features. We hypothesised that the ratios of spectral powers are more informative than the individual powers, and thus their use will increase the assessment accuracy. The use of the

ratio of absolute powers in the Alpha and Theta bands increased the accuracy by 10% for ages 37 and 39.

Overall, the highest accuracy was achieved by combining the standard spectral features with both the non-stationarity estimate and the Alpha/Theta ratio. The use of the combined set of features enabled achieving the accuracy of 80.7% for differentiation of the age groups 37 and 39. In comparison to using only the standard spectral powers, the gain in performance was 16%.

In the next chapter, we use the extracted features for assessment of a larger range of PCA groups. The chapter will address the problems of handling multiple classes.

Chapter 7

Classification systems

In this chapter we explore how the use of binary classification systems, can improve the accuracy of EEG maturity assessments with multiple age groups. We considered the EEG age classification task as a multiclass problem and trained the DT models to distinguish between the weeks of PCA. When classifying a broad range of PCA, the number of classes is increased, and it becomes more difficult to train a classification model to distinguish between them. It becomes even more problematic when the data from the neighbouring ages overlap because of variations in maturational patterns and uncertainty in PCA estimates. Another limitation of using the standard multiclass approach for maturity assessment is that it cannot take into account the prior information that the class labels, or newborns' ages in weeks, are naturally ordered.

We expect that converting the multiclass problem into a set of binary ones can can provide better performance of maturity assessments, as we discuss in more detail in Section 7.1. In Section 7.2 we present the accuracy of multiclass classification with 10 PCA groups. The accuracy is compared with that of the binary classification systems.

In Section 7.3 we describe experiments with the one-against-all classification, a conventional binarisation technique. We hypothesise that this technique will not achieve a better performance on our 10-class EEG problem, because its training will be affected by imbalance of data.

In Section 7.4 we explore the pairwise classification. We hypothesise that the pairwise system, shown promising for problems with multiple classes, will outperform the one-against-all and multiclass techniques. However, the combination of multiple binary classifiers makes the pairwise assessments more difficult to interpret. Therefore, in Section 7.5 we propose and test a meta-tree classifier which combines the decisions taking into account the prior information about the structure of maturity assessment problem. We hypothesise that the meta-

tree classifier will achieve a performance comparable to that of the pairwise one. At the same time the assessments will be made by fewer binary classifiers, and so the results will be easier to interpret. We summarise the performances of the techniques and conclude the chapter in Section 7.6.

7.1 EEG age classification

An important property of the age classification problem is the natural sequential ordering of class labels which represent the different PCA groups. The conventional multiclass approaches treating the ordered labels as nominal ignore ordering information, which could be used to improve the results. Therefore, it has been shown that such approaches are incapable of providing the best performance for problems with ordered classes (Frank and Hall, 2001). In cases when the class labels are continuous values, regression is typically used. However, in our case the weeks PCA are only coarse estimates, as the brain maturity may change during the week. Therefore the use of regression would be an *ad hoc* approach (Frank and Hall, 2001).

The main difference between the regression and ordinal classification problems is that in the latter the distances between consecutive ranks may not be constant (Baccianella et al., 2010). In other words, the values of all the features do not vary consistently with the labels. In the previous chapter, we saw that not all the EEG features have a linear relationship with PCA. We might assume that the features that are less consistent with PCA over a large range of ages are less informative. However such features may still be useful for separating some localised age groups.

Under such conditions, when the features do not show a linear relationship with the labels, it becomes problematic to define the cost function for regression (Mccullagh, 1980). In (Dembczynski et al., 2008) it was found that regression algorithms often performed poorly on ordinal problems.

Another characteristic of the maturity assessment problem is the large overlap between the samples from the neighbouring age groups. This is mostly caused by the normal variation of brain maturity patterns in the range of ± 1 or ± 2 weeks, the uncertainty in estimating the PCA, as well as the noise and artefacts in EEG. It would be desirable to find a classification model, which not only tries to assign most samples to their stated PCA group, but also minimises the error in the range of ± 1 or ± 2 weeks. For example, given an EEG recorded at 36 weeks, it would be "more wrong" to classify it as 40 weeks than as 37, and the overall results in a range of ± 1 or ± 2 weeks could be improved by taking into this information. The overlap between the age groups, uncertainty in PCA

estimates and the noise in EEG data also affect the between-class boundaries making them difficult to learn.

And an obvious difficulty with classifying a range of PCA groups is handling a large number of classes. It has been shown that the performance of multiclass systems tends to decrease with the number of classes (Uglov et al., 2008). This is because the larger the number of classes, the more difficult it is to learn the boundaries separating all the classes.

To handle a large number of classes, a multiclass problem can be transformed into a set of separate two-class problems. This approach is called binarisation. The binarisation often improves the performance on problems with multiple classes. This is because the boundaries between two classes are easier to learn(Hastie and Tibshirani, 1998; Friedman, 1996). The most widely used binarisation technique is the one-against-all classification transforming a problem with C classes into $(C-1)$ two-class problems of separating each of the classes from the rest. A limitation of his approach is that, for a large number of classes, the portions of training samples in the binary problems may be strongly imbalanced, making training of the classifiers problematic.

Alternatively, the pairwise or one-against-one classification has shown providing better performance in presence of multiple classes (Fürnkranz, 2002). The idea is to train $C(C-1)/2$ classifiers separating all pairs of classes, and then combine them to make the final decision. The pairwise system has been shown outperforming the multiclass approaches for large number of classes (Uglov et al., 2008), and shown to be promising when samples of different classes overlap because of noise and variations (Schetinin and Schult, 2005). A drawback of using pairwise classification is that the combining of classifiers trained on different data may increase the uncertainty in outcomes. To mitigate this problem, instead of using all $C(C-1)/2$ pairs to make a decision for each input, $(C-1)$ classifiers can be selected and organised into a directed acyclic graph (Platt et al., 2000; Bishop, 2007).

Problems with ordered labels were attempted to be solved using a modified one-against-all system as well as the pairwise one (Frank and Hall, 2001; Fürnkranz, 2002). The performances of the approaches have been counted for exact match between labels and outcomes. However, the performances in a range of ± 1 and ± 2 classes (in our case weeks PCA) are yet to be explored. In this chapter, we explore how EEG maturity assessments can be improved by using binarisation. We also explore whether the performances will be improved in the range of ± 1 and ± 2. Particularly, we expect that pairwise classification will provide better results because of its ability to handle a large number of classes and its robustness to noise and overlap in data.

Next, we show experiments with multiclass Bayesian classification of EEG in 10 PCA groups. The performance of the multiclass approach in the range of 0, ±1 and ±2 weeks will be compared to those of the binarisation approaches.

7.2 Multicategorical classification

The settings for running the Bayesian classification were made as follows. The number of DTs sampled in the burn-in phase was 100,000, and in a post burn-in phase 10,000. During the post burn-in phase each 10th model was collected in order to reduce the correlation between DT models. The pruning factor was set to five. The proposal variance was 1.0, and probabilities of making moves of birth, death, change-variable, and change-rule were set to 0.15, 0.15, 0.1, and 0.6, respectively. Under the above settings, the rate of acceptance of DT models during the integration was around 0.23 in both phases. The average DT included 66 nodes.

The EEG data included 952 recordings from newborns aged between 36 and 45 weeks of PCA. Each group included approximately 100 patients. All the recordings were additionally processed with an artefact rejection technique removing samples with abnormal amplitude deviation, as described in Chapter 4. The EEG recordings were also automatically tested on the presence of level of artefacts detected as described in Chapter 4.

The EEG were represented by the 36 standard spectral features as well as the new features describing the non-stationarity of an EEG recording estimated with the technique described in Chapter 6. The new features included the segment rate and 10-bin histograms of the pseudo-stationary intervals ranging from 2 to 20 sec. Additionally, the ratio of absolute powers in Alpha and Theta bands was included.

The results of age classification within the 10-fold cross-validation are presented in Table 7.1. The table shows that the performance in terms of the exact match of weeks is 30.1±12.5% and in the range ±2 weeks it is 85.5±0.8%. The ranges are similar to those that can be obtained for expert assessment (Parmelee et al., 1968) as discussed in more details in Chapter 8.

The confusion matrix for the multiclass system is given in Table 7.2.

We expect that the performances can be further improved by using two-class classification systems. Next, we briefly describe the one-against-all classification technique and show results of experiments on the age classification problem.

Table 7.1: Performance in the intervals of 0, ±1 and ±2 weeks and entropy of multiclass classification

Exact match	±1 week	±2 weeks	Entropy, bits
30.1±12.5	65.5±11.6	85.1±8.2	0.209±0.011

Table 7.2: Confusion matrix for the multiclass assessment

		Predicted									
		1	2	3	4	5	6	7	8	9	10
Actual	1	47	23	12	4	3	1	1	1	0	0
	2	34	29	18	9	5	3	2	0	0	0
	3	20	31	17	12	5	3	2	1	1	0
	4	8	8	12	14	17	13	5	3	2	1
	5	4	5	6	11	25	17	15	6	1	2
	6	1	5	2	12	15	25	15	9	4	3
	7	0	2	2	6	16	13	17	18	14	12
	8	0	1	2	3	8	7	19	16	18	25
	9	0	0	0	0	4	14	11	11	33	30
	10	0	0	0	2	2	1	4	4	23	64

7.3 One-against-all classification

The most widely used approach to transform the multiclass problems is the one-against-all classification (Bishop, 2007). Within this technique, C binary classifiers are trained to distinguish each of the C classes from the rest. A test sample is evaluated by C classifiers, and the class whose probability is highest is assigned to the sample. As the binary between-class boundaries are easier to learn than the multiclass ones, this technique is expected to improve the classification performance.

On the other hand, a problem with this technique is that portions of samples in the binary classifiers become imbalanced as the number of classes increases. The imbalance may lead to poor fitting of the classifiers, because the errors for the smaller class can become ignored. Consequently, we hypothesise that the one-against-all technique will not provide a better performance when the number of classes is large (Fürnkranz, 2002).

To test the hypothesis, we evaluate the performance of one-against-all technique on a maturity assessment problem with 10 PCA groups. To integrate the Bayesian classification with the one-against-all technique, a DT ensemble is collected for each of the C binary classifiers. The posterior probabilities of each of the C classes are obtained from the collected ensembles, and the class with the highest probability is assigned to each sample.

7.3.1 Experiments

The one-against-all technique was run on the same data and with the same settings as the multiclass one. Table 7.3 shows the average performances of age classification counted in the different ranges of weeks within the 10-fold cross-validation. We can see that the performance for exact match is comparable to that of the multiclass approach; however, in the range of ± 1 and ± 2 weeks the performances are slightly lower (by 2 and 1.5%, respectively).

Table 7.3: Performance of one-against-all classification in the intervals of 0, ± 1 and ± 2 weeks.

0 WEEKS	± 1 WEEK	± 2 WEEKS
29.9±9.2	63.5±7.8	83.5±8.4

7.3.2 Conclusion and discussion on section

The techniques converting a multiclass problem into a set of binary problems are typically expected to improve the performance. A commonly used technique is the one-against-all classification. However, in our experiments, the one-against-all technique did not improve the performance of age classification of 10 PCA groups. One explanation to this may be that the learning of binary classifiers was affected by data size imbalance. Fürnkranz (2002) has similarly observed that the one-against-all approach performs worse than the multiclass on a set of problems with 10 classes.

Frank and Hall (2001) proposed to modify the one-against-all technique specifically to improve to performance for problems with ordered classes. They suggested that for such problems the binary classifiers can be trained to distinguish each class from the neighbouring ones. An advantage of this approach is that data imbalance can be avoided. This technique will be explored in further work. In the next section we briefly describe an alternative approach, the pairwise classification, and show results of experiments on EEG data.

7.4 Pairwise classification

The idea behind the pairwise approach is to independently train two-class models separating all possible pairs of classes and combine them to approximate the between-class boundaries. Given that each model is trained on data of two classes, the pairwise technique is not affected by data imbalance when the number of classes is increased. Furthermore, it has been shown (Fürnkranz, 2002)

that pairwise classification outperforms the multiclass and one-against-all approaches for problems with ordered labels. The performances were counted for direct match between the labels and classification.

In this section, we hypothesise that pairwise classification will outperform the multiclass and one-against-all techniques on EEG maturity assessment. In our experiments we will explore the performances for the exact match as well as in the range of ± 1 and ± 2 weeks.

7.4.1 Implementation

Each of the binary classifiers learns to divide the samples from each pair of classes. Therefore, for C classes, we need to train $C(C-1)/2$ binary classifiers. The outcomes of the classifiers that deal with one class are combined into one group, so that the number of the groups corresponds to the number of classes. For each class, $(C-1)$ outcomes are combined.

To combine the outputs, different approaches have been proposed. An obvious approach is to combine the pairwise votes of the $(C-1)$ classifiers voting for either of the two classes and assign a test sample to the class that has most votes (Friedman, 1996). A drawback of this approach is that the voting ignores the class posterior probability estimates provided by the classifiers. To take into account the probability estimates, the outcomes of the binary classifiers can be combined into the final class posterior probabilities (Hastie and Tibshirani, 1998). This method, aiming to approximate the desired probabilities for each input, requires additional computations.

Alternatively, we can treat the outputs as class membership values and sum the outputs for each class to make the final decision. For each input, the class with the highest value will be assigned (Uglov et al., 2008; Schetinin and Schult, 2005). Within this technique, we can estimate the desired probabilities by normalising the summed outputs for each class by the sum of the outputs for all K classes.

Thus the class posterior probability for class c_k can be estimated as follows (Friedman, 1996).

$$Pr(y = c_k|x) = \frac{Pr(x|y = c_k)Pr(y = c_k)}{\sum_{l=1}^{K} Pr(x|y = c_l)Pr(y = c_l)} \tag{7.1}$$

In (Fürnkranz, 2002) it was proposed to integrate the pairwise system with ensemble classification. Within this technique, for each of the two-class problems, an ensemble of classifiers is trained, and then the decisions counted over each ensemble are combined. We can use a similar approach to integrate the pairwise system with Bayesian classification and and collect an ensemble of

Bayaesian DTs solving a two-class problem for each of the $C(C-1)/2$ classifiers. To make the final decisions, the posterior probabilities integrated over the ensembles of DTs will be combined, summing $(C-1)$ posterior probabilities for each class.

7.4.2 Experiments

The pairwise classification was run on the same data and with the same settings as the multiclass. The probabilities obtained from the two-class ensembles were combined using the voting technique and the estimation of class posterior probabilities as described in the previous subsection.

Although Bayesian averaging over pairwise DTs required learning 45 ensembles to classify 10 age groups, the computational time was comparable to that needed to run BMA over multiclass DTs. This is because the binary DTs, containing 6 nodes on average, were much shorter and easier to learn than multiclass DTs with more than 60 nodes.

Table 7.4 shows the average performances for both techniques in the ranges of 0, ± 1 and ± 2 weeks. We can see that combining of probabilities provides better performance for exact match while in the larger range the performances of both techniques are comparable. In comparison with the multiclass approach, pairwise approach performs better in the range of 0, ± 1 and ± 2 weeks by 3.4%, 4.5% and 2.4%, respectively. The confusion matrix for pairwise system with the combined probabilities is given in Table 7.5.

The estimates of class posterior probabilities enable us to count the entropy of the pairwise classification. The average entropy was 0.312 ± 0.005, around 33% larger than that of the multiclass classification. This means that the uncertainty in assessments is significantly larger.

Table 7.4: Performances of pairwise classification in the intervals of 0, ± 1 and ± 2 weeks, with the techniques of voting and combining probabilities

Voting			Probabilities		
Exact match	± 1 week	± 2 weeks	Exact match	± 1 week	± 2 weeks
30.9 ± 0.9	69.7 ± 0.8	87.8 ± 0.6	33.5 ± 0.8	70.0 ± 0.9	87.5 ± 0.9

7.4.3 Conclusion and discussion on section

We hypothesised that pairwise classification, previously shown promising to handle problems with a large number of classes and with ordered labels, will outperform the multiclass and one-against-all techniques on EEG maturity assessment.

Table 7.5: Confusion matrix for the pairwise system with combined probabilities

		Predicted									
		1	2	3	4	5	6	7	8	9	10
Actual	1	49	31	4	2	4	1	1	0	0	0
	2	27	48	17	2	6	0	0	0	0	0
	3	19	31	25	8	6	1	1	0	1	0
	4	5	15	14	18	17	6	6	2	0	0
	5	3	11	4	13	36	12	11	1	1	0
	6	0	5	7	12	23	21	15	4	3	1
	7	0	2	4	3	13	22	20	12	15	9
	8	0	0	5	2	8	17	12	16	22	17
	9	0	0	0	2	3	8	12	17	30	31
	10	0	0	0	1	3	2	5	10	23	56

In our experiments, the pairwise approach outperformed the multiclass one in the range of 0, ± 1 and ± 2 weeks by 3.4%, 4.5% ans 2.4%, respectively. The one-against-all technique was outperformed by 3.6%, 6.5% and 4%. However the entropy increased by 33% compared to that of the multiclass technique.

As discussed by Hastie and Tibshirani (1998), pairwise classification may be affected by the following problem. Suppose we have trained 45 binary classifiers for a 10-class problem. However, for each class $C : C = (1, 10)$, only 9 of the 45 classifiers were trained on samples from this class C. Most of the classifiers were trained on samples of other classes, and thus their outputs for the samples of class C may be ambiguous, especially if they were trained on samples similar to those of class C but not belonging to this class. These outputs, when they are combined, may affect the classification performance. Especially, this may be a problem if some classes have very different meanings, although their data are similar. One example is classification of handwritten digits, if, say, digits 1 and 7 or 0 and 9 look similar.

In case of maturity assessment, the classes are ordered, and the samples that are most similar are expected to be close in ages. For example, consider that samples of age groups 36 and 37 weeks are hard to distinguish. If an EEG recorded at 36 weeks is assigned to 37 weeks, this result reflects the EEG maturity, which can be slightly delayed or accelerated relatively to the labelled age.

However, combining the classifiers trained on classes other than C may increase the classification entropy for samples of C. Consider a sample of 36 weeks being presented to classifiers trained on samples of 38/41 weeks. Typically, 36 weeks will be more similar to 38 than to 41, so the posterior probability of 38 weeks will become increased. Other classifiers may in a similar way contribute to probabilities of classes other than C, affecting the entropy. Another reason

for the increased entropy is the uncertainty in the outcomes of classifiers trained to distinguish between the neighbouring weeks.

Another problem that needs to be explored is the interpretation of the pairwise DTs. In general, to interpret the DT ensembles, one DT providing maximum posterior probability can be selected (Schetinin et al., 2007). Obviously, shorter DTs are favourable for interpretation. In our experiments, the binary DTs consist of 6 nodes on average, whereas an average multiclass DT consists of 60 nodes. However, interpretation of the pairwise DT ensembles is complicated because each sample is processed by 45 ensembles of binary DTs.

One approach to improving the interpretation is to organise the binary classifiers into a directed acyclic graph, similarly as proposed for DAGSVM (Platt et al., 2000). Within this approach, only $C-1$ classifiers need to be evaluated for each sample, depending on the path in the graph. Still, some of the classifiers will be evaluating samples of unseen classes, and so their outputs may increase the uncertainty in assessments.

Alternatively, to obtain more interpretable results and avoid ambiguous classification, binary classifiers can be organised in a hierarchical structure to dichotomise data iteratively. In the next section, we explore this approach.

7.5 Meta-tree

Instead of training the classifiers on each pair of classes and combining all outcomes, we can make advantage of the prior information that EEGs recorded at consecutive post-conceptional ages can be naturally merged into age groups. Specifically, we can first learn to classify EEGs into groups which merge several weeks PCA and then split these groups further. We refer to this approach as a meta-tree. Within this approach, each data sample is evaluated only by those classifiers which are selected based on the outcomes at previous levels.

We hypothesise that the meta-tree will provide a performance comparable to that of the pairwise classification. We also expect that meta-tree assessments with hierarchical dependences between classifiers will be easier to interpret.

One problem is, however, that errors made at the first splits will be carried to the subsequent levels. We hypothesise that introducing double checking of difficult samples will provide an opportunity to correct these errors. First, we test a meta-tree structured for six PCA weeks. Having explored the performance, we run experiments with the 10 age groups.

7.5.1 Implementation

Similarly to the common DT, the meta-tree splits data iteratively; for each input, the next split is chosen depending on the outcome of the previous split. However, within our implementation the structure of the meta-tree is predefined and each split contains a Bayesian DT ensemble. The data are split starting with two age groups. Each of group contains a half of the age groups. At the last iteration, a pair of two consecutive post-conceptional weeks are distinguished.

The ensembles are trained on age groups assigned to the corresponding split, and each data sample is evaluated only by those DT ensembles which are selected based on the outcomes. Thus the problems of ambiguous classification will be likely avoided.

We expect that the meta-tree can provide better interpretation of the maturity assessments in comparison to the pairwise approach. While for the pairwise approach each test sample is evaluated by $C(C-1)/2$ classifiers, within the hierarchical dichotomisation the number of classifiers will be at most equal to the number of levels of the binary tree. Moreover, each of the classifiers has an interpretable impact on the decision, as it can be seen from its position in the meta-tree.

One problem with hierarchical dichotomisation of data, however, is that classification errors made at the first splits are carried over to the subsequent ones. With each new level the errors are accumulated. Aiming to mitigate this problem, we introduce splits enabling the misclassified samples to be rechecked. We hypothesise that such rechecking will correct the classification errors made at the first splits of the meta-tree. In the following subsection, we test the concept of meta-tee classification on EEG maturity assessment problem for six PCA groups.

7.5.2 Experiments with six age groups

The structure of a meta-tree for 6 classes is shown in Fig. 7.1. This structure has 3 levels of hierarchy and includes 7 classifiers, each of which is an ensemble of Bayesian DTs trained to distinguish between the PCA ages grouped into class A and class B. On the first level, an ensemble of DTs divides all data samples into two groups, that is separates the samples of classes 1, 2 and 3 from samples of classes 4, 5 and 6. On the next level, the classes with lower probabilities are excluded, and the data are split further into the remaining classes.

Note the classifier 2 separating classes 1 and 2 from classes 3 and 4 as well as the classifier 3 separating classes 3 and 4 from classes 5 and 6. Samples of the classes 3 and 4 could easily be confused by classifier 1 as they lay close to the

between-class boundary. By including these classes on the Level 2, we enable the classification to be rechecked.

In the experiments we used a dataset of 630 EEG in six PCA groups from 36-41 weeks. The Bayesian averaging for all the classifiers was run with the same settings as used for the pairwise technique in the previous section. The average DT on Level 1 included 10 nodes, on Level 2 – 6 nodes, and on Level 3 – 4 nodes.

For comparison, we also run the pairwise classification for the same six age groups. The average performances and entropies of the pairwise classification and the meta-tree were counted within the 3-fold cross validation. The performance of the pairwise classification was 32.5±6.0, whereas that of the meta-tree was 31.8±6.5, and thus the average performances were comparable for both methods. The average performances are given for exact age match only; the counting of outcomes over a range of weeks does not make sense for a six-class problem, in which the number of outcomes is limited.

The class posterior probabilities within the meta-tree can be estimated as the portion training samples of each class falling into the terminal splits. The entropy of the tree classification can be counted from the probabilities, and this will be explored in future work.

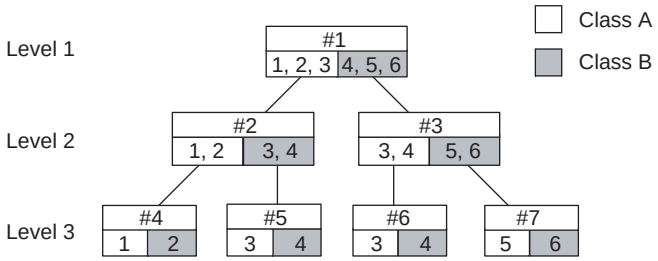

Figure 7.1: Structure of meta-tree for 6 classes with rechecking for samples of classes 3 and 4.

We have hypothesised that the rechecking of classification outcomes for classes 3 and 4 enables correcting the the errors made at the first split and so improves the classification performance. To test this hypothesis, we explore the numbers of test samples of each of the 6 ages assigned to the classes A and B by the splits of the meta-tree. Fig. 7.2 shows the numbers counted for 210 test samples in one of the three data folds. The performance on this fold was 33.0%. The counts of samples that were correctly assigned to class A or B in each split

are shown with white bars whereas the misclassified samples are shown with grey.

Looking at the split 1, we can see that 9 samples of age 4 are wrongly assigned to class A, and 10 samples of age 3 – to class B. We expect that classification of these samples will be corrected at terminal splits 5 and 6. However, looking at the numbers of samples assigned to age group 4 by the split 5, we see that only samples of the true class 4 are classified correctly. The other samples are distributed among the classes 1, 2 and 3. Likewise, only one sample of age 3 is classified at split 6.

From the above observations, it seems that the rechecking did not correct the classification errors made at the Level 1, contrary to our hypothesis. An explanation to this may be that the maturity-related patterns of the misclassified samples were more similar patterns of other ages, and so the repeated classification could not change the outcomes. It is possible that the rechecking only increases the possibility of classification errors. In this case, we can hypothesise that, alternatively, a simplified meta-tree without the rechecking would provide better results.

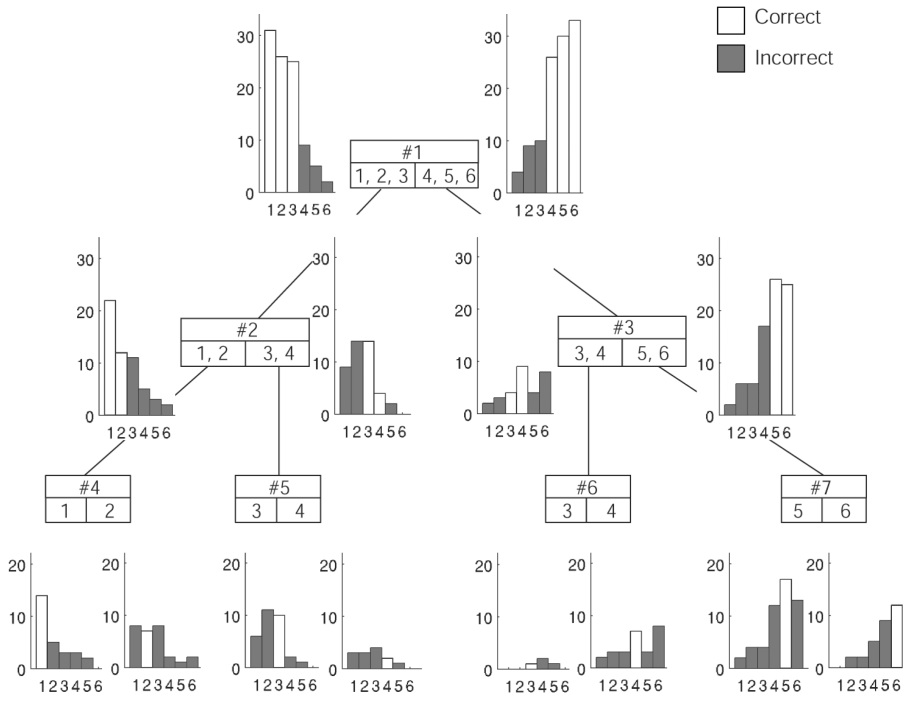

Figure 7.2: Numbers of test samples falling in the splits of meta tree.

To test this hypothesis, we implement a simpler structures of the meta-tree with no rechecking at Level 2. The structure of this meta-tree is shown in Fig. 7.3. Tested on the same data, this meta-tree provided the performance of 34.0±8.2%, 2% higher than that obtained with the rechecking.

However, one possible drawback of using this structure is that the data sizes are imbalanced for classes A and B in the splits 2 and 3. The imbalance may lead to higher error rates for age groups 3 and 4, which are in the smaller classes. Therefore, the performance may be improved further by using a three-class classifier for these splits. This modification will be explore in future work. Next, we show results of experiments with the simple meta-tree for the 10-class maturity assessment problem.

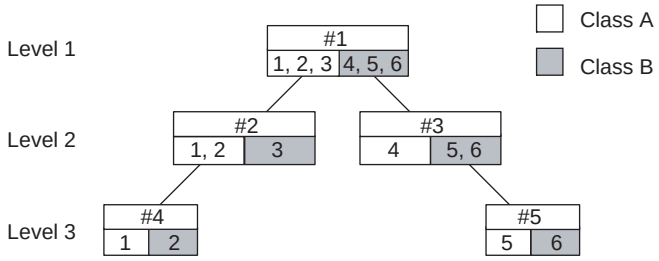

Figure 7.3: Structure of simple meta-tree for 6 classes without rechecking.

7.5.3 Experiments with ten age groups

The experiments with the meta-tree classification of 10 age groups were run on the same data as those with the multiclass, pairwise and one-against-all approaches. The meta-tree for 10 age groups is shown in Fig. 7.4. In cases when the numbers of ages were unequal in classes A and B, the grouping was done arbitrary. We can see that the meta-tree has four levels, and so each test sample is processed by at most four ensembles within this structure. The average DTs in the ensembles of the first, second, third and fourth levels included 12, 8, 4 and 3 nodes, respectively.

Table 7.6 shows the average performances within the 10-fold cross-validation over the PCA intervals. For the exact match, the performance is 2% lower than that of the pairwise classification. The imbalance of data in some of the splits may be the cause may cause the meta-tree to perform worse than pairwise classification on the exact match of weeks. In case of imbalance, the samples of the smaller class may tend to be assigned to the larger class, and this causes the errors for the exact match. Nevertheless, in the ranges ±1 and ±2 weeks the performances of the meta-tree and pairwise classification are comparable.

The meta-tree also outperforms the multiclass classification by approximately 3% on average. The confusion matrix for the meta-tree is given in Table 7.7.

Table 7.6: Performance of meta-tree classification in the intervals of 0, ±1 and ±2 weeks.

0 WEEKS	±1 WEEK	±2 WEEKS
31.5±6.5	70.6±8.5	88.3±6.2

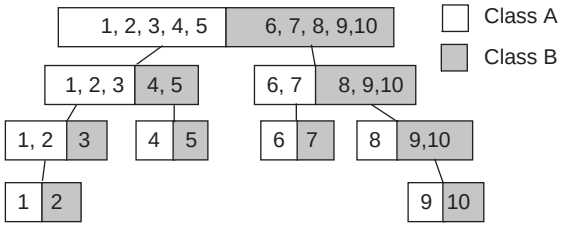

Figure 7.4: Structure of meta tree for 10 age groups.

Table 7.7: Confusion matrix for the meta-tree

		PREDICTED									
		1	2	3	4	5	6	7	8	9	10
ACTUAL	1	47	37	1	2	2	1	2	0	0	0
	2	27	59	59	6	4	3	1	0	0	0
	3	15	47	9	9	11	6	1	2		1
	4	4	23	5	18	14	7	8	3	1	0
	5	3	12	2	11	20	26	12	3	3	0
	6	0	6	3	11	17	28	11	9	6	0
	7	0	3	2	2	11	27	11	12	24	8
	8	0	0	3	1	6	13	9	21	31	15
	9	0	0	0	1		8	13	16	43	22
	10	0	0	0	1	2	3	4	3	43	44

7.5.4 Conclusion and discussion on section

We proposed to organise the binary classifiers in a hierarchical meta-tree structure to dichotomise data iteratively. We hypothesised that the meta-tree will provide the performance comparable to that of the pairwise approach, and at the same time enable better interpretation of maturity assessments. In our experiments, we found that the performance of meta-tree in the range of ±1 and ±2 weeks is comparable to that of the pairwise classification ($p > 0.18$,$p > 0.64$).

Each decision can be interpreted using at most four short DTs whose contributions to the classification can be understood from their position in the meta-tree.

However, the meta-tree slightly underperformed the pairwise classification on the exact match of weeks. One reason for this may be the imbalance in data for the binary classifiers dealing with data from an odd number of age groups. To improve the results, in cases of imbalance, three-class classification could be used instead of the binary to make the numbers of samples per class similar. Alternatively, we can take advantage of the ordered nature of the maturity assessment problem and omit from the training the class that is further from the between-class boundary, similarly as in the ordinal classification technique proposed by Frank and Hall (2001). These points will be researched in future work.

7.6 Chapter discussion and conclusion

The brain maturity classification task is characterised by large number of classes, sequential ordering of class labels, and large overlap between samples of neighbouring age groups. For such applications, the conventional multiclass approaches may perform poorly, because the decision boundary separating all the classes becomes difficult to learn. We hypothesised that the performance of brain maturity assessments can be improved by using a classifier system which is more suitable given these characteristics of data.

Binarisation techniques, such as the one-against-all and pairwise classification, have been shown outperforming the multiclass approaches by splitting a multiclass problem into a set of binary ones which are easier to solve. We hypothesised that pairwise classification, specifically, would provide better performance, as it has been shown effectively handling problems with ordered labels and with large number of classes whose data samples overlap. We also expected that performances in the range of ± 1 and ± 2 weeks can be improved.

In our experiments with Bayesian maturity assessment of 10 PCA groups, the pairwise classification outperformed the multiclass approach in the range of 0, ± 1 and ± 2 weeks by 3.5%, 4.5% and 2.4%, respectively. The p values of these improvements, computed with the Mann-Whitney U test, were $p < 0.14, p < 0.09$ and $p < 0.19$, respectively. The pairwise classification also outperformed the conventional one-against-all binarisation by 4.5%, 6.5% and 4.0%, $p < 0.09, p < 0.01$ and $p < 0.1$. Thus, the improvement for ± 1 week was statistically significant.

However, a drawback of the pairwise approach is that multiple binary classifiers trained on all pairs of classes are combined for the decision. The combination of multiple classifiers makes interpretation of assessments difficult and

increases uncertainty in results. To provide more interpretable results and reduce the uncertainty, we proposed to train binary classifiers to split the data iteratively and organised the classifiers in a meta-tree. The performance of the proposed meta-tree classifier was shown comparable to that of the pairwise classification in the ranges of ± 1 and ± 2 weeks. Within the meta-tree, each test sample was evaluated by 4 binary classifiers whose contribution was clearly defined.

Fig. 7.5 compares the performances of the methods in the ranges of 0, ± 1 and ± 2 weeks. We can see that the pairwise (PW) and the meta-tree (MT) techniques outperform the multiclass (MC) and one-again-all (1/all) classification in all ranges. Overall, the results support our hypothesis that binarisation improves the accuracy of maturity assessments. hypothesised The meta-tree outperforms the multiclass approach by 1.2%, 5.1%, and 3.2% on average, $p < 0.68, p < 0.05$ and $p < 0.11$. Thus, the improvement was statistically significant for the interval of ± 1 week.

The median performance of the meta-tree in the range of 0 weeks is, however, 4% lower than that of the pairwise classification. This lower performance may be caused by imbalance of class samples in the binary classifiers dealing with data from an odd number of age groups. We hypothesise that the performance may be improved if the imbalance is reduced, and will test this hypothesis in future work.

An alternative way of improving the performance of the multiclass DT in the range of ± 1 and ± 2 weeks is by appropriately setting cost of misclassification during training. This would enable penalising those outcomes which are further from the labelled PCA and allow more tolerance to misclassification in a close range of ages. This approach will be explored in future work.

Another question left for further work is the estimation of class posterior probabilities within the meta-tree technique. A possible approach is to estimate the probabilities as the portion of training samples of each class falling into the terminal splits of the meta-tree. The entropy of the meta-tree classification can then be counted from the probabilities.

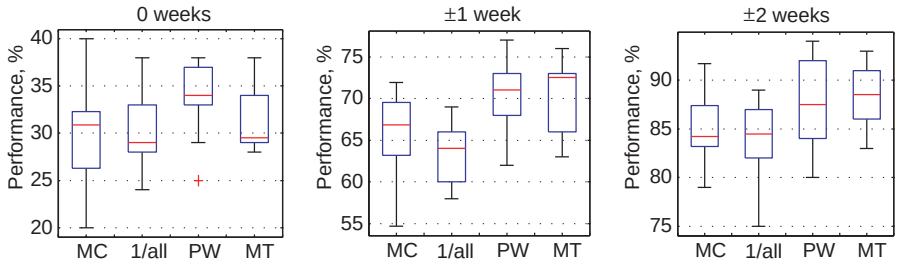

Figure 7.5: Performances in the intervals of 0, ±1 and ±2 weeks for the classification techniques: multiclass (MC), one-against-all (1/all), pairwise (PW) and meta-tree (MT).

Chapter 8

Results of maturity assessments

In this chapter, we explore the accuracy of Bayesian assessments over the typical intervals of ± 1 and ± 2 weeks PCA. We compare the accuracy with that of expert assessments obtained for similar age groups.

In Sections 8.1 and 8.2, we describe the dataset used in our experiments and show the results of Bayesian age classification with 10 PCA groups. Then, in Section 8.3, we show the relative importances of the standard spectral features and the most informative new features described in Chapter 6. Having estimated the feature importances, we employ the refining of Bayesian DT ensemble, as described in Chapter 5, to discard the DTs which use weak EEG features. We expect that the refining will improve the accuracy of maturity assessments and decrease the uncertainty. We will also explore how the refining affects the class posterior distribution

Having found the cases of matched and mismatched maturity assessments, in Section 8.6, we will explore the shape of the class posterior distribution calculated over DT models for a given PCA. We expect that the shapes will be different for cases of matching and mismatching assessments. We hypothesise that when PCA matches EEG estimate, the distribution shape tends to be symmetrical as the areas of interests are mainly located around one age category. On the contrary, for the mismatching cases, the distribution becomes rather asymmetrical as the areas of interests are spread over different age categories. We will test this hypothesis on the EEG.

Finally, in Section 8.7 to provide interpretation of the assessment results, we will use the information from EEG database to test hypotheses about possible reasons for the mismatch in assessments. Specifically, we investigate how the

cases of dysmature assessments are related to patients' apnoea risk and to preterm birth. We conclude the chapter in Section 8.8.

8.1 Data

In our experiments we used 952 EEG recordings from newborns aged from 36 to 45 weeks of PCA. Each of the 10 age groups has been made including around 100 recordings. The electrode movement artefacts causing a significant change or shift in EEG amplitude were removed as described in Chapter 4.

For our experiments, the EEG features were made consisting of two groups, basis and extension ones. The 36 features of the basis group represent the relative and absolute average powers in the six frequency bands computed for the two electrodes and their sum. The basis features were first computed in each 10 sec segment, and then averaged over all segments in a recording. The averaging helps suppress variations and artefacts as shown in Chapter 5.

The features of the extension group represent the discontinuity of an EEG recording. The discontinuity is assessed as the distribution of the pseudo-stationary segments in EEG. The extension group included the total segment rate and 10 bins of the histogram of the segment lengths ranging from 2 to 20 sec. Finally, we added the ratio of absolute spectral powers in Theta and Alpha bands. The extraction of these new features has been described in Chapter 6. The total number of features in the extension group was 12. The two feature groups together included 48 EEG features computed in 10 sec EEG epochs.

8.2 Bayesian classification

The settings for running the Bayesian classification were made as follows. The number of DTs sampled in the burn-in phase was 100,000, and in a post burn-in phase 10,000. During the post burn-in phase each 10th model was collected in order to reduce the correlation between DT models. The pruning factor was set to five. The proposal variance was 1.0, and probabilities of making moves of birth, death, change-variable, and change-rule were set to 0.15, 0.15, 0.1, and 0.6, respectively. Under the above settings, the rate of acceptance of DT models during the integration was around 0.23 in both phases. In the burn-in phase, the log-likelihood as well as the size of DT became stabilised on average after 10,000 samples, as can be seen in in Fig. 8.1, so that the remaining 90,000 samples were drawn from an approximately stationary Markov Chain.

The performance and uncertainty of the DT ensemble collected in the post burn-in phase were evaluated within a 10-fold cross-validation. The average

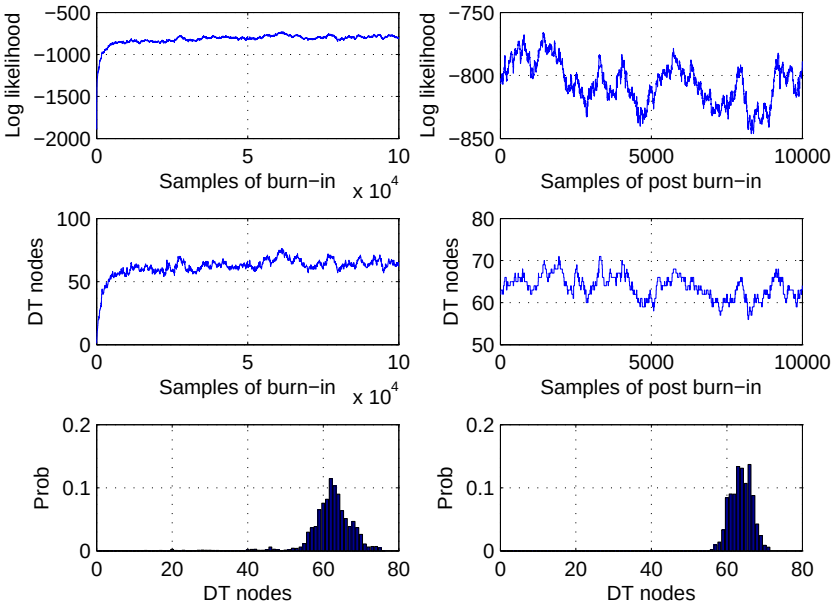

Figure 8.1: Log-likelihood, number of DT nodes and distribution of DT sizes during the burn-in and post burn-in phases.

performance was 30.6% and the the 2σ interval was 12.8% and the uncertainty in terms of entropy of the ensemble was 0.209±0.011.

8.3 Importance of EEG features

Although the 48 features from the basis and extension groups have been found informative for maturity assessment, there is no prior information about the most important features or a feature combination in the context of DT models. As discussed in Chapter 5, in the absence of the prior information about the importance of EEG features, the results of Bayesian classification will likely suffer from the lack of detailed exploration of a multidimensional space of model parameters. To improve the results, we can obtain posterior information on EEG feature importance, and use this information to refine the Bayesian DT ensemble. Additionally, the information on EEG feature importance will assist EEG experts in interpretation of the assessments.

As discussed in Chapter 5, using DT models for classification within the Bayesian methodology allows us to count the importance of the EEG features in

terms of the posterior probabilities of their use in DT ensemble. Fig. 8.2 shows the posterior probabilities of using all 48 EEG features in the basis (upper plot) and extension (lower plot) groups. The probabilities are averaged over the 10 folds.

First, we see the importance of the features ranges between 0.0025 (AbsSubdeltaC3T3) and 0.078 (Theta/Alpha Ratio). Second, we observe that not all the features of the basis group are equally important, only 12 out of the 36 features are of the importance greater than 0.02. In contrast, the importance of all the features of the extension group is higher than that.

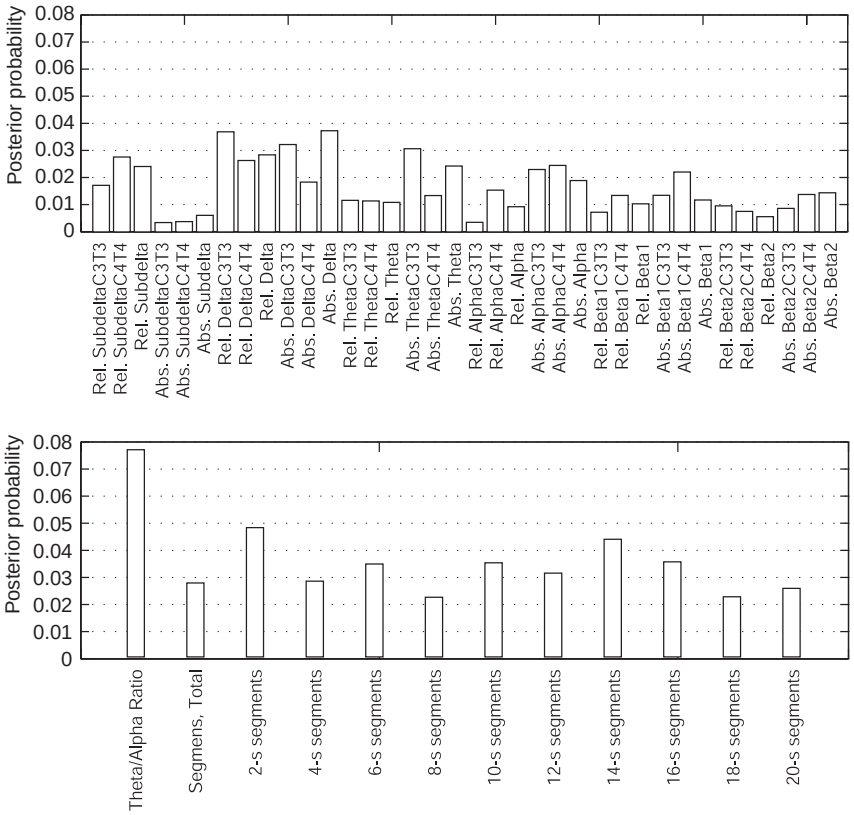

Figure 8.2: Importance (posterior probabilities) of 48 EEG features characterising the relative and absolute spectral powers (upper plot) and the Theta/Alpha ratio and EEG non-stationarity (lower plot).

8.4 Refining the ensemble

The above Fig. 8.2 shows that the probabilities (or importance) of the given features vary in a wide range. Some of the features with low importance are probably weak to make a distinguishable contribution to the classification. As described in Chapter 5, we can hypothesise that discarding DTs using such weak features will improve the performance within the proposed method.

According to this method, we found a set of 8 weakest features with probabilities below 0.006, and discarded the DTs using these features from the ensemble. After discarding, it was found that the performance median slightly increased from 31% to 34% when the threshold was changed to 0.003 and 8 weak features were defined. The uncertainty counted in terms of entropy of an ensemble is slightly decreased from 0.209±0.011 to 0.208±0.012. Further increasing the threshold to 0.006 lead to discarding 13 weak features without a significant drop in the performance. Clearly, the removal 13 out of 48 features makes the DT ensemble shorter and easier for interpretation.

The box plots in Fig. 8.3 show the average numbers of DT splits for the original and refined ensembles as well as for the discarded DTs. The original ensemble included 10,000 DTs, but after refining it included 5,800 DTs. We see that the median number of splits in the discarded set of DTs is 66.3, which is higher than that in the original ensemble. The median number of splits in the refined ensemble decreases to 65.5, as the portion of larger DTs has been removed. In the next subsections we will explore the accuracy of the refined ensemble of DTs.

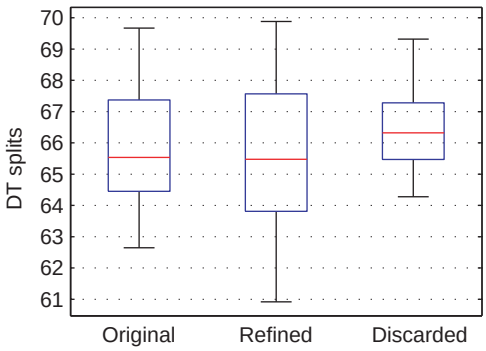

Figure 8.3: Average number of DT splits in the ensembles.

8.5 Performance of EEG assessment

Having obtained an ensemble of DT models, we can count the performance of the Bayesian assessment of brain maturity. The performance was counted within the three intervals: exact match, ±1 weeks, and ±2 weeks. The performances within each of these intervals were 30.1%, 65.5%, and 85.1%, respectively. As discussed in Chapter 1, the neurological assessment of newborn brain maturity is mainly made within ±2 weeks of PCA.

The performance of Bayesian assessments in the interval of ±2 weeks is comparable to that of the expert assessment reported in (Parmelee et al. 1968). These results, however, have been obtained on EEG in different age ranges and with different sample sizes: we used 952 recordings from newborns aged 36-45 weeks, whilst the experts have assessed only 47 recordings at ages 30 to 43 weeks.

Table 8.1 shows the spread of age classifications over the given age groups from 36 to 45 weeks of PCA. The table columns present the numbers of classifications fallen into the age groups ranged from -6 to +7 weeks. Thus Column 0 shows the numbers of classifications fallen into the actual age groups (exact matches), Column 1 shows the number of classifications fallen into an age group which is less than the actual age group by one week, Column 2 – less than by two weeks, etc.

Table 8.1: Spread of age classifications

	Mismatch (weeks)														
PCA	-6	-5	-4	-3	-2	-1	**0**	1	2	3	4	5	6	7	Total
36							**46**	24	11	5	3	2	0	1	92
37						32	**33**	16	10	5	3	1	0	0	100
38					18	29	**18**	15	5	3	2	1	1	0	92
39				8	9	10	**15**	17	13	5	3	2	1		83
40			4	5	6	11	**26**	16	14	6	2	2			92
41		1	5	2	11	15	**26**	15	10	3	3				91
42	0	2	2	6	16	12	**18**	17	15	12					100
43	1	1	3	10	9	19	**14**	19	23						99
44	0	0	3	14	12	10	**35**	29							103
45	2	2	1	3	5	25	**62**								100
Total	3	6	18	48	86	163	**293**	168	101	39	16	8	2	1	952

The Total column presents the number of EEG recordings in each age group. This column shows that the numbers of recordings in each group are similar. The Total row shows that the numbers of age classifications fallen in the age groups ranged between -6 and +7.

Assessing brain maturity within some age ranges may be more difficult than within others, because the EEG features may have different informativeness within different age ranges. Therefore, it is interesting to compare the accuracies of Bayesian and expert assessments within similar age groups. Table 8.2 shows the performance of the expert assessment of EEG maturation described in (Parmelee et al., 1968) together with the performance of the Bayesian assessment calculated within the same five age groups from 39 to 43 weeks of PCA within the ranges ± 1 and ± 2 weeks. We used the 465 recordings (in the above five age groups), whilst the experts have assessed only 27 recordings.

Table 8.2: Performances of expert and Bayesian assessments

Interval, weeks	Expert, %	Bayesian classification, %
± 1	59.5	53.7
± 2	77.3	80.8

Such a difference in the sample sizes does not allow us to compare the results directly. Nevertheless, we observe that the Bayesian assessment within ± 2 week interval, on average, slightly outperforms the expert assessment. It is also interesting that Bayesian assessments were made on EEG date from two electrodes, whereas the experts used 8 electrodes and additional polysomnographic channels.

It is important to note that an EEG assessment obtained within the Bayesian methodology is provided with an accurate estimate of the uncertainty. Below we describe our experiments and results in estimating the uncertainty for EEG assessment.

8.6 Estimation of uncertainty

In this section, we describe how the estimates of uncertainty obtained within the Bayesian assessment can assist experts to reduce possible errors. Having obtained an ensemble of DT models, first we calculate the desired estimates by using the original ensemble and then explore whether the estimates are improved by using the refined ensemble.

Second, we explore the class posterior probabilities obtained within the Bayesian assessment for patients assigned in different age groups. The assignments can be made matching or mismatching the stated PCA. We consider a mismatch of more than 2 weeks as the case of abnormal brain maturity, and therefore it is important to identify risk of the mismatch by analysing the posterior probability.

In our experiments, we used the ensemble of DTs obtained on the 857 cases to test the other 95 cases, roughly equally distributed over the 10 PCA groups. According to the spread of age classification given in Section 8.5, a few EEG assessments were found mismatching the PCA within the ± 2 week interval. We expect that the class posterior probability distribution obtained for a mismatched case differs from that obtained for a matched case. To test this hypothesis, we first look at two cases of 6^{th} class (41 weeks of PCA), one matching and the other mismatching the newborn's stated PCA.

Fig. 8.4 and 8.5 show the class posterior probabilities for these cases. Here, the left side plots show the class posterior probability distribution over the 10 classes within the 1σ intervals computed over all the DTs included in the original ensemble. For the matching case, the average probabilities of the classes 6 and 5 (weeks 41 and 40) are highest. This indicates normal brain maturity as these classes are within the interval of ± 1 weeks. We can also see that the 1σ intervals of the classes 6 and 5 do not overlap those of the other classes, and therefore the uncertainty in assessment is low. Contrary, for the mismatching case, the 1σ intervals of the probabilities of most classes are overlapping, and the uncertainty is much higher. The probability of class 1 is maximal; however, the probability of class 7 is second highest, and it is comparable within the 1σ interval. The probabilities indicate that the recording contains a mixture of dismature patterns and those appropriate for the age.

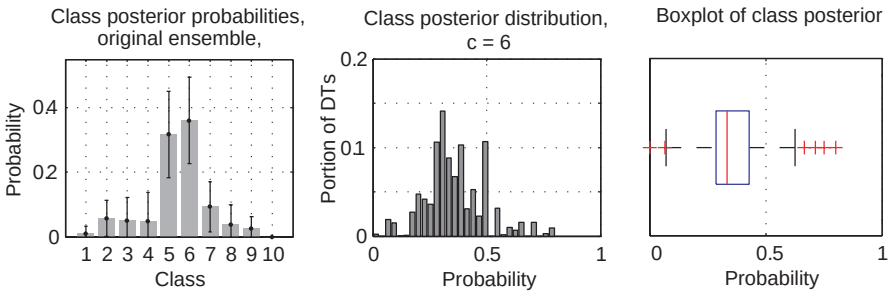

Figure 8.4: Probability distributions estimated for matching case.

These two cases illustrate the use of the class posterior probability distributions obtained within the Bayesian methodology for estimating the uncertainty. We observed that the distribution counted for the newborn's stated PCA (shown in the middle plots) becomes asymmetrical when an EEG assessment mismatches the PCA. The uncertainty can be quantitatively represented, and the shape asymmetry can be visually recognized.

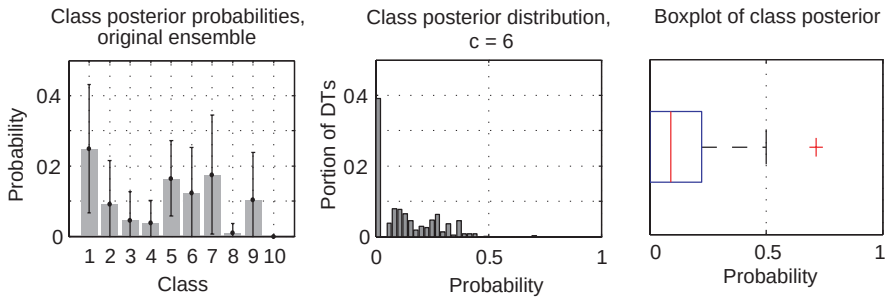

Figure 8.5: Probability distributions estimated for mismatching case.

The asymmetry of the shape of a distribution can be quantified by its skewness. We observed that the class posterior probability distributions for the stated ages tend to be skewed in cases of mismatched assessments. Fig. 8.6 compares the asymmetry of class posterior distributions in cases of matched and mismatched assessments in a set of 96 patients. We can see from the boxplots that the asymmetry (in terms of skewness) tends to increase with the magnitude of mismatch.

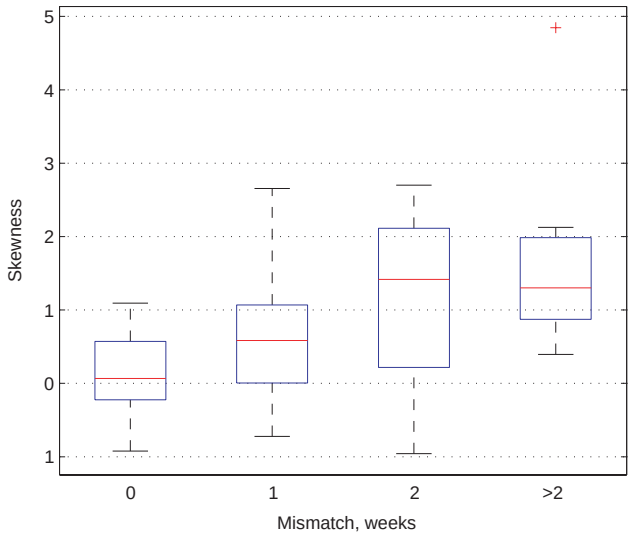

Figure 8.6: Skewness of the class posterior distributions in cases of matched and mismatched assessments

As described in the previous subsection, the refined ensemble of DT models has improved the performance of EEG assessment and reduced the entropy, and therefore we can observe the corresponding changes in the class posterior distributions. Fig. 8.7 shows these probabilities obtained with the refined ensemble for the above two cases.

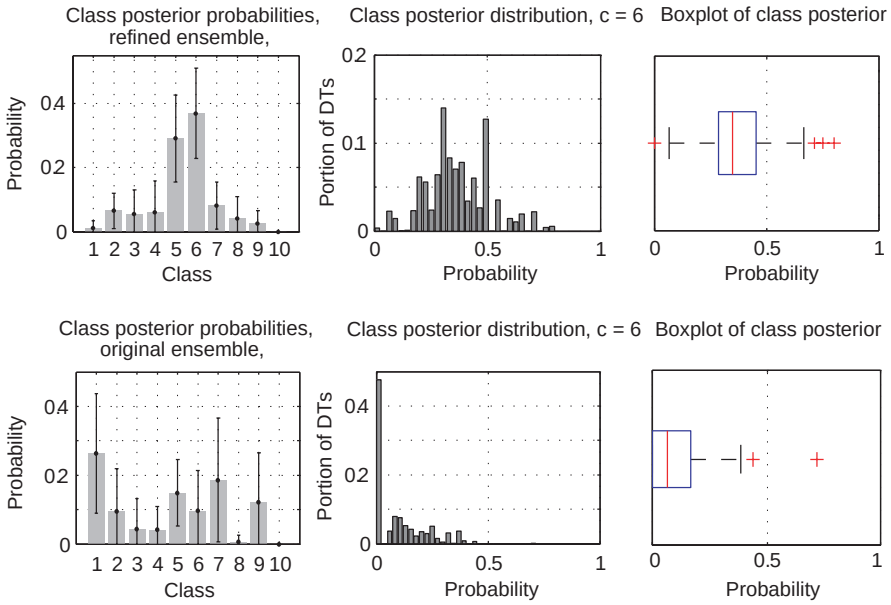

Figure 8.7: Probability distributions estimated with the refined ensemble of DTs for matching (upper plot) and mismatching (lower plot) cases.

The comparison shows that the average probabilities of classes 5 and 6 shown in the upper plot of Fig. 8.7 are slightly higher than those shown in Fig. 8.4 while their intervals are slightly smaller. As a result, using the refined ensemble decreases the uncertainty of EEG assessment. Comparing the posterior distribution obtained with the refined ensemble for the mismatching case shown in the lower plots of Fig. 8.7 and those obtained with the original ensemble (Fig. 8.5), we observe a similar decrease in the uncertainty of EEG assessment.

8.7 Causes of mismatched assessments

In the previous section, we found the cases for which the Bayesian assessments were mismatched by at least two weeks. It is interesting to explore the hypotheses about the causes of this mismatch, or brain dysmaturity. The first hypothesis is that the dysmature assessments are linked with high risk of ap-

noea, as previously shown in (Holthausen et al., 1999) on a smaller subset of the EEG data. The second hypothesis is that very pre-term birth is associated with dysmaturity observed when the newborns reach full-term age (Scher et al., 1990, 2003a; Conde et al., 2005).

To test the hypotheses, first we compare apnoea risk indexes of newborns with the matching and the mismatching assessments. Second, we explore whether the rate of mismatched assessments is higher for the very pre-term newborns than that for babies born at later gestational ages.

8.7.1 Apnoea risk

Each EEG recording in our dataset has been accompanied by an apnoea index, counted as the number of apnoea episodes per hour. The episodes, defined as periods of stopped breathing lasting at least 3 sec, have been detected automatically based on monitoring of respiration and oxygen saturation.

Having found the 811 recordings for which the assessments were matched within the interval of ± 2 weeks and 141 recording for which the assessments exceeded this interval, we compare the distributions of apnoea indexes in both these groups to see whether there exists a relationship between the mismatched assessments and higher apnoea index. Fig. 8.8 shows the distributions for matching and mismatching assessments. Although the sample sizes are different, we can see that both distributions have a similar shape. The indexes range between 0 and 100 and the median values are around 16 in both groups. Contrary to the hypothesis, it seems that no relationship exists between the dysmaturity and apnoea index, as the distributions are similar. To verify this we use hypothesis testing.

As the distributions are not normal, the t-test is not suitable, and we use the non-parametric KS-test to verify the null hypothesis that the apnoea indexes of both groups are from the same distribution. The two-sample KS-test could not reject the null hypothesis providing $p > 0.98$. The hypothesis was also tested on the 293 cases of exactly matching and 141 cases of dysmature assessments, and could not be rejected with $p > 0.71$. Thus, our data do not support the hypothesis that mismatched assessments are associated with high apnoea risk.

8.7.2 Very pre-term birth

For each EEG recording, the patients gestational age has been noted in the database alongside PCA. We use the information on gestational age to test the hypothesis that mismatched assessments are more likely for the very pre-term newborns than for babies born at later ages. We explore the mismatched

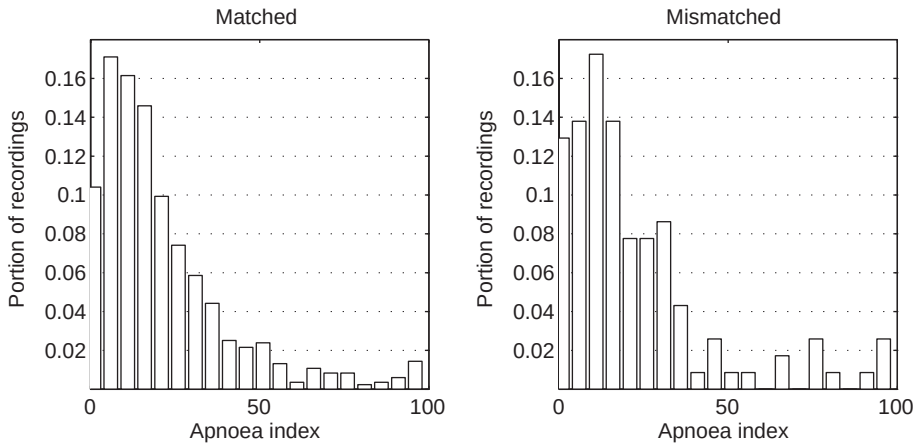

Figure 8.8: Distributions of apnoea indexes for cases with matched and mismatched assessments.

assessments between 41 and 45 weeks, when the full-term EEG patterns are expected to be fully developed.

In this PCA range, we identified two gestational age groups: 84 babies were born very pre-term, that is before 33 weeks gestational age, the remaining 409 babies were born aged 33 or more weeks. In the very pre-term group, 21 recordings was assessed as dysmature, whereas in the group with older gestational ages there were 55 such recordings. Table 8.3 summarises this data in a so-called 2×2 contingency table. Analysis of this data with a standard X^2 test with 1 degree of freedom showed a statistically significant difference $p < 0.02$. Thus the data support the hypothesis that mismatched assessments at full-term age are more likely for very pre-term newborns than for babies born after 32 weeks gestation.

Table 8.3: Numbers of matched and mismatched assessments with the different gestational age

Gestational age	Matched	Mismatched	Total
$<$ 33 weeks	63	21	84
$\geq$ 33 weeks	354	55	409
Total	417	76	493

8.8 Chapter discussion and conclusions

We explored the accuracy of Bayesian assessments of EEG maturity and showed that the accuracy was comparable to that obtained by experts. Although the

EEG recordings used in our experiments were different, the assessment accuracies in the intervals of ± 1 and ± 2 weeks PCA were in a similar range.

Having obtained an ensemble of Bayesian DT, we estimated the importances of the EEG features in terms of the frequencies of their use in the ensemble. It was found that the new features described in Chapter 7 had higher importances than 66% of the standard spectral features.

We also expected to obtain accurate estimation of class posterior distribution within the Bayesian assessment to provide experts with the exhaustive information on risk in EEG assessment of the newborn's brain. In our experiments, we showed that the Bayesian assessment of the posterior probabilities are accurate and can be used for evaluating the risk of possible errors.

The results of Bayesian assessments were improved by using the refining technique described in Chapter 6. We showed that the refining reduced the uncertainty of assessments and this effect was observed in the ensemble entropy counted over all test data as well in class posterior probabilities shown for individual patients.

Finally to provide interpretation of assessment results, we tested two hypotheses abut possible causes of the mismatched assessments. We found that the assessed dysmaturity was not related to high apnoea risk, contrary to results reported by (Holthausen et al., 1999). However a statistically significant relationship was found between the mismatch and very pre-term birth. The lag in maturation of pre-term newborns has been observed previously by EEG experts (Scher et al., 1990, 2003a; Conde et al., 2005). This finding supports the validity of our assessment technology.

Chapter 9

Conclusions

We studied how brain maturity of newborns can be automatically assessed from sleep EEG. It was proposed to assess the brain maturity within the methodology of Bayesian averaging, which in theory provides the most accurate assessments along with the estimates of uncertainty enabling experts to take into account the full information about the risk of decision making. Such information is particularly important when assessing the EEG signals which are highly variable and corrupted by artefacts.

The work presents the first results on automated assessment of the brain maturity made in the typical intervals of newborns stated age. The assessment technology was tested on 952 EEG recordings of newborns aged 36 to 45 weeks after conception, and the accuracy was comparable to that achieved by EEG experts manually analysing EEG maturational patterns. Moreover, we showed that maturation of newborns aged 36 to 45 weeks can be assessed from two-channel EEG, without the conventionally used multiple channels and polysomnogram (Chapter 8).

The use of decision tree models within Bayesian averaging enabled selecting the EEG features most important for the assessment. The feature selection becomes important when data are represented by multiple features the prior information on which is unavailable. In our case, The EEG data were represented by multiple EEG features, and it was expected that some of the features were making a weak contribution to the assessments.

It was hypothesised that the use of weak features within Bayesian averaging over decision trees unnecessarily increases a model parameter space, which needs to be explored in detail to achieve proportional sampling from areas of interests. The larger the number of weak features, the greater is the number of models using these features, and the greater is their negative impact on results of Bayesian classification.

We expected that discarding of models using weak features will reduce the negative influence, and proposed a technique for refining a decision tree ensemble from models which use the weak features. The proposed technique was shown capable of increasing the performance and decreasing the ensemble uncertainty. At the same time, this technique enabled finding a subset of the most important EEG features (Chapter 5).

To further improve the accuracy of assessments, we extracted the new EEG features complementing the standard spectral powers. Based on the clinical observations that discontinuity is the most important maturational characteristic, we explored the conventional and new techniques for extracting features representing EEG discontinuity. Specifically, we proposed to estimate EEG discontinuity as the "non-stationarity". The new feature, counted as the rate of pseudo-stationary EEG intervals, was shown outperforming the conventional discontinuity estimates. Used in combination with the standard spectral features, the new feature improved the accuracy of assessments by 6% on average (Chapter 6).

When classifying a large range of newborns' ages, the accuracy of conventional techniques may be negatively affected by an increase in the number of classes. We hypothesised that converting a multiclass problem into a set of binary ones will improve the accuracy of assessments with multiple classes. We also hypothesised that the assessment accuracy can be improved by taking into account the prior information that the class labels, or newborns' ages in weeks, are naturally ordered.

In the experiments, the pairwise binarisation technique improved the assessment accuracy. However, a weakness of this technique is that classification decisions are made by combining outputs from multiple binary classifiers. This means that decisions become difficult to interpret. To simplify interpretation of maturity assessments, we proposed a meta-tree classifier which provides a performance comparable to that of the pairwise binarisation, while using for each decision only a few binary classifiers, whose contribution is clearly defined within the meta-tree structure (Chapter 7).

A general problem of newborn EEG analysis is that the EEG signals are weak and easily corrupted by muscle and technical artefacts which affect the clinical interpretation. Therefore experts need to detect and mark the artefacts to be excluded from analysis. The results may be subjective or inconsistent between different experts. The inconsistencies in artefact removal may affect the accuracy of Bayesian assessments. This motivated us to hypothesise that the accuracy can be improved by removing the artefacts automatically. To test this hypothesis, we compared the accuracy of assessments after expert and automated removal of artefacts. We showed that the accuracy achieved after the

automated removal is comparable or slightly better than that achieved after the removal of the artefacts marked by experts. We also showed that the negative influence of EEG artefacts on the maturity assessment can be suppressed by averaging over multiple short epochs of EEG (Chapter 4).

Finally, having evaluated the assessment accuracy on EEG recorded at 36 to 45 weeks, we explored the patient cases for which the Bayesian assessments were mismatched with the stated ages. We found that mismatched assessments at full-term age were more likely for newborns born very pre-term. This finding agrees with observations of EEG experts that maturational patterns of very pre-term newborns may be altered (Chapter 8).

9.1 Future work

- *Recognition of EEG patterns.* In this thesis, we assessed brain maturity from EEG features extracted from the whole recordings. This approach does not take into account the patterns and waves which experts typically analyse, such as patterns of the quiet and active sleep states, delta brushes and Theta/Alpha bursts. We expect that extraction of new features describing these patterns and waves could provide additional information to improve the accuracy of assessments.

  The first step to obtain such information is to detect the quiet and active sleep states and count the EEG features in each of these states. The preliminary results (Schetinin et al., 2011) have shown that features extracted from the quiet sleep state are more informative than those from the active sleep, and the use of these features can improve the assessment accuracy.

- *Improving performance of the meta-tree.* To improve the accuracy of Bayesian maturity assessment on 10 PCA groups, or classes, we proposed a meta-tree classifier to split the multiclass problem into a set of binary ones. The meta-tree outperformed the conventional multiclass approach by approximately 3% on average. Yet, the performance of the meta-tree was found affected by data imbalance problem, and we expect that an improvement can be obtained by reducing the imbalance. One way of dealing with the imbalance is to take into account the natural ordering of classes and concentrate on training the binary classifiers to discriminate only the neighbouring age groups.

  Another open question is the estimation of class posterior probabilities within the meta-tree. A possible approach is to estimate the probabilities as the portion of training samples of each class falling into the terminal

splits of the meta-tree. The entropy of the meta-tree classification can be counted from the probabilities.

- *Introducing misclassification costs to improve results of ordered classification.* Taking into account the ordered nature of the maturity assessment problem, an alternative way of improving the performance is by appropriately setting cost of misclassification during training. This would enable penalising those outcomes which are far from the labelled PCA and allow more tolerance to misclassification in a close range of ages.

- *Extension to other PCA groups.* We have tested our technology of maturity assessments on EEG of newborns aged 36 to 45 weeks PCA. Assessment of maturation of newborns younger than 36 weeks is an important clinical problem. To make Bayesian assessments for this age group sufficient EEG data are required. It will be interesting to explore the importances of the extracted EEG features on this age group, for which the EEG maturational patterns are different, and so the feature importances may change.

- *Validation of assessments by experts.* The EEG data available for our research have been recorded from newborns in hospitals. Some of the newborns could have clinical conditions that caused their EEG maturity to be altered and to mismatch their stated PCA. Obviously, such alterations introduce noise in data, affecting accuracy of assessments. To validate our results, it would be valuable to obtain expert assessments of brain maturity for a subset of recordings from our dataset including cases for which the Bayesian assessments were matched and mismatched to PCA.

- *Improving informativeness of aEEG features.* aEEG is becoming an established technique in monitoring brain function of newborns. We found that time domain features, enabling the performance of maturity assessment to be improved, can be automatically extracted from the aEEG signal. Future work will explore how the informativeness of these features can be improved by removing EEG artefacts prior to applying the aEEG technique.

Bibliography

Agarwal, R., Gotman, J., Flanagan, D., and Rosenblat, B. (1998). Automatic EEG analysis during long-term monitoring in the ICU. *Electroencephalography and Clinical Neurophysiology*, 107(1):44–58.

Appel, U. and Brandt, A. V. (1983). Adaptive sequential segmentation of piecewise stationary time series. *Information Sciences*, 29(1):2756.

Aufrichtig, R., Pedersen, S. B., and Jennum, P. (1991). Adaptive segmentation of EEG signals. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 13, pages 453–454.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Feature selection for ordinal regression. In *The 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1748–1754, New York, NY, USA. ACM.

Barlow, J. S. (1985). Methods of analysis of nonstationary EEGs, with emphasis on segmentation techniques: a comparative review. *Journal of Clinical Neurophysiology*, 2(3):267–304.

Barlow, J. S., Creutzfeldt, O., Michael, D., Houchin, J., and Epelbaum, H. (1981). Automatic adaptive segmentation of clinical EEGs. *Electroencephalography and Clinical Neurophysiology*, 51(5):512–525.

Beckwith, L. and Parmelee, Jr., A. H. (1986). EEG patterns of preterm infants, home environment, and later IQ. *Child Development*, 57(3):777–789.

Bell, A. H., McClure, B. G., McCullagh, P. J., and McClelland, R. J. (1991a). Spectral edge frequency of the EEG in healthy neonates and variation with behavioural state. *Biology of the Neonate*, 60(2):69–74.

Bell, A. H., McClure, B. G., McCullagh, P. J., and McClelland, R. J. (1991b). Variation in power spectral analysis of the EEG with gestational age. *Journal of Clinical Neurophysiology*, 8(3):312–319.

Bihannic, A. L., K.Beauvais, Busnel, A., de Barace, C., and Furby, A. (2012). Prognostic value of EEG in very premature newborns. *Archives of Disease in Childhod - Fetal and Neonatal Edition*, 97(2):F106–F109.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer.

Blinowska, K. J., Czerwosz, L. T., Drabik, W., Franaszczuk, P. J., and Ekiert, H. (1981). EEG data reduction by means of autoregressive representation and discriminant analysis procedures. *Electroencephalography and Clinical Neurophysiology*, 51(6):650658.

Bliss (2012). Bliss: for babies born too soon, too small, too sick. http://www.bliss.org.uk.

Bodenstein, G., Schneider, W., and Malsburg, C. V. (1985). Computerized EEG pattern classification by adaptive segmentation and probability density function classification. Description of the method. *Computers in Biology and Medicine*, 15(5):297–313.

Boylan, G. B. (2008). *Neonatal Cerebral Investigation*, chapter Principles of EEG, pages 9–21. Cambridge University Press.

Boylan, G. B., Murray, D. M., and Rennie, J. M. (2008). *Neonatal Cerebral Investigation*, chapter The normal EEG and aEEG, pages 83–91. Cambridge University Press.

Brazier, M. A. and Casby, J. U. (1952). Crosscorrelation and autocorrelation studies of electroencephalographic potentials. *Electroencephalography and Clinical Neurophysiology*, 4(2):201 – 211.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall.

Buntine, W. (1998). Learning classification trees. *Statistics and Computing*, 2:6373.

Burdjalov, V. F., Baumgart, S., and Spitzer, A. R. (2003). Cerebral function monitoring: A new scoring system for the evaluation of brain maturation in neonates. *Pediatrics*, 112(4):855–861.

Castellanos, N. P. and Makarov, V. A. (2006). Recovering EEG brain signals: artifact suppression with wavelet enhanced independent component analysis. *Journal of Neuroscience Methods*, 158(2):300–312.

Chang, B. S., Schachter, S. C., and Schomer, D. L. (2005). *Atlas of Ambulatory EEG*. Elsevier.

Chipman, H., George, E., and McCullock, R. (1998). Bayesian CART model search. *Journal of American Statistics*, 93:935–960.

Clarencon, D., Renaudin, M., Gourmelon, P., Kerckhoeve, A., Caterini, R., Boivin, E., Ellis, P., Hille, B., and Fatome, M. (1996). Real-time spike detection in EEG signals using the wavelet transform and a dedicated digital signal processor card. *Journal of Neuroscience Methods*, 70(1):5–14.

Conde, J. R., de Hoyos, A., Martínez, E., Campo, C., Pérez, A., and Borges, A. (2005). Extrauterine life duration and ontogenic EEG parameters in preterm newborns with and without major ultrasound brain lesions. *Clinical Neurophysiology*, 116(12):2796–809.

Cooper, R., Binnie, C., and Schaw, J. C. (2003). *Clinical Neurophysiology: EEG, paediatric neurophysiology, special techniques and applications*, chapter EEG analysis. Elsevier Science.

Crowell, D. H., Jones, R. H., Kapuniai, L. E., and Leung, P. (1977). Autoregressive representation of infant EEG for the purpose of hypothesis testing and classification. *Electroencephalography and Clinical Neurophysiology*, 43(3):317–324.

Crowell, D. H., Kapuniai, L. E., and Jones, R. H. (1978). Autoregressive spectral estimates of newborn brain maturational level: Classification and validation. *Psychophysiology*, 15(3):204208.

Curzi-Dascalova, L., Figueroa, J. M., Eiselt, M., Christova, E., Virassamy, A., dAllest, A. M., Guimaraes, H., Gaultier, C., and Dehan, M. (1993). Sleep state organization in premature infants of less than 35 weeks gestational age. *Pediatric research*, 34(5):624–628.

Davis, H., Davis, P. A., Loomis, A. L., Harvey, E. N., and Hobart, G. (1938). Human brain potentials during the onset of sleep. *Journal of Neurophysiology*, 1(1):24–38.

Davis, H., Davis, P. A., Loomis, A. L., Harvey, E. N., and Hobart, G. (1939). Electrical reactions of the human brain to auditory stimulation during sleep. *Journal of Neurophysiology*, 2(6):500–514.

Dembczynski, K., Kotlowski, W., and Slowinski, R. (2008). Ordinal classification with decision rules. In *The 2007 ECML/PKDD international conference on Mining complex data*, MCD'07, pages 169–181, Berlin, Heidelberg. Springer-Verlag.

Denison, D., Holmes, C., Mallick, B., and Smith, A. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley.

Djordjevic, V., Reljin, N., Gerla, V., Lhotska, L., and Krajca, V. (2009). Feature extraction and classification of EEG sleep recordings in newborns. In *9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009*. IEEE.

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. In *The 17th International Conference on Machine Learning*, pages 223–230. Morgan Kaufmann Publishers.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley & Sons, Inc., New York, NY, USA, 2nd edition.

Dumermuth, G., Huber, P., Kleiner, B., and Gasser, T. (1970). Numerical analysis of electroencephalographic data. *Audio and Electroacoustics, IEEE Transactions on*, 18(4):404 – 411.

Estevez, P. A., Held, C., Holzmann, C., Perez, C., Perez, J., Heiss, J., Garrido, M., and Peirano, P. (2002). Polysomnographic pattern recognition for automated classification of sleepwaking states in infants. *Medical & Biological Engineering & Computing*, 40(1):105–113.

Frank, E. and Hall, M. (2001). A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, ECML'01, pages 145–156. Springer-Verlag.

Friedman, J. (1996). Another approach to polychotomous classification. Technical report, Department of Statistics, Stanford University.

Fürnkranz, J. (2002). Pairwise classification as an ensemble technique. In *Proceedings of the 13th European Conference on Machine Learning*, ECML'02, pages 97–110. Springer-Verlag.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo and Bayesian model determination. *Biometrika*, 82:711–732.

Hahn, J. S., Monyer, H., and Tharp, B. R. (1989). Interburst interval measurements in the EEGs of premature infants with normal neurological outcome. *Electroencephalography and Clinical Neurophysiology*, 73(5):410–418.

Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471.

Hellstrom-Westas, L., de Vries, L. S., and Rosen, I. (2008). *Atlas of amplitude-integrated EEGs in the newborn*. Informa Healthcare, London, 2nd edition.

Hellstrom-Westas, L., Rosen, I., de Vries, L. S., and Greisen, G. (2006). Amplitude-integrated EEG classification and interpretation in preterm and term infants. *NeoReviews*, 7(2).

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

Hoffman, C. S., Messer, L. C., Mendola, P., Savitz, D. A., Herring, A. H., and Hartmann, K. E. (2008). Comparison of gestational age at birth based on last menstrual period and ultrasound during the first trimester. *Paediatric and Perinatal Epidemiology*, 22(6):587–596.

Holthausen, K., Breidbach, O., Scheidt, B., and Frenzel, J. (1999). Clinical relevance of age-dependent EEG signatures in the detection of neonates at high risk for apnea. *Neuroscience Letters Volume*, 268(3):123–126.

Holthausen, K., Breidbach, O., Scheidt, B., and Frenzel, J. (2000). Brain dysmaturity index for automatic detection of high-risk infants. *Pediatric Neurology*, 22(3):187–191.

Inouye, T., Shinosaki, K., Sakamoto, H., Toi, S., Ukai, S., Iyama, A., Katsuda, Y., and Hirano, M. (1991). Quantification of EEG irregularity by use of the entropy of the power spectrum. *Electroencephalography and Clinical Neurophysiology*, 79(3):204–10.

Jakaite, L. and Schetinin, V. (2008). Feature selection for Bayesian evaluation of trauma death risk. In *The 14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics*, pages 123–126. Springer.

Janjarasjitt, S., Scher, M. S., and Loparo, K. (2008). Nonlinear dynamical analysis of the neonatal EEG time series: The relationship between neurodevelopment and complexity. *Clinical Neurophysiology*, 119(8):822836.

Jasper, H. H. and Andrews, H. L. (1938). Brain potentials and voluntary muscle activity in man. *Journal of Neurophysiology*, 1(2):87–100.

Jennekens, W., Ruijs, L. S., Lommen, C. M., Niemarkt, H. J., Pasman, J., van Kranen-Mastenbroek, V., Wijn, P. F., van Pul, C., and Andriessen, P. (2011). Automatic burst detection for the EEG of the preterm infant. *Physiological Measurement*, 32(10).

Jun, S. C., George, J. S., Kim, W., Pare-Blagoev, J., Plis, S., Ranken, D. M., and Schmidt, D. M. (2008). Bayesian brain source imaging based on combined meg/eeg and fmri using mcmc. *NeuroImage*, 40(4):1581 – 1594.

Kato, T., Okumura, A., Hayakawa, F., Tsuji, T., Natsume, J., and Watanabe, K. (2011). Evaluation of brain maturation in pre-term infants using conventional and amplitude-integrated electroencephalograms. *Clinical Neurophysiology*, 122(10):1967–1972.

Korotchikova, I., Connolly, S., Ryan, C. A., Murray, D. M., Temko, A., Greene, B. R., and Boylan, G. B. (2009). EEG in the healthy term newborn within 12 hours of birth. *Clinical Neurophysiology*, 120(6):1046–53.

Koszer, S. E., Moshe, S. L., and Holmes, G. I. (2006). *Clinical neurophysiology of infancy, childhood, and adolescence*, chapter Visual Analysis of the Neonatal Electroencephalogram. Elsevier.

Krajca, V., Petranek, S., Mohylova, J., Paul, K., Gerla, V., and Lhotska, L. (2009). Modeling the microstructure of neonatal EEG sleep stages by temporal profiles. In Lim, C. T., Goh, J. C. H., and Magjarevic, R., editors, *13th International Conference on Biomedical Engineering*, volume 23 of *IFMBE Proceedings*, pages 133–137. Springer Berlin Heidelberg.

Kropotov, J. (2009). *Quantitative EEG, event-related potentials and neurotherapy*. Elsevier.

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis*. Academic Press.

Kuncheva, L. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons.

Lederman, D. and Tabrikian, J. (2012). Classification of multichannel EEG patterns using parallel hidden Markov models. *Medical and Biological Engineering and Computing*, 50(4):319–328.

Lippe, S., Roy, M. S., Perchet, C., and Lassonde, M. (2007). Electrophysiological markers of visuocortical development. *Cerebral Cortex*, 17(1):100–107.

Lofhede, J., Thordstein, M., Lfgren, N., Flisberg, A., Rosa-Zurera, M., Kjellmer, I., and Lindecrantz, K. (2010). Automatic classification of background EEG activity in healthy and sick neonates. *Journal of Neural Engineering*, 7(1).

Lombroso, C. T. (1985). Neonatal polygraphy in full-term and premature infants: a review of normal and abnormal findings. *Journal of Clinical Neurophysiology*, 2(2):105–155.

Loomis, A. L., Harvey, E. N., and Hobart, G. A. (1937). Cerebral states during sleep, as studied by human brain potentials. *Journal of Experimental Psychology*, 21(2):127–144.

MacKay, D. J. C. (1998). Introduction to Monte Carlo methods. In Jordan, M. I., editor, *Learning in Graphical Models*, NATO Science Series, pages 175–204. Kluwer Academic Press.

Mccullagh, P. (1980). Regression models for ordinal data. *Journal of Royal Statistical Society B*, 42:109–142.

McEwen, J. A. and Anderson, G. B. (1975). Modeling the stationarity and gaussianity of spontaneous electroencephalographic activity. *IEEE Transactions on Biomedical Engineering*, 22(5):361–369.

Mizrahi, E. M., Hrachovy, R. A., and Kellaway, P. (2003). *Atlas of Neonatal Electroencephalography*. Lippincott Williams & Wilkins.

Niedermeyer, E. (2005). *Electroencephalography: basic principles, clinical applications, and related fields*, chapter Maturation of the EEG: Development of waking and sleep patterns, pages 209–234. Lippincott Williams & Wilkins, 5th edition.

Niemarkt, H. J., Andriessen, P., Pasman, J., Vles, J. S., Zimmermann, L. J., and Oetomo, S. B. (2008). Analyzing EEG maturation in preterm infants: The value of a quantitative approach. *Journal of Neonatal-Perinatal Medicine*, 1(3):131–144.

Niemarkt, H. J., Andriessen, P., Peters, C. H., Pasman, J., Zimmermann, L. J., and Bambang Oetomo, S. (2010). Quantitative analysis of maturational changes in EEG background activity in very preterm infants with a normal neurodevelopment at one year of age. *Early human development*, 86(4):219–224.

Niemarkt, H. J., Jennekens, W., Pasman, J., Katgert, T., van Pul, C., Gavilanes, A. W., Kramer, B. W., Zimmermann, L. J., Oetomo, S. B., and Andriessen, P. (2011). Maturational changes in automated EEG spectral power analysis in preterm infants. *Pediatric Research*, 70:529–534.

Nolan, H., Whelan, R., and Reilly, R. B. (2010). FASTER: Fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods*, 192(1):152–62.

Okumura, A., Hayakawa, M., Oshiro, M., Hayakawa, F., Shimizu, T., and Watanabe, K. (2010). Nutritional state, maturational delay on electroencephalogram, and developmental outcome in extremely low birth weight infants. *Brain and Development*, 32(8):613–618.

Olischar, M., Klebermass, K., Kuhle, S., Hulek, M., Kohlhauser, C., Rcklinger, E., Pollak, A., and Weninger, M. (2004). Reference values for amplitude-integrated electroencephalographic activity in preterm infants younger than 30 weeks gestational age. *PEDIATRICS*, 113(1):61–66.

Parmelee, Jr., A. H., Schulte, F. J., Akiyama, Y., Wenner, W. H., Schultz, M. A., and Stern, E. (1968). Maturation of EEG activity during sleep in premature infants. *Electroencephalography and Clinical Neurophysiology*, 24(4):319–329.

Parmelee, Jr., A. H., Wenner, W. H., Akiyama, Y., Schultz, M. A., and Stern, E. (1967). Sleep states in premature infants. *Developmental medicine and child neurology*, 9(1):70–77.

Paul, K., Krajca, V., Roth, Z., Melichar, J., and Petranek, S. (2003). Comparison of quantitative EEG characteristics of quiet and active sleep in newborns. *Sleep Medicine*, 4(6):543–552.

Piryatinska, A., Terdik, G., Woyczynski, W. A., Loparo, K., Scher, M. S., and Zlotnik, A. (2009). Automated detection of neonate EEG sleep stages. *Computer Methods and Programs in Biomedicine*, 95(1):31–46.

Platt, J., Cristianini, N., and Shawe-Taylor, J. (2000). Large margin DAGS for multiclass classification. In Solla, S. A., Leen, T. K., and Mueller, K. R., editors, *Advances in Neural Information Processing Systems 12*, pages 547–553.

Pressler, R., Bady, B., Binnie, C., Boylan, G. B., Connell, J. A., Lutschg, J., Oozeer, R. C., Prior, P. F., Scheffner, D., Suppiej, A., and Tedman, B. M. (2003). *Clinical Neurophysiology: EEG, paediatric neurophysiology, special techniques and applications*, chapter Neurophysiology of the neonatal period, pages 450–506. Elsevier Health Sciences.

Richards, J. E., Parmelee, Jr., A. H., and Beckwith, L. (1986). Spectral analysis of infant EEG and behavioral outcome at age five. *Electroencephalography and Clinical Neurophysiology*, 64(1):1–11.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer.

Robert, C., Gaudyb, J.-F., and Limogea, A. (2002). Electroencephalogram processing using neural networks. *Clinical Neurophysiology*, 113(5):694 – 701.

Rubin, M. A. (1938). The distribution of the alpha rhythm over the cerebral cortex of normal man. *Journal of Neurophysiology*, 1(4):313–323.

Sanei, S. and Chambers, J. A. (2007). *EEG Signal Processing*. Wiley-Interscience.

Scher, M. S. (1997). Neurophysiological assessment of brain function and maturation: A measure of brain adaptation in high risk infants. *Pediatric Neurology*, 16(3).

Scher, M. S. (2006). *Clinical neurophysiology of infancy, childhood, and adolescence*, chapter Electroencephalography of the Newborn: Normal Features. Elsevier.

Scher, M. S., Johnson, M. W., Ludington, S. M., and Loparo, K. (2011). Physiologic brain dysmaturity in late preterm infants. *Pediatric Research*, 70:524528.

Scher, M. S., Jones, R. H., Steppe, D. A., Cork, D. L., Seltman, H. J., and Banks, D. L. (2003a). Functional brain maturation in neonates as measured by EEG-sleep analyses. *Clinical Neurophysiology*, 114(5):875–882.

Scher, M. S., Ludington-Hoe, S., Kaffashi, F., Johnson, M. W., Holditch-Davis, D., and Loparo, K. (2009). Neurophysiologic assessment of brain maturation after an eight-week trial of skin-to-skin contact on preterm infants. *Clinical Neurophysiology*, 120(10):18121818.

Scher, M. S., M., S., Hatzilabrou, G. M., Greenberg, N. L., Cebulka, G., Krieger, D., Guthrie, R., and Sclabassi, R. J. (1990). Computer analyses of EEG-sleep in the neonate: methodological considerations. *Journal of Clinical Neurology*, 7(3):417–441.

Scher, M. S., Martin, J. G., Steppe, D. A., and Banks, D. L. (1994a). Comparative estimates of neonatal gestational maturity by electrographic and fetal ultrasonographic criteria. *Pediatric Neurology*, 11(3):214–8.

Scher, M. S., Steppe, D. A., Banks, D. L., Guthrie, R. D., and Sclabassi, R. J. (1995). Maturational trends of EEG sleep measures in the healthy preterm neonate. *Pediatric Neurology*, 12(4):314–322.

Scher, M. S., Steppe, D. A., Dokianakis, S. G., and Guthrie, R. D. (1994b). Maturation of phasic and continuity measures during sleep in preterm neonates. *Pediatric research*, 36(6):732–737.

Scher, M. S., Steppe, D. A., Salerno, D. G., Beggarly, M. E., and Banks, D. L. (2003b). Temperature differences during sleep between fullterm and preterm neonates at matched post-conceptional ages. *Clinical Neurophysiology*, 114(1):17–22.

Scher, M. S., Sun, M., Steppe, D. A., Banks, D. L., Guthrie, R. D., and Sclabassi, R. J. (1994c). Comparisons of EEG sleep state-specific spectral values between healthy full-term and preterm infants at comparable postconceptional ages. *Sleep*, 17(1):47–51.

Scher, M. S., Waisanen, H., Loparo, K., and Johnson, M. W. (2005). Prediction of neonatal state and maturational change using dimensional analysis. *Journal of Clinical Neurophysiology*, 22(3):159–165.

Schetinin, V., Fieldsend, J. E., Partridge, D., Coats, T. J., Krzanowski, W. J., Everson, R. M., Bailey, T. C., and Hernandez, A. (2007). Confident interpretation of bayesian decision tree ensembles for clinical applications. *IEEE Transactions on Information Technology in Biomedicine*, 11(3):312–319.

Schetinin, V., Fieldsend, J. E., Partridge, D., Krzanowski, W. J., Everson, R. M., and Bailey, T. C. (2004). The Bayesian decision tree technique with a sweeping strategy. In *The International Conference on Advances in Intelligent Systems*. IEEE Computer Society.

Schetinin, V., Fieldsend, J. E., Partridge, D., Krzanowski, W. J., Everson, R. M., Bailey, T. C., and Hernandez, A. (2006). Comparison of the Bayesian and randomized decision tree ensembles within an uncertainty envelope technique. *Journal of Mathematical Modelling and Algorithms*, 5:397–416.

Schetinin, V., Jakaite, L., and Schult, J. (2011). Informativeness of sleep cycle features in bayesian assessment of newborn electroencephalographic maturation. *Computer-Based Medical Systems, IEEE Symposium on*, 0:1–6.

Schetinin, V. and Schult, J. (2005). A neural-network technique to learn concepts from electroencephalograms. *Theory in Biosciences*, 124(1):41–53.

Scholz, F. W. and Stephens, M. A. (1987). K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, 82(399):918–924.

Sterman, M. B., McGinty, D., Harper, R. M., Hoppenbrouwers, T., and Hodgman, J. E. (1982). Developmental comparison of sleep EEG power spectral

patterns in infants at low and high risk for sudden death. *Electroencephalography and Clinical Neurophysiology*, 53(2):166–181.

Tharp, B. R. (1990). Electrophysiological brain maturation in premature infants: an historical perspective. *Journal of Clinical Neurophysiology*, 7(3):302–14.

Tharp, B. R., Scher, M. S., and Clancy, R. R. (1989). Serial EEGs in normal and abnormal infants with birth weights less than 1200 grams – a prospective study with long term follow-up. *Neuropediatrics*, 20(2):64–72.

Thornberg, E. and Thiringer, K. (1990). Normal pattern of the cerebral function monitor trace in term and preterm neonates. *Acta Paediatrica*, 79(1):2025.

Trujillo-Barreto, N. J., Aubert-Vazquez, E., and Valdes-Sosa, P. A. (2004). Bayesian model averaging in eeg/meg imaging. *NeuroImage*, 21(4):1300 – 1319.

Trujillo-Ortiz, A., Hernandez-Walls, R., Barba-Rojo, K., Cupul-Magana, L., and Zavala-Garcia, R. C. (2007). AnDarksamtest: Anderson-Darling k-sample procedure to test the hypothesis that the populations of the drawn groups are identical: A MATLAB file.

Turnbull, J., Loparo, K., Johnson, M. W., and Scher, M. S. (2001). Automated detection of trace alternant during sleep in healthy full-term neonates using discrete wavelet transform. *Clinical Neurophysiology*, 112(10):1893–1900.

Uglov, J., Jakaite, L., Schetinin, V., and Maple, C. (2008). Comparing robustness of pairwise and multiclass neural-network systems for face recognition. *EURASIP Journal of Advances in Signal Processing*, 2008.

Vakkuri, A., Yli-Hankala, A., Talja, P., Mustola, S., Tolvanen-Laakso, H., Sampson, T., and Vierti-Oja, H. (2004). Time-frequency balanced spectral entropy as a measure of anesthetic drug effect in central nervous system during sevoflurane, propofol, and thiopental anesthesia. *Acta Anaesthesiologica Scandinavica*, 48(2):145–53.

van de Velde, M., Ghosh, I. R., and Cluitmans, P. J. (1999). Context related artefact detection in prolonged EEG recordings. *Computer methods and programs in biomedicine*, 60(3):183–96.

van de Velde, M., van Erp, G., and Cluitmans, P. J. (1998). Detection of muscle artefact in the normal human awake EEG. *Electroencephalography and Clinical Neurophysiology*, 107(2):149158.

Victor, S., Appleton, R. E., Beirne, M., Marson, A. G., and Weindling, A. M. (2005). Spectral analysis of electroencephalography in premature newborn infants: Normal ranges. *Pediatric Research*, 57(3):336–341.

Viniker, D. A., Maynard, D. E., and Scott, D. F. (1984). Cerebral function monitor studies in neonates. *Clinical Electroencephalography*, 15(4):185–192.

Walter, W. G. (1936). The location of cerebral tumors by electroencephalography. *Lancet*, 2:305–308.

Walter, W. G. (1943). The distribution of the alpha rhythm over the cerebral cortex of normal man. *Electronic Engineering*, 16:236–238.

Webb, A., Copsey, K., and Cawley, G. (2011). *Statistical Pattern Recognition*. John Wiley & Sons.

West, C. R. (2006). *The role of spectral edge frequency monitoring in neonatal intensive care*. PhD thesis, The University of Auckland.

West, C. R., Harding, J. E., Williams, C. E., Nolan, M., and Battin, M. R. (2011). Cot-side electroencephalography for outcome prediction in preterm infants: observational study. *Archives of disease in childhood. Fetal and neonatal edition.*, 96(2):F108–13.

Wong, L. (2008). *Quantitative Continuity Feature for Preterm Neonatal EEG Signal Analysis*. PhD thesis, Computer Systems Engineering, The University of Auckland.

Wong, L. and Abdulla, W. (2008). Automatic detection of preterm neonatal EEG background states. In *The IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP 2008)*, page 421 424.