Prediction of Microsleeps from EEG using Bayesian Approaches

Mohammadreza Shoorangiz

A thesis presented for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering
at the
University of Canterbury,
Christchurch, New Zealand.

February 2018

ABSTRACT

Microsleeps are brief unintentional episodes of sleep-related loss of consciousness up to 15 s during an active task, such as driving. Such episodes of unresponsiveness are of particularly high importance in people who perform high-risk occupations requiring extended and unimpaired visuomotor performance, such as truck drivers, train drivers, pilots, and air-traffic controllers, where microsleeps can, and do, result in catastrophic accidents and fatalities. Therefore, prediction and even early detection of microsleeps has the potential to prevent sleep-related accidents and save lives.

The aim of this study was to investigate Bayesian methods in the prediction of microsleeps from the EEG. Bayesian methods were also investigated as a measure of improving performance in microsleep detection. This study used data from a previous study, 'Study A', in which 15 healthy non-sleep-deprived participants performed two 1-h sessions of 1-D continuous tracking task (CTT). Participants who experienced at least one definite microsleep during the two CTT sessions, i.e., 8 subjects, were included for the analyses. The original gold-standard was formed by integrating two independent measures: the video-rating from a human expert and tracking flat-spots.

In this study, we first refined the gold-standard to improve its accuracy and minimize potential false labels. A microsleep was defined as a non-tracking episode in conjunction with a deep-drowsy or lapse video rating, whereas a responsive episode was defined as a satisfactory tracking performance for at least of 5 s irrespective of video ratings. The remainder of the gold-standard was labelled uncertain and pruned out. We then proposed four Bayesian models for feature reduction. The first Bayesian model was variational Bayesian robust factor analysis (VBRFA), which is an extension of variational Bayesian FA (VBFA) that finds a robust latent space. This model was then extended to variational Baysian multi-subject RFA (VBMSRFA), which assumes independent mean and noise terms for individual subjects. Variational Bayesian hierarchical MSRFA (VBHMSRFA) finds a group level loading matrix and allows individual subjects to have slightly different loading matrices. Finally, variational Bayesian hierarchical multi-subject robust joint matrix factorization (VBHMSRJMF) incorporates the information of class labels while finding a less subject-variable latent space. All of these proposed methods used variational inference to approximate the posterior probabilities. Moreover, automatic relevance determination (ARD) motivated prior distributions were used in all models to automatically find the optimum number of latent variables.

Various features of EEG data were extracted and investigated, namely power spectral features (PSF) (192 features), power spectral features using individual alpha frequency (PSF-IAF) (192 features), multiple domain features (MDF) (176 features), wavelet mean squared features (WMSF) (80 features), wavelet log mean squared features (WLMSF) (80 features), and wavelet energy percentage features (WEPF) (80 features). These features were extracted from 2-, 5-, 10-s window lengths. In addition, four classifiers, i.e., linear discriminant analysis (LDA), linear support vector machine (SVM), tree augmented naïve Bayes (TAN), and variational Bayesian logistic regression (VBLR), were investigated to discriminate microsleeps. We found that aggregating features of the three window-lengths performed superior to individual single-window features. The proposed Bayesian feature reduction methods were applied to each feature set and the meta-features were extracted. These meta-features were then fed to a classifier to perform detection and prediction of microsleeps.

The best microsleep state detection performance was phi correlation coefficient (phi) – a performance measure for imbalanced datasets to quantify the correlation between actual and predicted labels – phi = 0.47 with an LDA and VBMSRFA meta-features of WLMSF (AUC-ROC = 0.95; AUC-PR = 0.49; GM = 0.83; Sn = 0.74; Pr = 0.38). On the other hand, the highest performance of the original features without any feature reduction/selection method, i.e., baseline performance, was $\varphi = 0.40$, achieved with a linear SVM and PSF (AUC-ROC = 0.95; AUC-PR = 0.49; GM = 0.79; Sn = 0.74; Pr = 0.38).

With a prediction time of $\tau = 1$ s for microsleep state prediction, the highest performance was $\varphi = 0.44$ which was achieved with an LDA classifier and VBMSRFA meta-features of WLMSF (AUC-ROC = 0.94; AUC-PR = 0.44; GM = 0.80; Sn = 0.69; Pr = 0.36). In contrast, the highest baseline performance was $\varphi = 0.35$ with a linear SVM and MDF (AUC-ROC = 0.92; AUC-PR = 0.42; GM = 0.76; Sn = 0.70; Pr = 0.35). Overall, VBMSRFA meta-features of WLMSF led to the highest performance for both detection and prediction of microsleep states.

The highest performance in terms of AUC-ROC and AUC-PR for microsleep onset detection was achieved with a linear SVM and PSF (AUC-ROC = 0.91; AUC-PR = 0.09; $\varphi = 0.08$; GM = 0.71; Sn = 0.79; Pr = 0.03). Notwithstanding, our proposed Bayesian methods achieved slightly higher GM and phi values. The highest phi of 0.10 (AUC-ROC = 0.89; AUC-PR = 0.05; GM = 0.78; Sn = 0.70; Pr = 0.03; Sp = 0.88) was achieved with a VBLR classifier and VBMSRFA meta-features of WLMSF. The highest GM, however, was 0.81 (AUC-ROC = 0.91; AUC-PR = 0.05; $\varphi = 0.09$; Sn = 0.77; Pr = 0.02; Sp = 0.85) with the same meta-features but with a linear SVM classifier.

With a prediction time of $\tau = 1$ s, the highest values of GM and phi for microsleep onset prediction were 0.80 and 0.08, respectively, and were achieved with a linear SVM and VBHMSRFA meta-features of MDF (Sn = 0.81; Pr = 0.01; Sp = 0.80). The highest performance of the baseline with the same prediction time was GM = 0.71 and $\varphi = 0.07$, achieved with a linear SVM and MDF (Sn = 0.76; Pr = 0.01; Sp = 0.74). Increasing the prediction time to $\tau = 10$ s, the highest performance for microsleep onset prediction was seen

with a linear SVM and VBHMSRFA meta-features of MDF (φ = 0.04; GM = 0.66; Sn = 0.66; Pr = 0.01; Sp = 0.71), while the highest performance of the baseline was with a linear SVM and WMSF (φ = 0.04; GM = 0.54; Sn = 0.57; Pr = 0.01; Sp = 0.73).

Our findings suggests that taking inter-subject variability into account can improve accuracy of microsleep detection and prediction by reducing the variability of classification threshold between training and test subjects. However, although our results indicate that microsleeps can be better detected and predicted by incorporation of Bayesian methods, the performances are still too low for real-life applications. Further investigations are needed to find a substantially improved microsleep prediction system for real-life applications.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Richard Jones and Dr. Steve Weddell, for their continuous support of my PhD study and research.

I would like to thank Richard for providing me the opportunity to undertake this project. His patience, motivation, enthusiasm, attention to details, and guidance helped me throughout research and writing of this thesis. I could not have imagined having a better mentor for my PhD study. I would also like to thank Steve for sharing his insightful comments, technical expertise, and encouragement.

I would like to acknowledge the support of the University of Canterbury which provided me a Doctoral Scholarship. I wish to acknowledge the travel grant I received from New Zealand Brain Research Institute to attend the IEEE Engineering in Medicine and Biology conference in Jeju, South Korea.

My sincere thanks goes to my colleagues and friends at New Zealand Brain Research Institute for their support, friendship, and energy. In particular, I would like to thank Dr. Sudhanshu Ayyagari and Mr. Bob Young for their energy, friendship, and constant encouragement.

Last but not the least, I would like to thank my family, my parents and brothers, for supporting me spiritually during the writing of this thesis and throughout my life in general.

CONTENTS

PREFACE

This thesis is submitted for the degree of Doctor of Philosophy in Electrical and Computer Engineering at the University of Canterbury. The research for this thesis was completed between October 2014 and October 2017 while I was enrolled in the Department of Electrical and Computer Engineering at the University of Canterbury. The work was carried out as part of Christchurch Neurotechnology Research Programme at the New Zealand Brain Research Institute and was supervised by Professor Richard Jones and Dr. Steve Weddell. I was supported by a University of Canterbury Doctoral Scholarship. The New Zealand Brain Research Institute kindly provided a travel grant for conference attendance.

PUBLICATIONS

Conference papers

- **Shoorangiz, R.**, Weddell, S., Jones, R. (2017), 'Bayesian Multi-Subject Factor Analysis to Predict Microsleeps from EEG Power Spectral Features', In *Proceedings of the 39th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4183–4186.
- **Shoorangiz, R.**, Weddell, S., Jones, R. (2016), 'Prediction of microsleeps from EEG: preliminary results', In *Proceedings of the 38th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4650–4653.

PRESENTATIONS

- July 2017, IEEE Engineering in Medicine and Biology conference (EMBC), Jeju Island, South Korea. Oral presentation.
- August 2016, IEEE Engineering in Medicine and Biology conference (EMBC), Orlando, Florida, United States. Poster presentation.

ABBREVIATIONS

ADASYN	Adaptive synthetic sampling
ANN	Artificial neural network
ARD	Automatic relevance determination
ASR	Artefact subspace reconstruction
AUC-PR	Area under the curve of precision recall
AUC-ROC	Area under the curve of receiver operating characteristic
BRFA	Bayesian robust factor analysis
BHMSRFA	Bayesian hierarchical multi-subject robust factor analysis
BHMSRJMF	Bayesian hierarchical multi-subject robust joint matrix factorization
BM	Definite behavioural microsleep
BMSRFA	Baysian multi-subject robust factor analysis
CTT	1-D continuous tracking task
DWT	Discrete wavelet transform
EEG	Electroencephalogram
EM	Expectation maximization
EMG	Electromyogram
EOG	Electrooculogram
ERP	Event-related potential
FA	Factor analysis
FN	False negetive
FP	False positive
GM	Geometric mean
i.i.d.	Independent and identically distributed
IAF	Individual alpha frequency
ICA	Independent component analysis

IWBW	Intensity-weighted bandwidth
IWMF	Intensity-weighted mean frequency
KFD	Katz fractal dimension
KL	Kullback-Leibler
kNN	K-nearest neighbours
LDA	Linear discriminant analysis
LOSO-CV	Leave-one-subject-out cross-validation
LSTM	Long short-term memory
MAP	Maximum a posteriori
MCL	Mean curve length
MCMC	Markov chain Monte Carlo
MDF	Multiple domain features
MLE	Maximum likelihood estimation
MSRFA	Multi-subject robust factor analysis
NHTSA	National Highway Traffic Safety Administration
NLE	Nonlinear energy
PCA	Principal component analysis
PFD	Petrosian fractal dimension
phi	Phi correlation coefficient
Pr	Precision
PSD	Power spectral density
PSF	Power spectral features
PSF-IAF	Power spectral features using individual alpha frequency
RACOG	Rapidly converging Gibbs sampler
RFA	Robust factor analysis
SMOTE	Synthetic minority over-sampling technique
Sn	Sensitivity
Sp	Specificity
SVM	Support vector machine
TAN	Tree augmented naïve Bayes
TN	True negetive
TP	True positive
VBFA	Variational Bayesian factor analysis

VBRFA	Variational Bayesian robust factor analysis
VBE	Variational Bayesian expectation
VBHMSRFA	Variational Bayesian hierarchical multi-subject robust factor analysis
VBHMSRJMF	Variational Bayesian hierarchical multi-subject robust joint matrix factorization
VBLR	Variational Bayesian logistic regression
VBM	Variational Bayesian maximization
VBMSRFA	Variational Bayesian multi-subject robust factor analysis
WEPF	Wavelet energy percentage features
WLMSF	Wavelet log mean squared features
WMSF	Wavelet mean squared features

Chapter 1

INTRODUCTION

1.1 OVERVIEW

Performing active and demanding tasks such as driving has become a part of our lives. These tasks are mostly accomplished without trouble since the information processing of the human brain is robust and flexible [Botvinick and Bylsma 2005]. However, failure to respond properly during an ongoing task due to a brief episode of sleep-related loss of consciousness, i.e., *microsleeps*, can lead to substantial errors. Microsleeps are usually harmless, as when a person momentarily falls asleep while reading a book or attending a lecture. However, microsleeps can result in catastrophic outcomes especially in the transport sector, where ‘nodding-off’ can result in serious accidents and fatalities [Higgins et al. 2017, Smith 2016, Watling 2014, Zhang et al. 2016a].

Microsleeps are brief episodes (up to 15 s) of unintentional sleep-related suspension of performance while performing an active task [Jones et al. 2010, Peiris et al. 2006]. They can occur without warning [Anund and Åkerstedt 2010] and are usually accompanied by behavioural changes including head nods, slow eye-closure, and increased duration of eye blinks [Jones et al. 2010].

It has been reported that fatigue is the cause of at least 9% of road accidents involving injuries and 16% of fatal crashes in New South Wales [Centre for Road Safety and Transport for NSW 2014]. A public poll results showed that 58.6% of drivers had driven while fatigued and drowsy, and, more importantly, 14.5% had fallen asleep at the wheel [Vanlaar et al. 2008]. According to National Highway Traffic Safety Administration (NHTSA), in 2014 a drowsy driver was involved in 846 fatalities in the United States [Higgins et al. 2017]. Moreover, crashes involving a drowsy driver totalled an estimated 83 000 per year in the United States (between 2005–2009), including nearly 1000 fatal crashes, 37 000 injury crashes, and 45 000 cases of property damage [Higgins et al. 2017, NHTSA 2011].

In 2015, the World Health Organization estimated that road traffic accidents are the ninth leading cause of death, claiming more than 1.2 million lives worldwide every year [World Health Organization 2015]. Furthermore, it estimated that road traffic accidents will become the seventh leading cause of death by 2030. The societal cost of drowsy driving in the United States has been estimated at $109 billion per year, including cost of societal harm and hospitalization [Higgins et al. 2017]. Drowsiness contributes to crashes in two ways: (1) the level of cognition drops and impairs the driver's reaction time and skills which leads to an increased accident risk [Larue et al. 2010, Lin et al. 2013, Wang et al. 2014a] and (2) the driver falls asleep at the wheel which leads to losing control [Higgins et al. 2017, Vanlaar et al. 2008].

Various studies have been done to determine who is more vulnerable and when is one more likely to have a microsleep [Innes et al. 2013, Peiris et al. 2006, Poudel et al. 2013, 2014]. These studies found that although sleep-deprived subjects have a higher tendency to have microsleeps, most normally-rested non-sleep-deprived subjects also experience microsleeps during a monotonous continuous task. In addition, long and irregular work shifts have been found to increase risk of fatigue [Baulk et al. 2007, Geiger-Brown et al. 2012, Härmä et al. 2002, Jay et al. 2008]. Poudel et al. [2014] found that 70% of 20 normally-rested subjects had frequent microsleeps during a 50-min continuous visuomotor task. Another study conducted on normally-rested non-sleep-deprived subjects concluded that 8 out of 15 subjects had definite microsleeps and that 6 of those had frequent microsleeps during two 1-h sessions of a 1-D continuous tracking task [Peiris et al. 2006]. Although sleep restriction increases the propensity to fall asleep, Innes et al. [2013] found no correlation between the number of microsleeps when normally-rested compared to when sleep-restricted.

1.2 MOTIVATION

According to the results of a national survey in the United States [Tefft 2010], 41% of drivers admitted having fallen asleep while driving at least once in their lifetime. Moreover, the proportion of drivers who had fallen asleep in the past year and past month were 11% and and 3.9%, respectively [Tefft 2010]. Several studies have revealed that most people, whether they are sleep-deprived or not, are vulnerable to having microsleeps [Buckley et al. 2016, Forsman et al. 2013, Innes et al. 2013, Poudel et al. 2014, Tefft 2014]. This raises major safety concerns, especially for those who perform high-risk occupations that require extended unimpaired visuomotor performance such as pilots, air-traffic controllers, truck drivers, and train drivers [Baas et al. 2000, Gander et al. 2014, Häkkänen and Summala 2001, Härmä et al. 2002, Lic and Summala 2000, Moller et al. 2006, NTSB 2016, Taneja 2007, Zhang and Chan 2014]. Hence, prediction of an imminent microsleep has the potential to save lives and prevent catastrophic accidents.

Detection of microsleeps has been the subject of several studies [Ayyagari et al. 2015, Davidson et al. 2007, Golz et al. 2007, Holub et al. 2015, Malla et al. 2010, Peiris et al. 2011]. However, although they were able to detect microsleeps, their performances were relatively low and impractical for real-life usage. In addition, detection of a microsleep occurs after the person has fallen asleep and thus might already be too late. Therefore, a microsleep prediction system is required to prevent the occurrence of microsleeps and consequently minimize accidents due to falling asleep at the wrong time.

The electroencephalogram (EEG) is a non-invasive approach to record brain activity and has a high temporal resolution [Freeman and Quiroga 2013]. The EEG has been widely used in the literature for estimation of drowsiness, inattention, vigilance level, and sleep stage [Akin et al. 2008, Bojić et al. 2010, Chen et al. 2015, Huang et al. 2016, 2008, Kurt et al. 2009, Larue et al. 2011, Subasi et al. 2005]. Moreover, it has been shown that the EEG can be used to identify reduced alertness and fatigue with relatively high accuracy [Lal and Craig 2005, Larue et al. 2015, Yeo et al. 2009]. Therefore, the EEG is expected to be a suitable source of information to predict imminent microsleeps with an acceptable temporal resolution.

With recent improvements, wireless EEG headsets such as Emotiv[1], Neurosky[2], and Cognionics[3] have made it easier to record brain activities, especially in real-life applications [Martinez-Leon et al. 2016, Mullen et al. 2015]. Knopp [2015] developed a platform Elapse, which is a wearable device capable of recording and processing multiple biosignals in real-time. Therefore, if imminent microsleeps could be predicted from the EEG, it would be feasible to develop a real-time EEG-based microsleep prediction system for real-life applications.

The EEG, however, has a few shortcomings such as low signal-to-noise ratio, high dimensionality, nonstationarity, and inter-subject and intra-subject variability [Freeman and Quiroga 2013, Haegens et al. 2014, Thomas and Vinod 2016, Tong and Thakor 2009, Wei et al. 2015]. Such shortcomings pose challenges for using EEG in machine learning, especially for generalizing a trained model to perform well for a new unseen subject. Bayesian methods, on the other hand, can explicitly model noise, recover the underlying signal, and even find a lower-dimension representation of data [Bishop 2006, Ghahramani 2015, Mohammadiha et al. 2013, Murphy 2012, Zhao et al. 2015a]. Moreover, Bayesian methods have been used for EEG signal processing [Georgieva et al. 2016, Puuronen and Hyvärinen 2014, Wipf and Nagarajan 2009], feature extraction [Kang and Choi 2014, Strobbe et al. 2014, Wu et al. 2015, 2011], and classification [Qian et al. 2017, Zhang et al. 2016b]. Therefore, it is expected that incorporating Bayesian methods into EEG-based microsleep prediction would lead to improved performance.

1.3 OBJECTIVES

Investigation of microsleep prediction was the overall objective of this thesis. The EEG was considered to contain information related to the imminent microsleep, which could be used for predicting and ultimately preventing microsleeps. More specifically, the objectives were as follows:

1. Conduct a review of the literature and discover the current knowledge of EEG-based approaches for the detection and prediction of microsleeps and lapses.

[1] www.emotiv.com

[2] www.neurosky.com/biosensors/eeg-sensor

[3] www.cognionics.com

2. Review the previously conducted Study A data (described in Section 2.5) and refine the gold-standard to improve its precision.

3. Identify and examine different features of EEG to find a reliable set of features for microsleep prediction.

4. Develop Bayesian feature reduction methods to maximally extract information to improve the performance of microsleep detection.

5. Investigate EEG patterns prior to microsleeps that can be used for prediction of imminent microsleeps.

6. Evaluate the performance of microsleep prediction using Study A data.

1.4 THESIS ORGANIZATION

This thesis is organized into 12 chapters. The current chapter provided an overview of the extent of the microsleep problem and the motivations for predicting these events. Chapters 2 and 3 provide a review of literature related to the key aspects of this research. Chapter 2 provides a literature review of EEG, microsleeps, and the approaches used for microsleep detection. Chapter 3 provides an overview of key concepts of Bayesian machine learning and inference models. Chapter 4 presents our aims and hypotheses for this research. Chapter 5 describes the steps undertaken to minimize EEG artefacts and refine the gold-standard to improve its precision. An overview of the microsleep prediction system including feature extraction, classification models, definition of microsleep state and event prediction, and evaluation procedures for the proposed methods is given in Chapter 6. Chapters 7–10 present our proposed Bayesian methods for feature reduction to improve microsleep prediction accuracy as well as the results achieved for microsleep detection and prediction. Chapter 7 presents an extension of variational Bayesian factor analysis (VBFA) with a robust latent space representation. Chapter 8 extends the proposed model of Chapter 7 to a multi-subject variant to reduce inter-subject variability. This is further extended to a hierarchical model of the loading matrix in Chapter 9. Chapter 10 extends the idea of Chapter 9 to exploit the gold-standard information at the time of feature reduction. A comparison of our proposed methods and a discussion is given in Chapter 11. Finally, the conclusion and key findings of this research, a critique of the current study, and ideas for future work are presented in Chapter 12. The original contributions of this thesis are provided in Chapters 5 and 7–11.

Chapter 2

ELECTROENCEPHALOGRAPHY AND MICROSLEEP DETECTION: A REVIEW

This chapter aims to (1) provide a brief overview of the electroencephalogram and its application in sleep-staging and drowsiness detection, (2) provide a brief overview of different lapse types, (3) review the literature of microsleep and lapse detection and their limitations, (4) and provide a description of experimental Study A.

2.1 ELECTROENCEPHALOGRAM

The electroencephalogram (EEG) is a measure of brain activity from different spatial locations on the scalp. These activities are in the form of electrical potential and reflect the underlying brain sources [Tong and Thakor 2009]. The EEG is usually recorded noninvasively from the scalp with multiple electrodes. The main advantages of the EEG are (1) noninvasive recording, (2) high temporal resolution, and (3) relative inexpensiveness. Hence, the EEG has been widely used for diagnostics and clinical settings as well as in research laboratories [Freeman and Quiroga 2013, Tong and Thakor 2009].

Recording the EEG is done by placing multiple electrodes on the scalp according to a specific spatial distribution. The 10–20 and 10–10 systems are commonly used and internationally recognized distributions of EEG electrodes to record brain activities [Oostenveld and Praamstra 2001, Tong and Thakor 2009], as shown in Figure 2.1. EEG electrodes are usually placed on the scalp with gel to increase the conductance. Acceptable impedance for the electrodes varies among researchers and textbooks, but is usually between 5 to 10 kΩ [Duffy et al. 1989, Freeman and Quiroga 2013, Tong and Thakor 2009].

A disadvantage of EEG is its susceptibility to artefact contamination [Daly et al. 2015]. The EEG has a very small amplitude (order of tens of µV) and hence can be easily contaminated with artefacts [Nolan et al. 2010]. Sources of artefacts can be intrinsic (e.g., eye-movement and muscle activity) or extrinsic (e.g., 50 Hz electrical line) [Duffy et al. 1989]. Independent component analysis (ICA) has been widely used in the literature to minimize artefacts, especially eye-movements and eye-blinks [Daly et al. 2013, Delorme et al. 2007, Klados et al. 2011]. Mullen et al. [2015] proposed artefact subspace reconstruction (ASR) to remove artefacts of EEG

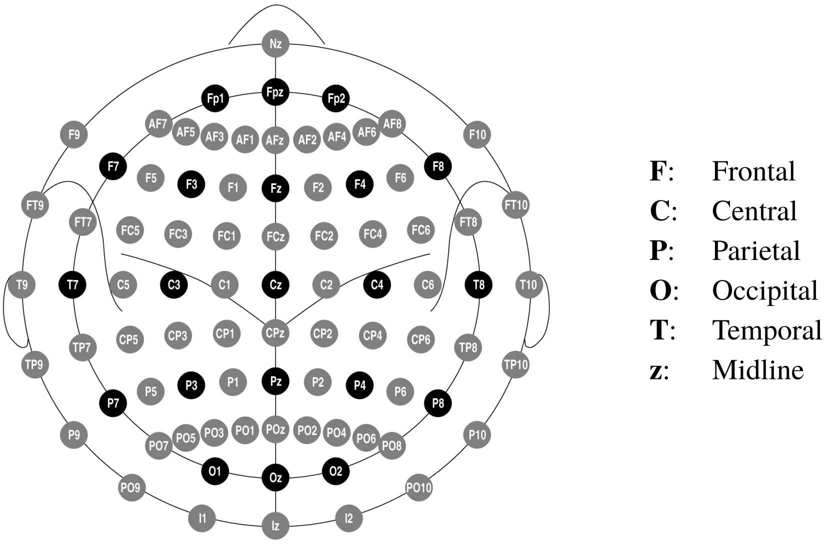

Nz
Fp1 Fpz Fp2
F9 AF7 AF5 AF3 AF1 AFz AF2 AF4 AF6 AF8 F10
F7 F5 F3 F1 Fz F2 F4 F6 F8
FT9 FT7 FC5 FC3 FC1 FCz FC2 FC4 FC6 FT8 FT10
T9 T7 C5 C3 C1 Cz C2 C4 C6 T8 T10
TP7 CP5 CP3 CP1 CPz CP2 CP4 CP6 TP8
TP9 TP10
P7 P5 P3 P1 Pz P2 P4 P6 P8
P9 PO7 PO5 PO3 PO1 POz PO2 PO4 PO6 PO8 P10
O1 Oz O2
PO9 PO10
I1 Iz I2
F: Frontal
C: Central
P: Parietal
O: Occipital
T: Temporal
z: Midline

Figure 2.1 Spatial distribution of the electrodes for EEG recording. The original 10–20 electrode positions are shown in black circles and the additional 10–10 extension electrodes are shown in grey. Reprinted from [Oostenveld and Praamstra 2001] with permission from Elsevier.

in real-time. ASR uses calibration data, i.e., a clean and artefact-free segment of data, and applies principal component analysis (PCA) to a sliding window of data. The principal components with a variance lower than a certain threshold, which is computed based on the calibration data, are removed and the data reconstructed from the remaining principal components [Mullen et al. 2013, 2015].

The EEG is divided into four standard clinical frequency bands, namely delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–35 Hz) [Duffy et al. 1989, Tong and Thakor 2009]. These frequency bands show different characteristics of cognition and mental state. Delta oscillations are increased with deep sleep. Theta rhythms are increased with drowsiness and lighter stages of sleep. The alpha band is usually seen in awake individuals, but is attenuated with eyes open. The beta band increases with alertness and excitement [Duffy et al. 1989, Freeman and Quiroga 2013, Tong and Thakor 2009].

2.1.1 Automated sleep-staging using EEG

Identification of stages of sleep can be clinically important. This is usually done manually using polysomnographic signals, although ratings are inconsistent among experts [Chapotot and Becq 2010, Stanley 1996]. Among physiological signals, the EEG has been found to be a good measure to identify stages of sleep [Asyali et al. 2007, Nicolaou and Georgiou 2011, Sheng et al. 2012, Weiss et al. 2011]. As a result, the EEG has been widely used for automatic sleep-staging and has shown compelling accuracies [Bajaj and Pachori 2013, Fell et al. 1996, Güneş et al. 2010, Khalighi et al. 2012, Şen et al. 2014].

Fell et al. [1996] investigated sleep staging using nonlinear and spectral features of EEG. They used an artificial neural network (ANN) classifier and found that nonlinear measures gave a better discrimination between stages I and II, whereas spectral features are better for distinguishing stage II from stages III and IV. Their highest overall classification accuracy was 77%.

Güneş et al. [2010] used frequency-domain features of 30-s EEG-epochs extracted with Welch's method [Welch 1967] to detect six sleep stages. They compared the performance of two classifiers, k-nearest neighbours (kNN) and C4.5 decision tree classifiers, and found that kNN with $k = 30$ performed the best. Their highest overall accuracy was 82.2%.

Şen et al. [2014] compared various features and classifiers for automatic sleep-staging. They extracted numerous features, including time domain, frequency domain, discrete wavelet transform (DWT), and entropy, from 30-s EEG-epochs. Different classifiers, such as support vector machine (SVM), ANN, C4.5 decision tree, and random forest, were used for the classification task, where the highest accuracy of 97% was achieved with a random-forest classifier.

These results indicate that the EEG contains substantial sleep-wake-related information which led to a high accuracy of 97% for automatic sleep-staging. Notwithstanding, the gold-standard for sleep-staging is usually identified by an expert from 30-s epochs of EEG. Hence, the results might be slightly biased.

2.1.2 Drowsiness detection using EEG

Subasi et al. [2005] investigated drowsiness detection from the EEG using wavelet decomposition of 5-s epochs. They applied an order-2 Daubechies DWT to each epoch of EEG and decomposed it to 5 levels. The wavelet approximation and details corresponding to the standard EEG frequency bands were then fed to an ANN classifier to detect one of the three labels: alert, drowsy, and sleep. Their overall accuracy was 95%. In a later study, Akin et al. [2008] added an electromyogram (EMG) signal from the chin and trained the ANN with both the EMG signal and wavelet decomposition of the EEG. The overall accuracy of their system was 99%.

Yeo et al. [2009] performed drowsy detection during a simulated driving session. Participants were asked to watch a video clip of a road-trip and perform as they were driving the car. A drowsiness gold-standard was generated based on an expert's rating of individual's blink frequency, blink duration, and EEG. An SVM classifier was trained on frequency-domain features of 10-s EEG segments. They were able to achieve an accuracy of 99.3% in identification of alertness and drowsiness of participants.

Chen et al. [2015] extracted nonlinear features of the wavelet decomposition of 8-s epochs of EEG to distinguish between alert and drowsy states. The EEG was recorded while participants were performing mental calculations in a cubicle. Similar to Yeo et al. [2009], the alert or drowsy states were defined by experts using eye-blink frequency and duration and the EEG. An extreme learning machine classifier was used and an accuracy of 97.3% achieved.

It is evident that the EEG contains information regarding arousal and wakefulness. However, although the accuracies of drowsiness detection and sleep-staging are high in the literature, gold-standards in both tasks were generated by an expert using EEG segments. A drawback of such gold-standard generation is that the behavioural cues and goal-directed performance have been discarded. Additionally, these gold-standards are not independent from the EEG itself since the labels were generated from EEG. Finally, although drowsiness detection might be of importance to reduce the likelihood of accidents by warning the driver, it does not necessarily detect microsleep or prevent drivers from momentarily falling asleep behind the wheel.

2.2 LAPSES OF RESPONSIVENESS

Lapses of responsiveness (‘*lapses*’) are short episodes of failure to respond during goal-directed tasks such as driving [Buckley et al. 2016, Geiger-Brown et al. 2012, Jones et al. 2010, Killeen 2013, Peiris et al. 2006, 2005, Weissman et al. 2006]. Lapses are different in terms of their association with the underlying cognitive mechanism and subject experience. Failure to respond in time is a type of lapse which leads to prolonged reaction times [Buckley et al. 2016, Unsworth and Robison 2016, Weissman et al. 2006]. Failure to respond correctly is another type of lapse resulting in response error [Finkbeiner et al. 2015]. Other lapses are complete phasic disruption of sensory-motor and cognitive performance [Jones et al. 2010, Peiris et al. 2006]. Behavioural cues such as slow eye-closure and head nodding are associated with some lapses [Peiris et al. 2006].

2.2.1 Classification of lapses

As mentioned in the previous section, different reasons and mechanisms exist for a temporary loss of responsiveness and therefore lapses are categorized as follows:

- ***Attention lapse*** is a brief loss of task-oriented attention resulting in a disruption of goal-directed behaviour. A momentary lapse in attention makes the individual respond slowly or even forget what he/she had intended to do [Buckley et al. 2016, Jones et al. 2010, Weissman et al. 2006]. Control of attentional focus is associated with frontal cortex which favours processing of relevant stimuli over irrelevant ones [Hopfinger et al. 2000, Weissman et al. 2006, Woldorff et al. 2004].

- ***EEG microsleep*** is defined as short disruption of performance identified by changes in the EEG power spectrum distribution, especially activity in theta band [Huang et al. 2008, Makeig et al. 2000, Williams et al. 1962]. Boyle et al. [2008] investigated EEG-microsleeps in participants with obstructive sleep apnoea and found measurable changes in driving performance. This thesis follows Boyle et al. [2008] for defining EEG-microsleep, which is an episode of 3 to 15 s disruption of performance identified from the EEG.

- ***Behavioural microsleep*** is an unintentional temporary loss of consciousness where the person seems to momentarily fall asleep. Behavioural microsleeps are associated with lower arousal levels and usually identified by behavioural cues including slow eye-closure,

> loss of facial-tone, and head nodding [Davidson et al. 2007, Jones et al. 2010, Peiris et al. 2006]. Behavioural microsleeps can be as short as 500 ms while the shortest episode of EEG-microsleep is 3 s. More importantly, behavioural microsleeps are identified by behavioural cues and task performance, whereas EEG-microsleeps are identified by burst of theta activity in the EEG. Moreover, Peiris et al. [2006] did not observe EEG theta bursts with behavioural microsleeps.

The focus of this thesis is on behavioural microsleeps, which are identified by the behaviour and goal-directed performance. For the sake of simplicity and readability, we use *microsleeps* to refer to behavioural microsleeps for the rest of this thesis.

2.3 MICROSLEEP AND LAPSE DETECTION

Golz et al. [2005, 2007] and Sommer et al. [2005] developed a microsleep detection system using 5 EEG channels and 2 electrooculogram (EOG) channels. Participants were asked to perform a driving simulation. To increase the likelihood of microsleeps, a monotonous task was intentionally selected. They used behavioural cues such as prolonged eye-closure and nodding-off and goal-directed performance such as driving incidents to identify microsleeps. Two sets of features, i.e., power spectral and delay vector variances, were extracted from 3-s segments of EEG and EOG. Power spectral features were calculated with a windowed periodogram of EEG segments and averaged over power in the standard EEG frequency bands. Delay vector variances, however, measure the nonlinearity and stochastic nature of the signal. They found that although fusion of both feature sets gave an accuracy of 88.8%, using only power spectral features achieved essentially the same accuracy at 88.0%, where accuracy was defined as the total number of correct classifications relative to the total number of instances. Furthermore, their highest accuracy was achieved with an SVM classifier. Although they achieved what appears to be high accuracy, the non-microsleep part of the data was selected to balance the two classes. This is acceptable for the training of classifiers but the testing phase must be performed on independent unselected data. Balancing the test data introduces bias to the performance measures since many of the ‘alert’ episodes have been removed, which substantially reduces opportunities for false detections. Furthermore, they concatenated data of all the subjects and then performed cross-validation. As a result, their reported performances are not true measures of performance and give little indication of the system’s ability to generalize to a new unseen test subject.

Krajewski et al. [2008] developed a microsleep detection system using speech processing. The experiment paradigm was similar to of Golz et al. [2007] where participants were performing a driving simulation. In addition to the driving task, the participants were instructed to engage in a verbal task similar to the communication between pilot and air-traffic controller. Frequency-domain features of speech were used to detect microsleep events. Using an SVM classifier, their highest accuracy was 86.1%. In a later study [Krajewski et al. 2009], the effect of various feature reduction methods were analysed. None of the feature reduction methods, either supervised or unsupervised, improved the performance. Although detection of microsleeps from speech is debatably practical in pilots and air-traffic controllers, it is not a viable solution for drivers as they are usually not talking. Also, when talking, they are less likely to have a microsleep, albeit more likely to have a lapse due to distracted attention. But a silent period could be due to an episode of microsleep, which would be missed with this system.

Lin et al. [2013] used a sustained-attention driving task to detect 'behavioural lapses' and assess the effectiveness of providing feedback during behavioural lapses. In this task [Huang et al. 2008, Lin et al. 2010], participants were asked to maintain lanes while random lane changes were induced. At the beginning of each session, 5-min calibration data were recorded to quantify the alert reaction-time. A behavioural lapse event for each individual was then arbitrarily defined as any event with a reaction-time of more than 3-times their respective alert reaction-time. Power spectral features were extracted from the ICA-decomposed EEG data. This study was extended to an online system to detect behavioural lapses and fatigues and provide feedback in real-time [Huang et al. 2016, Wang et al. 2014b]. Although their results indicated that providing feedback can reduce reaction times, detection accuracies were not included. A shortcoming of this method lies in the sustained-attention task itself, which is a discrete task. As a result, a lapse could not be identified until a lane deviation is induced and therefore the onset of a lapse remains unclear.

Peiris et al. [2006] used a 1-D continuous tracking task (CTT) to identify microsleeps and lapses. A continuous visuomotor task has a higher temporal accuracy and hence enabled them to identify the onset of lapses with a higher resolution. Davidson et al. [2007] used a long short-term memory (LSTM) recurrent neural network to detect lapses with a temporal resolution of 1 s. Using log-power spectral features, their performance in terms of phi correlation coefficient (φ given in Equation (6.21)) was 0.38. Peiris et al. [2011] pruned the data by removing noisy epochs and then used a stacked generalization of linear discriminant analysis (LDA) classifiers with power spectral and nonlinear features. They achieved a slightly higher phi of 0.39 with log-power spectral features. It is notable that both Davidson et al. and Peiris et al. used the leave-one-subject-out cross-validation (LOSO-CV) method to evaluate the performance of their methods and thus their reported performances provided much better estimates of the extent to which their methods can be generalized to new subjects. LaRocco [2015] applied SVM classifiers to this same pruned data which resulted in a decline of performance ($\varphi = 0.32$). Ayyagari et al. [2015] used a stack generalization of leaky echo-state neural networks and achieved a phi of 0.44 with the unpruned data and 0.51 with the pruned data. Nevertheless, although they were able to improve the performance of lapse detection system, the performance is still too low for real-life applications.

2.4 LIMITATIONS OF CURRENT EEG-BASED TECHNIQUES FOR MICROSLEEP DETECTION

Despite several works towards detection of lapses, a reliable method with an acceptable performance for real-life applications has yet to be found. Although some studies have reported high accuracies of lapse/microsleep detection, they cross-validated on concatenated data of multiple subjects without leaving an independent subject for testing phase [Golz et al. 2005, 2007, Sommer et al. 2005]. However, the EEG has subject-specific characteristics which have been used for biometric identification [DelPozo-Banos et al. 2015, Klonovs et al. 2013, Zhao et al. 2010]. As a result, computing performance measures on concatenated data of multiple subjects does not take the inter-subject variabilities into account. Thus, such performance measures will be biased and cannot be generalized to new unseen subjects.

A microsleep can only be detected at the onset of the event at the very earliest. However, this might be too late, as the individual is already non-responsive (e.g., while driving). As such, a microsleep prediction system is desired to be able to predict imminent microsleeps seconds before their occurrence. Golz et al. [2016] developed an EEG-based system to detect the onset of imminent microsleeps. They used power spectral and time-frequency features with SVM and ANN classifiers. Their highest detection accuracy of imminent microsleep onsets was 87.5%, achieved with an SVM and power spectral features. However, a major shortcoming of their study was that a cross-validation on the concatenated data of all subjects was used to evaluate performance. As a consequence, the training data and the test data were not independent and hence the performance is biased. Additionally, they only discriminated between microsleep and drowsy epochs and did not process the whole alert data of all subjects. Therefore, their reported performance cannot reflect their system's true predictive performance.

A major gap in the literature is evident for *prediction* of microsleeps. This study aims to address this issue by developing an EEG-based microsleep prediction system.

2.5 STUDY A

This research focuses mainly on the data of Study A. Study A is one of the datasets acquired by NeuroTech's Lapse Research Programme[1]. This dataset has been extensively used to detect lapses [Ayyagari et al. 2015, Ayyagari 2017, Davidson et al. 2007, LaRocco 2015, Peiris et al. 2011]. A description of this important study and its behavioural and EEG data is presented in this section from Peiris [2008].

2.5.1 Apparatus and procedure

Fifteen healthy non-sleep-deprived male volunteers with an average age of 26.5 years (18–36) were recruited for Study A. Restrictions were placed on age and gender to minimize sources

[1] `www.neurotech.org.nz`

of variation in the data. No current or previous neurological or sleep disorder was reported by the participants. Visual acuities (both eyes together) of the participants were 6/9 or better. The average sleep in the night prior to the test was 7.8 h (SD = 1.2 h, min = 5.1 h) to ensure participants were non-sleep-deprived. Participants performed two 1-h sessions of CTT, one week apart, while their physiological data, facial video, and task performance were recorded.

Participants performed the 1-D continuous tracking task (CTT) with an 8-s preview. The pseudo-random target had a period of 128 s and a bandwidth of 0.164 Hz, which was generated by summation of 21 sinusoids with random phases but frequencies evenly spaced at 0.00781 Hz intervals, shown in Figure 5.5. During the task, the target scrolled down the screen at a rate of 21.8 mm/s and the participants had a steering wheel (395 mm diameter) to follow the target with an arrow-shaped cursor. The position of steering wheel was recorded with a sampling frequency of 64 Hz which was used for analysing the tracking performance. Moreover, a camera 1 m from the participant recorded (25 frames per second) facial features while performing the CTT.

The EEG was recorded from 16 electrodes, namely Fp2, F4, C4, P4, O2, Fp1, F3, C3, P3, O1, F8, T4 (T8), T6 (P8), F7, T3 (T7), and T5 (P7), according to the 10–20 international system (see Figure 2.1). Additionally, horizontal and vertical EOG signals were recorded to facilitate removal of eye-artefacts. The EEG sampling frequency was 256 Hz. The reference and ground electrodes were placed on the forehead and the linked ears, respectively.

2.5.2 Original gold-standard

The behavioural gold-standard for Study A was originally formed by combining two independent measures, (1) the video ratings from a human expert and (2) tracking *flat-spots* found by an automated algorithm. An expert analysed the video recordings, facial cues, and marked them on a 6-scale basis, namely alert, distracted, forced eye-closure while alert, light drowsy, deep drowsy, and lapse. Independently, Peiris et al. [2006] developed an automated algorithm with relatively conservative thresholds to find the flat-spots. Based on the logical operator used to combine the two independent measures, two gold-standards were generated, *lapse index* and *definite behavioural microsleeps* (BM). The lapse index is the logical *OR* of the two independent measures and therefore a lapse was either a video-lapse and/or a tracking flat-spot. On the other hand, a BM was defined as the occurrence of both a tracking flat-spot *AND* a video-lapse. However, both of these gold-standards contained false information, as discussed in Section 5.3.

Chapter 3

RELATED BAYESIAN MACHINE LEARNING METHODS: A REVIEW

The aim of this chapter is to (1) provide an overview of Bayesian data analysis, (2) present variational inference, and (3) provide an overview of relevant Bayesian machine-learning algorithms.

3.1 BAYESIAN DATA ANALYSIS

A Bayesian approach is the process of using probabilities to describe the available data and infer the desired information [Eddy 2004]. An advantage of Bayesian methods is their ability to deal with uncertainty. Both measurement noise and finite amounts of data contribute to uncertainty in modelling [Bishop 2006]. Moreover, having many unknown parameters in a model introduces uncertainty about which parameters will perform well on new unseen data [Ghahramani 2015].

Classical non-probabilistic methods use point estimation, i.e., fixed values, for unknown parameters, whereas Bayesian methods use probability theory to describe uncertain parameters and integrate over all possible values of such paramters [Bishop 2006, Gelman et al. 2013, Ghahramani 2015]. Probability theory is a mathematical language which uses probability distributions to represent uncertain parameters and quantities [Bishop 2006]. After observing data, basic probability rules are used to update prior probability distributions to posterior distributions.

Assuming a prior probability distribution for the unknown parameters Θ, i.e., $p\left(\Theta \mid m\right)$ for the model m, the joint probability distribution $p\left(\mathbf{x}, \Theta \mid m\right)$ can be written as

$$p\left(\mathbf{x}, \Theta \mid m\right) = p\left(\mathbf{x} \mid \Theta, m\right) p\left(\Theta, m\right), \tag{3.1}$$

where $\mathbf{x}$ is a data vector and $p\left(\mathbf{x} \mid \Theta, m\right)$ is the likelihood of data $\mathbf{x}$ given the parameters Θ. Bayes' theorem [e.g., Bishop 2006] can be used to find the posterior distribution of the parameters Θ given the data $\mathbf{x}$,

$$p\left(\Theta \mid \mathbf{x}, m\right) = \frac{p\left(\mathbf{x} \mid \Theta, m\right) p\left(\Theta \mid m\right)}{p\left(\mathbf{x} \mid m\right)}, \tag{3.2}$$

where $p\left(\mathbf{x} \,\middle|\, m\right)$ is the model evidence or marginal likelihood and normalizes the posterior probability,

$$p\left(\mathbf{x} \,\middle|\, m\right) = \int p\left(\mathbf{x} \,\middle|\, \Theta, m\right) p\left(\Theta \,\middle|\, m\right) d\Theta. \tag{3.3}$$

Throughout this thesis, only one model was investigated at a time. Therefore, for the sake of simplicity, the model notation m is omitted from the rest of the probability distribution equations.

Given the observed data and a Bayesian model, posterior predictive analysis can be used to make predictions about new unseen data $\hat{\mathbf{x}}$ by using the posterior probability of the model parameters and integrating out all the variables except for the variables of interest [Ghahramani 2015],

$$p\left(\hat{\mathbf{x}} \,\middle|\, \mathbf{x}\right) = \int p\left(\hat{\mathbf{x}} \,\middle|\, \Theta, \mathbf{x}\right) p\left(\Theta \,\middle|\, \mathbf{x}\right) d\Theta. \tag{3.4}$$

A fully Bayesian treatment for inferring the parameters, however, is computationally challenging [Bishop 2006, Murphy 2012]. As the number of parameters and dimensions increase, the operation of integrating over all values of random variables becomes computationally more expensive. To address this issue, posterior probability can be summarized in a point estimate. Maximum *a posteriori* (MAP) estimates the mode (or median) of the posterior probability [Murphy 2012]. Despite the computationally appealing property of MAP, it has disadvantages. MAP does not contain any information regarding the uncertainty of values. Moreover, using MAP to find the predictive distribution can result in overfitting [Murphy 2012]. Approximate Bayesian inference algorithms, on the other hand, find an approximation to the true posterior probability.

Approximate Bayesian inference algorithms can be categorized into two main families, stochastic and deterministic approximate techniques [Sun 2013]. Numerical sampling methods such as Markov chain Monte Carlo (MCMC) fall into the stochastic approximate inference category. These methods randomly draw samples from the posterior distributions [Murphy 2012, Wu et al. 2016]. On the other hand, deterministic approximate methods such as variational inference find an approximation of the posterior with tractable distributions [Bishop 2006, Wu et al. 2016].

The advantage of stochastic inference methods lies in the fact that they are flexible for a wide range of distributions and they allow for faster programming. However, these methods can be very slow, especially with large-scale data [Bishop 2012, Sun 2013]. Moreover, monitoring the convergence of such methods is difficult [Sun 2013]. These disadvantages make them undesirable for large-scale real-time implementations. On the other hand, deterministic approximation methods are faster to converge [Bishop 2012, Sun 2013] and the convergence can be monitored [Murphy 2012, Sun 2013]. However, these methods use some assumptions about the posterior, such as factorized distributions, to make the inference fast and tractable. As a result, the approximate posterior might not converge to the true posterior [Bishop 2006, Sun

2013].

Examining the advantages and disadvantages of both of the approximate inference methods indicates that there is a compromise between speed and accuracy. Sampling methods can theoretically converge to the exact posterior distribution in the limit of infinite random samples [Bishop 2012] but are computationally demanding and slow. Deterministic methods, on the other hand, are generally faster, but almost never converge to the true posterior [Sun 2013]. The focus of this thesis is on prediction of microsleeps which requires fast computations. Therefore, deterministic methods have been chosen as the approximate inference method for the Bayesian models in this thesis.

Variational inference [e.g., Tzikas et al. 2008] is one of the deterministic approximation methods and has been widely used in the literature [e.g., Bishop 1999, Huang et al. 2007, Klami et al. 2013, McGrory and Titterington 2009, Wang 2007, Zhao et al. 2015a, 2016]. Variational inference approximates the posterior distribution with a simpler family of distributions [Bishop 2006, Sun 2013]. Laplace approximation is another method of deterministic posterior approximation [Murphy 2012, Sun 2013]. Laplace approximation uses the Taylor series to expand the negative log-posterior and approximate it with a Gaussian distribution [Friston et al. 2007, Murphy 2012]. However, Laplace approximation might not be accurate, especially when the posterior distribution is non-Gaussian [Friston et al. 2007]. Hence, variational inference is selected as the method of interest to approximate posterior probability of Bayesian models in this thesis.

3.2 VARIATIONAL BAYESIAN FRAMEWORK

The aim of variational inference is to approximate the true posterior $p\left(\Theta \,\middle|\, x\right)$ with an approximation distribution $q\left(\Theta\right)$, where Θ is the collection of all model parameters [Murphy 2012]. Kullback-Leibler (KL) divergence is used to quantify the difference between the true and approximate distributions [Murphy 2012],

$$\mathrm{KL}\left(q \parallel p\right) = \int q\left(\Theta\right) \ln\left(\frac{q\left(\Theta\right)}{p\left(\Theta \,\middle|\, x\right)}\right) d\Theta. \tag{3.5}$$

Minimizing Equation (3.5) directly is not straightforward. However, Equation (3.5) can be rewritten in terms of the log-evidence of the model [Bishop 2006],

$$\ln\left(p\left(x\right)\right) = \mathcal{L}\left(q\right) + \mathrm{KL}\left(q \parallel p\right), \tag{3.6}$$

where $\mathcal{L}\left(q\right)$ is the lower bound of the log-likelihood of data given by

$$\begin{aligned}\mathcal{L}\left(q\right) &= \int q\left(\Theta\right)\ln\left(\frac{p\left(\Theta,x\right)}{q\left(\Theta\right)}\right)d\Theta \\ &= \int q\left(\Theta\right)\ln\left(p\left(\Theta,x\right)\right)d\Theta - \int q\left(\Theta\right)\ln\left(q\left(\Theta\right)\right)d\Theta \\ &= \left\langle \ln\left(p\left(\Theta,x\right)\right)\right\rangle_{q(\Theta)} - \left\langle \ln\left(q\left(\Theta\right)\right)\right\rangle_{q(\Theta)}. \qquad (3.7)\end{aligned}$$

Maximizing Equation (3.7) is computationally less demanding than minimizing Equation (3.5) but these are equivalent. Coordinate ascent is the most widely-used algorithm for maximizing the lower-bound of the log-likelihood of data in variational inference [Wu et al. 2016]. Additionally, since all the expectations are taken with respect to the variational distributions, we drop the variational distribution notation from the expectation operators.

Variational inference can be extended to models with latent/hidden variables - i.e., variables which can't be measured directly but can be inferred from the observed data [Barber 2012, Bishop 2006, Murphy 2012]. This essentially is achieved by applying variational Bayesian expectation maximization (EM) to update parameters Θ and latent variables $\mathbf{Z}$. The variational Bayesian expectation (VBE) step updates the posterior distribution of latent variables $\mathbf{Z}$, assuming that the distribution of all parameters Θ are fixed. The variational Bayesian maximization (VBM) step updates the posterior distribution of the parameters Θ when the distribution of the latent variables $\mathbf{Z}$ are fixed. Iterating over the VBE and VBM steps increases the lower-bound of the log-evidence $\mathcal{L}$ which can be used to monitor the convergence of the variational inference. Variational inference has converged when the relative improvement of the $\mathcal{L}$ falls below a predefined threshold (e.g., 10^{-6}). Refer to Barber [2012], Beal [2003], Bishop [2006], Murphy [2012] for a comprehensive overview of variational inference.

3.3 RELATED BAYESIAN MACHINE LEARNING ALGORITHMS

Machine learning methods find the parameters of a flexible model in order to explain or fit the data. The term *learning* refers to optimizing the model parameters to minimize a cost function or maximize a utility function [Murphy 2012]. As mentioned in Section 3.1, however, uncertainty is inevitable due to finite and noisy data. Bayesian machine learning refers to building flexible models using the Bayesian probabilistic framework which has the advantage of explicitly modelling noise and uncertainty.

Bayesian methods are generally divided into two categories: *parametric* and *nonparametric*. Parametric Bayesian is a class of methods where the inference is limited to models with a finite set of parameters [Ghahramani 2012, Müller and Mitra 2013]. One of the challenges of Bayesian methods is to find a model flexible enough to capture all of the characteristics of the data [Ghahramani 2015]. This issue can be alleviated by using nonparametric methods which are highly flexible models. Counter-intuitively, nonparametric methods have an infinite number of parameters and hence their complexity tends to grow with more data [Ghahramani 2015, 2012]. An example of parametric methods is a linear classification method that finds a linear boundary to discriminate between classes, irrespective of the amount of training data. On the contrary, a nonparametric classification method can learn a nonlinear boundary which can become more complex with more training data [Ghahramani 2015].

Bayesian methods, both parametric and nonparametric, have been extensively used in the literature [e.g., Badillo et al. 2014, Kang and Choi 2014, Kim and Ghahramani 2006, Klami et al. 2013, Mukuta and Harada 2014, Wang 2007, Wu et al. 2015, 2011, Xu et al. 2009]. Although nonparametric Bayesian methods have a high degree of freedom, inferring the probabilities of interest can be computationally very expensive, especially when the number of data points increases [Ghahramani 2012, Hensman et al. 2015, Rasmussen 2004, Rasmussen and Williams 2006]. Due to the high computational demands of nonparametric Bayesian methods, this thesis is limited to parametric Bayesian methods which are less computationally demanding.

3.3.1 Bayesian principal component analysis

PCA is a generative model that finds a set of orthogonal principal components, usually lower-dimensional, to explain the observed data [Bishop 2006]. PCA has been widely used in the literature for feature reduction [Avendaño-Valencia et al. 2010, Chai et al. 2016, Yu et al. 2014], data compression [Ding et al. 2016, Sun et al. 2005, Wang et al. 2004], and data visualization [Bishop 2006, Jenssen 2013]. PCA uses a linear model to transform the original data into *principal components* but the number of components is not known a priori. Cross-validation [Hastie et al. 2009] is usually used to select the optimum number of principal components.

Tipping and Bishop [1999] proposed a probabilistic representation of PCA. They used EM to optimize parameters of the probabilistic PCA. However, probabilistic PCA does not automatically find the number of components and also is prone to overfitting. To overcome these issues, Bishop [1999] developed a Bayesian PCA by introducing prior probabilities over parameters of the probabilistic PCA, as shown in Figure 3.1. Assuming that N independent and identically distributed (i.i.d.) data have been observed, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and that each observation has D dimensions, Bayesian PCA is defined as [Bishop 1999]

$$p(\mathbf{Z}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n \mid \mathbf{0}, \mathbf{I}), \tag{3.8}$$

$$p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \mathbf{W}, \tau) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n \mid \mathbf{W}\mathbf{z}_n + \boldsymbol{\mu}, \tau^{-1}\mathbf{I}\right), \tag{3.9}$$

$$p(\mathbf{W} \mid \alpha) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{w}_k \mid \mathbf{0}, \alpha_k \mathbf{I}), \tag{3.10}$$

$$p(\boldsymbol{\mu}) = \mathcal{N}\left(\boldsymbol{\mu} \mid \mathbf{0}, \beta_{\mu}^{-1}\mathbf{I}\right), \tag{3.11}$$

$$p\left(\boldsymbol{\alpha}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,\middle|\, a_\alpha, b_\alpha\right), \tag{3.12}$$

$$p\left(\tau\right) = \mathcal{G}\left(\tau \,\middle|\, a_\tau, b_\tau\right), \tag{3.13}$$

where $K < D$ is the number of principal components (dimension of the latent space), $\mathbf{z}$ is a $K \times N$ matrix of latent variables, $\boldsymbol{\mu}$ is a D dimensional vector of the mean values, τ is the noise precision, $\mathbf{W}$ is a $D \times K$ loading matrix, $\boldsymbol{\alpha}$ is a K dimensional vector of hyperparameters over columns of $\mathbf{W}$, $\mathcal{N}$ is a normal distribution[1], and $\mathcal{G}$ is a Gamma distribution[2]. Hyperparameters $\boldsymbol{\alpha}$ control the complexity of their corresponding columns in the loading matrix. This type of prior distribution has been motivated by automatic relevance determination (ARD) [Bishop 1999, Mackay 1995]. Therefore, if the posterior probability of α_k is concentrated on larger values, the k^{th} column of the loading matrix $\mathbf{w}_k$ has a larger set of values of precision which results in smaller values of $\mathbf{w}_k$. Essentially, a column of the loading matrix is switched-off when the posterior of its corresponding hyperparameter is large enough.

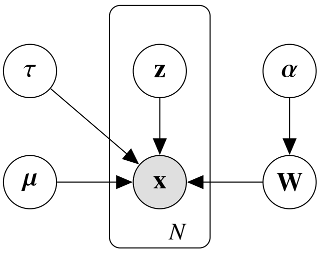


Figure 3.1 Graphical representation of Bayesian PCA adapted from [Bishop 1999].

Bishop [1999] used variational inference to find parameters of the Bayesian PCA model. Using randomly-generated data from a Gaussian distribution, he demonstrated that Bayesian PCA with ARD-motivated prior distributions on the loading matrix can automatically identify the optimum number of components.

Zhao and Jiang [2006] extended probabilistic PCA to use the Student-t distribution for the latent variables. They showed that using a Student-t distribution for the latent variables improves the robustness of probabilistic PCA. However, they did not use a Bayesian model. EM was used to maximize the log-likelihood of the evidence, but selecting the optimum number of components was done manually.

3.3.2 Bayesian factor analysis

Similar to PCA, factor analysis (FA) is a linear model which transforms data into a latent space [Bishop 2006]. The difference between FA and probabilistic PCA lies in the noise term. Probabilistic PCA assumes that the noise term has an isotropic covariance matrix, whereas FA

[1] $\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma\right) = |2\pi\Sigma|^{-1/2} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^{\top} \Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right)$

[2] $\mathcal{G}\left(x \,\middle|\, a, b\right) = \Gamma\left(a\right)^{-1} b^{a} x^{a-1} \exp\left(-bx\right)$

uses a diagonal covariance matrix for the noise term [Bishop 2006]. Ghahramani and Beal [2000] developed a Bayesian FA model employing ARD-motivated prior distributions over the loading matrix to automatically find the optimum number of components, as shown in Figure 3.2. The probabilities of the Bayesian FA model are similar to that of the Bayesian PCA with an exception of the noise term and likelihood of data, which are given by

$$p\left(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \mathbf{W}, \boldsymbol{\Psi}\right) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}_n \mid \mathbf{W}\mathbf{z}_n + \boldsymbol{\mu}, \boldsymbol{\Psi}^{-1}\right), \tag{3.14}$$

$$p\left(\boldsymbol{\Psi}\right) = \prod_{d=1}^{D} \mathcal{G}\left(\psi_d \mid a_{\psi}, b_{\psi}\right). \tag{3.15}$$

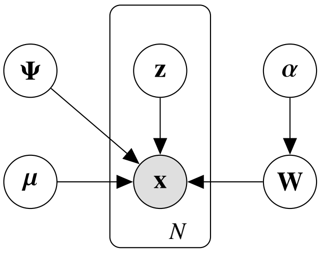


Figure 3.2 Graphical representation of Bayesian FA adapted from [Ghahramani and Beal 2000].

Ghahramani and Beal [2000] used variational inference to update the model parameters of Bayesian FA. The rotation problem is a disadvantage of the FA model where the loading matrix can not be identified uniquely [Bishop 2006, Zhao and Yu 2009]. Zhao and Yu [2009] imposed restrictions on the loading matrix to identify it uniquely. Notwithstanding, although the rotation problem affects the interpretation of latent variables, it does not change the form of the latent space [Bishop 2006]. Therefore, the rotation problem does not change the performance of classification using latent variables.

3.4 BAYESIAN METHODS IN BRAIN SIGNAL ANALYSIS AND IMAGING

Bayesian methods have been widely used in the brain-imaging literature [Alpert and Yuan 2009, Croce et al. 2016, de Rochefort et al. 2010, Hinne et al. 2013, Lucka et al. 2012, Wu et al. 2016, Zhang et al. 2015]. Since the focus of this thesis is on the EEG, we specifically concentrate on Bayesian methods in the EEG literature.

Bayesian methods in EEG source reconstruction have been extensively studied in the literature [Baillet and Garnero 1997, Belardinelli et al. 2012, Cortes et al. 2012, Costa et al. 2015, Daunizeau et al. 2006, Kiebel et al. 2008, López et al. 2014, Lucka et al. 2012, Phillips et al. 2005, Stahlhut et al. 2011, Trujillo-Barreto et al. 2008, Wipf and Nagarajan 2009, Zumer et al. 2007]. Most of these works used an FA model to find independent/uncorrelated sources of brain activity from the EEG. Zumer et al. [2007] developed a probabilistic approach to estimate source activity using knowledge of event timing and independence from noise and interference (SAKETINI). This model assumes that EEG trials are independent, evoked sources exist only after the stimulus has been presented, and background EEG activity and noise exist before and after the presentation of the stimulus. They used a variational Bayesian FA (VBFA) model on the pre-stimulus data to learn the background activity and noise. The sources were then learnt from the post-stimulus data using a variational method that assumes the estimated noise and background activity from the pre-stimulus data are fixed throughout the post-stimulus data. Stahlhut et al. [2011] used VBFA to simultaneously reconstruct EEG sources and a forward model. They used an ARD-motivated prior distribution to automatically identify the optimum number of EEG sources. These studies highlight the advantage of Bayesian methods for uncertain and noisy data, such as EEG data. EEG source reconstruction is an ill-defined problem where the potential sources severely outnumber the EEG channels [López et al. 2014, Lucka et al. 2012] and thus classical models are prone to overfit. Bayesian methods, on the other hand, average over all values of a parameter and hence are not prone to overfitting [Ghahramani 2015].

Wu et al. [2011] developed a hierarchical Bayesian model to extract spatio-temporal patterns of the EEG. Their model comprised two FA models with a common loading matrix where each FA model explained the data of one condition. The posterior probabilities were approximated using a variational inference. In a later study, Wu et al. [2015] expanded their model to probabilistic common spatial patterns. They provided both variational and MAP inferences. MAP inference was faster and less computationally demanding but was prone to overfitting. Variational inference, on the other hand, was computationally more expensive but automatically found the optimum number of components. Although these models have shown good performances for single-trial motor imagery data, it is unknown whether they will be suitable for the microsleep datasets, as responsive labels correspond to awake moments but the individual could be engaged in any mental task. However, the mental tasks are well defined in a motor imagery task.

Wu et al. [2014] incorporated a hierarchical Bayesian FA to estimate event-related potentials (ERPs). They developed a variational inference algorithm to infer the posterior probabilities of their Bayesian model. They were able to extract spatio-temporal information of ERPs while the optimum number of source components was automatically found with an ARD-motivated prior distribution.

Ko et al. [2009] used a Bayesian network for emotion recognition from the EEG. They extracted relative power of the theta, alpha, beta, and gamma bands from EEG as features. They used a generative Bayesian network classifier to distinguish between emotions. It is worth mentioning that Bayesian networks are directed acyclic graphs which can represent uncertain data with prior information regarding conditional independences among variables [Jebara 2003]. In another study, Yoon and Chung [2013] used probability distributions to generatively model power spectral features. They used a logistic regression with log-posterior probabilities to classify emotions.

It is evident that Bayesian methods have been used for different purposes in the EEG brain-imaging literature. Bayesian methods are capable of dealing with noise and uncertainty in the data which makes them an elegant choice for EEG processing. Moreover, the high-dimensionality of the EEG and the presence of correlations between EEG-electrodes increases the chance of overfitting for classical approaches, whereas Bayesian methods are less prone to overfitting problem. Therefore, using Bayesian methods for prediction/detection of microsleeps could lead to improvement in performance.

Chapter 4

AIMS AND HYPOTHESES

4.1 AIMS

The aim of this research was to predict behavioural microsleeps, both in terms of states and events, using EEG signals acquired from the scalp and Bayesian methods for data analysis.

4.2 HYPOTHESES

This research posed three questions and hypotheses.

4.2.1 Hypothesis 1 - Bayesian robust feature reduction

- **Question:** Various sources of noise and uncertainty exist in scalp EEG signals, which deteriorate the performance of microsleep detection/prediction systems. This raises the question: can microsleep detection/prediction performance be improved by the incorporation of a Bayesian feature reduction model which explicitly models noise and uncertainty?

- **Hypothesis:** Explicit modelling of noise and uncertainty will improve performance of the microsleep detection/prediction system.

- **Rationale:** Scalp EEG signals have been studied widely in the literature. One of the challenges of the EEG is its susceptibility to noise and artefacts. Despite immense research into removal of artefacts from the EEG, it is near impossible to realize a complete noise-free scalp EEG recording. As a result, extracted features from the EEG are to some extent noisy with an unknown noise intensity. Additionally, the high-dimensional nature of EEG can potentially degrade the performance of a brain-state classifier [Lemm et al. 2011]. Bayesian feature reduction methods, however, are capable of explicitly modelling noise and uncertainty while finding a lower-dimension representation of data [Zhao and Yu 2009]. Moreover, Bayesian methods can automatically infer the optimum number of lower-dimension components needed from the data and avoid overfitting [Nakajima et al. 2013, Wang 2007, Zhao and Yu 2009].

- **Significance:** Microsleeps are one of the major causes of road accidents. Prediction of an imminent microsleep and even detection of an ongoing microsleep with high accuracy could ultimately save lives.

4.2.2 Hypothesis 2 - Bayesian multi-subject robust feature reduction

- **Question:** Inter- and even intra-subject variabilities pose a challenge for finding a classification model that can generalize to new unseen subjects. Does incorporating a Bayesian method that finds a common latent-space among subjects improve the performance of microsleep detection/prediction?

- **Hypothesis:** Using a Bayesian model to find a common latent-space among all subjects will improve the performance of a microsleep detection/prediction system.

- **Rationale:** Inter-subject and intra-subject variabilities of the EEG have been shown to degrade the performance of EEG-based classification methods [Blankertz et al. 2007, Jatupaiboon et al. 2013, Matiko et al. 2015]. A short calibration segment of data from an individual subject is usually used to adapt a classifier to a specific subject. However, it is impractical to collect calibration data and retrain the classifier before every driving session. Bayesian methods, on the other hand, can incrementally update the parameters with new observed data [Chien and Chen 2009, Murphy 2012, Yu and Gales 2007]. Therefore, a Bayesian model can find the shared parameters among subjects at the time of training, then adapt to a new subject's data at the time of testing. Incorporating a Bayesian model to adapt to a new subject's space should improve the performance of an EEG-based microsleep detection/prediction system.

- **Significance:** Improving performance of the microsleep detection/prediction system can prevent catastrophic sleep-related accidents. A Bayesian multi-subject model can adapt to an individual's data without needing a calibration session, which makes the microsleep detection/prediction system more practical.

4.2.3 Hypothesis 3 - Imminent Microsleep prediction

- **Question:** Is it possible to predict imminent microsleep episodes using scalp EEG signals?

- **Hypothesis:** There are specific changes in the EEG before microsleeps which can be identified in real-time and used to predict imminent microsleeps.

- **Rationale:** EEG has been used to detect drowsiness [Akin et al. 2008, Chen et al. 2015, Kurt et al. 2009, Yeo et al. 2009], detect stage of sleep [Chapotot and Becq 2010, Güneş et al. 2010], and predict epileptic seizures [Kanemura et al. 2012, Schad et al. 2008]. EEG has also been used for prognosis of patients with cardiac arrest to detect subclinical seizures [Westhall et al. 2014]. It is evident that the EEG embodies a large amount of information. Therefore, EEG might contain specific patterns related to an imminent microsleep. Discovering and using these patterns might result in an EEG-based system capable of predicting imminent episodes of microsleep.

- **Significance:** Accurate prediction of imminent microsleeps can improve transportation safety. Warning feedback can be provided to individuals to prevent microsleeps.

Chapter 5

EEG PREPROCESSING AND GOLD-STANDARD REFINEMENT

5.1 INTRODUCTION

This thesis investigates detection and prediction of microsleep from the EEG. The data of a previous study described in Section 2.5, 'Study A', was used to evaluate our proposed methods. Peiris et al. [2006] found that only 8 out of 15 subjects had at least one definite microsleep during two 1-h sessions of CTT. Data from these 8 subjects were used to evaluate the proposed methods in this project.

An EEG-based microsleep prediction system uses features of EEG to predict an imminent microsleep. However, an EEG segment might have artefacts, such as muscle and eye-movement artefacts, and hence preprocessing is necessary. In addition, a behavioural gold-standard containing microsleep episodes is required to train and evaluate the prediction system. In this chapter, the preprocessing of EEG is presented in Section 5.2. It is then followed by the steps undertaken to refine the gold-standard including minimizing temporal displacement in the tracking task, analysing tracking performance, redefining microsleeps, and introducing the '*uncertain*' category.

5.2 EEG PREPROCESSING

This section provides the steps for preprocessing EEG data of Study A (Section 2.5). These steps were undertaken to remove various artefacts from the raw EEG and prepare it for feature extraction. A Hampel filter [Liu et al. 2004, Pearson 2002] with a window length of 15 EEG samples, i.e., 7 samples on each side, was used to identify the outliers exceeding 10 standard deviations, calculated using mean absolute deviation. These were then replaced with the local median (Figure 5.1a). A large threshold was intentionally chosen to ensure that only highly-deviated data points would be replaced. The data was then re-referenced to the common average [Tong and Thakor 2009] of all EEG derivations. Common average reference is a spatial filter in which the average activity of all EEG channels is deducted from individual electrodes [Yu et al. 2014]. Following this, the data was filtered with a zero-phase finite impulse response bandpass filter with cut-off frequencies of 0.5 to 45 Hz using the *firfilt*[1] package of EEGLAB[2] [Delorme and Makeig 2004].

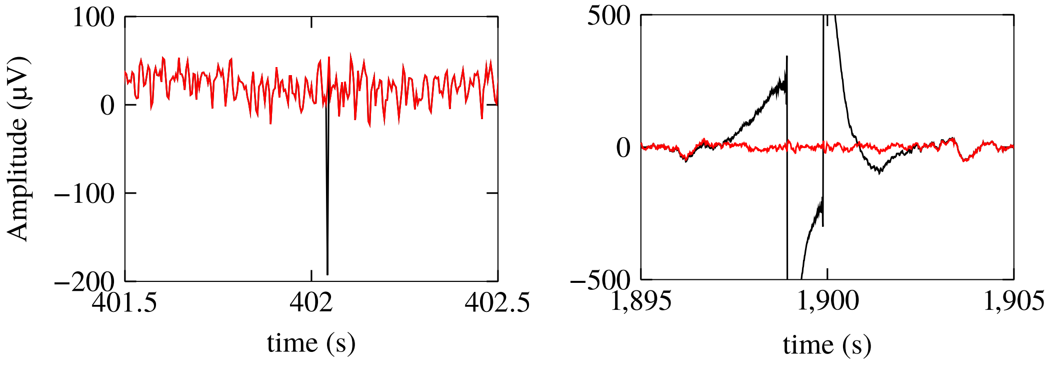
In this project, ASR [Mullen et al. 2013, 2015] was used to minimize artefacts of the EEG. ASR was chosen since Mullen et al. [2015] showed that ASR can be applied in real-time and it does not distort EEG, while other techniques such as ICA are computationally expensive [Albera et al. 2012]. ASR requires a clean "calibration" data to use as a base for removing artefacts from the rest of the data. In this project, the calibration data was extracted using a threshold on z-score of EEG data. To clean a noisy segment of data, a PCA was applied to the noisy segment and its principal components were extracted. Using the covariance matrix of the calibration data, the extracted components of the noisy data were projected to the calibration data's space. A threshold derived from the calibration data was then applied to remove the components which resulted in high-amplitude artefacts. The remaining components were then back-projected into channel space.

A visual inspection of the EEG was performed to manually identify and mark high amplitude artefacts, e.g., the electrode pop artefact. ASR was applied to a 2-min window around each of the high amplitude marked artefacts to reconstruct the background EEG with respect to the surrounding data. ASR uses a z-score threshold to find clean data for calibration. To limit reconstruction to highly deviated regions, a z-score threshold of 10 was found appropriate for all subjects and sessions, which recognized low amplitude EEG as calibration data. Figure 5.1b illustrates an EEG channel contaminated with two consecutive electrode pop artefacts and its reconstruction using ASR. The duration of reconstructed large artefacts for each subject and session is given in Table 5.1. It can be seen that the data of some subjects are quite noisy, which might have been due to (1) the set-up of the EEG electrodes, or (2) the subject's movement to stay awake. On average, 68.2 s (0.0–388.7 s) of the EEG data of each 1-h session was reconstructed from large artefacts. The large artefacts were reconstructed instead of pruned (as was done in Peiris et al. [2011]) so as to be able to use all the microsleep gold-standards. This process was done on the basis of EEG artefacts and was completely blind to the gold-standard.

Next, ASR was applied to a sliding window of 4 min with a step of 2 min, i.e., 50% overlap between successive epochs. The z-score threshold of ASR was set to 5 and the common data between two consecutive windows were averaged. This minimized most of the artefacts in the EEG, such as eye-blinks, jaw clenching, and movement artefacts. Canonical correlation analysis blind source separation [Clercq et al. 2006] was the last step to minimize the remaining muscle artefacts. Figure 5.2 shows a 5-s EEG epoch contaminated with eye-blink and muscle artefacts before and after artefact removal.

[1] http://home.uni-leipzig.de/biocog/content/widmann/eeglab-plugins

[2] https://sccn.ucsd.edu/eeglab/index.php



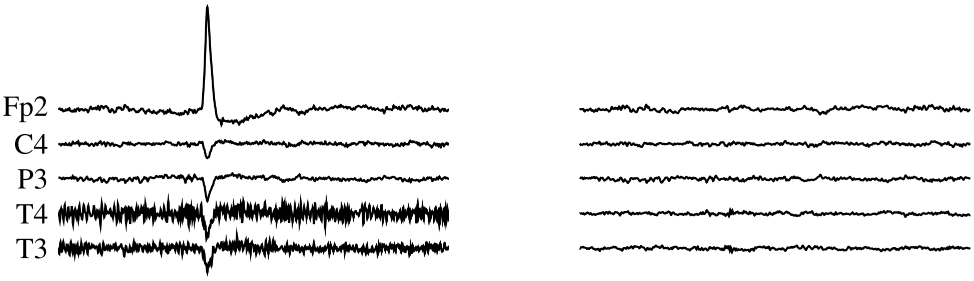
(a) Outlier removal using Hampel filter.

(b) Electrode pop artefact removal using ASR.

Figure 5.1 An example of the first two stages of EEG artefact removal: (a) using a Hampel filter to identify a single point outlier (black) and approximating it with the local median (red), and (b) an EEG channel contaminated with electrode pop artefacts (black) and its reconstruction (red) using ASR of the 2 min surrounding data.

Table 5.1 Duration of the reconstructed EEG for visually identified large artefacts.

	Duration (s)	
Subject	Session 1	Session 2
804	0.0	79.0
809	8.0	61.0
810	19.0	261.0
811	19.0	68.7
814	35.8	3.5
817	7.0	13.6
819	0.0	4.4
820	388.7	122.6



(a) Contaminated with artefacts

(b) Artefacts removed

Figure 5.2 An illustration of a 5-s segment of EEG, (a) contaminated with eye-blink and muscle artefacts, and (b) after artefact removal.

5.3 GOLD-STANDARD REFINEMENT

As mentioned in Section 2.5.2, the original gold-standard of Study A comprised two independent measures of (1) tracking performance analysis and (2) behavioural video ratings. The two measures were then combined using logical operators. Depending upon the logical operator used, two gold-standards were generated. A *Lapse index* was defined as either a video lapse rating *or* a tracking flat-spot, whereas a *definite behavioural microsleep* (BM) was defined as to having both video lapse rating *and* tracking flat-spot [Peiris et al. 2006, 2011]. This process, however, is prone to introducing false positives and/or false negatives into the gold-standard. For instance, a person might exhibit microsleep behavioural cues, such as eye-closure, leading to a video lapse rating from an expert, while the tracking performance is satisfactory. This situation introduces a false positive in the lapse index, while a BM would not be affected. On the other hand, a person might get a deep drowsy video rating from an expert, while they have in fact stopped tracking. In this case, the lapse index will remain unaffected, but the BM would suggest that the person is responsive, which would result in a false negative. To overcome these shortcomings and to improve the accuracy of the gold-standard, analysis of tracking performance was refined, microsleep criteria were redefined, and an *Uncertain* category was added to the gold-standard.

The CTT had an 8-s preview of the forthcoming target. The importance of the preview is two fold: (1) participants could anticipate the target and keep tracking with their eyes closed, and (2) a small lead or lag might have been introduced between the target and tracking, despite the instruction to track the target closely. In the first scenario, an expert might have assigned a video lapse rating due to prolonged eye closure while being unaware of a subject's satisfactory tracking performance. Continuous satisfactory tracking performance for a certain period of time indicates that the participant was responsive. In the second scenario, there might exist tracking errors that are solely due to lead/lag between the tracking and target. Therefore, the subject might have been tracking the target coherently with a slight delay but it might not be identified as satisfactory tracking due to the conservative tracking performance analysis. Notwithstanding, the existence of lead/lag between the tracking response and target does not affect identification of the microsleeps, but it affects the identification of the satisfactory tracking.

To minimize the effects of the preview of the CTT on the tracking performance analysis, the following two steps were performed to improve the synchrony of the target and tracking response. First, a cross-correlation between the 1-h tracking response and target was computed to find any global temporal displacement and to allow shifting the target up to 2 s to match the tracking response. Table 5.2 represents the global temporal displacement between the target and tracking response for each individual subject, with a positive value denoting a lead in the tracking task, i.e., a subject was tracking the future target. This shows that all the subjects were, on average, 0.15 s (0.03–0.48 s) ahead of the target. The target of each session was compensated accordingly.

The second step was to minimize local leads or lags. This was done by calculating the

Table 5.2 The consistent time shift between the tracking response and target throughout a whole session. A positive value indicates that the subject's performance was ahead of the target.

	Time shift (s)	
Subject	Session 1	Session 2
804	0.11	0.07
809	0.30	0.06
810	0.48	0.16
811	0.03	0.04
814	0.07	0.13
817	0.09	0.20
819	0.14	0.18
820	0.13	0.18
Average	0.15	

cross-correlation between the tracking response and the target of a sliding window of 10.5 s consisting of a 0.5 s region of interest and a 10-s adjacent window, i.e., 5 s on each side. Temporal adjustments up to 1 s were made to the target corresponding to the region of interest. Figure 5.3 depicts the target, tracking response, and adjusted target of a 10-s segment of the CTT. In this case, the subject's response was leading the target from 1305 to 1310 s, and hence the target's trace was modified to reduce the temporal displacement. A trend was observed that, on average across all 8 subjects, the response slightly led the target at the beginning of the study, lagged behind the target towards the middle of the session, then returned to zero lead/lag of the target towards the end. The average temporal compensations of all subjects over the time of the CTT is shown in Figure 5.4.

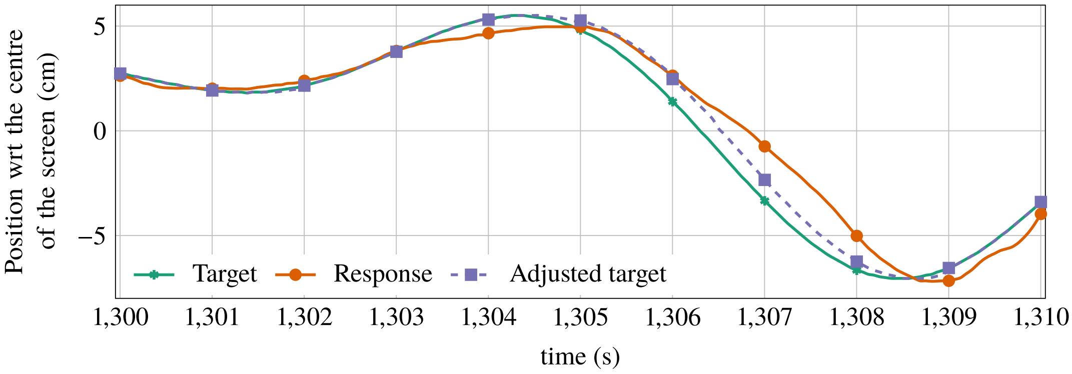


Figure 5.3 Adjusting the target to account for local temporal displacement in a subject's tracking performance.

After removing the potential temporal inconsistencies between the tracking response and the target, analysis of the tracking performance was performed by finding three tracking trajectories, namely *responsives*, *erroneous regions*, and *flat-spots*. At first, segments of the target with a

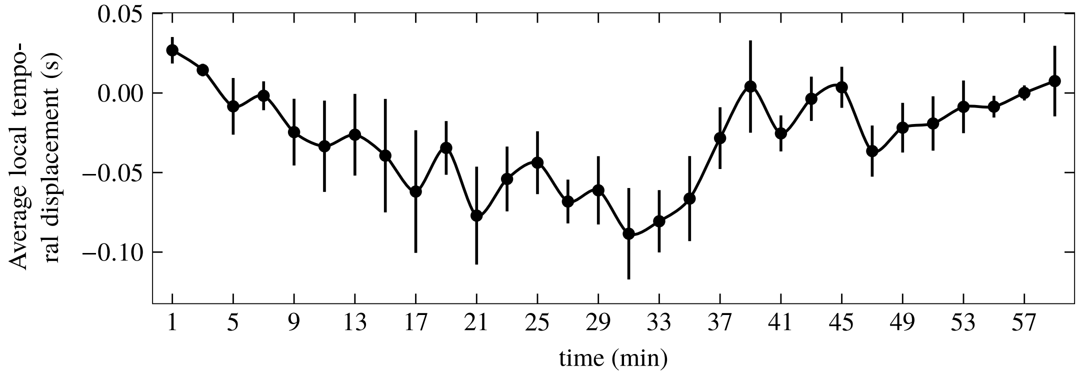


Figure 5.4 The local temporal displacement (mean ± SE) of all subjects over time.

slow velocity were temporarily removed from the analysis. Figure 5.5 depicts a cycle of the target of the CTT and its velocity, in which the velocity drops to zero 34 times in each cycle. Temporarily excluding slow velocity targets was necessary as it was inaccurate to estimate a subject's responsiveness when the target was moving slowly. The threshold of slow velocity was chosen as the 10$^{\text{th}}$ percentile of the peak velocity of the target, i.e., 2.6 mm/s. To process the remainder of the data, a sliding window of 1 s was used to find the fractions of the tracking response with a low mean absolute error with respect to the target. Since the velocity of the target was varying, a variable threshold was used for each window. To this end, 70% of the mean target velocity of each window was used as the threshold, and the windows with a lower mean absolute error than their respective thresholds were added to the responsives. Afterwards, the slow velocity episodes with a total duration of less than 2 s were added to the responsives if (1) the mean absolute tracking error was less than 11.6 mm, and (2) a window of 1.5 s on each side was considered responsive. Similarly, slow velocity episodes of up to 4 s were added to the responsives if the mean absolute tracking error was less than 15.4 mm and there was a 4-s responsive window on each side. Finally, responsive episodes with a duration of over 5 s were retained and the rest were pruned out.

A sliding window of 0.2 s was used to identify and examine both flat-spots and erroneous regions. Data corresponding to the windows with a mean absolute tracking error of over 30.8 mm were marked erroneous. However, the process of forming flat-spots imposed more conditions to find the portions of data with a slow response and a high tracking error. First, the velocity of the tracking response was processed. Slow tracking windows were defined as windows with a mean tracking velocity of less than (1) 10% of the mean velocity of the corresponding target or (2) 1.2 mm/s. Second, segments of the slow tracking data with a duration of at least 6 s were added to flat-spots without further processing. Lastly, slow tracking windows with a mean absolute tracking error over 40% of the mean absolute of the corresponding target displacement or 15.4 mm were also added to the flat-spots.

After marking both flat-spots and erroneous regions, a similar post-processing was applied. First, episodes shorter than 1 s were pruned out to remove potential false positives. Second, towards the end of each episode, the tracking error is decreasing for several samples, i.e., the

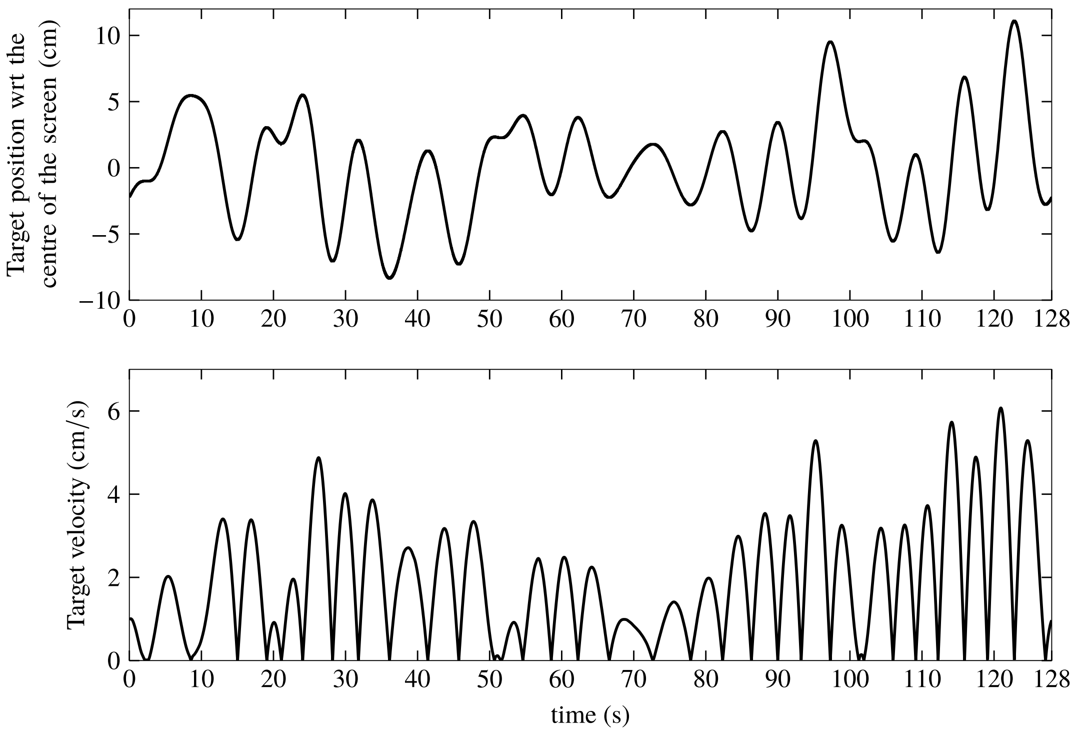


Figure 5.5 A cycle of the target of the CTT (top) and its corresponding velocity (bottom).

derivative of the tracking error is negative. These samples were also pruned out. This was necessary as it was possible for participants to observe the preview of the following 8 s of the target and anticipate it. Therefore, after encountering a flat-spot or erroneous region, a subject could have waited for the target to get closer to the tracking cursor before resuming tracking, as shown in Figure 5.6.

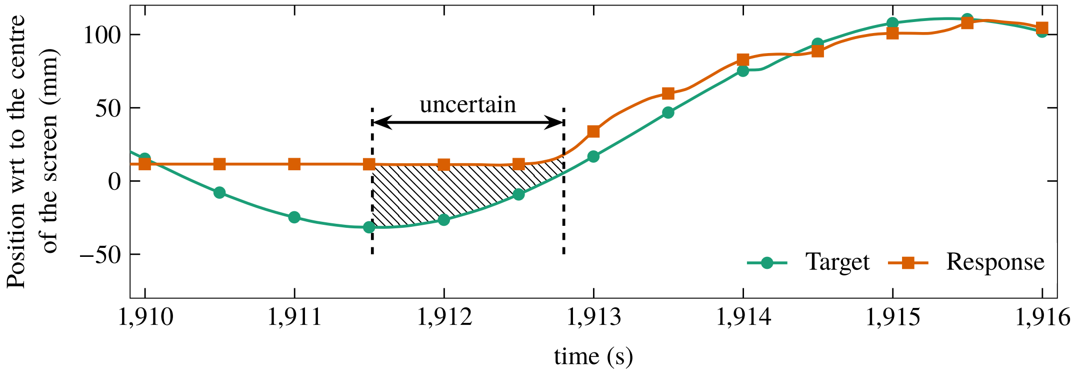
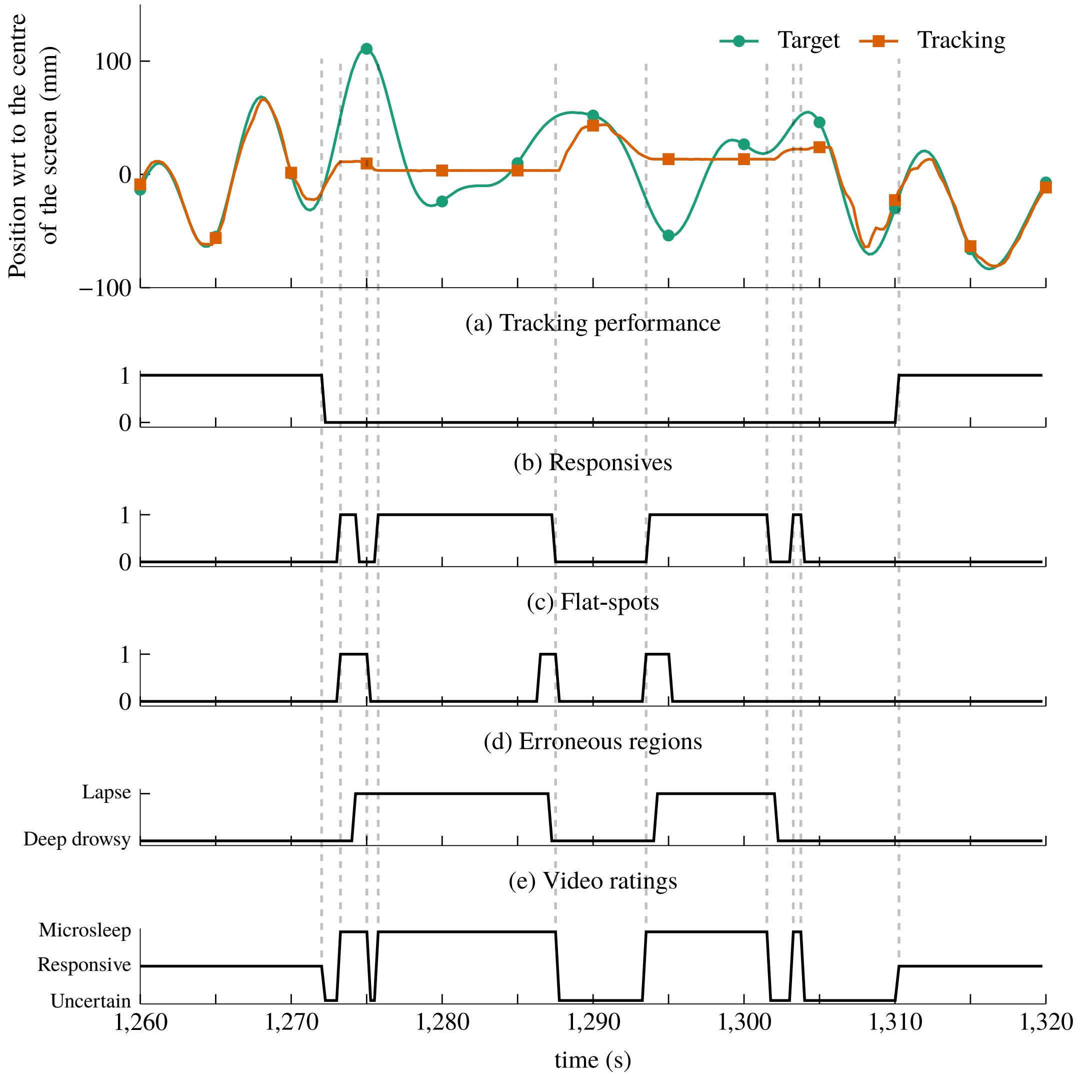


Figure 5.6 An uncertain region with a flat response and a decreasing tracking error which was followed by a satisfactory response.

To form a gold-standard, the tracking analysis was fused with the video ratings. Both tracking analysis and video ratings were decimated to 4 Hz to match the frequencies of the two measures. This leads to a temporal resolution of 0.25 s for the gold-standard. A responsive label was defined as satisfactory tracking performance, i.e, responsive regions of tracking, irrespective of the video ratings. A microsleep, however, was defined as an episode of *non-tracking*, i.e., the union of flat-spots and erroneous regions, in conjunction with a video-rating of deep drowsy or lapse. The video ratings were included in identification of microsleeps to avoid false positives (FPs). The remainder of the gold-standard was labelled *Uncertain* due to the lack of information, from our analysis, to accurately identify the state of responsiveness. Fusion of the tracking performance analysis and the video ratings to form the refined gold-standard is shown in Figure 5.7.



(a) Tracking performance

(b) Responsives

(c) Flat-spots

(d) Erroneous regions

(e) Video ratings

(f) Gold standards

Figure 5.7 Fusion of the tracking performance analysis and video ratings to obtain the refined gold-standard.

Refinement of the gold-standard can be summarized as:

1. Minimize the local and global temporal displacements between the tracking performance and target by slightly adjusting the target.

2. Find segments with at least 5 s of satisfactory tracking performance and label each as reponsive.

3. Find non-tracking episodes of tracking performance, i.e., flat-spots and erroneous regions.

4. Generate the gold-standard by integrating the analysis of tracking performance and video ratings by:

    - Directly adding the tracking responsives to the gold-standard.
    - Defining microsleeps as the conjunction of drowsy/lapse video ratings *and* non-tracking episodes.
    - Labelling the remainder of gold-standard as uncertains.

5.4 DESCRIPTIVE STATISTICS OF THE REFINED GOLD-STANDARD

It was observed that the duration of the non-tracking episodes increased with time-on-task until the middle of the session but the tracking performance improved in the second half of the session. This follows the video-rating patterns, in which the subjects were also drowsier in the middle of the sessions, and the tracking error patterns. Figure 5.8 depicts the average tracking error, duration of the non-tracking regions, and duration of the responsives and microsleeps across all subjects and sessions over time, in which each data point corresponds to a 128-s segment, i.e., a cycle of the pseudo-random target, and all measures are presented as an average over each segment.

Based on the refined gold-standard, the average number of microsleep events over all CTT 1-h sessions was 15.4/h (0–44/h). Due to the uncertain parts of the gold-standard, any microsleep preceded by a responsive episode was counted as an event, while uncertain labels were nulled out. There were uncertain episodes both preceded and followed by microsleeps or responsives and hence the number of microsleep events reported might be less than the actual number. Over both 1-h sessions of CTT, four subjects had more than 30 microsleep events and one had 60 microsleeps. Based on all subjects, the total duration of microsleeps in each 1-h session was 129.5 s (0–742.3 s). Moreover, the total duration of non-trackings in a 1-hour session was on average 143.0 s (0.0–742.9 s), indicating that although the duration of microsleeps were similar to of the non-trackings, the microsleeps were shorter due to consideration of video ratings. Table 5.3 presents the number of microsleep events, as well as total duration of responsives and microsleeps for individual subjects. Two participants – 810 and 811 – had no microsleeps in the first session. Additionally, subject 814 had only one microsleep in the first session. It is evident that the distribution of microsleeps was highly variable across the subjects. In addition, subject 820 had a poor tracking performance and thus it was difficult to identify microsleeps and responsives accurately. This led to a high percentage of uncertains in the gold-standard.

Interestingly, the total duration of microsleeps relative to the number of events was quite high. For instance, the total duration of microsleeps for the second session of subject 804 was

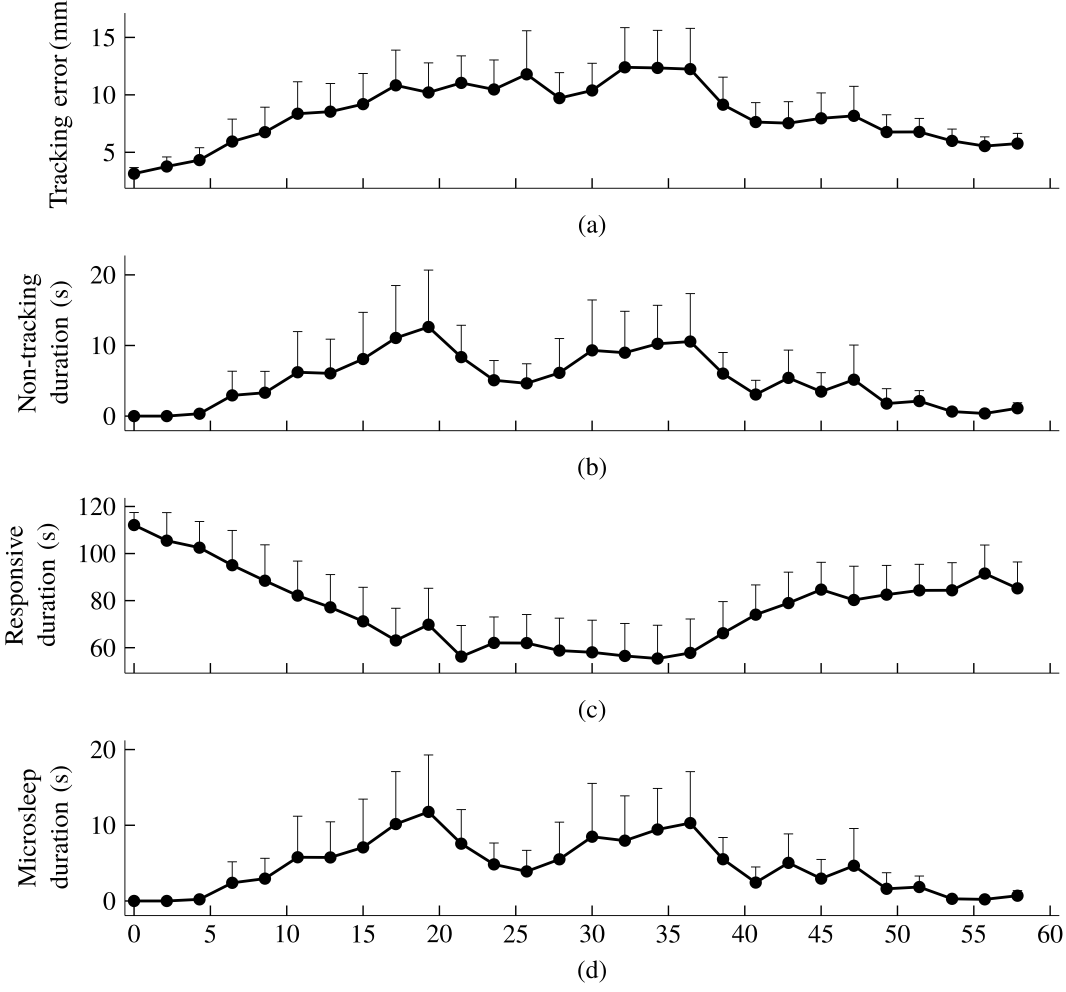


Figure 5.8 Tracking performance and video ratings across all subjects and both sessions (mean ± SE): (a) duration of responsive regions, (b) duration of non-tracking episodes, (c) average video ratings. Each data point corresponds to a cycle of periodic pseudo-random target, i.e., 128 s. Data points are placed to correspond to the start of a cycle.

742.3 s with only 16 microsleeps, leading to an average microsleep event duration of 46.4 s which should be considered as Sleep. However, counting the actual the number of events was impossible due to having uncertain labels. Therefore, using the CTT of Study A, it was infeasible to find the actual duration of individual microsleeps, i.e., an accurate estimation of start and end times of microsleeps.

5.5 COMPARISON OF REFINED AND ORIGINAL GOLD-STANDARDS

Two gold-standards were originally generated to identify lapses and behavioural microsleeps. However, both of those gold-standards were prone to errors. As mentioned in Section 5.3, lapses were defined as either tracking flat-spots, or lapse video-ratings, or both, whereas behavioural microsleeps were defined as occurrence of both tracking flat-spots and lapse video-ratings. Figure 5.9 depicts a 20-s tracking performance accompanied with the original gold-standards

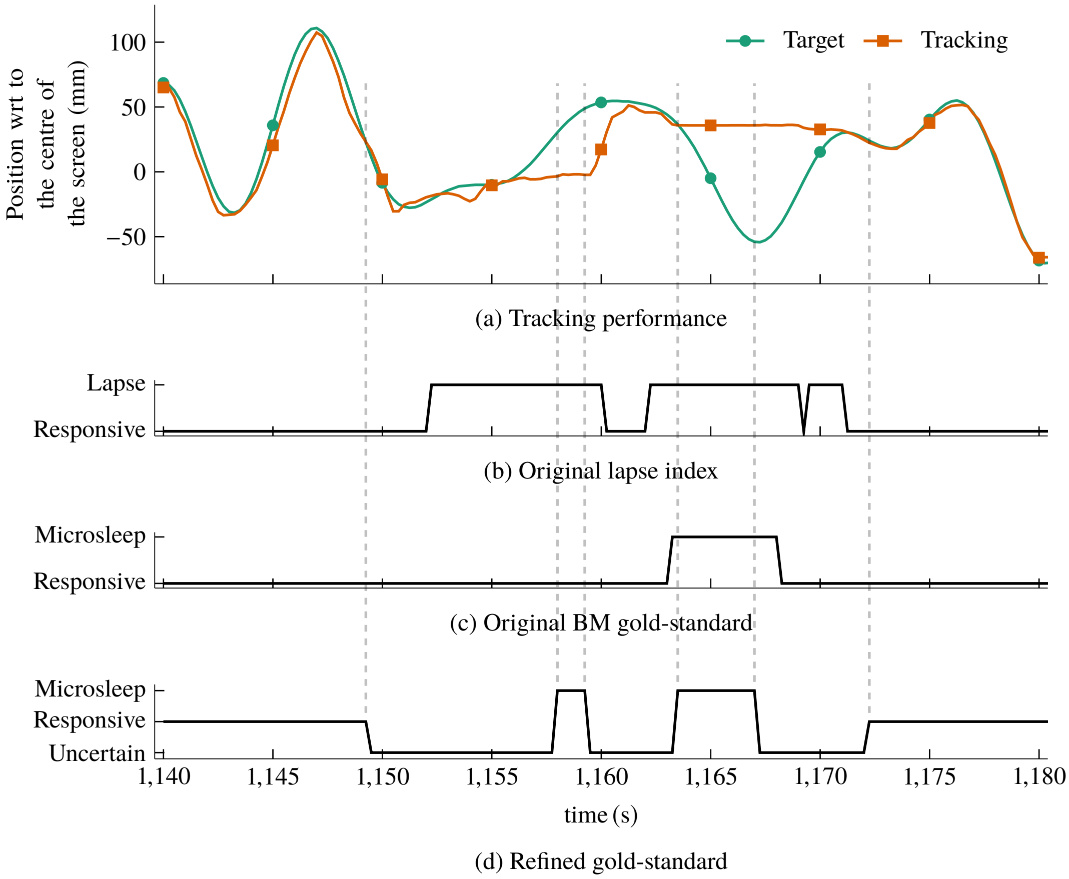
Table 5.3 The number of microsleep events and the total durations of microsleeps and responsives for individual subjects based on the refined gold-standard.

Subject ID	Number of microsleep events		Total duration of microsleeps (s)		Total duration of responsives (min)	
	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2
804	44	16	269.0	742.3	28.6	28.0
809	24	5	107.3	14.3	29.0	45.2
810	0	5	0.0	6.3	37.9	46.8
811	0	18	0.0	57.3	53.8	41.6
814	1	35	1.8	104.8	51.3	40.4
817	8	17	8.3	95.3	36.0	32.8
819	20	21	34.8	57.5	39.9	39.4
820	16	17	206.3	367.0	7.8	13.9
Total	247		2071.8		572.3	

as well as the refined version. It is evident that the lapse index assigned larger portion of data to the lapse class while the tracking performance might have been satisfactory. The BM gold-standard attributed fewer data to the microsleeps, although the tracking performance might have implied a lapse. Tables 5.4 and 5.5 present the descriptive statistics of both of the original gold-standards and show the substantial difference between the total duration of the responsives and microsleeps/lapses of the two gold-standards. Using the lapse index led to a total duration of 5399.7 s for the lapses (cf. 2225.8 s for microsleeps using BM gold-standard) and 870.0 min for the responsives (cf. 922.9 min for BM gold-standard). Since both of the original gold-standards used a binary system, a conservative identification of one class (e.g., microsleeps) might have led to wrong labels in the other class (e.g., responsives).

Table 5.4 The number of lapse events and the total duration of lapses and responsives for individual subjects based on the original lapse index gold-standard.

Subject ID	Number of lapse events		Total duration of lapses (s)		Total duration of responsives (min)	
	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2
804	183	92	691.6	1217.2	48.5	39.7
809	71	26	355.0	93.0	54.1	58.4
810	14	18	29.5	35.5	59.5	59.4
811	2	80	6.0	334.5	59.9	54.4
814	2	50	12.5	251.0	59.8	55.8
817	26	110	78.0	551.1	58.7	50.8
819	132	96	241.5	237.5	56.0	56.0
820	129	112	524.1	741.6	51.3	47.6
Total	1143		5399.7		870.0	



(a) Tracking performance

(b) Original lapse index

(c) Original BM gold-standard

(d) Refined gold-standard

Figure 5.9 An illustration of the differences between the original and refined gold-standards.

Table 5.5 The number of microsleep events and the total duration of microsleeps and responsives for individual subjects based on the original BM gold-standard.

Subject ID	Number of microsleep events		Total duration of microsleeps (s)		Total duration of responsives (min)	
	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2
804	73	116	183.5	735.6	56.9	47.7
809	29	5	144.5	22.5	57.6	59.6
810	0	4	0.0	7.0	60.0	59.9
811	0	24	0.0	50.5	60.0	59.2
814	1	35	2.5	149.5	60.0	57.5
817	0	59	0.0	179.0	60.0	57
819	13	28	19.0	72.5	59.7	58.8
820	65	95	193.5	466.1	56.8	52.2
Total	547		2225.8		922.9	

The refined gold-standard, however, benefits from the uncertain category which made it less likely to have misinterpreted labels. It is apparent that the total durations of the microsleeps identified in the refined gold-standard were closer to the original BM gold-standard, while the responsives were substantially smaller than both of the original gold-standards. As a result of our conservative analysis, responsives were limited to satisfactory tracking performances, microsleeps were limited to non-trackings accompanied with deep-drowsy or lapse video-ratings, and the remainder was labelled uncertain. Notwithstanding, there might be some uncertain segments of the gold-standard that could be attributed to responsives/microsleeps, but the tracking performance analysis was automated and designed to generate a highly conservative gold-standard to minimize potential errors.

For the rest of this thesis, the focus is on the refined gold-standard, and from herein the term ‘gold-standard’ refers to this.

5.6 SUMMARY

This chapter described the preprocessing steps of EEG. Multiple steps of preprocessing were performed on EEG data to remove various artefacts, including Hampel and band-pass filtering, re-referencing to the common average of all channels, applying ASR to high-amplitude visually-identified artefacts, applying ASR to a 4-min moving window, and minimizing the remaining muscle artefacts with canonical correlation analysis blind source separation.

Refinement of the gold-standard was also presented in this chapter and was undertaken so as to minimize errors in the gold-standard. This was done by introducing an *uncertain* category to avoid labelling data without adequate information. Satisfactory tracking performance for at least 5 s, irrespective of the video ratings, was used to identify responsives. Conversely, microsleeps were defined as a conjunction of non-trackings and deep-drowsy/lapse video ratings. Due to a lack of information, the rest of the data were marked uncertain. It was observed that the total durations of microsleeps in the refined gold-standard were relatively similar to the original BM gold-standard, but the responsives were substantially less.

Chapter 6

MICROSLEEP PREDICTION PROCEDURES: FEATURE EXTRACTION, CLASSIFICATION, AND PERFORMANCE EVALUATION

6.1 INTRODUCTION

An overview of a microsleep detection/prediction system is shown in Figure 6.1. The first stage is to collect EEG from the scalp of an individual, which is then fed to a preprocessing step to minimize various artefacts, as described in Section 5.2. Features of the preprocessed EEG are then extracted. A feature reduction method could be used to reduce the dimensionality of features. The (reduced) features are then fed to a classifier to identify imminent microsleeps. A warning can then be provided to the user prior to what would otherwise have been a microsleep. However, no feedback was provided to the users in this study. In addition, the focus of this thesis is to exploit Bayesian methods for the feature reduction step.

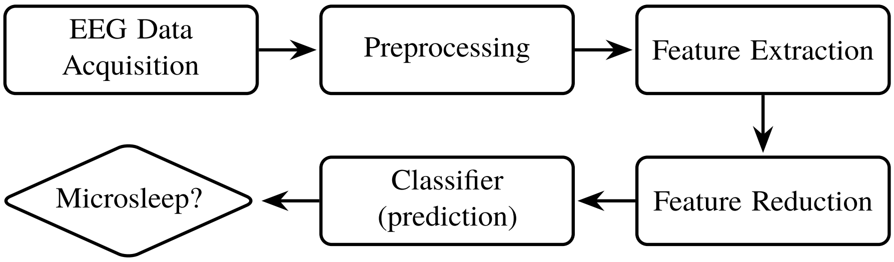


Figure 6.1 Overview of the microsleep prediction system.

6.2 FEATURE EXTRACTION

To extract features, the EEG was first segmented into 2, 5, and 10 s epochs. Using multiple EEG windows allow us to examine transient and tonic changes [Huang et al. 2008, Lin et al. 2010] to predict microsleeps. The sliding window of EEG segmentation was set to 0.25 s to match the temporal resolution of the gold-standard. Having a high temporal resolution is a key requirement to ensure our system is able to quickly identify microsleeps. Hence, a temporal resolution of 0.25 s for microsleep identification was used in this study. Various channel-wise features of each segment were extracted: power spectral features (PSF), power spectral features using individual alpha frequency (PSF-IAF), wavelet features, and multiple domain features (MDF).

6.2.1 Power spectral features

In a given signal, power spectral density (PSD) estimates the distribution of power over frequency components. PSD has been used in the literature to detect/predict the level of arousal and alertness [Chai et al. 2016, Golz et al. 2007, Huang et al. 2008, Jap et al. 2009, Jung et al. 1997, Lal and Craig 2005, Lin et al. 2010, 2013, Wang et al. 2014b]. Therefore, the PSD of EEG data was selected for further investigation.

Welch's modified periodogram [Welch 1967] was used to estimate the PSD of individual channels of an EEG epoch [Diez et al. 2008, Golz et al. 2007, Naderi and Mahdavi-Nasab 2010]. This method computes the PSD of a signal by averaging the periodogram of smaller overlapping windowed segments and as a result has a lower variance compared to a periodogram of the whole epoch [Freeman and Quiroga 2013, Tong and Thakor 2009]. The parameters of Welch's method were set to a 2-s segment with a 75% overlap between consecutive segments, in which a Hamming window was applied to each segment. Using these settings, Welch's method becomes a single periodogram when the length of the EEG epoch is 2 s.

Given an EEG epoch, 12 features of PSD for each individual electrodes were calculated by averaging the power across different frequency bands, as shown in Table 6.1, and transforming these to logarithmic-scale since these features had log-normal distributions. Features of all electrodes of an epoch were then concatenated to form a feature vector, i.e., PSF, leading to a total of $12 \times 16 = 192$ features.

Table 6.1 Frequency bands to calculate power spectral features (PSF) for an EEG electrode.

Feature	**Frequency Band** (Hz)
Delta	1.0 – 4.5
Theta	4.5 – 8.0
Alpha	8.0 – 12.5
Alpha1	8.0 – 10.5
Alpha2	10.5 – 12.5
Beta	12.5 – 25.0
Beta1	12.5 – 15.0
Beta2	15.0 – 25.0
Gamma	25.0 – 45.0
Gamma1	25.0 – 35.0
Gamma2	35.0 – 45.0
Overall	1.0 – 45.0

6.2.2 Power spectral features using individual alpha frequency

As mentioned in Section 6.2.1, the PSF were extracted from fixed frequency bands of EEG. However, it has been shown that the individual alpha frequency (IAF) of EEG has a large inter-subject variability [Haegens et al. 2014, Klimesch 1999, Posthuma et al. 2001]. In adult humans, the alpha frequency is the dominant EEG frequency band and changes with age, memory performance, and speed of information processing [Haegens et al. 2014, Klimesch 1999]. Therefore, using a set of fixed frequency bands for feature extraction might not take inter-subject variabilities into account. The IAF can be used to define a set of frequency bands, in which the IAF is the anchor point. Such frequency bands have been used for drowsiness detection from EEG [Qian et al. 2017].

To calculate IAF, EEG was segmented into 8-s windows. Since the occipital alpha frequency is the easiest to detect [Posthuma et al. 2001], the average of the O1 and O2 electrodes of individual epochs were used to estimate IAF. For a given epoch, PSD of the average signal (O1 and O2) was first computed using Welch's method with 4-s windows and 3-s overlaps. Then, IAF was estimated as the centre of gravity over the extended alpha range, i.e., 7–14 Hz, of the PSD [Goljahani et al. 2012, Klimesch 1999],

$$\mathrm{IAF} = \frac{\int_7^{14} f \times \mathrm{PSD}(f) df}{\int_7^{14} \mathrm{PSD}(f) df}. \tag{6.1}$$

Table 6.2 shows the frequency bands using IAF as the anchor point. After computation of the IAF for a time point (using its previous 8-s window), PSF-IAF of an electrode in the corresponding epoch was calculated as the mean power over the IAF-based frequency bands, and then was transformed to logarithmic-scale. Features of all electrodes of an epoch were then concatenated to form a feature vector, i.e., PSF-IAF, leading to a total of $12 \times 16 = 192$ features.

6.2.3 Wavelet features

The wavelet transform is a powerful method to find the time-frequency representation of a signal, especially for nonstationary signals such as EEG [Akin et al. 2008, Li et al. 2016, Tong and Thakor 2009]. The DWT is a multi-resolution approximation method which decomposes a signal into multiple sub-bands [Faust et al. 2015, Kumar et al. 2014, Sanei et al. 2007]. At each decomposition level, DWT decomposes the signal into lower and higher frequency components known as approximation and detail, respectively. Since the sampling rate of our EEG data is 256 Hz, applying a 5-level DWT results in standard EEG frequency bands, as shown in Figure 6.2.

The DWT has been used for sleep staging using EEG [Khalighi et al. 2012, Şen et al. 2014]. Both authors suggested that the Daubechies wavelet of order-4 (db4) performs better than others. Therefore, in this study, three feature sets were generated from db4 wavelet coefficients of individual electrodes for every epoch.

Table 6.2 Frequency bands to calculate power spectral features using individual alpha frequency (PSF-IAF) for an EEG electrode.

Feature	**Frequency Band** (Hz)
Delta	IAF-9.5 – IAF-6
Theta	IAF-6 – IAF-2.5
Alpha	IAF-2.5 – IAF+2
Alpha1	IAF-2.5 – IAF
Alpha2	IAF – IAF+2
Beta	IAF+2 – IAF+14.5
Beta1	IAF+2 – IAF+4.5
Beta2	IAF+5 – IAF+14.5
Gamma	IAF+14.5 – IAF+34.5
Gamma1	IAF+14.5 – IAF+24.5
Gamma2	IAF+24.5 – IAF+34.5
All Frequencies	IAF-9.5 – IAF+34.5

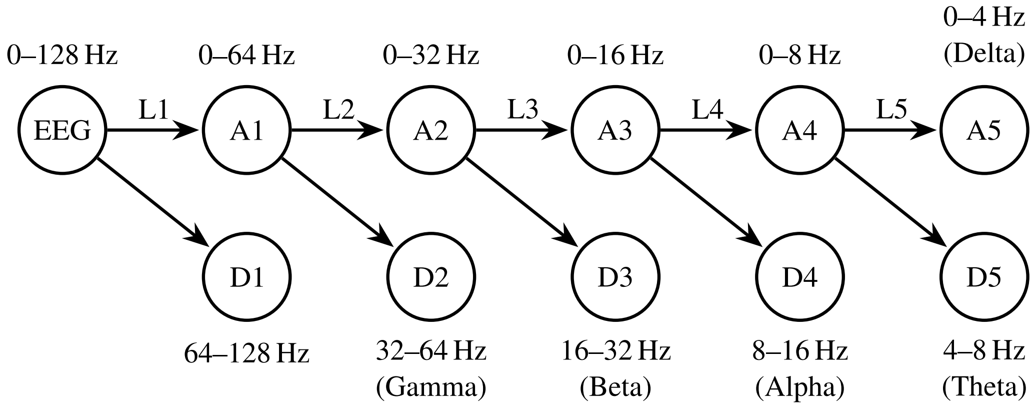


Figure 6.2 Wavelet 5 level decomposition for an EEG signal and its respective bands.

1) **Wavelet mean squared features (WMSF)**: is the collection of the mean-squared wavelet-coefficients of the sub-bands corresponding to EEG frequency bands, i.e., A5, D2, D3, D4, and D5. This generates 80 features per EEG epoch.

2) **Wavelet log mean squared features (WLMSF)**: is similar to WMSF with all the features transformed to logarithmic scale. Hence, it totals 80 features per EEG epoch.

3) **Wavelet energy percentage features (WEPF)**: is the percentage of the energy of each sub-band relative to its overall energy. The energy of each sub-band can be calculated with $||\mathbf{x}||_2^2$, where **x** is the wavelet coefficients of the corresponding sub-band. Since the summation of all sub-bands adds up to 100%, only 5 sub-bands were required to avoid linear dependencies. This totals 80 features per EEG epoch.

6.2.4 Multiple domain features

Multiple domain features (MDF) are generated from various features of time/frequency domains of EEG to form a feature matrix. These features were selected for further investigation because of their performance in detection of alertness/drowsiness and sleep staging [Bojić et al. 2010, Chapotot and Becq 2010, Chen et al. 2015, Şen et al. 2014]. Multiple domain features (MDF) were formed by concatenation of the following features for all EEG electrodes:

- **Hjorth parameters** are a set of three time-domain features describing a single channel of EEG [Navascués and Sebastián 2009, Rodríguez-Bermúdez et al. 2013, Vidaurre et al. 2009]. Three Hjorth features are *activity*, *mobility*, and *complexity*. Hjorth activity (HA) is the variance of an EEG signal, i.e., signal power, and represents the width of the signal. Hjorth mobility (HM) estimates the mean frequency of the EEG. Hjorth complexity (HC) estimates the bandwidth of the EEG by computing the mobility of the first derivative of EEG relative to the mobility of the EEG itself. The Hjorth parameters are calculated as [Şen et al. 2014, Vidaurre et al. 2009]

$$HA = \sigma_0^2, \tag{6.2}$$

$$HM = \frac{\sigma_1}{\sigma_0}, \tag{6.3}$$

$$HC = \frac{\sigma_2 \sigma_0}{\sigma_1^2}, \tag{6.4}$$

  where σ_0 is the standard deviation of the signal and σ_1 and σ_2 are the standard deviations of the first and second order derivative of the signal.

- **Petrosian fractal dimension (PFD)** is the simplest approximation of fractal dimension and computes a measure of signal complexity [Pavithra et al. 2014, Şen et al. 2014, Upadhyay et al. 2015]. PFD simplifies the computation of fractal dimension by transforming the signal to a binary representation and approximating the fractal dimension with the number of sign changes,

$$\text{PFD} = \frac{\log_{10}(N_{\text{EEG}})}{\log_{10}\left(\frac{N_{\text{EEG}}}{N_{\text{EEG}}+0.4N_{\Delta}}\right)}, \tag{6.5}$$

  where N_{EEG} is the number of sample points of the EEG signal and N_{Δ} is the number of sign changes of the signal.

- **Katz fractal dimension (KFD)** is another method of estimating fractal dimension of a signal [Paramanathan and Uthayakumar 2008, Pavithra et al. 2014, Polychronaki et al. 2010]. KFD is more accurate than PFD but is computationally more expensive. KFD is calculated as [Polychronaki et al. 2010]

$$\mathrm{KFD} = \frac{\log\left(N_{\mathrm{EEG}} - 1\right)}{\log\left(N_{\mathrm{EEG}} - 1\right) + \log\left({}^{d}/_{L}\right)}, \tag{6.6}$$

where N_{EEG} is the number of EEG-signal points, d is the diameter, and L is the curve length. Assuming that $\mathbf{x} = \left\{x_1, x_2, \ldots, x_{N_{\mathrm{EEG}}}\right\}$ is the sequence of EEG signal, the diameter and curve length are

$$L = \sum_{n=2}^{N} \left|x_n - x_{n-1}\right|, \tag{6.7}$$

$$d = \max_{n}\left(\left|x_n - x_{n-1}\right|\right). \tag{6.8}$$

- **Mean curve length (MCL)** is an approximation of KFD [Şen et al. 2014]. Assuming that $\mathbf{x} = \left\{x_1, x_2, \ldots, x_N\right\}$ is an EEG signal, MCL is calculated by

$$\mathrm{MCL} = \frac{1}{N}\sum_{n=1}^{N} \left|x_n - x_{n-1}\right|. \tag{6.9}$$

- **Hurst exponent** is a measure of long-range self-similarity within a time-series [Geng et al. 2011, Sheng et al. 2012, Yuan et al. 2011]. The Hurst exponent can have a value of 0–1, where a value of 0.5 corresponds to random data. Assuming that the signal is $\mathbf{x} = \left\{x_1, x_2, \ldots, x_N\right\}$, deviation of the first k data points from the mean of the first n data points is

$$W_k = \sum_{t=1}^{k} x_t - \frac{k}{n}\sum_{t=1}^{n} x_t, \quad \begin{array}{c} 1 \le k \le n \\ 1 \le n \le N \end{array}. \tag{6.10}$$

The range $\mathrm{R}(n)$ is defined as the maximum difference between the deviations of the first n-points,

$$\mathrm{R}(n) = \max(0, W_1, \ldots, W_n) - \min(0, W_1, \ldots, W_n),\ 1 \le n \le N. \tag{6.11}$$

The Hurst exponent is then given by

$$H \times n + C_H = \frac{\log\left({}^{\mathrm{R}(n)}/_{\mathrm{S}(n)}\right)}{\log\left(n\right)},\ 1 \le n \le N, \tag{6.12}$$

where C_H is a finite constant independent of n and $\mathrm{S}(n)$ is the empirical standard deviation of the first n points. The Hurst exponent can be computed by fitting a line to the right hand side of Equation (6.12).

- **Nonlinear energy (NLE)**, also known as mean Teager energy, is a feature of EEG that has been widely used for epileptic seizure prediction [Greene et al. 2008]. The NLE estimates instantaneous energy of a signal and particularly identifies transient changes such as sleep spindles and seizure spikes [Imtiaz et al. 2013]. It has also been used for automatic sleep staging and ranked among top features [Şen et al. 2014]. Assuming that the signal is $\mathbf{x} = \left\{x_1, x_2, \ldots, x_N\right\}$, NLE is calculated by

$$\mathrm{NLE} = \frac{1}{N} \sum_{n=2}^{N-1} \left(x_n^2 - x_{n-1} \times x_{n+1}\right). \tag{6.13}$$

- **Spectral entropy** identifies the complexity or regularity of the EEG [Fell et al. 1996, Greene et al. 2008]. To calculate spectral entropy, the probability distribution of the signal is approximated by its PSD. The spectral entropy is then calculated by

$$H_s = -\frac{1}{N_f} \sum_{f=f_l}^{f_u} \mathrm{PSD}(f) \log\left(\mathrm{PSD}(f)\right), \tag{6.14}$$

  where N_f is the number of frequency bins, and f_l and f_u are the lower and upper frequency limits, respectively. In this study, the lower and upper frequencies were set to 0.5 Hz and 45 Hz, respectively.

- **Intensity-weighted mean frequency (IWMF)**, also known as gravity frequency, finds the weighted average frequency of a signal relative to its PSD [Chen et al. 2015, Greene et al. 2008, Yeo et al. 2009]. Having the PSD of a signal, the IWMF can be calculated as

$$\mathrm{IWMF} = \frac{\sum_f f \times \mathrm{PSD}(f)}{\sum_f \mathrm{PSD}(f)}. \tag{6.15}$$

- **Intensity-weighted bandwidth (IWBW)**, also known as frequency variability, is defined as variance of the frequency [Chen et al. 2015, Greene et al. 2008, Yeo et al. 2009]. Using IWMF and PSD, the calculation of IWBW is

$$\mathrm{IWBW} = \sqrt{\frac{\sum_f \mathrm{PSD}(f)\left(\mathrm{IWMF} - f\right)^2}{\sum_f \mathrm{PSD}(f)}}. \tag{6.16}$$

Calculating these features for individual EEG electrodes, i.e., 11 features per electrode, and concatenating them, i.e., constructing an MDF feature vector, leads to a total of $11 \times 16 = 176$ features per epoch.

6.3 CLASSIFICATION MODELS FOR MICROSLEEP DETECTION AND PREDICTION

After generating feature sets, as described in Section 6.2, the sets were individually fed to a classifier to perform the prediction task. This section presents a description of classifiers used for the prediction of microsleeps.

6.3.1 Linear discriminant analysis

LDA is a classification method that tries to distinguish two or more classes using a hyperplane [Hastie et al. 2009, Murphy 2012]. LDA has been widely used in the brain-imaging literature to discriminate different brain states [Lemm et al. 2011, Lotte et al. 2007]. The hyperplane aims to minimize the inter-class variances and maximize the distance between class means [Xanthopoulos et al. 2012]. LDA assumes that data of all classes are normally distributed with the same covariance matrices. Fitting a classification model to the data is done with maximum likelihood estimation (MLE). In this process, the mean of each class is set to the empirical mean of data of that class. The covariance matrix is set to a weighted average of empirical covariance of data from all classes (refer to Hastie et al. [2009] and Xanthopoulos et al. [2012] for comprehensive coverage).

Although MLE is a simple and appealing approach to fit a classification model, it can result in overfitting [Murphy 2012]. To minimize overfitting of LDA, a ridge regularization, i.e., shrinkage, can be added to the covariance matrix [Hastie et al. 2009],

$$\hat{\Sigma} = \lambda \, \mathrm{diag}\left(\hat{\Sigma}_{mle}\right) + (1 - \lambda)\hat{\Sigma}_{mle}, \tag{6.17}$$

where $\hat{\Sigma}_{mle}$ is the solution of MLE for covariance matrix using training data, $\hat{\Sigma}$ is the regularized covariance matrix, λ controls the regularization, and $\mathrm{diag}\left(\mathbf{X}\right)$ is a matrix with diagonal elements of $\mathbf{X}$. $\lambda = 0$ simplifies the system to the MLE, whereas $\lambda = 1$ results in a diagonal covariance matrix.

Throughout this thesis, to select the best value of $\lambda \in \left\{0, 0.1, \ldots, 1\right\}$ for an LDA, a 5-fold cross-validation on the training data was performed. To minimize overfitting, this procedure was done before training each LDA classifier.

6.3.2 Linear support vector machine

The SVM is a widely used classification method which finds a maximum margin decision boundary [Bishop 2006, Hearst et al. 1998, Kumar et al. 2014, Quitadamo et al. 2017, Yeo et al. 2009, Zhang and Hua 2015]. Margin is defined as the smallest distance between the data points and the decision boundary. SVM can make use of a kernel trick to form a nonlinear classifier, but employing a kernel increases computational complexity [Lawrence and Schölkopf 2001]. The computational complexity of a kernel classifier can be as high as $O(N^3)$, where N is the number of training instances [Lawrence and Schölkopf 2001]. Therefore, as the amount of training data increases, the computational complexity of a kernel method becomes a critical issue. In addition, although nonlinear SVMs are more flexible to find a separation boundary compared to linear classifiers, they are more likely to overfit to training data because of their flexibility [Hastie et al. 2009]. A kernel SVM (e.g., polynomial kernel) classifier requires a cross-validation to find an appropriate regularization parameter. Additionally, a kernel has a set of parameters which requires cross-validation to select the optimum values [Hastie et al. 2009]. The selection of the kernel parameters and the regularization parameter leads to a nested cross-validation which, for a highly demanding classifier with a large amount of data, can take days, or even months, to complete. Due to time constraints, repeating the same procedure for different feature sets and different test subjects would be infeasible for this project. Hence, the focus of this research was limited to a linear SVM. Refer to Bishop [2006] and Murphy [2012] for complete coverage of SVMs.

Ridge regularization was used to reduce generalization error, i.e., overfitting. This is a compromise between minimizing training error (maximizing the margin) and minimizing $\lambda\|\mathbf{w}\|_2^2/2$, where $\|\mathbf{x}\|_2$ is the l^2-norm of the vector $\mathbf{x}$, $\mathbf{w}$ is the classifier's weight vector, and λ is the regularization coefficient [Jebara 2003, Zhang and Yang 2003]. To select the regularization coefficient, a 5-fold cross-validation on training data was performed to select a value of $\log\left(\lambda\right) \in \left\{-8, -7, \ldots, 1\right\}$ with the lowest cross-validation error. Similar to the LDA training procedure, the regularization coefficient was found for every linear SVM at the training stage throughout this thesis.

6.3.3 Variational Bayesian logistic regression

Logistic regression is another linear classification method and has been widely used [Bagley et al. 2001, Dreiseitl and Ohno-Machado 2002, Hastie et al. 2009]. Logistic regression is a discriminative approach that directly models data with a conditional probability [Murphy 2012]. However, it is prone to severe overfitting. Variational Bayesian logistic regression (VBLR) has been developed to prevent overfitting without a need for cross-validation [Bishop 2006, Drugowitsch 2013]. VBLR uses a hierarchical Bayesian structure with ARD motivated prior distributions over weight coefficients, as shown in Figure 6.3. ARD determines the relevance of each feature and applies a separate prior, i.e., regularization, to individual features, which can effectively switch-off irrelevant features [Drugowitsch 2013]. Refer to Bishop [2006] and Drugowitsch [2013] for detailed overview of VBLR.

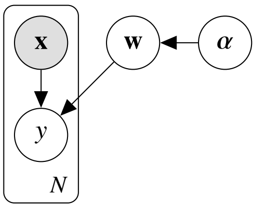


Figure 6.3 Graphical representation of variational Bayesian logistic regression (VBLR), where **x** is the feature vector, y is the class label, **w** is the weight vector, and α is the prior over **w**.

6.3.4 Tree augmented naïve Bayes

Tree augmented naïve Bayes (TAN) is a Bayesian classifier that relaxes the conditional independence assumption of Naïve Bayes [Barber 2012, Blanco et al. 2005, Jiang et al. 2012]. It constructs a tree structure to model the conditional dependencies in which each feature depends on one other feature, as shown in Figure 6.4. Such structure is learnt with the Chow-Liu algorithm [Chow and Liu 1968], while MLE is used to estimate the conditional probability distribution of features [Bielza and Larrañaga 2014, dos Santos et al. 2011]. Pérez et al. [2006] proposed an extension of TAN with conditional Gaussian probability distributions for continuous data. Madden [2009] showed that a TAN has similar accuracy to a general Bayesian network, while its computational complexity is much lower. Therefore, this classifier was selected for further investigation in microsleep prediction. Refer to Jensen and Nielsen [2007] for a more detailed overview.

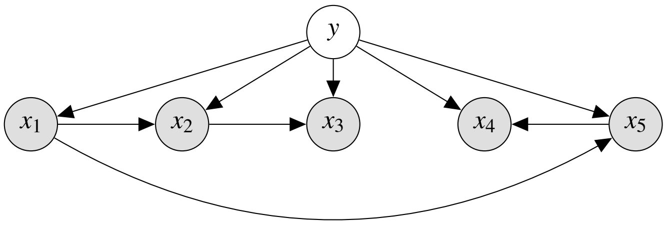


Figure 6.4 An example structure of tree augmented naïve Bayes (TAN) for a 5-features dataset.

6.3.5 Learning from imbalanced data

In terms of class distribution, the microsleep dataset is intrinsically imbalanced. Imbalanced data, however, have an adverse impact on the standard learning algorithms of most classifiers [Chawla 2010, He and Garcia 2009, Sun et al. 2009]. Sampling and cost-sensitive methods are two state-of-the-art solutions to the imbalance learning problem [He and Garcia 2009, Sun et al. 2009]. Sampling methods alter the data to find a balanced representation of data. This could be achieved with over-sampling of the minority class, under-sampling of the majority class, or a combination of both [He and Garcia 2009, López et al. 2013]. Cost-sensitive methods, on the other hand, use different cost values for misclassification of minority and majority classes [He and Garcia 2009, López et al. 2013, Thai-Nghe et al. 2010].

Synthetic minority over-sampling technique (SMOTE) [Chawla et al. 2002] is a well-known and powerful over-sampling method [He and Garcia 2009, López et al. 2013]. SMOTE uses the minority class and generates synthetic data to balance the dataset. At first, it uses the kNN method to find the k nearest data points of each data instance (k is usually 5). Then, one of the k nearest points is randomly selected and, finally, a synthetic point between the two data instances is randomly generated.

Adaptive synthetic sampling (ADASYN) [He et al. 2008] is an extension of SMOTE in which more synthetic data points are generated near the boundary of two classes. ADASYN finds the k nearest points in the whole dataset for each instance of the minority class. It then uses SMOTE to over-sample the minority class, where the number of generated synthetic data for each minority data is relative to the number of majority instances around it.

Rapidly converging Gibbs sampler (RACOG) [Das et al. 2015] is an over-sampling method which fits a probability distribution to the data and then samples data from the fitted distribution. RACOG uses Chow-Liu dependence tree [Chow and Liu 1968] to approximate probability distribution of the data. It then uses a Gibbs sampler to sample data from the joint probability of data.

6.4 PERFORMANCE EVALUATION

Evaluating the performance of various methods for predicting microsleeps is an essential part of this research. The rest of this section describes the definition of the two types of microsleep prediction, i.e., state and onset predictions. Then the validation method and performance measures used in this study are described.

6.4.1 Microsleep state prediction

A microsleep is a brief instance of sleep causing loss of consciousness up to 15 s [Jones et al. 2010]. The discrete gold-standard can be used as an indicator of the state of responsiveness. Therefore, we can predict the state of responsiveness at every step. The importance of state prediction is two-fold: (1) having the uncertain labels makes it difficult to identify all microsleep events and (2) predicting the microsleep state is initially equivalent to onset prediction but, if the onset is not detected, the prediction becomes that of one or more of the following states in the current microsleep event.

Features of a T-s segment of EEG are used to predict the state of responsiveness at τ s ahead. This process is repeated every 0.25 s to predict the states of the gold-standard, i.e., responsives and microsleeps, with a high temporal resolution. Notwithstanding, the uncertain labels of the gold-standard were nulled out since there was not enough information to accurately label them. Figure 6.5 depicts a schematic of the microsleep state prediction procedure, where the highlighted points are used for training and testing of the state prediction system.

6.4.2 Microsleep onset prediction

The ultimate goal is to predict imminent microsleeps. Although state prediction identifies the state of responsiveness continuously, it might fail to predict the onset of a microsleep. It is, however, difficult to identify microsleep onsets due to existence of the uncertain labels. In this research, the onset of a microsleep is defined as the first instance of microsleep state after a period of responsive state, without considering uncertain labels.

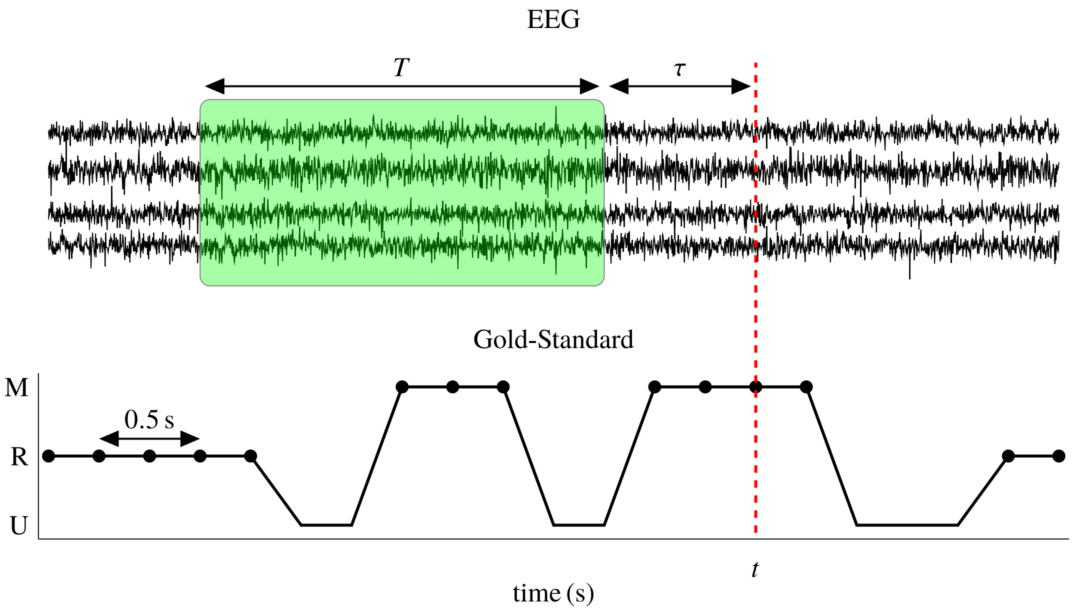


Figure 6.5 The schematic of microsleep state prediction at time t. The gold-standard corresponds to responsives (R), microsleeps (M), and uncertains (U). Extracted features of the EEG highlighted box are used to predict the state of responsiveness at time t.

For onset prediction, all of the responsive states as well as the microsleep onsets were used as the gold-standard. Similar to microsleep state prediction, the features of a T-s segment of EEG are used to predict the onset of a microsleep. A continuous prediction, with steps of 0.25 s, allows us to predict microsleep onsets with a high temporal resolution. The schematic of microsleep onset prediction is shown in Figure 6.6, where the highlighted points are used for training and testing of the onset prediction system.

6.4.3 Validation and performance measures

It is important to have a subject-independent evaluation of prediction performance of our methods. This is achieved by performing leave-one-subject-out cross-validation (LOSO-CV) for performance evaluation, as follows:

1. Reserve one subject as the independent *test* subject.
2. Use the other 7 subjects to train the classifiers.
3. Apply a 5-fold cross-validation on the training data (7 subjects) to find regularization coefficients, if needed.
4. Feed the test subject to the classifiers and obtain the performance measures.
5. Repeat steps 1–4 until all subjects have been used as a test subject once.
6. Find the overall performances by averaging the performance measures for all 8 subjects.

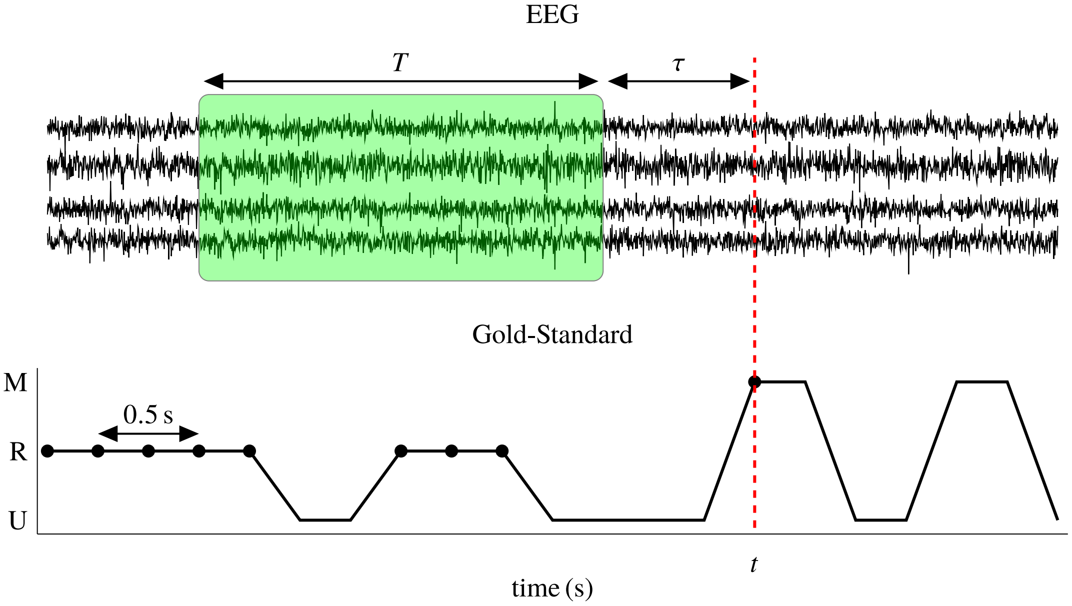


Figure 6.6 Schematic of microsleep event prediction at time t. The gold-standard corresponds to responsives (R), microsleeps (M), and uncertains (U). Extracted features from the EEG highlighted box are used to predict the onset of microsleep at time t.

Several measures were used to evaluate the performance of microsleep prediction systems. The basic measures are sensitivity (Sn), specificity (Sp), and precision (Pr),

$$\text{Sn} = \frac{TP}{TP + FN}, \tag{6.18}$$

$$\text{Sp} = \frac{TN}{TN + FP}, \tag{6.19}$$

$$\text{Pr} = \frac{TP}{TP + FP}, \tag{6.20}$$

where TP is true positives, TN is true negetives, FP is false positives, and FN is false negetives. Both microsleep datasets of this study, i.e., state and onset predictions, are highly imbalanced. Therefore, three additional measures, namely geometric mean (GM), phi, and F-measure, were also calculated, as they are widely used to evaluate the performance of imbalance learning problems [Hor et al. 2013, López et al. 2013, Sun et al. 2009, Vihinen 2012]. These measures are given by

$$\text{phi} = \varphi = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}}, \tag{6.21}$$

$$\text{GM} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}, \tag{6.22}$$

$$\text{F-measure} = \frac{(1 + \beta)^2 \times \text{Sn} \times \text{Pr}}{\beta^2 \times \text{Sn} + \text{Pr}}, \tag{6.23}$$

where β is a user-specified coefficient indicating the relative importance of precision and sensitivity. Phi is also known as Matthews correlation coefficient (MCC). In addition, several studies have used curve-based performance evaluation metrics for imbalanced learning problems [Folleco et al. 2009, Gong and Kim 2017, He and Garcia 2009, López et al. 2013, Saito and Rehmsmeier 2015, Sun et al. 2009]. Therefore, two curve-based measures, i.e., area under the curve of receiver operating characteristic (AUC-ROC) and area under the curve of precision recall (AUC-PR), were also calculated in this study.

In the literature, it has been argued that more than a single performance metric is required to assess an imbalanced learning problem such as microsleep prediction [He and Garcia 2009, Vihinen 2012, 2013]. This is due to the shortcomings and potential biases of individual performance metrics. For instance, precision does not provide any information regarding the number of false negatives, whereas sensitivity does not contain information on false positives. However, F-measure, GM, and phi metrics provide informations about different combinations of the contingency table. Phi and F-measure offer insight on the functionality of a classifier, whereas GM provides information about the balance between sensitivity and specificity [He and Garcia 2009, López et al. 2013, Sun et al. 2009, Vihinen 2012].

Sensitivity to the imbalance distribution is another potential issue. It has been shown that precision, phi, F-measure, and AUC-PR are sensitive to the imbalance ratio of the data [He and Garcia 2009, Saito and Rehmsmeier 2015]. However, the imbalance ratio within microsleep datasets is highly variable between subjects. This might lead to performance measure inaccuracies when LOSO-CV is applied and subject-independent performance metrics are computed.

In this research, we report multiple performance metrics to gain a better understanding of various aspects of each algorithm. However, F-measure, with $\beta = 1$, and phi were found to be highly correlated ($\rho = 0.98$) and thus only the phi metric is reported.

6.5 BASELINE PERFORMANCES

As discussed in Section 5.3, the gold-standard was refined in this study. As a result, the previous studies in the literature [Ayyagari et al. 2015, Ayyagari 2017, Davidson et al. 2007, LaRocco 2015, Peiris 2008, Peiris et al. 2011] are not directly comparable to our results. Therefore, a series of methods were used to find a set of performances with the refined gold-standard. The best of these performances, i.e., baseline, was then used to assess the performance of our proposed methods. This section details the various methods used to find the best baseline performance.

As the focus of this project was on integrating Bayesian methods in the feature-reduction step, various alternative feature-reduction methods were also applied to each feature set. These methods were PCA, Bayesian PCA, FA, and VBFA. In addition, the effects of two feature-selection methods on performance were examined. These feature selection methods are: (1) greedy forward feature-selection algorithm based on mutual information [Battiti 1994, Kwak and Choi 2002] and (2) greedy forward feature-selection based on Hellinger distance [Yin et al.

2013].

Contrary to our expectations, the original features had similar or superior performances compared to the reduced-dimension feature sets in almost all cases. Therefore, the original features without any feature-reduction/selection methods were used for baseline performances throughout this thesis.

In addition, baseline methods were processed with the cost-sensitive and the three oversampling methods mentioned in Section 6.3.5. Our results showed small differences between the imbalance methods. The cost-sensitive method had superior performances most of the time. Furthermore, sampling methods were generally slower than cost-sensitive learning. Hence, the cost-sensitive method was chosen to address the imbalance learning problem for the rest of this thesis.

6.6 SUMMARY

This chapter provided an outline of the procedures used in microsleep prediction systems, including feature extraction, classification methods, two definitions of microsleep prediction, and the performance evaluation steps. Feature sets of the EEG extracted for microsleep prediction were power spectral features (PSF), power spectral features using individual alpha frequency (PSF-IAF), multiple domain features (MDF), wavelet mean squared features (WMSF), wavelet log mean squared features (WLMSF), and wavelet energy percentage features (WEPF). Moreover, a brief description of different classifiers chosen for microsleep prediction was provided. These classifiers were linear discriminant analysis (LDA), linear support vector machine (SVM), variational Bayesian logistic regression (VBLR), and tree augmented naïve Bayes (TAN). Microsleep onset and state prediction were explained. Finally, procedures to evaluate overall performance as well as the performance metrics were described.

Chapter 7

VARIATIONAL BAYESIAN ROBUST FACTOR ANALYSIS

7.1 INTRODUCTION

As mentioned in Chapter 6, an EEG-based microsleep detection/prediction system requires the collected EEG data to be processed in several stages. An overview of preprocessing of the EEG-data was provided in Section 5.2. Various features of the EEG data were then extracted for further processing, as described in Section 6.2. Due to the high-dimensionality of the feature sets, it was considered that applying a feature reduction method and finding a lower-dimension representation of the data could achieve a higher performance. As mentioned in Section 6.5, contrary to our expectation, the original features without any feature reduction/selection method, resulted in the same or higher performances for most of the feature sets and hence were considered as the baseline performance. However, this lack of increased performance with feature reduction/selection method is likely to be due to noise and/or inter-subject variability of the EEG data. Although EEG noise, artefacts, and artefact removal have been investigated in several studies [Clercq et al. 2006, Daly et al. 2015, 2013, Delorme et al. 2007, Klados et al. 2011, Mullen et al. 2015, Nolan et al. 2010], they are unlikely to have achieved fully-cleaned data by eliminating all of the artefacts. Removing artefacts improves the signal to noise ratio of EEG, but it does not guarantee a noiseless signal. As a result of EEG noise, extracted features from EEG data will also contain noise which can deteriorate the performance of microsleep detection/prediction.

The current chapter and Chapters 8–10 present our proposed Bayesian feature-reduction methods to improve the performance of microsleep detection and investigate microsleep prediction. As mentioned in Section 3.3.2, the Bayesian FA model uses a diagonal noise term which assigns different noises for individual features (cf. Bayesian PCA which assumes an isotropic noise term). Since EEG data is collected with multiple electrodes, the noise level of individual electrodes can be different. Hence, our proposed methods extend Bayesian FA model for feature-reduction.

After presenting each method in this chapter and Chapters 8–10, microsleep detection and prediction performance were investigated with the proposed method. Each chapter includes the comparison between the performance of the proposed method and the baselines. A comprehensive discussion is then provided in Chapter 11 providing a critical comparison of classifiers, feature sets, and proposed methods.

To address the sensitivity of VBFA to noise, we developed a variational Bayesian robust FA (VBRFA) model that uses a Student-t distribution for each latent variable. For the remainder of this chapter, Section 7.2 presents an overview of FA model. Section 7.3 describes our model and its underlying assumptions. The variational formulation of VBRFA is presented in Section 7.4. This is then followed by the results and discussion of microsleep prediction using VBRFA.

7.2 FACTOR ANALYSIS

FA is a matrix factorization method which is used to find a compressed representation of higher-dimensional data. It has been widely used in literature for feature reduction, feature extraction, and visualization [Nakajima et al. 2013, Zhao and Yu 2009]. FA is a linear Gaussian model and finds a set of independent latent variables and a factor loading matrix to explain the correlations of the data. Let the observed D-dimensional data be $\mathbf{X} = \left\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right\}$. The FA model of $\mathbf{x}_n$ can be mathematically expressed as

$$\mathbf{x}_n = \mathbf{W}\mathbf{z}_n + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_n, \tag{7.1}$$

$$\mathbf{z}_n \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right), \tag{7.2}$$

$$\boldsymbol{\varepsilon}_n \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Psi}^{-1}\right), \tag{7.3}$$

where $\mathbf{z}_n$ is a K-dimensional ($K < D$) vector of latent variables (factors), $\mathbf{W}$ is a $D \times K$ loading matrix, $\boldsymbol{\mu}$ is a D-dimensional mean vector, $\boldsymbol{\varepsilon}_n$ is a D-dimensional noise vector, $\mathbf{I}$ is an identity matrix, and $\boldsymbol{\Psi}$ is a diagonal matrix. The latent variables are assumed to be drawn from independent normal-distributions with zero means and unit variances. Similarly, the noise terms are assumed to have independent and zero-mean normal distributions.

For a given data vector $\mathbf{x}_n$, the marginal probability can be calculated by integrating over all possible values of the latent variables $\mathbf{z}_n$, which results in a normal distribution given by [Zhao and Yu 2009]

$$\begin{aligned} p\left(\mathbf{x}_n \,\middle|\, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) &= \int p\left(\mathbf{x}_n \,\middle|\, \mathbf{z}_n, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) p\left(\mathbf{z}_n\right) d\mathbf{z}_n \\ &= \mathcal{N}\left(\mathbf{x}_n \,\middle|\, \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \boldsymbol{\Psi}\right). \end{aligned} \tag{7.4}$$

The EM method can be used to find the MLE of $\mathbf{W}$ and $\boldsymbol{\Psi}$, while the MLE of $\boldsymbol{\mu}$ is the empirical mean of the observations. However, finding the proper dimension of the latent variables for an FA model is problematic and can lead to overfitting, as when K is chosen too high, or underfitting, when K is selected too low. This shortcoming was resolved by employing a Bayesian treatment to find the latent space dimensionality using ARD and therefore to avoid overfitting [Beal 2003, Bishop 2006, Zhao and Yu 2009]. This was done by introducing prior probabilities over the model parameters $\boldsymbol{\theta} = \left\{\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}^{-1}\right\}$. Since inferring the model is analytically intractable, variational inference was applied to approximate the posterior probabilities. Although VBFA is a powerful method, it uses Gaussian latent-variables. The Gaussian distribution is known to be sensitive to outliers and noise [Gelman et al. 2013, Wu et al. 2009].

7.3 BAYESIAN ROBUST FACTOR ANALYSIS

As mentioned in Section 7.2, VBFA uses independent Gaussian distributions to represent latent variables. However, it was shown that Gaussian distribution is sensitive to noise [Gelman et al. 2013, Wu et al. 2009]. This shortcoming of VBFA is likely to prevent finding a robust representation of data in lower-dimension space especially when applied to an inherently noisy and uncertain data such as EEG. Therefore, the objective of RFA is to find a robust lower-dimensional representation of data which is less sensitive to outliers. To this end, we assume that latent variables have independent Student-t distributions. The latter is a heavy-tailed alternative of normal distribution and has been widely used in various applications to increase the robustness of models [Huang et al. 2017, Nguyen and Wu 2012, Sundar et al. 2012, Tipping and Lawrence 2005, Wu et al. 2009, Zhu et al. 2013]. The probability distribution of Student-t is mathematically expressed as [Bishop 2006]

$$St\left(z \mid \mu, \lambda, \nu\right) = \frac{\Gamma\left(\left(\nu+1\right)/2\right)}{\Gamma\left(\nu/2\right)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda\left(z-\mu\right)^2}{\nu}\right]^{-\left(\nu+1\right)/2}, \tag{7.5}$$

where Γ is the Gamma function[1], $\nu > 0$ is the degrees of freedom, λ is the inverse scale, and μ is the mean. This is equivalent to introducing a Gamma distribution as the prior of the precision of a normal distribution and integrating the precision variable out, which is [Bishop 2006, Murphy 2012]

$$St\left(z \mid \mu, \lambda = a/b, \nu = 2a\right) = \int \mathcal{N}\left(z \mid \mu, \tau^{-1}\right) \mathcal{G}\left(\tau \mid a, b\right) d\tau, \tag{7.6}$$

where τ is the precision of the normal distribution, and a and b are the shape and rate parameters of the Gamma distribution, respectively.

Incorporating Equation (7.6) into the VBFA model results in a model with latent variables that have Student-t distribution as their marginal probabilities. This is achieved by assuming independent Gaussian latent variables with unknown precision values and introducing Gamma distributions as the priors over precision values. Assume that the D dimensional observed data $\mathbf{X} = \left\{\mathbf{X}_1, \ldots, \mathbf{X}_S\right\}$ have been collected from S sessions $\left(s = 1, \ldots, S\right)$ and session s has N_s i.i.d. observations $\mathbf{X}_s = \left\{\mathbf{x}_1, \ldots, \mathbf{x}_{N_s}\right\}$. The goal is to find a lower-dimensional ($K < D$) latent

[1] $\Gamma\left(x\right) = \int_0^\infty z^{x-1} e^{-z} dz$

space such that

$$\mathbf{x}_{s,n} = \mathbf{W}\mathbf{z}_{s,n} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{s,n}, \tag{7.7}$$

$$\mathbf{z}_{s,n} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{-1}\right), \tag{7.8}$$

$$\lambda_k \sim \mathcal{G}\left(a_\lambda, b_\lambda\right), \tag{7.9}$$

where $\boldsymbol{\Lambda}$ is a positive diagonal matrix with diagonal elements of $\left\{\lambda_1, \ldots, \lambda_K\right\}$. The prior over mean ($\boldsymbol{\mu}$) and noise ($\boldsymbol{\varepsilon}$) vectors is assumed to be a set of independent normal-Gamma distributions, given by

$$p\left(\boldsymbol{\mu}, \boldsymbol{\Psi}\right) = \prod_{d=1}^{D} \mathcal{N}\left(\mu_d \,\middle|\, 0, \left(\beta_0 \psi_d\right)^{-1}\right) \mathcal{G}\left(\psi_d \,\middle|\, a_\psi, b_\psi\right), \tag{7.10}$$

where μ_d and ψ_d^{-1} are mean and noise variance of d^{th} feature, respectively, and β_0, a_ψ, and b_ψ are hyperparameters. Each column of the loading matrix $\mathbf{W}$ corresponds to a potential latent variable. Therefore, regularizing the columns of the loading matrix leads to automatic selection of appropriate dimensionality of the latent space. To this end, a set of ARD inspired hierarchical distributions were introduced for each column of the loading matrix as

$$p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right) = \prod_{k=1}^{K} p\left(\mathbf{w}_k \,\middle|\, \alpha_k\right) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k \,\middle|\, \mathbf{0}, \left(\alpha_k \mathbf{I}\right)^{-1}\right), \tag{7.11}$$

$$p\left(\boldsymbol{\alpha}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,\middle|\, a_\alpha, b_\alpha\right), \tag{7.12}$$

where α_k is the precision of the column k of the loading matrix $\mathbf{w}_k$. As a result, as the posterior probability of α_k moves towards the larger values, the variance of the $\mathbf{w}_k$ drops and therefore its values get smaller and shift towards zero. This effectively turns off extra latent variables and selects the appropriate number of components needed. Figure 7.1 depicts the probabilistic graphical representation of RFA.

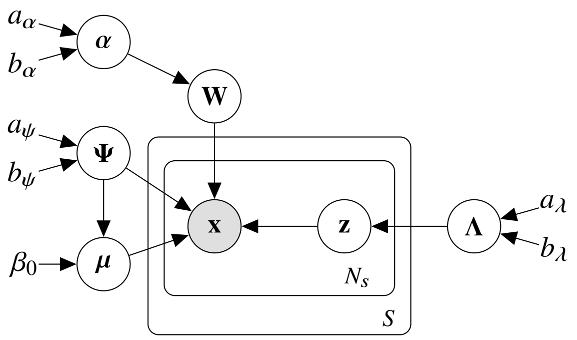


Figure 7.1 Graphical model representation of the Bayesian robust factor analysis (RFA) model.

7.4 VARIATIONAL INFERENCE

7.4.1 Training phase

Developing a full Bayesian treatment of the proposed model is analytically intractable. Therefore, a variational EM was used to approximate the posterior distributions of the model parameters, hyperparameters, and latent variables from the training data. The variational posterior was assumed to have a factorized form as

$$q\left(\Theta\right) = q\left(\boldsymbol{\Lambda}\right) q\left(\boldsymbol{\mu} \mid \boldsymbol{\Psi}\right) q\left(\boldsymbol{\Psi}\right) q\left(\mathbf{W}\right) q\left(\boldsymbol{\alpha}\right) \prod_{s=1}^{S} \prod_{n=1}^{N_s} q\left(\mathbf{z}_{s,n}\right), \tag{7.13}$$

where $\Theta = \left\{\boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{W}, \boldsymbol{\alpha}, \mathbf{Z}\right\}$ is the collection of all the unknowns, i.e., model parameters, hyperparameters, and latent variables. This was the only assumption to approximate the posterior probabilities [Bishop 2006]. Following Equation (3.7) and substituting the variational distribution given by Equation (7.13), the lower bound of the marginal log-likelihood $\mathcal{L}$ has the form of

$$\begin{aligned}\mathcal{L} = &\left\langle \ln\left(\frac{p\left(\boldsymbol{\alpha}\right)}{q\left(\boldsymbol{\alpha}\right)}\right)\right\rangle + \left\langle \ln\left(\frac{p\left(\mathbf{W} \mid \boldsymbol{\alpha}\right)}{q\left(\mathbf{W}\right)}\right)\right\rangle + \left\langle \ln\left(\frac{p\left(\boldsymbol{\Psi}\right)}{q\left(\boldsymbol{\Psi}\right)}\right)\right\rangle \\ &+ \left\langle \ln\left(\frac{p\left(\boldsymbol{\mu} \mid \boldsymbol{\Psi}\right)}{q\left(\boldsymbol{\mu} \mid \boldsymbol{\Psi}\right)}\right)\right\rangle + \left\langle \ln\left(\frac{p\left(\boldsymbol{\Lambda}\right)}{q\left(\boldsymbol{\Lambda}\right)}\right)\right\rangle + \sum_{s=1}^{S}\sum_{n=1}^{N_s} \left\langle \ln\left(\frac{p\left(\mathbf{z}_{s,n} \mid \boldsymbol{\Lambda}\right)}{q\left(\mathbf{z}_{s,n}\right)}\right)\right\rangle \\ &+ \sum_{s=1}^{S}\sum_{n=1}^{N_s} \left\langle \ln\left(p\left(\mathbf{x}_{s,n} \mid \mathbf{W}, \boldsymbol{\mu}, \mathbf{z}_{s,n}, \boldsymbol{\Psi}\right)\right) \right\rangle,\end{aligned} \tag{7.14}$$

where the last term is the objective of data fitting. Applying the variational method, the variational posterior distributions are given by (see Appendix A for the derivation)

$$q\left(\boldsymbol{\alpha}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \mid \tilde{a}_{\alpha}, \tilde{b}_{\alpha,k}\right), \tag{7.15}$$

$$q\left(\boldsymbol{\mu} \mid \boldsymbol{\Psi}\right) = \prod_{d=1}^{D} \mathcal{N}\left(\mu_d \,\middle|\, \tilde{\mu}_d, \left(\beta_{\mu}\psi_d\right)^{-1}\right), \tag{7.16}$$

$$q\left(\boldsymbol{\Psi}\right) = \prod_{d=1}^{D} \mathcal{G}\left(\psi_d \mid \tilde{a}_{\psi}, \tilde{b}_{\psi,d}\right), \tag{7.17}$$

$$q\left(\boldsymbol{\Lambda}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\lambda_k \mid \tilde{a}_{\lambda}, \tilde{b}_{\lambda,k}\right), \tag{7.18}$$

$$q\left(\mathbf{W}\right) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{w}_{d,\cdot}^{\top} \,\middle|\, \tilde{\mathbf{w}}_{w,d}, \tilde{\Sigma}_{w,d}\right), \tag{7.19}$$

$$q\left(\mathbf{Z}_s\right) = \prod_{n=1}^{N_s} \mathcal{N}\left(\mathbf{z}_{s,n} \mid \tilde{\mathbf{m}}_{z,s,n}, \tilde{\Sigma}_{z}\right), \tag{7.20}$$

where $\mathbf{w}_{d,.}$ is a row vector and corresponds to the d^{th} row of $\mathbf{W}$. The variational posterior parameters are given by

$$\tilde{\Sigma}_z = \Big(\langle \mathbf{W}^\top \boldsymbol{\Psi} \mathbf{W} \rangle + \langle \boldsymbol{\Lambda} \rangle\Big)^{-1}, \tag{7.21}$$

$$\tilde{\mathbf{m}}_{z,s,n} = \tilde{\Sigma}_z \langle \mathbf{W}^\top \rangle \Big(\langle \boldsymbol{\Psi} \rangle \mathbf{x}_{s,n} - \langle \boldsymbol{\Psi} \boldsymbol{\mu} \rangle\Big), \tag{7.22}$$

$$\tilde{a}_\alpha = a_\alpha + \frac{D}{2}, \tag{7.23}$$

$$\tilde{b}_{\alpha,k} = b_\alpha + \frac{\langle \mathbf{w}_k^\top \mathbf{w}_k \rangle}{2}, \tag{7.24}$$

$$\beta_\mu = \sum_{s=1}^{S} N_s + \beta_0, \tag{7.25}$$

$$\tilde{m}_{\mu,d} = \frac{1}{\beta_\mu} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left(x_{s,n,d} - \langle \mathbf{w}_{d,.} \rangle \langle \mathbf{z}_{s,n} \rangle\right), \tag{7.26}$$

$$\tilde{a}_\psi = a_\psi + \sum_{s=1}^{S} \frac{N_s}{2}, \tag{7.27}$$

$$\begin{aligned}\tilde{b}_{\psi,d} = {} & b_{\psi,d} - \frac{\beta_\mu}{2} \tilde{m}_{\mu,d}^2 + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \Bigg(x_{s,n,d}^2 - 2 x_{s,n,d} \langle \mathbf{w}_{d,.} \rangle \langle \mathbf{z}_{s,n} \rangle \\ & + \operatorname{tr}\Big(\langle \mathbf{w}_{d,.}^\top \mathbf{w}_{d,.} \rangle \langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^\top \rangle \Big) \Bigg),\end{aligned} \tag{7.28}$$

$$\tilde{\Sigma}_{w,d} = \left(\langle \operatorname{diag}(\boldsymbol{\alpha}) \rangle + \langle \psi_d \rangle \sum_{s=1}^{S} \sum_{n=1}^{N_s} \langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^\top \rangle \right)^{-1}, \tag{7.29}$$

$$\tilde{\mathbf{m}}_{w,d} = \tilde{\Sigma}_{w,d} \langle \psi_d \rangle \sum_{s=1}^{S} \sum_{n=1}^{N_s} \Big(\langle \mathbf{z}_{s,n} \rangle \left(x_{s,n,d} - \langle \mu_d \rangle \right) \Big), \tag{7.30}$$

$$\tilde{a}_\lambda = a_\lambda + \frac{1}{2} \sum_{s=1}^{S} N_s, \tag{7.31}$$

$$\tilde{b}_{\lambda,k} = b_\lambda + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \langle z_{s,n,k}^2 \rangle. \tag{7.32}$$

Iterating over Equations (7.21)–(7.32) is a variational EM in which, at each iteration, the variational posterior distribution of the latent variables are updated using Equations (7.21) and (7.22) (VBE-step) and then the posterior probabilities of the other parameters are updated using Equations (7.23)–(7.32) to maximize the lower bound of the marginal log-likelihood $\mathcal{L}$ (VBM-step). The lower bound of the marginal log-likelihood can be used to monitor the convergence of the VBRFA model and stop the variational EM when the relative improvement of $\mathcal{L}$ drops below a predefined threshold.

To remove the redundant latent variables, the factor corresponding to the highest value of $\langle \alpha \rangle$ is temporarily removed at each iteration and the lower bound of the marginal log-likelihood of the data is calculated. The component is removed if the lower bound $\mathcal{L}$ improved, otherwise is retained. This step is done to increase the computational speed, as the ARD would make the weights of extra latent variables very small but would not remove them [Bishop 2006, Zhao et al. 2015b].

Initialization of the variational formulation is done by setting the dimension of latent space to $K = D - 1$, the initial value of $\tilde{\mathbf{m}}_\mu$ to the empirical mean of the data, and using the first K components of the PCA coefficients and latent variables of the concatenated data as the initial values for $\tilde{\mathbf{M}}_w$ and $\tilde{\mathbf{M}}_z$, respectively. Although initializing with PCA might not lead to the global optimum, the results can be reproduced if needed. The hyperparameters of the prior distributions $\{a_\alpha, b_\alpha, a_\psi, b_\psi, a_\lambda, b_\lambda, \beta_0\}$ are initialized to a small value, e.g., 10^{-6}, to make the prior distributions uninformative. These hyperparameters could then be optimized at variational EM iterations to increase the lower bound of the marginal log-likelihood $\mathcal{L}$. By calculating the derivative of the $\mathcal{L}$ with respect to β_0, a closed-form updating equation can be found, which is

$$\beta_0 = \frac{D}{D\beta_\mu + \tilde{a}_\psi \sum_{d=1}^{D} \left(\tilde{m}_{\mu,d}^2 / \tilde{b}_{\psi,d}\right)}. \tag{7.33}$$

However, solving the derivative of the $\mathcal{L}$ for the hyperparameters of the Gamma distributions, i.e., $\{a_\alpha, b_\alpha, a_\psi, b_\psi, a_\lambda, b_\lambda\}$, does not yield closed-form solutions and therefore iterative optimization methods are required. These hyperparameters can be updated by iterating over and solving the followings:

$$\Psi\left(a_\alpha\right) = \ln\left(b_\alpha\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\Psi\left(\tilde{a}_\alpha\right) - \ln\left(\tilde{b}_{\alpha,k}\right)\right), \tag{7.34}$$

$$\Psi\left(a_\psi\right) = \ln\left(b_\psi\right) + \frac{1}{D}\sum_{d=1}^{D}\left(\Psi\left(\tilde{a}_\psi\right) - \ln\left(\tilde{b}_{\psi,d}\right)\right), \tag{7.35}$$

$$\Psi\left(a_\lambda\right) = \ln\left(b_\lambda\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\Psi\left(\tilde{a}_\lambda\right) - \ln\left(\tilde{b}_{\lambda,k}\right)\right), \tag{7.36}$$

$$b_\alpha^{-1} = \frac{1}{a_\alpha K}\sum_{k=1}^{K}\frac{\tilde{a}_\alpha}{\tilde{b}_{\alpha,k}}, \tag{7.37}$$

$$b_\psi^{-1} = \frac{1}{a_\psi D}\sum_{d=1}^{D}\frac{\tilde{a}_\psi}{\tilde{b}_{\psi,d}}, \tag{7.38}$$

$$b_\alpha^{-1} = \frac{1}{a_\lambda K}\sum_{k=1}^{K}\frac{\tilde{a}_\lambda}{\tilde{b}_{\lambda,k}}, \tag{7.39}$$

where Ψ is the digamma function[1]. Since the convergence speed of hyperparameters optimization is slow, we only updated the hyperparameters every 10 iterations. The pseudo-code of VBRFA is presented in Algorithm 7.1.

[1] $\Psi\left(x\right) = \frac{d}{dx}\ln\left(\Gamma\left(x\right)\right)$

Algorithm 7.1 The training algorithm of variational Bayesian robust factor analysis (VBRFA).

```
procedure Initializing
    $K = D - 1$
    $a_\alpha = b_\alpha = a_\psi = b_\psi = a_\lambda = b_\lambda = \beta_0 = 10^{-6}$
    $\tilde{a}_\psi = a_\psi$, $\tilde{\mathbf{b}}_\psi = b_\psi$, $\tilde{a}_\lambda = a_\lambda$, $\tilde{\mathbf{b}}_\lambda = b_\lambda$, $\tilde{a}_\alpha = a_\alpha$, $\tilde{\mathbf{b}}_\alpha = b_\alpha$, $\beta_\mu = \beta_0$
    $\tilde{\Sigma}_{w,d} = \mathbf{I}, \forall d \in \{1, \ldots, D\}$
    Set $\tilde{\mathbf{m}}_w$ to the first $K$ components of the PCA coefficients of the concatenated dataset.
    Set $\tilde{\mathbf{m}}_\mu$ to the empirical mean of the concatenated dataset.
    $RelTol = 10^{-6}$, $MaxIter = 1000$
for $iter = 1$ to $MaxIter$ do
    procedure VBE-step
        for $s = 1$ to $S$ do
            for $n = 1$ to $N_s$ do
                Update expectations of the latent variables using Equations (7.21) and (7.22)
    procedure VBM-step
        Update the variational parameters using Equations (7.23)–(7.32)
    procedure Update hyperparameters
        if Reminder($iter$, 10) is 0 then
            Update $\beta_0$ using Equation (7.33)
            Iterate over Equations (7.34)–(7.39) to until convergence.
    procedure Stopping criteria
        Calculate the lower bound of the marginal log-likelihood.
        if $\frac{\mathcal{L}(iter) - \mathcal{L}(iter-1)}{|\mathcal{L}(iter)|} < RelTol$ then
            Stop                                                        ▷ Converged
    procedure Pruning components
        Temporarily remove the component corresponding to the highest $\langle \alpha \rangle$.
        Calculate the lower bound of the marginal log-likelihood.
        if $\mathcal{L}$ (after pruning) $>$ $\mathcal{L}$ (before pruning) then      ▷ The component can be removed.
            $\mathcal{L}(iter) = \mathcal{L}$ (after pruning)
            Remove the component.
        else                                            ▷ The component can not be removed.
            $\mathcal{L}(iter) = \mathcal{L}$ (before pruning)
            Keep the component.
```

7.4.2 Testing phase

In this research, we used the latent variables as meta-features for a classification task, i.e., detection and prediction of microsleeps. Therefore, it is desirable to find an estimation of the latent variables for new and unseen data. A simple approach is to estimate the MAP of the latent variables based on the approximated variational posteriors from the training data without an update. This is achieved by

$$\mathbf{m}^*_{z,n} = \tilde{\Sigma}_z \tilde{\mathbf{M}}_w^\top \,\mathrm{diagmat}\left(\begin{bmatrix} \tilde{a}_\psi/\tilde{b}_{\psi,1} \\ \vdots \\ \tilde{a}_\psi/\tilde{b}_{\psi,D} \end{bmatrix}\right)\left(\mathbf{x}_n - \tilde{\mathbf{m}}_\mu\right) \tag{7.40}$$

where ‘diagmat’ is a diagonal matrix with the specified diagonal elements. This method, however, might not be accurate as the posterior distributions are fixed. A fully Bayesian treatment is needed to find the marginal posterior distributions of the latent variables given a new dataset. A variational EM is applied to find an approximation of the posterior probabilities that maximizes the log-likelihood of the posterior predictive distribution of a given data point. The latter is given by

$$\begin{aligned} p\left(\mathbf{x}_n \mid \mathcal{D}_{\text{train}}\right) &= \int p\left(\mathbf{x}_n \mid \mathbf{W}, \boldsymbol{\mu}, \mathbf{z}_n, \boldsymbol{\Psi}\right) p\left(\mathbf{W}, \alpha, \boldsymbol{\mu}, \boldsymbol{\Psi}, \boldsymbol{\Lambda} \mid \mathcal{D}_{\text{train}}\right) p\left(\mathbf{z}_n \mid \boldsymbol{\Lambda}\right) d\Theta \\ &= \int p\left(\mathbf{x}_n \mid \mathbf{W}, \boldsymbol{\mu}, \mathbf{z}_n, \boldsymbol{\Psi}\right) q\left(\mathbf{W}\right) q\left(\boldsymbol{\mu}, \boldsymbol{\Psi}\right) q\left(\boldsymbol{\Lambda}\right) p\left(\mathbf{z}_n \mid \boldsymbol{\Lambda}\right) d\Theta, \end{aligned} \tag{7.41}$$

where $\mathcal{D}_{\text{train}}$ is the training data. Similar to the training variational formulation, the parameters of the variational posterior approximations are given by

$$\hat{\Sigma}_z^n = \left(\langle\boldsymbol{\Lambda}\rangle + \langle\mathbf{W}^\top\boldsymbol{\Psi}\mathbf{W}\rangle\right)^{-1}, \tag{7.42}$$

$$\hat{\mathbf{m}}_z^n = \hat{\Sigma}_z^n\langle\mathbf{W}\rangle^\top\langle\boldsymbol{\Psi}\rangle\left(\mathbf{x}_n - \langle\boldsymbol{\mu}\rangle\right), \tag{7.43}$$

$$\hat{\Sigma}_{w,d}^n = \left(\langle\psi_d\rangle\langle\mathbf{z}_n\mathbf{z}_n^\top\rangle + \tilde{\Sigma}_{w,d}^{-1}\right)^{-1}, \tag{7.44}$$

$$\hat{\mathbf{m}}_{w,d}^n = \hat{\Sigma}_{w,d}^n\left(\tilde{\Sigma}_{w,d}^{-1}\,\tilde{\mathbf{m}}_{w,d} + \langle\psi_d\rangle\langle\mathbf{z}_n\rangle\left(x_{n,d} - \langle\mu_d\rangle\right)\right), \tag{7.45}$$

$$\hat{\beta}_\mu = 1 + \beta_\mu, \tag{7.46}$$

$$\hat{m}_{\mu,d} = \frac{1}{\hat{\beta}_\mu}\left(\beta_\mu\tilde{m}_{\mu,d} + x_{n,d} - \langle\mathbf{w}_{d,.}\rangle\langle\mathbf{z}_n\rangle\right), \tag{7.47}$$

$$\hat{a}_\psi = a_\psi + \frac{1}{2}, \tag{7.48}$$

$$\begin{aligned} \hat{b}_{\psi,d} = \tilde{b}_{\psi,d} &+ \frac{\beta_\mu}{2}\left(\tilde{m}_{\mu,d}\right)^2 - \frac{\hat{\beta}_\mu}{2}\left(\hat{m}_{\mu,d}\right)^2 \\ &+ \frac{1}{2}\left(x_{n,d}^2 - 2x_{n,d}\langle\mathbf{w}_{d,.}\rangle\langle\mathbf{z}_n\rangle + \mathrm{tr}\left(\langle\mathbf{z}_n\mathbf{z}_n^\top\rangle\langle\mathbf{w}_{d,.}^\top\mathbf{w}_{d,.}\rangle\right)\right), \end{aligned} \tag{7.49}$$

$$\hat{a}_\lambda = \tilde{a}_\lambda + \frac{1}{2}, \tag{7.50}$$

$$\hat{b}_{\lambda,k} = \tilde{b}_{\lambda,k} + \frac{1}{2}\langle z_{n,k}^2\rangle, \tag{7.51}$$

where all the expectations are taken with respect to the test variational distributions. Each iteration of estimation of the variational parameters increases the log-likelihood of the predictive probability, starting from the training variational parameters, and usually converges within 3 to

4 iterations. Moreover, the updating equations for the posterior parameters of the predictive probability are similar to the ones of the training phase, except for α, with the assumption that the test data is included in the training data. Consequently, as the number of training instances increases, Equations (7.40) and (7.43) converge to the same values. The posterior probabilities of the test data can be computed in parallel, which is a result of the i.i.d. assumption.

7.5 RESULTS AND DISCUSSION

We applied the proposed VBRFA to various microsleep features to find lower-dimension sets of meta-features. The latter was then used to train and test classifiers to predict microsleeps. For each feature set, VBRFA was applied in two ways. First, the training data of all subjects were concatenated and then VBRFA was applied, which is referred to as *VBRFA-1* meta-feature set. If there existed a consistent common component across all the subjects, it was expected that applying feature reduction to concatenated data of all subjects would exploit it. Second, a VBRFA was applied to the data of individual training subjects, resulting in 7 feature reduction models. Then, the meta-features of all the feature reduction models were aggregated to create a larger meta-feature set, which is referred to as *VBRFA-2*. Using a separate model for each individual subject allows the meta-features to capture subject-specific patterns. In addition, we aggregated both of the VBRFA-1 and VBRFA-2 meta-feature sets into a larger meta-feature set which is referred to as *VBRFA-3*. Since the VBRFA-3 meta-feature set contains the other two meta-features, it was expected to achieve a higher performance.

Figure 7.2 shows an example of lower bound convergence for the proposed VBRFA model. For this example, a concatenated dataset comprised PSF of 7 subjects was used. Original data had 192 features. After 79 iterations, VBRFA found that 19 was the optimum number of latent variables to explain the original data. The training process took 63 s. After convergence, the value of lower bound can be used as an approximation to the marginal likelihood of data given the feature reduction model. A higher value of lower bound indicates better fit to data. However, because our aim was to detect and predict microsleeps, the values of lower bounds of different models were not compared.

7.5.1 Detection and prediction of microsleep states

We first examined the effect of EEG window length for feature extraction on detection performance. This was done by fixing the classifier to a single LDA and a $\tau = 0$ s, i.e., detection. Using these conditions, the performances of state detection of microsleeps for the three variants of VBRFA meta-features are shown in Table 7.1. AUC-ROC and AUC-PR were chosen as performance measures because they are threshold free and therefore more suitable to compare models. It was observed that features extracted from a 2-s window length had lower performances across all feature sets. However, EEG window lengths of 5 and 10 s had similar performances.

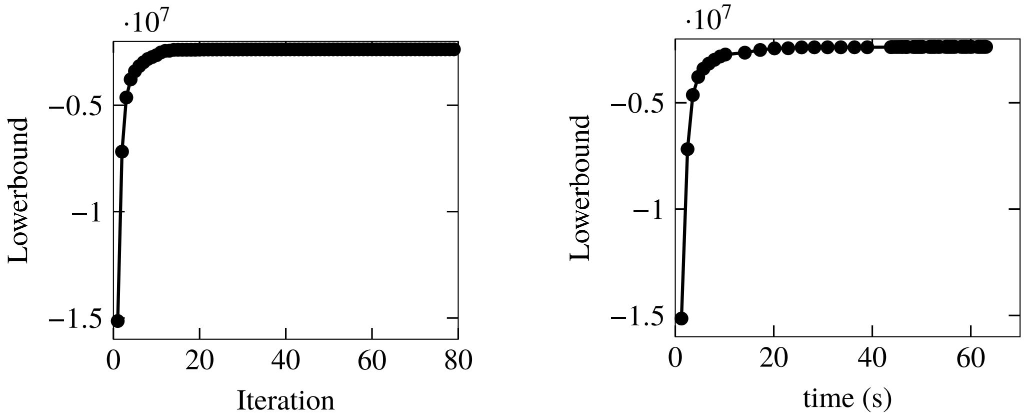


Figure 7.2 An example convergence curve for the lower bound ($\mathcal{L}$) of proposed VBRFA model.

Notwithstanding, the WEPF feature set extracted from 10-s windows of EEG had higher average performances relative to its shorter EEG windows.

Table 7.1 Performance (mean ± SE) of microsleep state detection, i.e., $\tau = 0$ s, and an LDA classifier on different VBRFA meta-feature sets. A bold value indicates the highest performances of each feature set and italics indicate the highest of overall. Two-tail Wilcoxon signed-rank tests were performed between the performances with VBRFA meta-features and baseline features and significant improvements were identified.

Feature set	Feature type	2-s EEG window		5-s EEG window		10-s EEG window	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.86±0.03	0.31±0.12	0.88±0.03	**0.36±0.12**	0.86±0.04	**0.36±0.12**
	VBRFA-1	0.81±0.04	0.27±0.11	0.83±0.06	0.31±0.13	0.84±0.05	0.29±0.12
	VBRFA-2	0.86±0.03	0.28±0.11	0.88±0.02	0.31±0.12	**0.89±0.03**	0.34±0.12
	VBRFA-3	0.85±0.03	0.28±0.11	**0.89±0.02**	0.31±0.12	**0.89±0.02**	0.33±0.12
PSF	Baseline	0.89±0.02	0.34±0.12	***0.92±0.01***	***0.41±0.12***	***0.92±0.02***	0.37±0.12
	VBRFA-1	0.83±0.04	0.29±0.11	0.90±0.02	0.37±0.12	0.90±0.03	0.36±0.12
	VBRFA-2	0.87±0.02	0.32±0.12	***0.92±0.02***	0.39±0.13	0.91±0.02	0.38±0.12
	VBRFA-3	0.88±0.02	0.33±0.12	***0.92±0.02***	***0.41±0.13***	***0.92±0.02***	0.39±0.12
PSF-IAF	Baseline	0.88±0.02	0.34±0.12	0.90±0.02	0.37±0.13	0.91±0.02	0.35±0.12
	VBRFA-1	0.83±0.04	0.28±0.12	0.89±0.03	0.34±0.13	0.89±0.04	0.33±0.12
	VBRFA-2	0.87±0.03	0.32±0.12	0.90±0.02	0.37±0.12	0.91±0.02	0.37±0.12
	VBRFA-3	0.88±0.02	0.34±0.12	0.91±0.02	**0.38±0.12**	***0.92±0.01***	0.36±0.12
WMSF	Baseline	0.87±0.03	0.35±0.12	**0.91±0.02**	**0.38±0.14**	**0.91±0.02**	0.37±0.13
	VBRFA-1	0.84±0.04	0.31±0.12	0.85±0.05	0.35±0.13	0.84±0.06	0.34±0.13
	VBRFA-2	0.87±0.03	0.32±0.12	**0.91±0.02**	0.35±0.12	**0.91±0.02**	0.35±0.12
	VBRFA-3	0.87±0.03	0.32±0.12	0.90±0.02	0.35±0.12	**0.91±0.02**	0.36±0.12
WLMSF	Baseline	0.88±0.03	0.36±0.13	0.91±0.02	***0.41±0.13***	0.91±0.02	0.37±0.13
	VBRFA-1	0.83±0.05	0.31±0.13	0.88±0.03	0.33±0.13	0.89±0.03	0.33±0.13
	VBRFA-2	0.88±0.03	0.35±0.13	***0.92±0.02***	***0.41±0.13***	0.91±0.02	0.36±0.13
	VBRFA-3	0.87±0.03	0.32±0.12	0.90±0.02	0.35±0.12	0.91±0.02	0.36±0.12
WEPF	Baseline	0.74±0.03	0.21±0.10	0.78±0.03	0.22±0.11	0.81±0.02	0.23±0.11
	VBRFA-1	0.74±0.04	0.15±0.06	0.74±0.03	0.15±0.07	0.80±0.03	0.17±0.06
	VBRFA-2	0.79±0.03*	0.23±0.10$^{\sim}$	0.82±0.03*	0.25±0.12	0.84±0.02	0.27±0.11
	VBRFA-3	0.87±0.03*	0.32±0.12$^{\sim}$	0.90±0.02**	0.35±0.12**	**0.91±0.02***	**0.36±0.12**

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

In terms of the VBRFA meta-feature sets, VBRFA-1 had the lowest average performance compared to other meta-features, whereas VBRFA-2 and VBRFA-3 had relatively similar mean performances which were also comparable to of the baselines. Two-tail Wilcoxon signed-rank tests were used to compare the performance of VBRFA meta-features versus baselines, see Section 6.5 for a description of the baseline performances. Significant improvements relative to the baseline were only found in WEPF.

Although most of the performances of VBRFA meta-features were not significantly superior, the numbers of meta-features were substantially lower than those of the baselines, as shown in Table 7.2. This can be interpreted as the lower-dimensions found by VBRFA have similar information to their corresponding baseline features, which is most likely a result of finding an uncorrelated latent space. Since the average performance of VBRFA-1 was low, while the average performances of VBRFA-2 and VBRFA-3 were similar and comparable to the baseline, for the rest of this chapter we only report the performances of the baseline and VBRFA-3.

Table 7.2 Average number of VBRFA meta-features.

Feature set	Feature type	Number of features			
		2 s EEG window	5 s EEG window	10 s EEG window	Baseline
MDF	VBRFA-1	22.9	24.8	27	176
	VBRFA-2	70	73	85	
	VBRFA-3	92.9	97.8	112	
PSF	VBRFA-1	16.4	19.9	22.5	192
	VBRFA-2	56	72	86	
	VBRFA-3	72.4	91.9	108.5	
PSF-IAF	VBRFA-1	17.4	19.4	21	192
	VBRFA-2	63	73	88	
	VBRFA-3	80.4	92.4	109	
WMSF	VBRFA-1	8	9.1	10	80
	VBRFA-2	39	43	48	
	VBRFA-3	47	52.1	58	
WLMSF	VBRFA-1	9.5	11.1	11.9	80
	VBRFA-2	34	41	49	
	VBRFA-3	43.5	52.1	60.9	
WEPF	VBRFA-1	10.3	10.8	10.8	80
	VBRFA-2	48	49	47	
	VBRFA-3	58.3	59.8	57.8	

We then combined the features extracted from various EEG windows to find a more complete representation of EEG characteristics. Thus, for any time point in the gold-standard, features corresponding to the previous 2, 5, and 10-s EEG segments were extracted and concatenated into one feature vector. It was expected that the extracted features from a short EEG window would characterize transient behaviours, while features of a long EEG window would correspond more to tonic changes. Therefore, aggregating features of multiple EEG windows was expected to improve the performance of microsleep detection. Table 7.3 shows the microsleep state detection with an LDA classifier. As expected, substantial improvements relative to single-window features were observed indicating that both the tonic and transit dynamics of EEG have information regarding microsleeps. Moreover, with an LDA classifier, the VBRFA-3 meta-features of PSF-IAF had the highest detection performance with AUC-ROC = 0.95, AUC-PR = 0.47, φ = 0.39, and GM = 0.80. Notwithstanding, the performances of VBRFA-3 meta-features were mostly comparable with the baseline features, with an exception of WEPF, while the number of features were lower.

Table 7.3 Performance (mean ± SE) of microsleep state detection using aggregated VBRFA-3 meta-features of multiple EEG windows with an LDA classifier. A bold value indicates the highest performances within individual feature sets, whereas italics indicate the highest among all feature sets.

Feature set	Feature type	Microsleep detection performance					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	0.90±0.02	**0.40±0.13**	**0.71±0.08**	**0.34±0.10**	**0.62±0.10**	0.33±0.12
	VBRFA-3	**0.92±0.02**	0.38±0.12	0.69±0.08	0.32±0.08	0.59±0.11	**0.34±0.12**
PSF	Baseline	**0.94±0.01**	0.43±0.12	0.74±0.06	0.36±0.08	0.68±0.11	***0.36±0.12***
	VBRFA-3	**0.94±0.02**	**0.46±0.12**	**0.77±0.06**	**0.38±0.08**	**0.71±0.10**	0.35±0.11
PSF-IAF	Baseline	0.94±0.01	0.44±0.12	0.76±0.05	0.37±0.08	0.70±0.10	***0.36±0.12***
	VBRFA-3	***0.95±0.01***	***0.47±0.12***	***0.80±0.04***	***0.39±0.08***	***0.74±0.09***	0.34±0.11
WMSF	Baseline	**0.92±0.02**	**0.40±0.14**	0.66±0.10	**0.33±0.09**	0.57±0.12	**0.31±0.11**
	VBRFA-3	**0.92±0.02**	0.39±0.13	**0.67±0.11**	**0.33±0.09**	**0.60±0.12**	0.30±0.11
WLMSF	Baseline	**0.94±0.01**	**0.44±0.13**	**0.76±0.07**	**0.36±0.10**	**0.70±0.11**	**0.31±0.12**
	VBRFA-3	0.92±0.02	0.39±0.13	0.67±0.11	0.33±0.09	0.60±0.12	0.30±0.11
WEPF	Baseline	0.84±0.02	0.27±0.12	**0.70±0.03**	0.25±0.07	**0.67±0.08**	0.25±0.13
	VBRFA-3	**0.92±0.02**	**0.39±0.13**	0.67±0.11	**0.33±0.09**	0.60±0.12	**0.30±0.11**

A single LDA has a simple structure and is computationally fast. But it might not be the best classifier for detection and prediction of microsleeps. Therefore, three additional classifiers, namely VBLR, TAN, and linear SVM, were also used for microsleep state detection/prediction (described in Section 6.3). Table 7.4 shows the detection performance of different classifiers across feature sets. It is evident that the linear SVM marginally outperformed in most cases, although LDA and VBLR had comparable performances. TAN had the lowest performance among classifiers. PSF-IAF and PSF had the highest AUC-ROC performances, but PSF had the highest AUC-PR. The remaining feature sets had relatively similar performances except for WEPF which had the lowest performance.

The prediction of microsleep states was then investigated by increasing the delay between the gold-standard and its corresponding EEG segment. The prediction time τ was set between 0 to 1 s with step of 0.25 s. Figure 7.3 depicts the average AUC-PR of different classifiers, feature sets, and prediction times. As expected, a drop in performance was observed with increasing prediction time. PSF had the highest detection (τ = 0 s) performance of the feature sets, whereas WLMSF and MDF had higher prediction performances at τ = 1.0 s. In addition, VBRFA-3 meta-features had similar or slightly lower performances compared to the baseline features, except for WEPF and WLMSF, where VBRFA-3 meta-features of WEPF had significantly higher performances in most cases. Furthermore, the linear SVM outperformed other classifiers

Table 7.4 Performance of different classifiers for microsleep state detection with aggregated VBRFA-3 meta-features of multiple EEG windows. A bold value indicates the highest performance among selected classifiers and an italic value indicates the highest overall.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.90	0.40	0.93	**0.48**	0.90	0.42	0.77	0.22
	VBRFA-3	0.92	0.38	**0.94**	0.46	0.91	0.40	0.77	0.22
PSF	Baseline	0.94	0.43	0.94	***0.49***	**0.94**	0.44	0.70	0.21
	VBRFA-3	0.94	0.46	***0.95***	0.48	0.94	0.47	0.81	0.30
PSF-IAF	Baseline	0.94	0.44	***0.95***	**0.47**	0.93	0.45	0.73	0.26
	VBRFA-3	***0.95***	**0.47**	***0.95***	0.46	0.94	0.44	0.80	0.29
WMSF	Baseline	0.92	0.40	**0.93**	0.42	0.92	0.39	0.77	0.22
	VBRFA-3	0.92	0.39	**0.93**	**0.43**	0.91	0.39	0.77	0.23
WLMSF	Baseline	**0.94**	0.44	0.93	**0.45**	0.92	**0.45**	0.76	0.25
	VBRFA-3	0.92	0.39	0.93	0.43	0.91	0.39	0.77	0.23
WEPF	Baseline	0.84	0.27	0.86	0.29	0.85	0.28	0.82	0.23
	VBRFA-3	0.92	0.39	**0.93**	**0.43**	0.91	0.39	0.77	0.23

in most cases.

Figure 7.4 shows other performance measures of linear SVM for microsleep state prediction for τ = 0–1 s. Surprisingly, microsleep state prediction performance in terms of phi and GM was higher with the baseline PSF and PSF-IAF, compared to their corresponding VBRFA-3 meta-features, although their AUC-PR were relatively similar. Baseline PSF had the highest detection performance, i.e., $\varphi = 0.40$ and GM = 0.79. With a prediction time of $\tau = 1.0$ s, however, baseline MDF had the best performance, i.e., $\varphi = 0.35$ and GM = 0.76. Using VBRFA-3 meta-features, PSF-IAF had the highest detection performance of microsleep state with a single LDA classifier, i.e., GM = 0.80 and $\varphi = 0.39$. The same combination of meta-feature set and classifier also had the highest performance with a prediction time of $\tau = 1.0$ s, i.e., GM = 0.72 and $\varphi = 0.34$.

As expected, a decline in performance with prediction time τ was observed, which was similar for both baseline features and VBRFA-3 meta-features. Notwithstanding, a faster decline in performance was observed with VBRFA-3 meta-features as compared to baseline features. This might indicate that the lower-dimension representation of features using VBRFA, i.e., VBRFA-3 meta-features, has slightly less information than the original features.

The features and models that achieved the highest performances in terms of phi and GM were selected for sensitivity and precision analysis. The average sensitivity, specificity, and precision of the baseline PSF using a linear SVM were 0.74, 0.89, and 0.38, respectively. Similarly, using an LDA with PSF-IAF meta-features resulted in an average sensitivity of 0.74, specificity of 0.89, and precision of 0.34. Interestingly, the average sensitivity and specificity of both models are very close, while the precision varied substantially. This might be due to the inter-subject variability of imbalance ratios, where a false microsleep prediction in a subject with a lower number of microsleeps has a higher impact on precision. Nevertheless, despite relatively high AUC-ROC of both models, the sensitivity was moderate with low precision. This indicates a high number of false positives with a few false negatives.

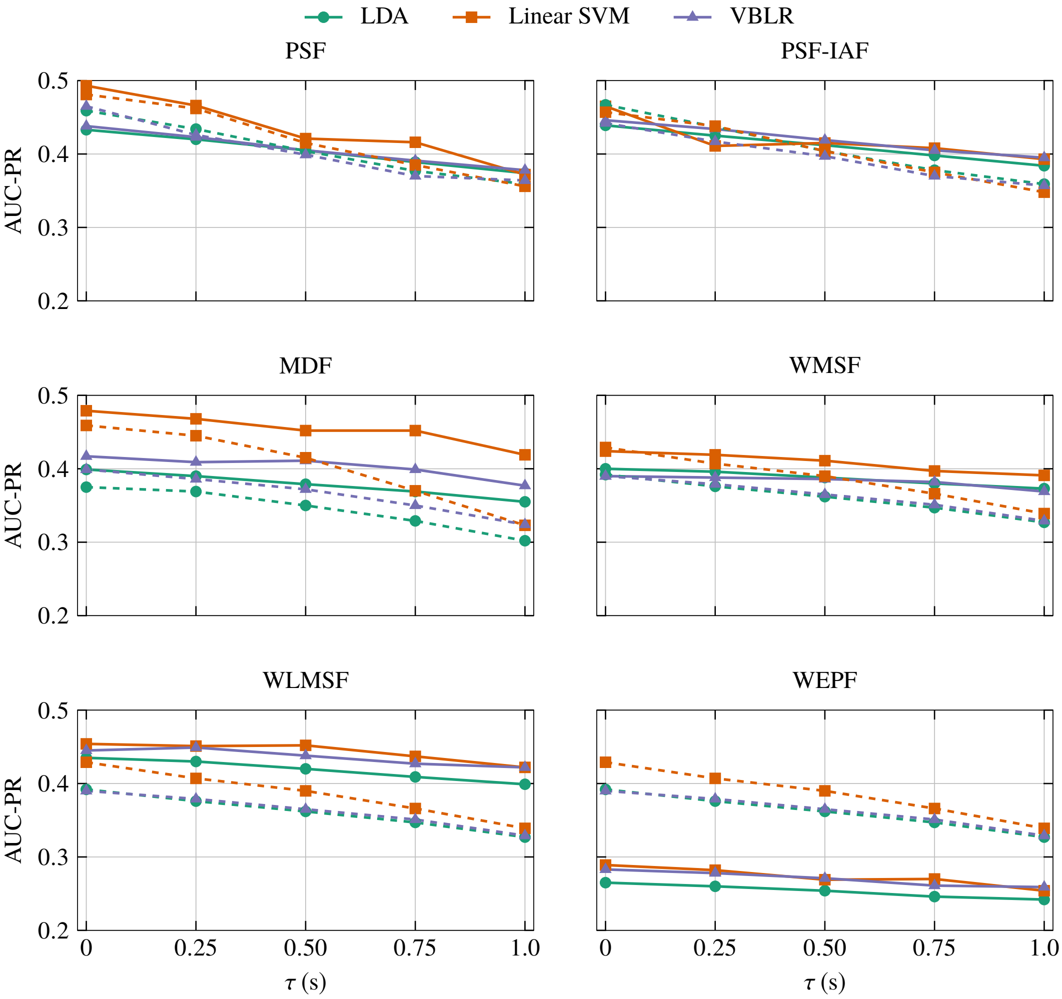


Figure 7.3 Performance of microsleep state prediction in terms of AUC-PR using VBRFA-3 meta-features and different classifiers versus prediction time for τ = 0–1 s. Solid lines correspond to the baseline features and dashed lines correspond to the VBRFA-3 meta-features.

7.5.2 Detection and prediction of microsleep onset

The ultimate goal is to predict and identify imminent microsleep events. Due to higher performance of aggregated features extracted from multiple EEG windows, the results of this section are limited to the aggregated features of baseline and VBRFA-3. Table 7.5 shows the results of microsleep onset detection at $\tau = 0$ s for various feature sets. In terms of AUC-ROC, VBRFA-3 meta-features had slightly higher averages in five of the feature sets with the highest value of 0.89. The highest AUC-PR was achieved with VBRFA-3 meta-features of PSF-IAF, i.e., AUC-PR = 0.05. Interestingly, the highest performance of microsleep onset detection in terms of phi was achieved with the baseline MDF (φ = 0.09), although the baseline MDF AUC-ROC, AUC-PR, and GM were slightly lower than other feature sets.

The average performance of onset detection using a single LDA with VBRFA-3 meta-features and baseline features were relatively similar. Table 7.6 shows the average performance of microsleep onset detection with different classifiers. The linear SVM classifier marginally

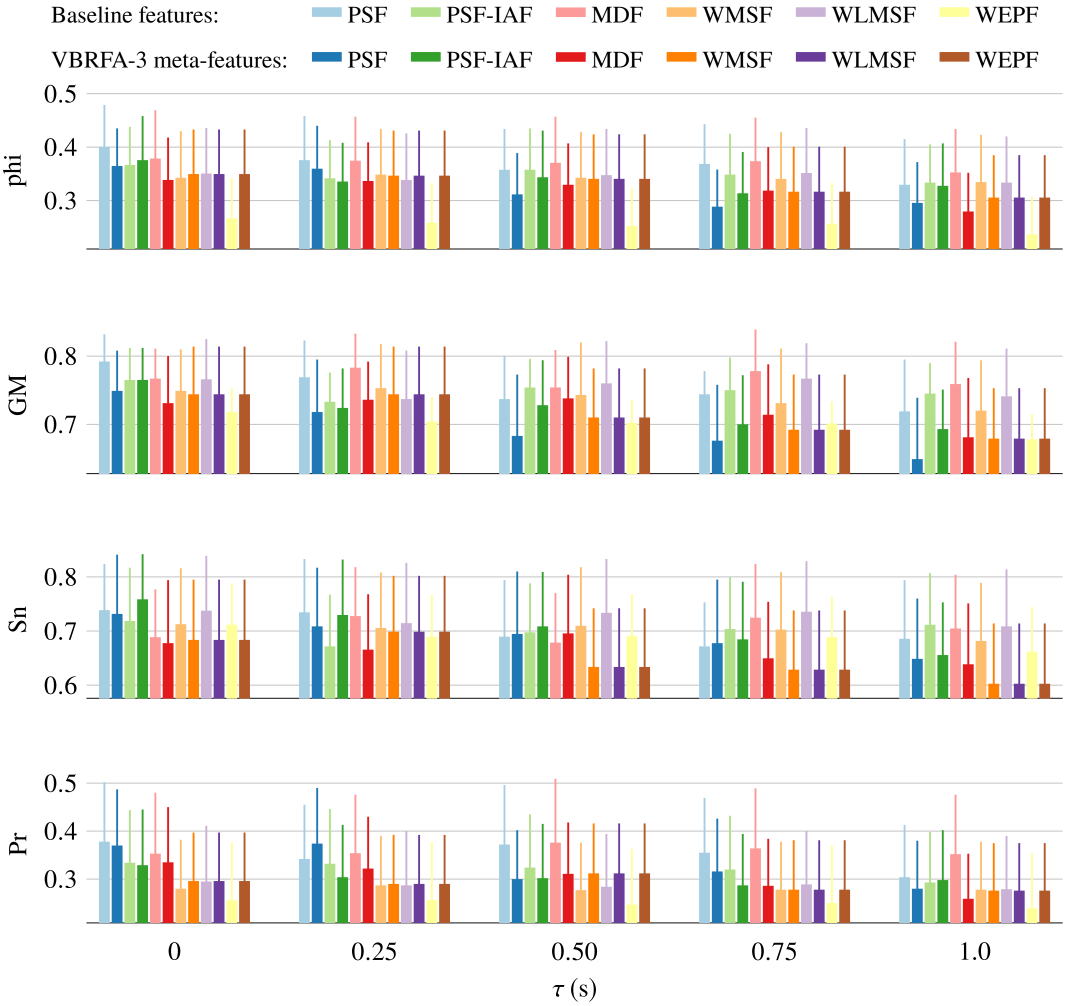


Figure 7.4 Performance (mean ± SE) of microsleep state prediction using a single linear SVM classifier and VBRFA-3 meta-features for τ = 0–1.0 s.

outperformed other classifiers, and the TAN classifier had the lowest performance in all cases. It is interesting that the performance of linear SVM was higher with baseline features, whereas LDA and VBLR had higher performances with VBRFA-3 meta-features in terms of AUC-ROC.

The highest performance of microsleep onset detection was achieved with baseline PSF and a linear SVM classifier (AUC-ROC = 0.91; AUC-PR = 0.09). Other feature sets had slightly lower performances, with the exception of baseline WEPF which had substantially lower performances.

Figure 7.5 represents the performance of microsleep onset prediction in terms of AUC-ROC with different prediction times up to τ = 10 s. As expected, a decline in performance with increased prediction time τ was observed. In addition, the performance of TAN with VBRFA-3 meta-features was mostly higher than TAN with baseline features. The linear SVM had the highest performance in almost all cases and LDA was the second best classifier. With τ = 0 s, the best performance of linear SVM was achieved with baseline PSF (AUC-ROC = 0.91;

Table 7.5 Performance (mean ± SE) of microsleep onset detection using aggregated VBRFA-3 meta-features of multiple EEG windows with an LDA classifier. A bold value indicates the highest performances in individual feature sets, whereas an italic indicates the highest among all feature sets. Significant improvements were identified with two-tail Wilcoxon signed-rank tests.

Feature set	Method	Microsleep onset detection performance					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	**0.87±0.02**	**0.03±0.01**	**0.69±0.05**	***0.09±0.02***	**0.55±0.08**	***0.02±0.01***
	VBRFA-3	0.86±0.02	0.02±0.01	0.64±0.07	0.07±0.02	0.49±0.09	***0.02±0.01***
PSF	Baseline	0.87±0.01	**0.04±0.03**	**0.71±0.04**	**0.07±0.01**	0.63±0.09	***0.02±0.01***
	VBRFA-3	***0.89±0.01***$^{\sim}$	**0.04±0.02**	**0.71±0.04**	**0.07±0.01**	0.66±0.09	0.01±0.00
PSF-IAF	Baseline	0.87±0.01	0.04±0.03	***0.73±0.04***	**0.08±0.01**	0.66±0.09	***0.02±0.01***
	VBRFA-3	**0.88±0.01****	***0.05±0.02***	***0.73±0.03***	0.07±0.01	**0.68±0.08**	0.01±0.00
WMSF	Baseline	0.88±0.02	0.03±0.02	**0.69±0.07**	**0.08±0.02**	**0.60±0.11**	***0.02±0.01***
	VBRFA-3	***0.89±0.02***	**0.04±0.02**	0.63±0.10	0.07±0.02	0.55±0.12	***0.02±0.01***
WLMSF	Baseline	0.88±0.02	***0.05±0.03***	***0.73±0.05***	**0.07±0.01**	***0.70±0.10***	0.01±0.00
	VBRFA-3	***0.89±0.02***	0.04±0.02	0.63±0.10	**0.07±0.02**	0.55±0.12	***0.02±0.01***
WEPF	Baseline	0.73±0.04	0.01±0.01	**0.63±0.05**	0.04±0.01	**0.60±0.11**	0.01±0.00
	VBRFA-3	***0.89±0.02****	**0.04±0.02**	**0.63±0.10**	**0.07±0.02**	0.55±0.12	***0.02±0.01***

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table 7.6 Performance of different classifiers for microsleep onset detection with aggregated VBRFA-3 meta-features. A bold value indicates the highest performance among selected classifiers, whereas an italic value indicates the highest performance among individual feature sets. Significant improvements were identified with two-tail Wilcoxon signed-rank tests.

Feature set	Method	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.87	0.03	***0.91***	**0.06**	0.82	0.05	0.72	0.01
	VBRFA-3	**0.86**	0.02	**0.86**	**0.06**	0.83	0.03	0.76	0.01
PSF	Baseline	0.87	0.04	***0.91***	***0.09***	0.83	0.04	0.70	0.01
	VBRFA-3	0.89$^{\sim}$	0.04	**0.90**	**0.06**	0.85	0.03	0.84*	0.02*
PSF-IAF	Baseline	0.87	0.04	***0.91***	**0.07**	0.81	0.04	0.70	0.01
	VBRFA-3	0.88**	0.05	**0.90**	**0.06**	0.85*	0.03	0.83**	0.02$^{\sim}$
WMSF	Baseline	0.88	0.03	**0.90**	**0.05**	0.80	0.02	0.74	0.01
	VBRFA-3	**0.89**	0.04	**0.89**	**0.06**	0.85**	0.04*	0.74	0.01
WLMSF	Baseline	0.88	0.05	**0.90**	**0.06**	0.83	0.04	0.79	0.02
	VBRFA-3	**0.89**	0.04	**0.89**	**0.06**	0.85	0.04	0.74	0.01
WEPF	Baseline	0.73	0.01	**0.78**	**0.02**	0.74	0.01	0.76	0.01
	VBRFA-3	**0.89***	0.04	**0.89****	**0.06**$^{\sim}$	0.85*	0.04*	0.74	0.01

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

AUC-PR = 0.09), while LDA had the highest performance with VBRFA-3 meta-features of WEPF (AUC-ROC = 0.89; AUC-PR = 0.04). With a prediction time of $\tau = 10$ s, however, a TAN classifier with VBRFA-3 meta-features of PSF had the highest AUC-ROC of 0.75, and a linear SVM with baseline WLMSF features had the second best performance (AUC-ROC = 0.74).

Since a linear SVM outperformed other classifiers in most cases, the rest of the results of microsleep state prediction with VBRFA meta-features are limited to those of a linear SVM classifier. Figure 7.6 shows the rest of performance measures for microsleep onset prediction

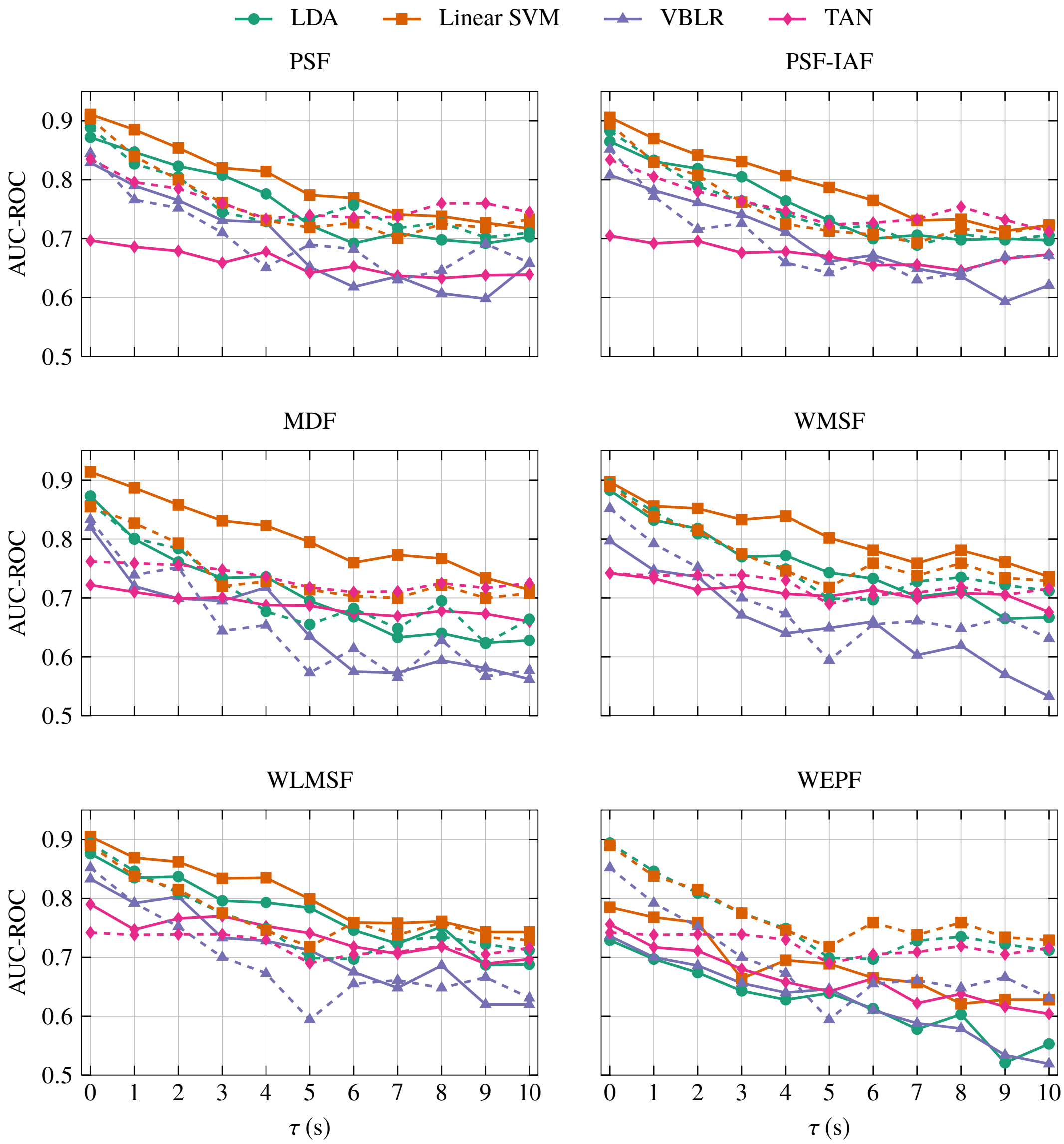


Figure 7.5 AUC-ROC of microsleep onset prediction using different classifiers versus prediction time τ up to 10 s. Solid lines correspond to the baseline features and dashed lines correspond to the VBRFA-3 meta-features.

with different values of τ. It was observed that all performance measures dropped substantially by increasing the prediction time. The precision of most feature sets dropped below 0.01 around a prediction time of $\tau = 2$ s, which then reached a plateau state. However, the sensitivities declined with the increment of τ with a maximum of 0.79 with a prediction time of $\tau = 0$ s using baseline PSF and a minimum of 0.53 with $\tau = 6$ s using the baseline WEPF. In addition, the maximum phi and GM were achieved with baseline MDF and a prediction time of $\tau = 0$ s ($\varphi = 0.08$; GM = 0.78), whereas the baseline PSF and WEPF had the lowest GM (0.45) and phi (0.01), respectively.

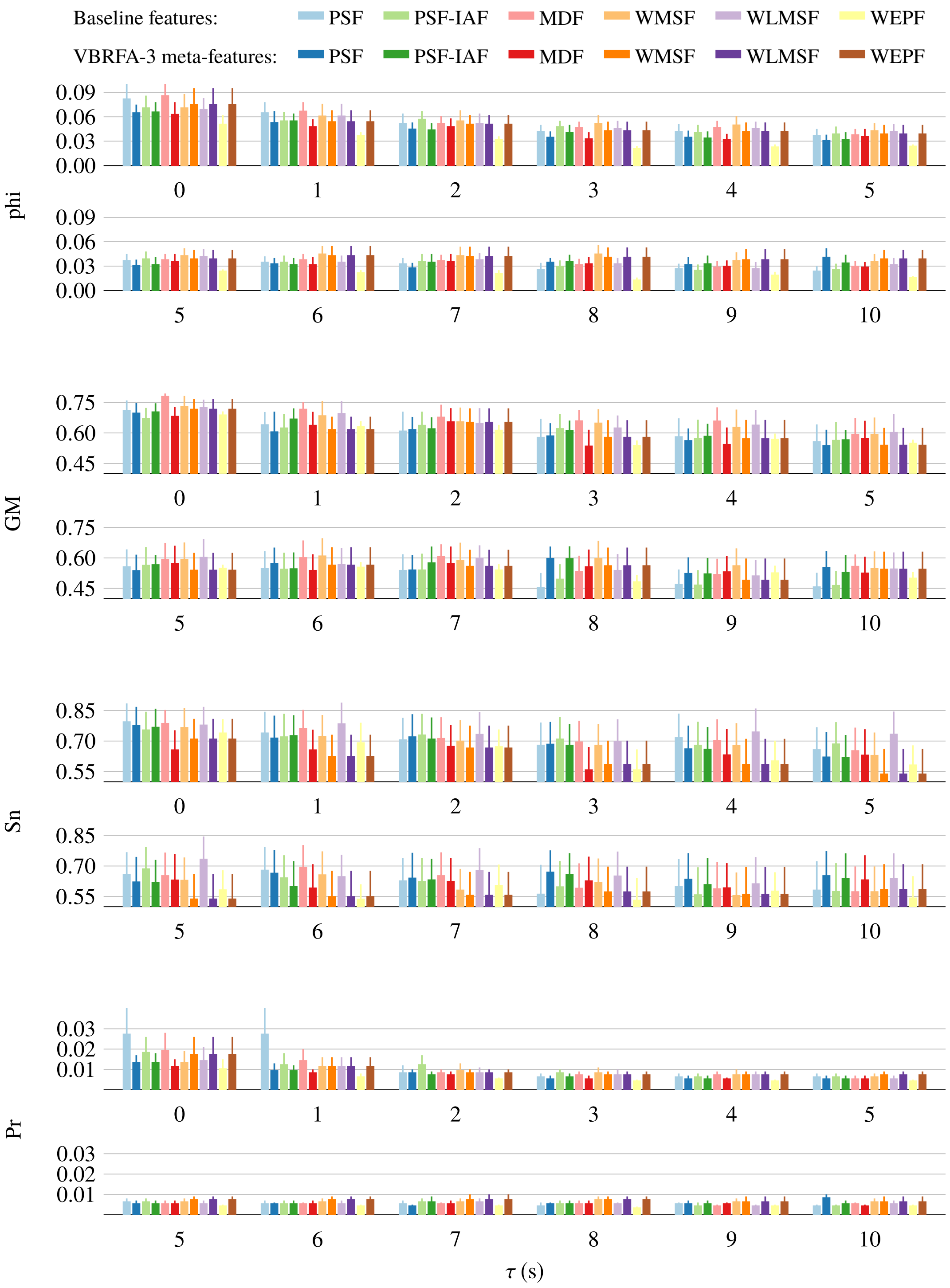


Figure 7.6 Performance (mean ± SE) of microsleep onset prediction using a single linear SVM classifier and VBRFA-3 meta-features for τ = 0–10 s.

7.6 SUMMARY

This chapter presented VBRFA, a Bayesian model for feature reduction. VBRFA uses a robust latent space to reduce sensitivity to noise. In addition, it finds the optimum dimension of latent space using an ARD-motivated set of prior distributions. Variational inference for the training and testing phases were also derived. This was then applied to the detection and prediction of microsleeps and results were reported.

The advantage of using VBRFA for microsleep state detection/prediction was that it reduced dimensionality (as shown in Table 7.2), but statistically significant performance improvements were only achieved with WEPF. Using the aggregated features of multiple EEG windows and an LDA classifier, VBRFA-3 meta-features of PSF-IAF had the highest performances for $\tau = 0\,\mathrm{s}$, i.e., AUC-ROC = 0.95, AUC-PR = 0.47, GM = 0.80, and $\varphi = 0.39$. Increasing the prediction time to $\tau = 1.0\,\mathrm{s}$ led to a substantial performance drop, i.e., AUC-ROC = 0.91 (4.2% drop), AUC-PR = 0.36 (23.4% drop), GM = 0.72 (10.0% drop), and $\varphi = 0.34$ (12.8% drop).

Microsleep onset detection had higher performances with baseline features compared to the VBRFA-3 meta-features, except for WEPF. For microsleep onset detection, baseline PSF had the highest AUC-PR (AUC-ROC = 0.91; AUC-PR = 0.09), while baseline MDF had the highest GM (AUC-ROC = 0.91; GM = 0.78). Using VBRFA-3 meta-features, however, the highest AUC-PR was achieved with WEPF (AUC-ROC = 0.89; AUC-PR = 0.06), while PSF-IAF had the highest GM (AUC-ROC = 0.88; GM = 0.73). Comparing the performances of microsleep onset detection using VBRFA-3 meta-features and baseline features showed that, in a few cases, the performance of baseline features were significantly higher than VBRFA-3 meta-features. Notwithstanding, VBRFA-3 meta-features of WEPF performed better than their respective baseline features.

A rapid decline of performance was observed with increasing prediction time, as shown in Figures 7.5 and 7.6, with the highest AUC-ROC of 0.91 for detection ($\tau = 0\,\mathrm{s}$) dropped to 0.72 for $\tau = 10\,\mathrm{s}$ (using baseline PSF with a linear SVM classifier). The highest AUC-ROC with onset prediction time of $\tau = 10\,\mathrm{s}$ was achieved with VBRFA-3 meta-features of PSF and a TAN classifier (AUC-ROC = 0.75; AUC-PR = 0.01), while the highest GM and phi were achieved with the combination of VBRFA-3 meta-features and a TAN classifier (AUC-ROC = 0.73; GM = 0.63; $\varphi = 0.04$).

Although the performance measures in terms of AUC-ROC and GM were moderate with short prediction times, model precisions were too low. This indicates that the onset detection and prediction systems had many FPs compared to the total number of microsleep onsets, which is impractical in real life scenarios.

Chapter 8

VARIATIONAL BAYESIAN MULTI-SUBJECT ROBUST FACTOR ANALYSIS

8.1 INTRODUCTION

A robust Bayesian feature reduction method, VBRFA, was presented and discussed in Chapter 7. Despite the ability of VBRFA to find a compact meta-feature space, microsleep prediction performance did not improve over the baseline in most cases. This may have been due to inter-subject variability, which has been shown to deteriorate the performance of testing on new subjects [Gerven et al. 2009, Wei et al. 2017]. Additionally, EEG has been used as a biometric identification in several studies and has shown a high accuracy for such a task [DelPozo-Banos et al. 2015, Klonovs et al. 2013, Thomas and Vinod 2016, Zhao et al. 2010]. This shows the extent of variability of EEG among subjects which is likely to prevent finding a generalized lower-dimension representation of data for all subjects.

The aim of the work in this chapter was to extend VBRFA to find a less variable inter-subject latent space. It was expected that by reducing the inter-subject variability of meta-features, the performance of microsleep prediction would outperform the baseline features. Section 8.2 presents our proposed multi-subject extension of VBRFA, i.e., variational Baysian multi-subject RFA (VBMSRFA), and its underlying procedure. The variational formulation of VBMSRFA for both training and testing phases are given in Section 8.3. Finally, Section 8.4 presents the results and discussion of microsleep prediction using VBMSRFA.

8.2 BAYESIAN MULTI-SUBJECT ROBUST FACTOR ANALYSIS

BMSRFA aims to find a smaller inter-subject-variable lower-space representation of data among multiple subjects. Reducing subject variability can potentially lead to a more accurate detection/prediction system. To reduce inter-subject variability, BMSRFA lets subjects share information via a common loading matrix while each subject has its own mean and noise terms. This model assumes that the underlying procedure of the phenomenon of interest, i.e., microsleeps, is similar between the subjects, while the mean and noise of the features vary between subjects, and even between sessions. The source of the variation can be impedance of EEG electrodes, room noise, head shape, or even underlying brain activity, to name a few.

Assume $\mathbf{X}_s = \left\{\mathbf{x}_{s,1}, \ldots, \mathbf{x}_{s,N_s}\right\}$ is the data collected from subject/session s. Assuming independent mean and noise terms for subject s, the reconstruction of data from the latent variables can be written as

$$\mathbf{x}_{s,n} = \boldsymbol{\mu}_s + \mathbf{W}\mathbf{z}_{s,n} + \boldsymbol{\varepsilon}_{s,n}. \tag{8.1}$$

This formulation finds a latent space that is potentially more consistent among all subjects. Assigning independent normal-Gamma distributions as prior probabilities of the mean and noise terms of each subject, leads to

$$p\left(\boldsymbol{\mu}, \boldsymbol{\Psi}\right) = \prod_{s=1}^{S} \prod_{d=1}^{D} \mathcal{N}\left(\mu_{s,d} \,\middle|\, m_{\mu,d}, \left(\beta_0 \psi_{s,d}\right)^{-1}\right) \mathcal{G}\left(\psi_{s,d} \,\middle|\, a_\psi, b_\psi\right), \tag{8.2}$$

where $\mathbf{m}_\mu$ is the mean hyperparameter of the prior distribution of the mean. Using a common hyperparameter over the mean of all subjects and updating it allows us to find a better suited prior for the test phase. This is desirable since the model updates the posterior probability of the test subject's mean and noise terms in addition to finding the latent variables. The prior probabilities for the rest of the parameters and hyperparameters are identical to the RFA. Figure 8.1 depicts a graphical representation of the BMSRFA.

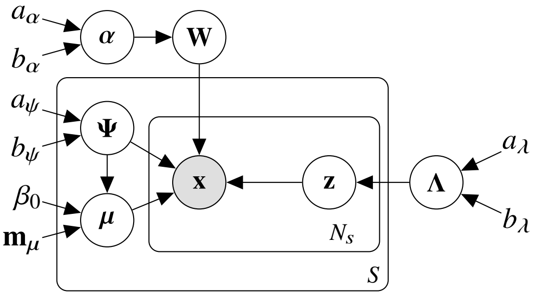


Figure 8.1 Graphical model representation of the Baysian multi-subject robust factor analysis (BMSRFA).

8.3 VARIATIONAL INFERENCE

8.3.1 Training phase

Similar to VBRFA, finding an analytical closed-form solution to the posterior distribution is intractable. Thus, a variational Bayesian method was employed to approximate the posterior probability. To this end, the variational distribution is assumed to be factorized as

$$q\left(\Theta\right) = q\left(\mathbf{W}\right) q\left(\alpha\right) q\left(\boldsymbol{\Lambda}\right) \prod_{s=1}^{S} \left(q\left(\boldsymbol{\mu}_s, \boldsymbol{\Psi}_s\right) \prod_{n=1}^{N_s} q\left(\mathbf{z}_{s,n}\right) \right), \tag{8.3}$$

where $\Theta = \big(\mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}\big)$ is the set of model parameters and latent variables. Using Equation (3.7), the lower bound of the evidence log-likelihood is given by

$$\begin{aligned}\mathcal{L} = &\left\langle \ln\left(\frac{p(\boldsymbol{\Lambda})}{q(\boldsymbol{\Lambda})}\right)\right\rangle + \left\langle \ln\left(\frac{p(\boldsymbol{\alpha})}{q(\boldsymbol{\alpha})}\right)\right\rangle + \left\langle \ln\left(\frac{p(\mathbf{W}\,|\,\boldsymbol{\alpha})}{q(\mathbf{W})}\right)\right\rangle \\ &+ \sum_{s=1}^{S}\left\{\left\langle \ln\left(\frac{p(\boldsymbol{\mu}_s\,|\,\boldsymbol{\Psi}_s)}{q(\boldsymbol{\mu}_s\,|\,\boldsymbol{\Psi}_s)}\right)\right\rangle + \left\langle \ln\left(\frac{p(\boldsymbol{\Psi}_s)}{q(\boldsymbol{\Psi}_s)}\right)\right\rangle\right\} \\ &+ \sum_{s=1}^{S}\sum_{n=1}^{N_s}\left\{\left\langle \ln\left(\frac{p(\mathbf{z}_{s,n}\,|\,\boldsymbol{\Lambda})}{q(\mathbf{z}_{s,n})}\right)\right\rangle + \left\langle \ln\Big(p\left(\mathbf{x}_{s,n}\,|\,\mathbf{W}, \boldsymbol{\mu}_s, \mathbf{z}_{s,n}, \boldsymbol{\Psi}_s\right)\Big)\right\rangle\right\},\end{aligned} \tag{8.4}$$

where the objective is to approximate the posterior distributions to maximize $\mathcal{L}$. This is achieved by finding the approximate posterior distributions of the $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ with respect to each individual subject/session data, while the posterior distributions of $\boldsymbol{\alpha}$, $\boldsymbol{\Lambda}$, and $\mathbf{W}$ are approximated with respect to the data of all the subjects. In the VBE step, all the parameters are kept fixed and the variational parameters of the latent variables maximizing $\mathcal{L}$ are given by

$$\tilde{\Sigma}_{z,s} = \Big(\langle \mathbf{W}^\top \boldsymbol{\Psi}_s \mathbf{W}\rangle + \langle \boldsymbol{\Lambda}\rangle\Big)^{-1}, \tag{8.5}$$

$$\tilde{\mathbf{m}}_{z,s,n} = \tilde{\Sigma}_{z,s}\langle \mathbf{W}^\top\rangle\langle \boldsymbol{\Psi}_s\rangle\Big(\mathbf{x}_{s,n} - \langle \boldsymbol{\mu}_s\rangle\Big), \tag{8.6}$$

where the variational distribution of the latent variables are given by

$$q(\mathbf{Z}) = \prod_{s=1}^{S}\prod_{n=1}^{N_s}\mathcal{N}\left(\mathbf{z}_{s,n}\,\middle|\,\tilde{\mathbf{m}}_{z,s,n}, \tilde{\Sigma}_{z,s}\right). \tag{8.7}$$

The two terms of Equation (8.5) control the complexity and uncertainty of the latent variables. When the data of a subject is noisy, the first term, $\langle \mathbf{W}^\top \boldsymbol{\Psi}_s \mathbf{W}\rangle$, drops and the uncertainty increases. But a redundant component k would have a high value of $\langle \lambda_k\rangle$ which effectively switches it off for all the subjects.

After updating the variational parameters of the latent variables, the VBM step is performed. This aims to update the variational distributions of the model parameters which are given by

$$q(\boldsymbol{\alpha}) = \prod_{k=1}^{K}\mathcal{G}\left(\alpha_k\,\middle|\,\tilde{a}_\alpha, \tilde{b}_{\alpha,k}\right), \tag{8.8}$$

$$q(\boldsymbol{\Psi}) = \prod_{s=1}^{S}\prod_{d=1}^{D}\mathcal{G}\left(\psi_{s,d}\,\middle|\,\tilde{a}_{\psi,s}, \tilde{b}_{\psi,s,d}\right), \tag{8.9}$$

$$q(\boldsymbol{\mu}\,|\,\boldsymbol{\Psi}) = \prod_{s=1}^{S}\prod_{d=1}^{D}\mathcal{N}\left(\mu_{s,d}\,\middle|\,\tilde{m}_{\mu,s,d}, (\beta_{\mu,s}\psi_{s,d})^{-1}\right), \tag{8.10}$$

$$q(\mathbf{W}) = \prod_{d=1}^{D}\mathcal{N}\left(\mathbf{w}_{d,\cdot}^\top\,\middle|\,\tilde{\mathbf{m}}_{w,d}, \tilde{\Sigma}_{w,d}\right), \tag{8.11}$$

$$q\left(\mathbf{\Lambda}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\lambda_k \mid \tilde{a}_\lambda, \tilde{b}_{\lambda,k}\right), \tag{8.12}$$

where $\mathbf{w}_{d,.}$ corresponds to the d^{th} row of the loading matrix $\mathbf{W}$ and the parameters of the variational distribution are updated by

$$\begin{aligned}
\beta_{\mu,s} &= N_s + \beta_0, && (8.13)\\
\tilde{m}_{\mu,s,d} &= \frac{1}{\beta_{\mu,s}}\left(\beta_0 m_{\mu,d} + \sum_{n=1}^{N_s}\left(x_{s,n,d} - \langle \mathbf{w}_{d,.}\rangle\langle \mathbf{z}_{s,n}\rangle\right)\right), && (8.14)\\
\tilde{a}_{\psi,s} &= a_\psi + \frac{N_s}{2}, && (8.15)\\
\tilde{b}_{\psi,s,d} &= b_\psi + \frac{\beta_0}{2}m_{\mu,d}^2 - \frac{\beta_{\mu,s}}{2}\tilde{m}_{\mu,s,d}^2\\
&\quad + \frac{1}{2}\sum_{n=1}^{N_s}\left(x_{s,n,d}^2 - 2x_{s,n,d}\langle \mathbf{w}_{d,.}\rangle\langle \mathbf{z}_{s,n}\rangle + \operatorname{tr}\left(\langle \mathbf{w}_{d,.}^\top \mathbf{w}_{d,.}\rangle\langle \mathbf{z}_{s,n}\mathbf{z}_{s,n}^\top\rangle\right)\right), && (8.16)\\
\tilde{\Sigma}_{w,d} &= \left(\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right)\rangle + \sum_{s=1}^{S}\langle \psi_{s,d}\rangle\sum_{n=1}^{N_s}\langle \mathbf{z}_{s,n}\mathbf{z}_{s,n}^\top\rangle\right)^{-1}, && (8.17)\\
\tilde{\mathbf{m}}_{w,d} &= \tilde{\Sigma}_{w,d}\sum_{s=1}^{S}\langle \psi_{s,d}\rangle\sum_{n=1}^{N_s}\left(\langle \mathbf{z}_{s,n}\rangle\left(x_{s,n,d} - \langle \mu_{s,d}\rangle\right)\right), && (8.18)\\
\tilde{a}_\lambda &= a_\lambda + \frac{1}{2}\sum_{s=1}^{S}N_s, && (8.19)\\
\tilde{b}_{\lambda,k} &= b_\lambda + \frac{1}{2}\sum_{s=1}^{S}\sum_{n=1}^{N_s}\langle z_{s,n,k}^2\rangle, && (8.20)\\
\tilde{a}_\alpha &= a_\alpha + \frac{D}{2}, && (8.21)\\
\tilde{b}_{\alpha,k} &= b_\alpha + \frac{\langle \mathbf{w}_k^\top \mathbf{w}_k\rangle}{2}, && (8.22)
\end{aligned}$$

where tr $\left(\mathbf{A}\right)$ is the trace function which finds the summation of diagonal elements of matrix $\mathbf{A}$. Equation (8.17) shows that the contribution of each subject to the covariance of the loading matrix is proportional to the precision of each subject's data. The value of $\tilde{b}_{\alpha,k}$ depends on the second norm of the k^{th} column of the loading matrix. This acts as a regularization on the loading matrix and penalizes the high values.

Iterating over VBE and VBM steps is guaranteed not to decrease the lower bound of the evidence log-likelihood. The hyperparameters can also be optimized to increase the lower bound of the marginal log-likelihood. Taking the derivative of $\mathcal{L}$ with respect to each element of $\mathbf{m}_\mu$ leads to

$$m_{\mu,d} = \frac{\sum_{s=1}^{S}\langle \psi_{s,d}\rangle\langle \mu_{s,d}\rangle}{\sum_{s=1}^{S}\langle \psi_{s,d}\rangle}. \tag{8.23}$$

The hyperparameter $\mathbf{m}_\mu$ is common between all the subjects and, intuitively, its value is a weighted average of the posterior mean of all the subjects relative to their noise terms. The updating formula for β_0 can be calculated in the same manner, as

$$\begin{aligned}
\beta_0^{-1} &= \frac{1}{SD}\sum_{s=1}^{S}\sum_{d=1}^{D}\left\langle \psi_{s,d}\left(\mu_{s,d} - m_{\mu,d}\right)^2\right\rangle \\
&= \frac{1}{SD}\sum_{s=1}^{S}\sum_{d=1}^{D}\left(\left\langle\psi_{s,d}\right\rangle\left(\left\langle\mu_{s,d}\right\rangle - m_{\mu,d}\right)^2 + \beta_{\mu,s}^{-1}\right), \qquad (8.24)
\end{aligned}$$

where the values of $\mathbf{m}_\mu$ are first updated using Equation (8.23).

Updating the Gamma hyperparameters, i.e., $\left\{a_\alpha, b_\alpha, a_\psi, b_\psi, a_\lambda, b_\lambda\right\}$, can be done by taking the derivative of the $\mathcal{L}$ with respect to the hyperparameters, that is

$$\Psi\left(a_\alpha\right) = \ln\left(b_\alpha\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\Psi\left(\tilde{a}_\alpha\right) - \ln\left(\tilde{b}_{\alpha,k}\right)\right), \qquad (8.25)$$

$$\Psi\left(a_\psi\right) = \ln\left(b_\psi\right) + \frac{1}{SD}\sum_{s=1}^{S}\sum_{d=1}^{D}\left(\Psi\left(\tilde{a}_{\psi,s}\right) - \ln\left(\tilde{b}_{\psi,s,d}\right)\right), \qquad (8.26)$$

$$\Psi\left(a_\lambda\right) = \ln\left(b_\lambda\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\Psi\left(\tilde{a}_\lambda\right) - \ln\left(\tilde{b}_{\lambda,k}\right)\right), \qquad (8.27)$$

$$b_\alpha^{-1} = \frac{1}{a_\alpha K}\sum_{k=1}^{K}\frac{\tilde{a}_\alpha}{\tilde{b}_{\alpha,k}}, \qquad (8.28)$$

$$b_\psi^{-1} = \frac{1}{a_\psi SD}\sum_{s=1}^{S}\sum_{d=1}^{D}\frac{\tilde{a}_{\psi,s}}{\tilde{b}_{\psi,s,d}}, \qquad (8.29)$$

$$b_\alpha^{-1} = \frac{1}{a_\lambda K}\sum_{k=1}^{K}\frac{\tilde{a}_\lambda}{\tilde{b}_{\lambda,k}}. \qquad (8.30)$$

At each iteration, we computed the lower bound of the evidence log-likelihood in order to monitor the convergence. The variational EM was stopped when the relative improvement of $\mathcal{L}$ was below a threshold, i.e., 10^{-6}. Moreover, the component corresponding to the largest value of $\langle\boldsymbol{\alpha}\rangle$ was temporarily excluded at each iteration and the lower bound was examined. An improvement of the lower bound after excluding the component was used as a measure to permanently remove the component. Algorithm 8.1 presents the pseudo code of the VBMSRFA.

8.3.2 Testing phase

The training phase finds the posterior probability approximations of the Bayesian model given the training data. As a result, the shared model parameters $\left\{\mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}\right\}$ are approximated using the training data of multiple subjects. Given new data after the training phase, the next step is to determine the posterior probability of latent variables. However, the mean and noise terms of the subjects are assumed to be independent of each other. Therefore, these parameters of a test

Algorithm 8.1 The training algorithm of variational Bayesian multi-subject robust factor analysis (VBMSRFA).

```
procedure INITIALIZING
    $K = D - 1$, $\mathbf{m}_\mu = \mathbf{0}$, $a_\alpha = b_\alpha = a_\psi = b_\psi = a_\lambda = b_\lambda = \beta_0 = 10^{-6}$.
    $\tilde{a}_{\psi,s} = a_\psi$, $\tilde{\mathbf{b}}_{\psi,s} = b_\psi$, $\tilde{a}_\lambda = a_\lambda$, $\tilde{\mathbf{b}}_\lambda = b_\lambda$, $\tilde{a}_\alpha = a_\alpha$, $\tilde{\mathbf{b}}_\alpha = b_\alpha$, $\beta_{\mu,s} = \beta_0$.
    $RelTol = 10^{-6}$, $MaxIter = 1000$.
    for $s = 1$ to $S$ do
        Set $\tilde{\mathbf{m}}_{\mu,s}$ to the empirical mean of the subject $s$ data.
    $\tilde{\Sigma}_{w,d} = \mathbf{I}, \forall d \in \{1, \ldots, D\}$.
    Set $\tilde{\mathbf{m}}_w$ to the first $K$ components of the PCA coefficients of the concatenated demeaned datasets.
for $iter = 1$ to $MaxIter$ do
    procedure VBE
        for $s = 1$ to $S$ do
            for $n = 1$ to $N_s$ do
                Update expectations of the latent variables using Equations (8.5) and (8.6).
    procedure VBM
        for $s = 1$ to $S$ do
            Update the subject specific variational parameters using Equations (8.13)–(8.16).
        Update the shared variational parameters using Equations (8.17)–(8.22).
    procedure UPDATE HYPERPARAMETERS
        if Reminder($iter$, 10) is 0 then
            Update $\mathbf{m}_\mu$ and $\beta_0$ using Equations (8.23) and (8.24).
            Update Gamma hyperparameters by iterating over Equations (8.25)–(8.30).
    procedure STOPPING CRITERIA
        Calculate the lower bound of the marginal log-likelihood.
        if $\frac{\mathcal{L}(iter)-\mathcal{L}(iter-1)}{|\mathcal{L}(iter)|} < RelTol$ then
            Stop.                                                  ▷ Converged
    procedure PRUNING LATENT VARIABLES
        Temporarily remove the component corresponding to the highest $\langle\alpha\rangle$.
        Calculate the lower bound of the marginal log-likelihood.
        if $\mathcal{L}$ (after pruning) > $\mathcal{L}$ (before pruning) then    ▷ The component can be removed.
            $\mathcal{L}(iter) = \mathcal{L}$ (after pruning) and remove the component.
        else                                           ▷ The component can not be removed.
            $\mathcal{L}(iter) = \mathcal{L}$ (before pruning) and keep the component.
```

subject have to be estimated from the test subject's own data. Assuming that the test data arrives sequentially and over time, we have utilized an incremental variational inference to approximate the posterior probability distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ of a test subject. Given a new observation $\mathbf{x}_n$, the objective is to maximize

$$\begin{aligned} p\left(\mathbf{x}_n \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathcal{D}_{\text{train}}\right) &= \int p\left(\mathbf{x}_n \,\middle|\, \mathbf{z}_n, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) p\left(\mathbf{W} \,\middle|\, \mathcal{D}_{\text{train}}\right) p\left(\mathbf{z}_n \,\middle|\, \boldsymbol{\Lambda}\right) \\ &\qquad p\left(\boldsymbol{\Lambda} \,\middle|\, \mathcal{D}_{\text{train}}\right) p\left(\boldsymbol{\mu}, \boldsymbol{\Psi} \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right) d\Theta \\ &\approx \int p\left(\mathbf{x}_n \,\middle|\, \mathbf{z}_n, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) q\left(\mathbf{W}\right) p\left(\mathbf{z}_n \,\middle|\, \boldsymbol{\Lambda}\right) \\ &\qquad q\left(\boldsymbol{\Lambda}\right) p\left(\boldsymbol{\mu}, \boldsymbol{\Psi} \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right) d\Theta, \end{aligned} \tag{8.31}$$

where Θ is a set of the model parameters and the probability distributions of $p\left(\mathbf{W} \,\middle|\, \mathcal{D}_{\text{train}}\right)$ and $p\left(\boldsymbol{\Lambda} \,\middle|\, \mathcal{D}_{\text{train}}\right)$ are approximated by their respective variational distributions from the training phase, i.e., $q\left(\mathbf{W}\right)$ and $q\left(\boldsymbol{\Lambda}\right)$. For each observation, we use a factorized variational distribution to approximate the posterior of all the parameters as

$$q^*\left(\Theta^{(n)}\right) = q^*\left(\mathbf{W}\right) q^*\left(\mathbf{z}_n\right) q^*\left(\boldsymbol{\Lambda}\right) q^*\left(\boldsymbol{\mu}^{(n)}, \boldsymbol{\Psi}^{(n)}\right), \tag{8.32}$$

where $\Theta^{(n)}$ corresponds to the model parameters after observing the n^{th} data point. After observing a new data point, the variational distributions are updated, where the prior probabilities over the mean and noise terms are set to their variational distributions from the last step. This can be represented by modifying Equation (8.31) to

$$\begin{aligned} p\left(\mathbf{x}_n \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathcal{D}_{\text{train}}\right) &\approx \int p\left(\mathbf{x}_n \,\middle|\, \mathbf{z}_n, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) q\left(\mathbf{W}\right) p\left(\mathbf{z}_n \,\middle|\, \boldsymbol{\Lambda}\right) \\ &\qquad q\left(\boldsymbol{\Lambda}\right) q^*\left(\boldsymbol{\mu}^{(n-1)}, \boldsymbol{\Psi}^{(n-1)}\right) d\Theta. \end{aligned} \tag{8.33}$$

The variational parameters of the n^{th} observation can be updated by

$$\hat{\Sigma}_{w,d}^{(n)} = \left(\langle\psi_d\rangle\langle\mathbf{z}_n\mathbf{z}_n^\top\rangle + \tilde{\Sigma}_{w,d}^{-1}\right)^{-1}, \tag{8.34}$$

$$\hat{\mathbf{m}}_{w,d}^{(n)} = \hat{\Sigma}_{w,d}^{(n)}\left(\tilde{\Sigma}_{w,d}^{-1}\tilde{\mathbf{m}}_{w,d} + \langle\psi_d\rangle\langle\mathbf{z}_n\rangle\left(x_{n,d} - \langle\mu_d\rangle\right)\right), \tag{8.35}$$

$$\hat{\Sigma}_z^{(n)} = \left(\langle\boldsymbol{\Lambda}\rangle + \langle\mathbf{W}^\top\boldsymbol{\Psi}\mathbf{W}\rangle\right)^{-1}, \tag{8.36}$$

$$\hat{\mathbf{m}}_z^{(n)} = \hat{\Sigma}_z^{(n)}\langle\mathbf{W}\rangle^\top\langle\boldsymbol{\Psi}\rangle\left(\mathbf{x}_n - \langle\boldsymbol{\mu}\rangle\right), \tag{8.37}$$

$$\hat{a}_\lambda^{(n)} = \tilde{a}_\lambda + \frac{1}{2}, \tag{8.38}$$

$$\hat{b}_{\lambda,k}^{(n)} = \tilde{b}_{\lambda,k} + \frac{1}{2}\langle z_{n,k}^2\rangle, \tag{8.39}$$

$$\beta_\mu^{(n)} = 1 + \beta_\mu^{(n-1)}, \tag{8.40}$$

$$\hat{m}_{\mu,d}^{(n)} = \frac{1}{\beta_\mu^{(n)}}\left(\beta_\mu^{(n-1)}\hat{m}_{\mu,d}^{(n-1)} + x_{n,d} - \langle\mathbf{w}_{d,.}\rangle\langle\mathbf{z}_n\rangle\right), \tag{8.41}$$

$$a_\psi^{(n)} = a_\psi^{(n-1)} + \frac{1}{2}, \tag{8.42}$$

$$
\begin{aligned}
b_{\psi,d}^{(n)} = b_{\psi,d}^{(n-1)} &+ \frac{\beta_\mu^{(n-1)}}{2}\left(\hat{m}_{\mu,d}^{(n-1)}\right)^2 - \frac{\beta_\mu^{(n)}}{2}\left(\tilde{m}_{\mu,d}^{(n)}\right)^2 \\
&+ \frac{1}{2}\left(x_{n,d}^2 - 2x_{n,d}\langle \mathbf{w}_{d,\cdot}\rangle\langle \mathbf{z}_n\rangle + \operatorname{tr}\left(\langle \mathbf{z}_n\mathbf{z}_n^\top\rangle\langle \mathbf{w}_{d,\cdot}^\top \mathbf{w}_{d,\cdot}\rangle\right)\right). \qquad (8.43)
\end{aligned}
$$

Iterating over Equations (8.34)–(8.43) maximizes the lower bound of the log marginal probability of the new observed data point. When the variational parameters have converged, the MAP of the latent variables is used as meta-features for the classification tasks.

8.4 RESULTS AND DISCUSSION

8.4.1 Detection and prediction of microsleep states

Table 8.1 shows the results of microsleep state detection with an LDA classifier versus EEG window lengths used for feature extraction. With this setup, the highest AUC-ROC and AUC-PR were achieved with VBMSRFA meta-features of WLMSF (AUC-ROC = 0.94; AUC-PR = 0.45), which had a substantial improvement over baseline features. Furthermore, the performance of microsleep state detection with VBMSRFA meta-features was slightly higher than that of baseline features in most cases. It was observed that most of the highest performances for each feature set were achieved with features extracted from 5-s EEG windows.

Table 8.1 Performance (mean ± SE) of microsleep state detection using VBMSRFA meta-features versus baseline features with an LDA classifier. A bold value indicates the highest performance of each feature set and italics indicate the highest overall. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements of VBMSRFA relative to the baseline.

Feature set	Feature type	2-s EEG window		5-s EEG window		10-s EEG window	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.86±0.03	0.31±0.12	0.88±0.03	0.36±0.12	0.86±0.04	0.36±0.12
	VBMSRFA	**0.91±0.02**$^{\sim}$	0.36±0.11	**0.91±0.02**	**0.41±0.11**	0.88±0.03	0.38±0.11
PSF	Baseline	0.89±0.02	0.34±0.12	0.92±0.01	0.41±0.12	0.92±0.02	0.37±0.12
	VBMSRFA	0.92±0.01	0.37±0.12	**0.93±0.01**	**0.42±0.12**	0.92±0.02	0.39±0.12
PSF-IAF	Baseline	0.88±0.02	0.34±0.12	0.90±0.02	**0.37±0.13**	**0.91±0.02**	0.35±0.12
	VBMSRFA	0.90±0.03	**0.37±0.13**	0.86±0.04	**0.37±0.12**	0.90±0.03	**0.37±0.13**
WMSF	Baseline	0.87±0.03	0.35±0.12	0.91±0.02	0.38±0.14	0.91±0.02	0.37±0.13
	VBMSRFA	0.88±0.03*	0.36±0.13	**0.92±0.02**	**0.40±0.14**	0.90±0.02	0.37±0.13
WLMSF	Baseline	0.88±0.03	0.36±0.13	0.91±0.02	0.41±0.13	0.91±0.02	0.37±0.13
	VBMSRFA	0.93±0.01	0.42±0.12$^{\sim}$	***0.94±0.01***	***0.45±0.12***	0.91±0.02	0.41±0.12$^{\sim}$
WEPF	Baseline	0.74±0.03	0.21±0.10	0.78±0.03	0.22±0.11	**0.81±0.02**	0.23±0.11
	VBMSRFA	**0.81±0.03***	**0.26±0.10**$^{\sim}$	**0.81±0.04**	**0.26±0.12**	**0.81±0.04**	0.25±0.12

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

The number of VBMSRFA meta-features for individual feature sets is given in Table 8.2. Although the numbers of meta-features were smaller than for the baselines, they were substantially higher than for the meta-features of VBRFA (Table 7.2). However, although applying VBMSRFA to WLMSF extracted from 5-s EEG windows led to an average of 79 meta-features (cf. 52.1 for VBRFA-3), the performances of VBMSRFA were substantially higher than for VBRFA, i.e., AUC-ROC = 0.94 (cf. 0.92 for VBRFA), AUC-PR = 0.45 (cf. 0.41 for VBRFA), $\varphi = 0.44$ (cf. 0.33 for VBRFA; $p = 0.023$), and GM = 0.83 (cf. 0.74 for VBRFA). Interestingly, using VBRFA-3 meta-features of the WEPF to predict microsleeps outperformed VBMSRFA. This suggests that there might be subject-independent useful information that can be removed by VBMSRFA.

Table 8.2 Average number of VBMSRFA meta-features.

Feature set	Number of features			
	2-s EEG window	5-s EEG window	10-s EEG window	Baseline
MDF	143	151.6	152.8	176
PSF	138	154.4	165.1	192
PSF-IAF	150.3	155.1	169.5	192
WMSF	79	79	79	80
WLMSF	79	79	79	80
WEPF	68.6	74.5	76.1	80

The aggregation of features extracted from multiple EEG windows led to the highest baseline performances. This same procedure was applied to the VBMSRFA meta-features for individual feature sets. The features extracted from each EEG window length were transformed to VBMSRFA meta-features, leading to three VBMSRFA models corresponding to 2, 5, and 10-s EEG segments. After computing meta-features for each EEG window length, the meta-features of different window lengths were aggregated into a larger meta-feature matrix. The use of parallel independent models was due to computational convenience. For example, the aggregation of multiple EEG windows of PSF leads to 576 features and thus an initial latent space dimension of $K = 575$, where each dimension has a covariance matrix $\tilde{\Sigma}_{w,d} \in \mathbb{R}^{K \times K}$ that needed to be inverted at every iteration of variational inference. This is computationally expensive and hence multiple parallel methods were used. The performance of microsleep state detection using the aggregated meta-features and an LDA classifier is shown in Table 8.3. This showed an average AUC-ROC improvement of 2.7% (0–6.17%) and AUC-PR improvement of 13.1% (2.50–27.02%) with aggregated VBMSRFA meta-feature sets compared to their respective best single-window ones.

Table 8.4 shows the performance of different classifiers for microsleep state detection using aggregated VBMSRFA meta-features. The highest detection performances were achieved with the LDA and linear SVM classifiers, while the VBLR had slightly lower performances. The TAN classifier had the lowest performance among all classifiers. The highest performance was achieved with VBMSRFA meta-features of WLMSF (AUC-ROC = 0.95; AUC-PR = 0.49). Conversely, the meta-features of WEPF had the lowest performance among all feature sets. The remaining VBMSRFA meta-features had relatively similar performances. Since the aggregated meta-features of multiple EEG windows showed superior performances, their results are reported in the remainder of this chapter.

Performance of microsleep state prediction was investigated by increasing the prediction

Table 8.3 Performance (mean ± SE) of microsleep state detection using aggregated VBMSRFA meta-features extracted from multiple EEG windows and an LDA classifier. A bold value indicates the highest performance in an individual feature set, whereas italics indicate the highest among all feature sets. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements of VBMSRFA features compared to the baseline.

Feature set	Method	Microsleep state detection performance					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	0.90±0.02	0.40±0.13	**0.71±0.08**	0.34±0.10	**0.62±0.10**	0.33±0.12
	VBMSRFA	**0.93±0.01**	**0.44±0.11**	0.66±0.07	**0.36±0.08**	0.51±0.10	***0.42±0.13***
PSF	Baseline	0.94±0.01	0.43±0.12	0.74±0.06	0.36±0.08	0.68±0.11	***0.36±0.12***
	VBMSRFA	***0.95±0.01***	**0.46±0.12**	**0.81±0.03**	**0.46±0.09****	**0.70±0.06**	**0.41±0.12**
PSF-IAF	Baseline	**0.94±0.01**	0.44±0.12	**0.76±0.05**	0.37±0.08	**0.70±0.10**	0.36±0.12
	VBMSRFA	**0.94±0.02**	**0.47±0.12**	**0.76±0.05**	**0.43±0.09**	0.61±0.07	**0.41±0.13**
WMSF	Baseline	**0.92±0.02**	0.40±0.14	0.66±0.10	0.33±0.09	0.57±0.12	0.31±0.11
	VBMSRFA	**0.92±0.02**	**0.41±0.13**	**0.70±0.09**	**0.36±0.10**	**0.59±0.11**	**0.35±0.12**
WLMSF	Baseline	0.94±0.01	0.44±0.13	0.76±0.07	0.36±0.10	0.70±0.11	0.31±0.12
	VBMSRFA	***0.95±0.01***	***0.49±0.12***	***0.83±0.03***	***0.47±0.09*****	***0.74±0.05***	**0.38±0.12**
WEPF	Baseline	0.84±0.02	0.27±0.12	**0.70±0.03**	0.25±0.07	**0.67±0.08**	0.25±0.13
	VBMSRFA	**0.86±0.03**	**0.32±0.12**	0.63±0.07	**0.26±0.09**	0.47±0.10	**0.29±0.13**

Wilcoxon signed-rank test: ** $p < 0.01$

Table 8.4 Performance of different classifiers for microsleep state detection with aggregated VBMSRFA meta-features. A bold value indicates the highest performance among selected classifiers and an italic value is the highest overall.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.90	0.40	0.93	**0.48**	0.90	0.42	0.77	0.22
	VBMSRFA	0.93	0.44	**0.94**	0.46	0.92	0.42	0.72	0.20
PSF	Baseline	0.94	0.43	0.94	***0.49***	0.94	0.44	0.70	0.21
	VBMSRFA	***0.95***	0.46	0.93	0.40	0.94	0.41	0.67	0.17
PSF-IAF	Baseline	0.94	0.44	***0.95***	**0.47**	0.93	0.45	0.73	0.26
	VBMSRFA	0.94	**0.47**	0.92	0.42	0.92	0.40	0.62	0.21
WMSF	Baseline	0.92	0.40	**0.93**	0.42	0.92	0.39	0.77	0.22
	VBMSRFA	0.92	0.41	**0.93**	**0.43**	0.92	**0.43**	0.74	0.20
WLMSF	Baseline	0.94	0.44	0.93	0.45	0.92	0.45	0.76	0.25
	VBMSRFA	***0.95***	***0.49***	***0.95***	***0.49***	0.94	0.48	0.79	0.29
WEPF	Baseline	0.84	0.27	0.86	0.29	0.85	0.28	0.82	0.23
	VBMSRFA	0.86	0.32	**0.88**	0.33	**0.88**	**0.34**	0.68	0.20

time τ from 0 to 1 s, in steps of 0.25 s, and computing the performance of different models. The AUC-PR of different classifiers is shown in Figure 8.2. It was observed that an LDA classifier outperformed others with VBMSRFA meta-features of PSF and PSF-IAF. Furthermore, the performance of the three classifiers, i.e., LDA, linear SVM, and VBLR, was relatively similar when VBMSRFA meta-features of WLMSF were used. However, a linear SVM classifier performed better on VBMSRFA meta-features of MDF, whereas VBLR performed better on WMSF and WEPF meta-features. Overall, VBMSRFA meta-features of PSF, WMSF, and WLMSF had superior performances compared to the baselines. With $\tau = 0$ s, the highest

AUC-PR was achieved with the VBMSRFA meta-features of WLMSF and the baseline PSF (AUC-PR = 0.49), where the classifiers were LDA and linear SVM.

On increasing prediction time to $\tau = 1.0$ s, the highest AUC-ROC and AUC-PR were achieved with VBMSRFA meta-features of WLMSF, as shown in Figure 8.2, where the three linear classifiers, LDA, linear SVM, and VBLR, had similar performances (AUC-ROC = 0.94; AUC-PR = 0.45). On the other hand, the highest performance of baseline features at $\tau = 1.0$ s was also achieved with WLMSF (AUC-ROC = 0.91; AUC-PR = 0.42). This suggests that WLMSF has more information regarding the state of responsiveness of a subject with a longer prediction time. Also, applying VBMSRFA to WLMSF to find a less subject-variable latent space improved the performance of microsleep state prediction.

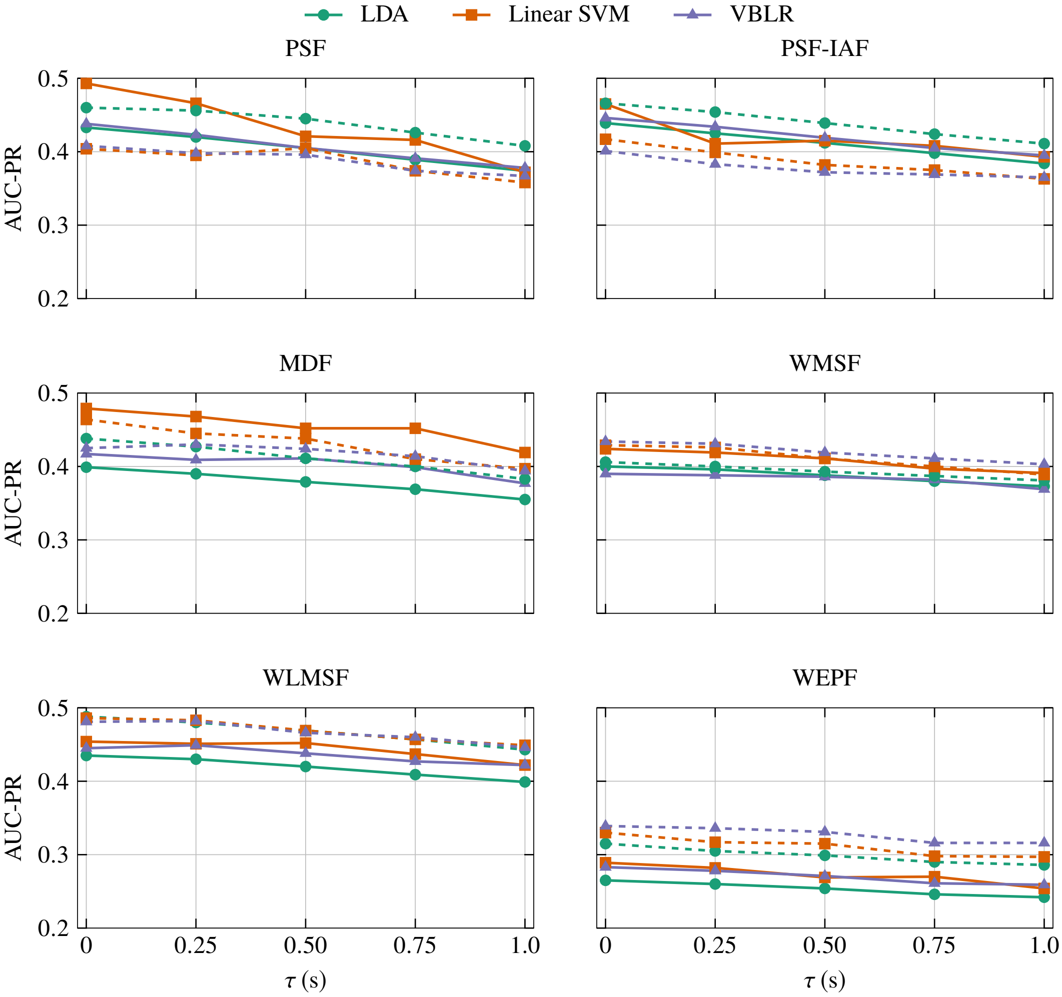


Figure 8.2 AUC-PR of microsleep state prediction using different classifiers versus prediction time τ for τ = 0–1 s. Solid lines correspond to the performance of a classifier with baseline features, whereas dashed lines correspond to their respective performance with VBMSRFA meta-features.

Performance of the prediction of microsleep states in terms of phi is reported in Table 8.5, where two-tailed Wilcoxon signed-rank tests were performed to find significant improvements of each classifier with VBMSRFA meta-features relative to their respective baseline. It was observed that using VBMSRFA meta-features increased phi in many cases. Substantial improvements of phi were achieved with LDA classifiers and VBMSRFA meta-features of PSF-IAF and PSF across prediction times of $\tau = 0$ to 1 s. The average improvement of phi with a combination of LDA and PSF-IAF meta-features across all prediction times relative to the best of baselines was 15.6% (14.3–16.7%). Similarly, the average improvement of phi across prediction times for VBMSRFA meta-features of PSF was 21.0% (16.2–24.3%). Using VBMSRFA meta-features of WEPF, VBLR performed better, in terms of phi, than other classifiers, with an average improvement of 11.7% (8.0–16.0%) compared to the baseline. Improvements of phi with VBMSRFA meta-features of WMSF and WLMSF were consistent for three of the classifiers, namely LDA, linear SVM, and VBLR. In this regard, using VBMSRFA meta-features of WMSF with linear SVM classifiers across all prediction times improved phi by an average of 5.6% (2.9–10.5%). The highest overall improvement of phi across all prediction times was achieved with VBMSRFA meta-features of WLMSF and LDA classifiers, i.e., 29.6% (27.8–31.4%). In addition, using VBMSRFA meta-features of WLMSF significantly improved performance in terms of phi for three classifiers, i.e., LDA, linear SVM, and VBLR, across all prediction times. The highest performances of MDF, however, were achieved with the baseline.

Surprisingly, the performance of TAN was substantially lower with VBMSRFA meta-features than that of the baseline. One reason might be the fact that VBMSRFA finds a set of independent latent variables (meta-features) while TAN represents the data using conditional dependencies. Therefore, TAN might model the data with false dependencies that consequently lead to a substantial decline in performance.

The LDA classifier had comparatively good results, both in terms of AUC-PR and phi, and hence the remainder of the results of this section for microsleep state prediction using VBMSRFA are limited to the LDA classifier. Performance of LDA classifiers in terms of sensitivity, precision, and GM for various feature sets across prediction times are shown in Figure 8.3. It was observed that using VBMSRFA meta-features improved GM consistently for three feature sets – PSF, WMSF, and WLMSF – across all prediction times. Sensitivity, however, revealed little difference between baseline features and VBMSRFA meta-features. Thus, using VBMSRFA meta-features has increased the specificity and lowered the number of false positives, which is evident in the improvement in precision. With $\tau = 0$ s, the highest performance was achieved with VBMSRFA meta-features of WLMSF (GM = 0.83; $\varphi = 0.47$; Pr = 0.38; Sn = 0.74), whereas the best of baseline was achieved with a linear SVM classifier and PSF (GM = 0.79; $\varphi = 0.40$; Pr = 0.38; Sn = 0.74). Similarly, VBMSRFA meta-features of WLMSF had the highest performance with prediction time of $\tau = 1.0$ s (GM = 0.80; $\varphi = 0.44$; Pr = 0.36; Sn = 0.69), whereas the best of baseline performance was achieved with a linear SVM classifier and MDF (GM = 0.76; $\varphi = 0.35$; Pr = 0.35; Sn = 0.70).

Conversely, applying VBMSRFA to MDF and WEPF, and using those meta-features with LDA classifiers to predict microsleep states resulted in a poorer performance. Excluding the TAN classifier, which had the poorest performances among all classifiers, the lowest detection performances were with the VBMSRFA meta-features and the baseline of WEPF and LDA

Table 8.5 Phi of different classifiers for microsleep state prediction with aggregated VBMSRFA meta-features for τ = 0–1 s. A bold value indicates the highest performance among selected classifiers and an italic value indicates the highest overall. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements with VBMSRFA meta-features relative to their respective baselines.

Feature set	τ (s)	LDA		Linear SVM		VBLR		TAN	
		Baseline	VBMSRFA	Baseline	VBMSRFA	Baseline	VBMSRFA	Baseline	VBMSRFA
MDF	0.00	0.34	0.36	0.38	**0.39**	0.36	0.36	0.28	0.17
	0.25	0.33	0.34	**0.37**	0.36	0.35	0.36	0.28	0.17
	0.50	0.33	0.34	**0.37**	0.35	0.36	0.35	0.27	0.17
	0.75	0.32	0.33	**0.37**	0.34	0.36	0.35	0.27	0.17
	1.00	0.31	0.32	**0.35**	0.33	0.34	0.34	0.26	0.16
PSF	0.00	0.36	**0.46****	0.40	0.40	0.37	0.39	0.21	0.05
	0.25	0.35	**0.45****	0.37	0.40	0.36	0.39	0.20	0.06
	0.50	0.34	**0.44****	0.36	0.39	0.35	0.38	0.20	0.06
	0.75	0.34	**0.43***	0.37	0.36	0.36	0.36	0.20	0.06
	1.00	0.33	**0.41***	0.33	0.36	0.34	0.36	0.19	0.05
PSF-IAF	0.00	0.37	**0.43**	0.36	0.39	0.36	0.37	0.26	0.08
	0.25	0.37	**0.43**	0.34	0.38	0.36	0.36	0.25	0.09
	0.50	0.36	**0.42**	0.36	0.36	0.36	0.36	0.25	0.07
	0.75	0.35	**0.40**	0.35	0.35	0.34	0.35	0.24	0.06
	1.00	0.34	**0.39**	0.33	0.34	0.34	0.34	0.23	0.06
WMSF	0.00	0.33	0.36	0.34	**0.38****	0.34	0.37$^{\sim}$	0.27	0.21
	0.25	0.33	0.36	0.35	**0.37***	0.34	0.36*	0.27	0.21
	0.50	0.33	**0.36**	0.34	**0.36**	0.33	**0.36**	0.27	0.21
	0.75	0.32	0.35	0.34	0.35	0.34	**0.36**	0.27	0.21
	1.00	0.32	0.34	0.33	0.34	0.33	**0.35**$^{\sim}$	0.26	0.21
WLMSF	0.00	0.36	**0.47****	0.35	0.45**	0.36	0.46**	0.24	0.22
	0.25	0.36	**0.46****	0.34	0.45*	0.35	0.45*	0.24	0.22
	0.50	0.35	**0.46****	0.35	0.44**	0.34	0.44*	0.24	0.21
	0.75	0.35	**0.45***	0.35	0.43*	0.34	0.44**	0.24	0.21
	1.00	0.34	**0.44***	0.33	0.43*	0.34	0.43**	0.23	0.21
WEPF	0.00	0.25	0.26	0.26	0.28	0.26	**0.29**	0.21	0.11
	0.25	0.24	0.26	0.26	0.27	0.26	**0.29**	0.21	0.11
	0.50	0.24	0.25	0.25	0.27	0.25	**0.29**	0.21	0.11
	0.75	0.23	0.25	0.25	0.25	0.25	**0.27**	0.20	0.12
	1.00	0.23	0.24	0.23	0.25	0.24	**0.27**	0.19	0.11

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

classifiers, i.e., GM = 0.63 for VBMSRFA (cf. 0.70 for baseline), φ = 0.26 for VBMSRFA (cf. 0.25 for baseline), Pr = 0.29 for VBMSRFA (cf. 0.25 for baseline), and Sn = 0.47 for VBMSRFA (cf. 0.67 for baseline).

8.4.2 Detection and prediction of microsleep onsets

Detection and prediction of imminent microsleeps requires accurate identification of the onset of a microsleep. This section provides the results of microsleep onset detection/prediction using aggregated VBMSRFA meta-features and compares the performances with the baselines. Since aggregated features showed higher performances with detection and prediction of microsleep states, we limit the results of this section to aggregated features.

The results of microsleep onset detection (τ = 0 s) using LDA classifiers and VBMSRFA aggregated meta-features are shown in Table 8.6. Two-tail Wilcoxon signed-rank tests were

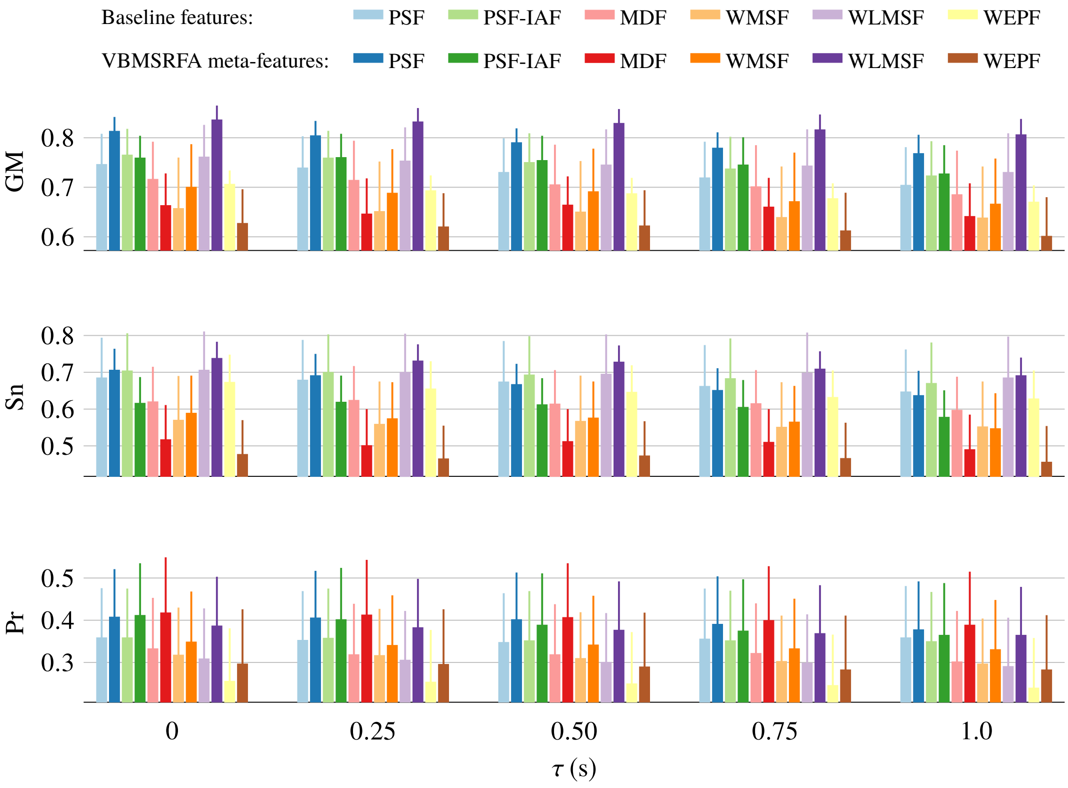


Figure 8.3 Performance (mean ± SE) of microsleep state prediction using VBMSRFA meta-features with an LDA classifier for τ = 0–1 s.

performed to identify significant improvements with respect to the baseline features with the same setup. There was slight improvements with VBMSRFA meta-features compared to baseline features, except for the WEPF in which GM and phi slightly declined. The highest improvement of GM was achieved with PSF (7.8%), while phi improved by 28.6% with PSF and WLMSF.

Performance of microsleep onset detection in terms of AUC-ROC and AUC-PR using other classifiers is presented in Table 8.7. The highest AUC-PR of 0.09 achieved with VBMSRFA meta-features was substantially lower than of 0.09 with the baseline features. Moreover, performance of the linear SVM with VBMSRFA meta-features was inferior to the baseline in all feature sets, except for WLMSF where the performance was similar. Notwithstanding, the performances of LDA and VBLR in terms of AUC-ROC improved in nearly all feature sets, compared to the baselines. With $\tau = 0$ s, a linear SVM performed better with baseline features, whereas VBMSRFA meta-features with an LDA had higher performances in most cases.

Figure 8.4 shows the AUC-ROC of microsleep onset prediction over prediction times up to 10 s. As expected, performance declined with prediction time τ irrespective of feature sets. VBMSRFA meta-features had lower performances than the best baseline performance. In addition, WEPF had the lowest AUC-ROC among all features. Unlike the baseline, LDA had mostly superior AUC-ROC with VBMSRFA meta-features among the classifiers. The best detection AUC-ROC of VBMSRFA meta-features was 0.91 and was achieved with a linear

Table 8.6 Performance (mean ± SE) of microsleep onset detection using VBMSRFA aggregated meta-features and an LDA classifier. A bold value indicates the highest performance in individual feature sets, and an italic value indicates the highest overall performance.

Feature set	Feature type	Microsleep onset detection performance					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	0.87±0.02	**0.03±0.01**	**0.69±0.05**	***0.09±0.02***	**0.55±0.08**	0.02±0.01
	VBMSRFA	***0.90±0.01***	**0.03±0.01**	0.68±0.06	***0.09±0.02***	0.54±0.08	***0.03±0.01***
PSF	Baseline	0.87±0.01	0.04±0.03	0.71±0.04	0.07±0.01	0.63±0.09	**0.02±0.01**
	VBMSRFA	**0.89±0.02**	***0.06±0.03****	**0.77±0.03**	***0.09±0.01***	***0.70±0.07***	**0.02±0.01**
PSF-IAF	Baseline	0.87±0.01	0.04±0.03	0.73±0.04	0.08±0.01	0.66±0.09	**0.02±0.01**
	VBMSRFA	**0.89±0.02****	***0.06±0.02***	**0.77±0.04**	***0.09±0.01***	**0.68±0.07**	**0.02±0.01**
WMSF	Baseline	0.88±0.02	0.03±0.02	0.69±0.07	0.08±0.02	0.60±0.11	***0.02±0.01***
	VBMSRFA	**0.89±0.02**	**0.04±0.02**	**0.72±0.08**	***0.09±0.02****	**0.62±0.10**	**0.02±0.01**
WLMSF	Baseline	0.88±0.02	**0.05±0.03**	0.73±0.05	0.07±0.01	***0.70±0.10***	0.01±0.00
	VBMSRFA	***0.90±0.01***	**0.05±0.02**	***0.78±0.03***	***0.09±0.02*****	0.69±0.05	**0.02±0.01**
WEPF	Baseline	**0.73±0.04**	**0.01±0.01**	**0.63±0.05**	**0.04±0.01**	**0.60±0.11**	**0.01±0.00**
	VBMSRFA	**0.73±0.04**	**0.01±0.00**	0.59±0.06	0.03±0.01	0.49±0.11	**0.01±0.00**

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table 8.7 Performance of different classifiers for microsleep onset detection with VBMSRFA aggregated meta-features. A bold value indicates the highest performance among classifiers, whereas an italic value indicates the highest performance overall.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.87	0.03	***0.91***	**0.06**	0.82	0.05	0.72	0.01
	VBMSRFA	**0.90**	**0.03**	0.82	**0.03**	0.86	**0.03**	0.72	0.01
PSF	Baseline	0.87	0.04	***0.91***	***0.09***	0.83	0.04	0.70	0.01
	VBMSRFA	**0.89**	**0.06**	**0.89**	**0.06**	0.84	0.03	0.75	0.01
PSF-IAF	Baseline	0.87	0.04	***0.91***	**0.07**	0.81	0.04	0.70	0.01
	VBMSRFA	**0.89**	**0.06**	0.85	0.05	0.85	0.04	0.72	0.01
WMSF	Baseline	0.88	0.03	**0.90**	**0.05**	0.80	0.02	0.74	0.01
	VBMSRFA	**0.89**	**0.04**	0.88	0.03	0.83	0.02	0.69	0.01
WLMSF	Baseline	0.88	0.05	**0.90**	**0.06**	0.83	0.04	0.79	0.02
	VBMSRFA	0.90	**0.05**	***0.91***	**0.05**	0.85	0.04	0.78	0.02
WEPF	Baseline	0.73	0.01	**0.78**	**0.02**	0.74	0.01	0.76	0.01
	VBMSRFA	0.73	**0.01**	**0.75**	**0.01**	**0.75**	**0.01**	0.70	**0.01**

SVM and WLMSF (cf. 0.91 for baseline with PSF and linear SVM). Similarly, the VBMSRFA meta-features of WLMSF had the highest AUC-ROC of 0.72 with a $\tau = 10\,\mathrm{s}$ (cf. 0.74 for baseline with a linear SVM and WMSF). On the other end of the spectrum, the lowest detection AUC-ROC of VBMSRFA meta-features was 0.69 with WMSF and a TAN classifier (cf. 0.70 for baseline with a TAN and PSF). The combination of VBLR and WEPF had the lowest AUC-ROC for both VBMSRFA meta-features (0.55) and the baseline (0.52).

Figure 8.5 shows the performance of microsleep onset prediction in terms of phi. Interestingly, the phi values for VBMSRFA meta-features were mostly superior to the baseline, especially with shorter prediction times. LDA and linear SVM had similar phi values with VBMSRFA meta-features of PSF and WLMSF. For the rest of the VBMSRFA meta-features, LDA had higher performances than other classifiers. On the other hand, VBLR and TAN classifiers had

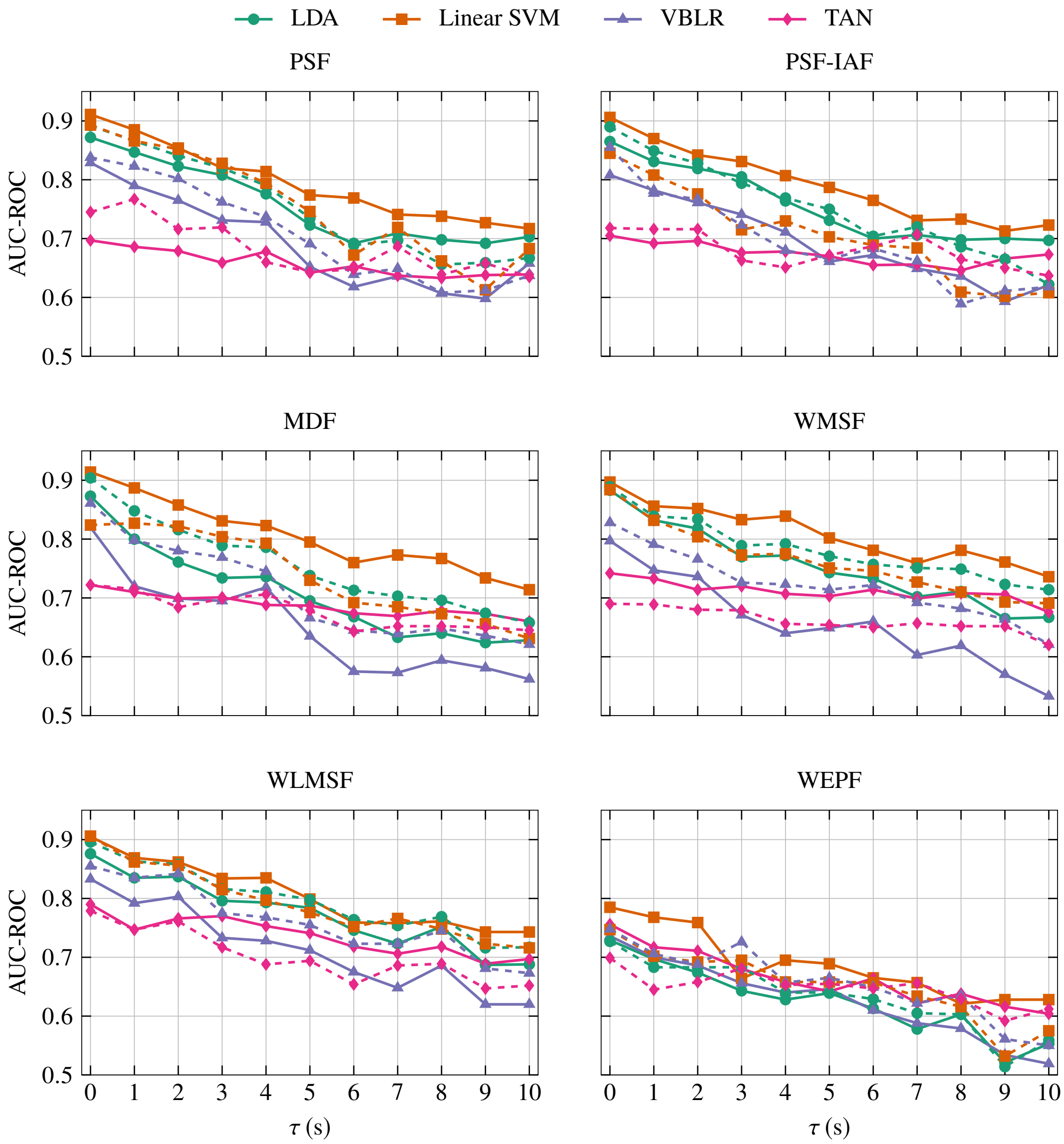


Figure 8.4 AUC-ROC of microsleep onset prediction using different classifiers versus prediction time for $\tau = 0$–10 s. Solid lines correspond to the baseline features, whereas dashed lines correspond to the VBMSRFA meta-features.

the lowest performances. The VBLR classifier outperformed TAN with shorter prediction times but the opposite was observed with PSF and PSF-IAF meta-features. Furthermore, comparing phi values with AUC-ROC values confirms that higher values of AUC-ROC does not guarantee higher phi values. This is in line with the findings of Zou et al. [2016]. Moreover, improvement of phi value without improvements in AUC-ROC indicates that a classification threshold might be subject dependent and thus a reduction of subject variability could increase phi value without necessarily increasing AUC-ROC.

The other performance measures – GM, sensitivity, and precision – of microsleep onset prediction using an LDA classifier are shown in Figure 8.6. It was observed that using VBMSRFA meta-features of PSF-IAF and WLMSF slightly improved GM compared to baseline. Conversely, the GM of microsleep onset prediction with VBMSRFA meta-features of WEPF was lower

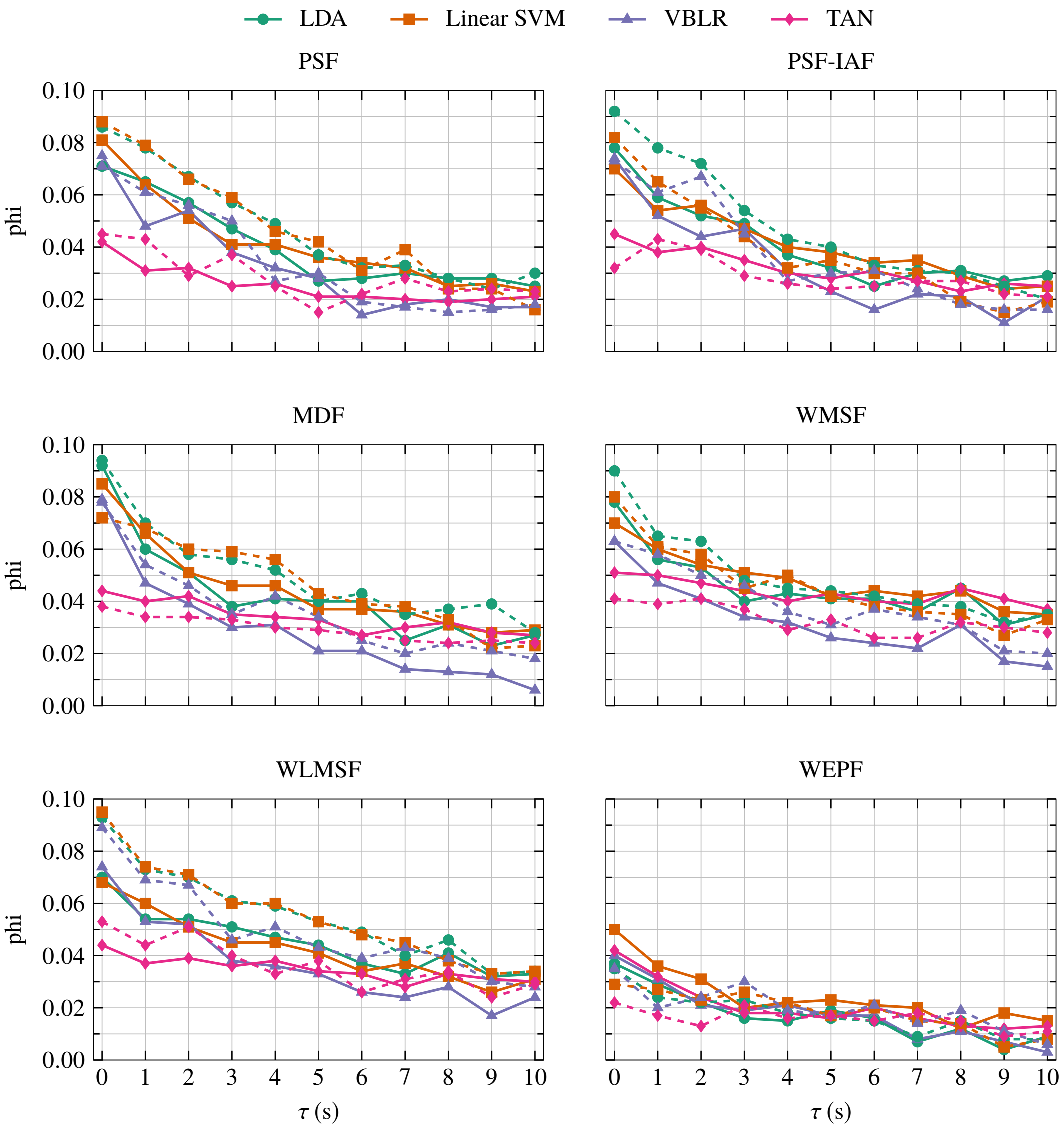


Figure 8.5 Phi of microsleep onset prediction using different classifiers versus prediction time for τ = 0–10 s. Solid lines correspond to the baseline features, whereas dashed lines correspond to the VBMSRFA meta-features.

than the baseline. Furthermore, the GM of VBMSRFA meta-features of WMSF were relatively similar to the baseline. The highest GM for microsleep onset detection was achieved with VBMSRFA meta-features of WLMSF (0.78 for LDA cf. 0.81 for linear SVM), while the highest GM of the baseline was 0.78 with a linear SVM and MDF. When the prediction time was τ = 10 s, the highest GM was 0.62, achieved with VBMSRFA meta-features of WLMSF, whereas the highest baseline GM of 0.55 was achieved with MDF. Furthermore, the VBMSRFA meta-features of WLMSF consistently performed slightly better across all prediction times.

There was no consistent improvement of sensitivity for microsleep onset prediction with VBMSRFA meta-features. However, slight improvements in GM with VBMSRFA led to slight improvements in precision, indicating a lower number of false positives. With τ = 0 s, the best sensitivity was achieved with a linear SVM and VBMSRFA meta-features of WLMSF

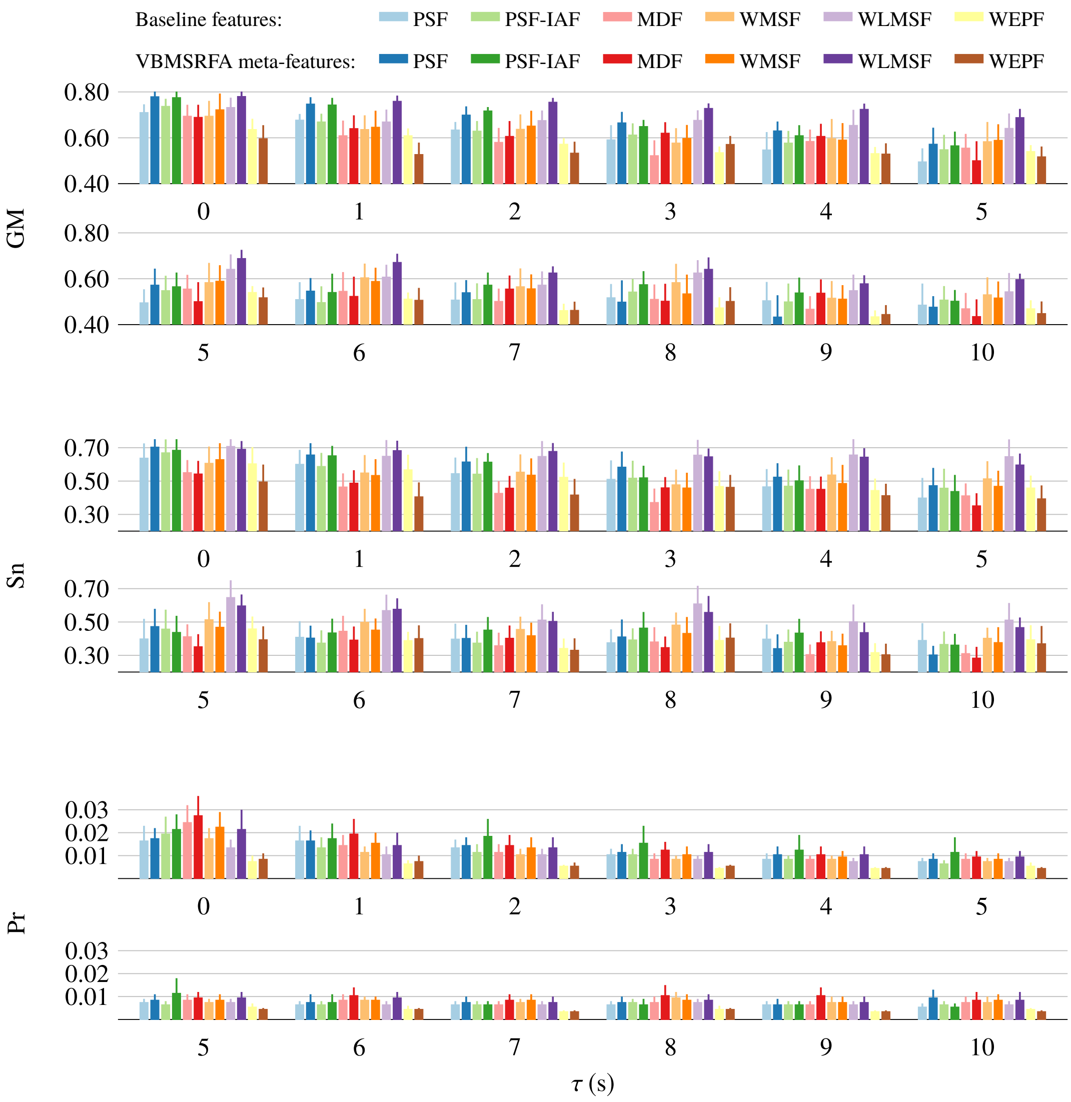


Figure 8.6 Performance (mean ± SE) of microsleep onset prediction using VBMSRFA meta-features and an LDA classifier for τ = 0–10 s.

(Sn = 0.77; Pr = 0.02), which was slightly lower than those of the baseline PSF (Sn = 0.79; Pr = 0.03). Although the sensitivity was moderate, the precision was too low.

8.5 SUMMARY

VBMSRFA was presented in this chapter, which is a Bayesian model for multi-subject feature reduction. VBMSRFA uses a robust latent space to reduce the noise sensitivity and allow each subject to have its own mean and noise terms. ARD-motivated prior distribution is used to automatically infer the optimum dimension of latent space. Variational inference for the training and testing phases was derived and presented. VBMSRFA was used for microsleep detection and prediction of states and onsets.

In terms of microsleep state detection and prediction, using VBMSRFA meta-features of PSF and WLMSF significantly improved phi, and improvements with PSF-IAF were also substantial. These improvements were achieved with LDA classifiers. The highest phi for microsleep state detection was 0.47 with WLMSF (cf. 0.40 for baseline with PSF). Increasing the prediction time to $\tau = 1.0$ s led to the highest phi of 0.44 with WLMSF (cf. 0.35 for baseline with MDF).

For microsleep onset detection, the baseline PSF had the highest AUC-PR of 0.09 (cf. 0.06 for VBMSRFA meta-features of PSF and PSF-IAF). Although the AUC-ROC did not improve with VBMSRFA meta-features, substantial improvement in GM was observed. The highest GM for microsleep onset detection was 0.81 with a linear SVM and VBMSRFA meta-features of WLMSF (cf. 0.78 for baseline with MDF). With a prediction time of $\tau = 10$ s, the VBMSRFA meta-features of WLMSF also had the highest GM of 0.62 (cf. 0.44 for baseline with MDF). Since the sensitivity of microsleep onset prediction with baseline features and VBMSRFA meta-features was relatively similar while the latter had higher GM values, the number of false positives of VBMSRFA was lower than that of the baseline which suggests a slightly higher precision with VBMSRFA meta-features.

Chapter 9

VARIATIONAL BAYESIAN HIERARCHICAL MULTI-SUBJECT ROBUST FACTOR ANALYSIS

9.1 INTRODUCTION

A Bayesian multi-subject feature reduction method, i.e., VBMSRFA, was presented in Chapter 8. However, VBMSRFA uses a fixed loading matrix for all the subjects. Such assumption however may be inaccurate since EEG is highly subject dependent, as mentioned in Chapter 8. Therefore, the aim of this chapter is to extend VBMSRFA by adding a hierarchical step allowing subjects to share a group-level loading matrix while each subject has its own loading matrix. The novel proposed method – variational Bayesian hierarchical multi-subject robust factor analysis (VBHMSRFA) – simultaneously finds individual mean, noise, and loading matrices while all subjects share a group-level loading matrix. At the testing phase, VBHMSRFA adapts the mean and noise terms for the test subject, while the group-level loading matrix is integrated out.

The Bayesian structure and the underlying assumptions of VBHMSRFA are presented in Section 9.2. It is then followed by the variational formulation and inference of the proposed method for training and testing phases, Section 9.3. Finally, the results and discussion of microsleep detection and prediction using VBHMSRFA meta-features of various feature sets are presented in Section 9.4.

9.2 BAYESIAN HIERARCHICAL MULTI-SUBJECT ROBUST FACTOR ANALYSIS

VBMSRFA extended the VBRFA to model multi-subjects assuming that the mean and noise terms of the individual subjects are independent. This can be further extended by letting subjects share information via a group loading matrix but have their own individual differences. Let $\mathbf{W}_s$ be the loading matrix of the subject s, the representation of the n^{th} data of subject s can be expressed as

$$\mathbf{x}_{s,n} = \mathbf{W}_s \mathbf{z}_{s,n} + \boldsymbol{\mu}_s + \boldsymbol{\varepsilon}_{s,n}, \tag{9.1}$$

$$\boldsymbol{\varepsilon}_{s,n} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Psi}_s^{-1}\right). \tag{9.2}$$

The mean and noise variables $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ are assumed to be independent for each subject, similar to BMSRFA. The prior probabilities over these variables are

$$p\left(\boldsymbol{\mu}_s\right) = \prod_{d=1}^{D} \mathcal{N}\left(\mu_{s,d} \,\middle|\, m_{\mu,d}, \psi_{s,d}^{-1}\right), \tag{9.3}$$

$$p\left(\boldsymbol{\Psi}_s\right) = \prod_{d=1}^{D} \mathcal{G}\left(\psi_{s,d} \,\middle|\, a_\psi, b_\psi\right). \tag{9.4}$$

Furthermore, the loading matrix of each subject, $\mathbf{W}_s$, is slightly different from the loading matrix of other subjects. However, the information is shared between subjects with a group-level loading matrix $\mathbf{M}_w$, as

$$\mathbf{W}_{s,k} = \mathbf{M}_{w,k} + \boldsymbol{\zeta}_{s,k}, \tag{9.5}$$

$$\boldsymbol{\zeta}_{s,k} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\alpha}^{-1}\right), \tag{9.6}$$

where $\boldsymbol{\zeta}_{s,k}$ is the variation of k^{th} column of the loading matrix of the subject s from the group-level loading matrix. To make the model fully Bayesian, the prior distributions over the columns of $\mathbf{M}_w$ are chosen as normal distributions,

$$p\left(\mathbf{M}_w \,\middle|\, \alpha\right) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{m}_{w,k} \,\middle|\, \mathbf{0}, \left(\beta_w \alpha_k \mathbf{I}\right)^{-1}\right). \tag{9.7}$$

A set of independent Gamma distributions are selected as the prior probabilities over $\boldsymbol{\alpha}$,

$$p\left(\boldsymbol{\alpha}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,\middle|\, a_\alpha, b_\alpha\right). \tag{9.8}$$

An ARD-motivated prior, using $\boldsymbol{\alpha}$, was used to regularize the group-level loading matrix and promote group-level sparsity. In addition, using $\boldsymbol{\alpha}$ to regularize the individual-level loading matrices results in a similar regularization for all subjects. This can be shown by finding the joint probability of the group-level and individual-level loading matrices and integrating out the group-level loading matrix, that is

$$\begin{aligned} p\left(\mathbf{W}_s \,\middle|\, \alpha\right) &= \int p\left(\mathbf{W}_s \,\middle|\, \mathbf{M}_w, \alpha\right) p\left(\mathbf{M}_w \,\middle|\, \alpha\right) d\mathbf{M}_w \\ &= \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_{s,k} \,\middle|\, \mathbf{0}, \left(\frac{\beta_w}{1+\beta_w}\alpha_k\right)^{-1}\right). \end{aligned} \tag{9.9}$$

Thus, extra latent variables will be automatically pruned out from both the individual and group-level matrices. Similar to the BRFA and BMSRFA, the prior distribution of the latent variables is chosen as

$$p\left(\mathbf{z}_{s,n}\right) = \mathcal{N}\left(\mathbf{z}_{s,n} \,\middle|\, \mathbf{0}, \boldsymbol{\Lambda}\right), \tag{9.10}$$

$$p\left(\boldsymbol{\Lambda}\right)=\prod_{k=1}^{K}\mathcal{G}\left(\lambda_k\,\middle|\,a_\lambda,b_\lambda\right),\tag{9.11}$$

leading to a marginal Student-t distribution for the latent variables. A graphical model representation of the Bayesian hierarchical MSRFA (BHMSRFA) is shown Figure 9.1.

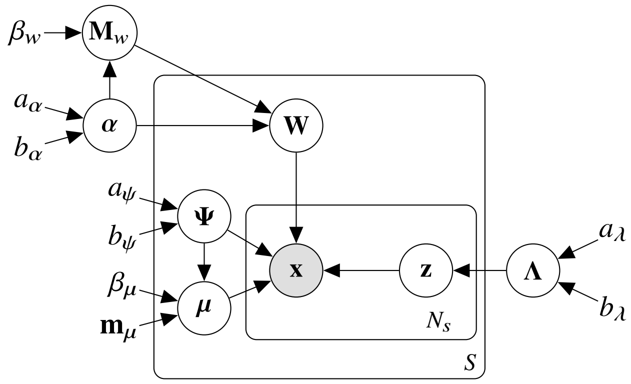


Figure 9.1 Graphical model representation of Bayesian hierarchical multi-subject robust factor analysis (BHMSRFA) model.

9.3 VARIATIONAL INFERENCE

9.3.1 Training phase

An analytical Bayesian inference of BHMSRFA is intractable. A variational inference is therefore used to approximate the posterior distribution of the model parameters from the training data. The variational distribution is assumed to be factorized as

$$q\left(\Theta\right)=q\left(\mathbf{M}_w\,\middle|\,\alpha\right)q\left(\alpha\right)q\left(\boldsymbol{\Lambda}\right)\prod_{s=1}^{S}\left(q\left(\mathbf{W}_s\right)q\left(\boldsymbol{\mu}_s\,\middle|\,\boldsymbol{\Psi}_s\right)q\left(\boldsymbol{\Psi}_s\right)\prod_{n=1}^{N_s}q\left(\mathbf{z}_{s,n}\,\middle|\,\boldsymbol{\Lambda}\right)\right),\tag{9.12}$$

where $\Theta=\left\{\mathbf{M}_w,\alpha,\boldsymbol{\Lambda},\mathbf{W},\boldsymbol{\mu},\boldsymbol{\Psi}\right\}$ is the set of all model parameters. Using Equations (3.7) and (9.12), the lower bound of the evidence log-likelihood can be written as

$$\begin{aligned}\mathcal{L}=&\left\langle\ln\left(\frac{p\left(\boldsymbol{\Lambda}\right)}{q\left(\boldsymbol{\Lambda}\right)}\right)\right\rangle+\left\langle\ln\left(\frac{p\left(\alpha\right)}{q\left(\alpha\right)}\right)\right\rangle+\left\langle\ln\left(\frac{p\left(\mathbf{M}_w\,\middle|\,\alpha\right)}{q\left(\mathbf{M}_w\,\middle|\,\alpha\right)}\right)\right\rangle\\&+\sum_{s=1}^{S}\left\{\left\langle\ln\left(\frac{p\left(\boldsymbol{\mu}_s\,\middle|\,\boldsymbol{\Psi}_s\right)}{q\left(\boldsymbol{\mu}_s\,\middle|\,\boldsymbol{\Psi}_s\right)}\right)\right\rangle+\left\langle\ln\left(\frac{p\left(\boldsymbol{\Psi}_s\right)}{q\left(\boldsymbol{\Psi}_s\right)}\right)\right\rangle+\left\langle\ln\left(\frac{p\left(\mathbf{W}_s\,\middle|\,\mathbf{M}_w,\alpha\right)}{q\left(\mathbf{W}_s\right)}\right)\right\rangle\right\}\\&+\sum_{s=1}^{S}\sum_{n=1}^{N_s}\left\{\left\langle\ln\left(\frac{p\left(\mathbf{z}_{s,n}\,\middle|\,\boldsymbol{\Lambda}\right)}{q\left(\mathbf{z}_{s,n}\right)}\right)\right\rangle+\left\langle\ln\left(p\left(\mathbf{x}_{s,n}\,\middle|\,\mathbf{W}_s,\boldsymbol{\mu}_s,\mathbf{z}_{s,n},\boldsymbol{\Psi}_s\right)\right)\right\rangle\right\},\end{aligned}\tag{9.13}$$

where the latent variables and the group-level loading matrix $\mathbf{M}_w$ are regularized based on all the subjects, the mean and noise terms are regularized per subject, and the individual loading matrices are regularized based on the both the subject and group-level data. A variational EM is utilized to maximize the lower bound of the marginal data log-probability. The VBE step updates the parameters of the variational distribution of the latent variables, using

$$\tilde{\Sigma}_{z,s} = \Big(\big\langle \mathbf{W}_s^\top \mathbf{\Psi}_s \mathbf{W}_s \big\rangle + \langle \mathbf{\Lambda} \rangle\Big)^{-1}, \tag{9.14}$$

$$\tilde{\mathbf{m}}_{z,s,n} = \tilde{\Sigma}_{z,s} \big\langle \mathbf{W}_s^\top \big\rangle \big\langle \mathbf{\Psi}_s \big\rangle \Big(\mathbf{x}_{s,n} - \langle \boldsymbol{\mu}_s \rangle\Big), \tag{9.15}$$

where the variational distribution of the latent variables is given by

$$q\left(\mathbf{Z}\right) = \prod_{s=1}^{S} \prod_{n=1}^{N_s} \mathcal{N}\left(\mathbf{z}_{s,n} \,\middle|\, \tilde{\mathbf{m}}_{z,s,n}, \tilde{\Sigma}_{z,s}\right). \tag{9.16}$$

After updating the distribution parameters of the latent variables, the VBM step updates the variational distribution of the rest of parameters to maximize the lower bound of the evidence log-likelihood. This can be achieved by updating the variational parameters using

$$\tilde{\beta}_{\mu,s} = N_s + \beta_\mu, \tag{9.17}$$

$$\tilde{m}_{\mu,s,d} = \frac{1}{\tilde{\beta}_{\mu,s}} \left(\beta_\mu m_{\mu,d} + \sum_{n=1}^{N_s} \Big(x_{s,n,d} - \langle \mathbf{w}_{s,d,.} \rangle \langle \mathbf{z}_{s,n} \rangle\Big)\right), \tag{9.18}$$

$$\tilde{a}_{\psi,s} = a_\psi + \frac{N_s}{2}, \tag{9.19}$$

$$\begin{aligned} \tilde{b}_{\psi,s,d} = b_\psi + \frac{\beta_\mu}{2} m_{\mu,d}^2 - \frac{\tilde{\beta}_{\mu,s}}{2} \tilde{m}_{\mu,s,d}^2 + \frac{1}{2} \sum_{n=1}^{N_s} \Bigg(x_{s,n,d}^2 \\ - 2x_{s,n,d} \langle \mathbf{w}_{s,d,.} \rangle \langle \mathbf{z}_{s,n} \rangle + \mathrm{tr}\left(\big\langle \mathbf{w}_{s,d,.}^\top \mathbf{w}_{s,d,.} \big\rangle \big\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^\top \big\rangle\right) \Bigg), \end{aligned} \tag{9.20}$$

$$\tilde{\Sigma}_{w,s,d} = \left(\big\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \big\rangle + \langle \psi_{s,d} \rangle \sum_{n=1}^{N_s} \big\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}{}^\top \big\rangle\right)^{-1}, \tag{9.21}$$

$$\tilde{\mathbf{w}}_{s,d} = \tilde{\Sigma}_{w,s,d} \left(\sum_{n=1}^{N_s} \langle \mathbf{z}_{s,n} \rangle \Big(\langle \psi_{s,d} \rangle x_{s,n,d} - \langle \psi_{s,d} \mu_{s,d} \rangle\Big) + \big\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \big\rangle \big\langle \mathbf{m}_{\mathbf{w},d,.}^\top \big\rangle\right), \tag{9.22}$$

$$\tilde{\beta}_w = S + \beta_w, \tag{9.23}$$

$$\tilde{\mathbf{m}}_{\mathbf{w},k} = \tilde{\beta}_w^{-1} \left(\sum_{s=1}^{S} \langle \mathbf{w}_{s,k} \rangle\right), \tag{9.24}$$

$$\tilde{a}_\alpha = a_\alpha + \frac{SD}{2}, \tag{9.25}$$

$$\tilde{b}_{\alpha,k} = b_\alpha + \frac{\sum_{s=1}^{S} \big\langle \mathbf{w}_{s,k}^\top \mathbf{w}_{s,k} \big\rangle - \tilde{\beta}_w \tilde{\mathbf{m}}_{\mathbf{w},k}^\top \tilde{\mathbf{m}}_{\mathbf{w},k}}{2}, \tag{9.26}$$

$$\tilde{a}_\lambda = a_\lambda + \frac{1}{2} \sum_{s=1}^{S} N_s, \tag{9.27}$$

$$\tilde{b}_{\lambda,k} = b_\lambda + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \big\langle z_{s,n,k}^2 \big\rangle, \tag{9.28}$$

where the variational distributions are given by

$$q\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right) = \prod_{s=1}^{S}\prod_{d=1}^{D} \mathcal{N}\left(\mu_{s,d} \,\middle|\, \tilde{m}_{\mu,s,d}, \left(\tilde{\beta}_{\mu,s}\psi_{s,d}\right)^{-1}\right), \tag{9.29}$$

$$q\left(\boldsymbol{\Psi}\right) = \prod_{s=1}^{S}\prod_{d=1}^{D} \mathcal{G}\left(\psi_{s,d} \,\middle|\, \tilde{a}_{\psi,s}, \tilde{b}_{\psi,s,d}\right), \tag{9.30}$$

$$q\left(\mathbf{W}\right) = \prod_{s=1}^{S}\prod_{d=1}^{D} \mathcal{N}\left(\mathbf{w}_{s,d,.}^{\top} \,\middle|\, \tilde{\mathbf{w}}_{s,d}, \tilde{\Sigma}_{w,s,d}\right), \tag{9.31}$$

$$q\left(\mathbf{M}_{w} \,\middle|\, \alpha\right) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{m}_{w,k} \,\middle|\, \tilde{\mathbf{m}}_{w,k}, \left(\tilde{\beta}_{w}\alpha_{k}\mathbf{I}\right)^{-1}\right), \tag{9.32}$$

$$q\left(\boldsymbol{\alpha}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_{k} \,\middle|\, \tilde{a}_{\alpha}, \tilde{b}_{\alpha,k}\right), \tag{9.33}$$

$$q\left(\boldsymbol{\Lambda}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_{k} \,\middle|\, \tilde{a}_{\lambda}, \tilde{b}_{\lambda,k}\right). \tag{9.34}$$

The derivation of the variational distributions are given in Appendix C.

Iterating over VBE and VBM steps is guaranteed not to decrease the evidence log-likelihood lower bound. The improvement of the latter can be used to monitor the convergence. In this research, the stopping criteria was set to a threshold of 10^{-6} for relative improvement of the lower bound of the evidence log-likelihood.

The hyperparameters $\left\{\mathbf{m}_{\mu}, \beta_{\mu}, \beta_{w}, a_{\alpha}, b_{\alpha}, a_{\psi}, b_{\psi}, a_{\lambda}, b_{\lambda}\right\}$ can also be updated to increase the lower bound of the evidence log-likelihood. Taking the derivative of the lower bound $\mathcal{L}$ with respect to $m_{\mu,d}$ leads to the updating equation,

$$m_{\mu,d} = \frac{\sum_{s=1}^{S}\left(\langle\psi_{s,d}\rangle\tilde{m}_{\mu,s,d}\right)}{\sum_{s=1}^{S}\langle\psi_{s,d}\rangle}, \tag{9.35}$$

which updates the prior mean using a weighted average of the posterior means relative to their uncertainties. Similarly, an updating equation for β_{μ} can be found as

$$\beta_{\mu}^{-1} = \frac{1}{SD}\sum_{s=1}^{S}\left(D\tilde{\beta}_{\mu,s}^{-1} + \sum_{d=1}^{D}\langle\psi_{s,d}\rangle\left(\tilde{m}_{\mu,s,d} - m_{\mu,d}\right)^{2}\right), \tag{9.36}$$

where $m_{\mu,d}$ has been updated using Equation (9.35). Following the same procedure for β_{w}, an updating equation can be found as

$$\beta_{w}^{-1} = \frac{1}{KD}\left(\tilde{\beta}_{w}^{-1} + \sum_{k=1}^{K}\left\langle\alpha_{k}\tilde{\mathbf{m}}_{\mathbf{w},k}^{\top}\tilde{\mathbf{m}}_{\mathbf{w},k}\right\rangle\right). \tag{9.37}$$

While the derivative of $\mathcal{L}$ with respect to the hyperparameters of the Gamma distributions does not lead to an analytical solution, a coordinate descent optimization can be used to update them. Similar to VBMSRFA, updating hyperparameters of the Gamma distributions can be achieved by fixed-point iterating over

$$\Psi\left(a_\alpha\right) = \ln\left(b_\alpha\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\Psi\left(\tilde{a}_\alpha\right) - \ln\left(\tilde{b}_{\alpha,k}\right)\right), \tag{9.38}$$

$$\Psi\left(a_\psi\right) = \ln\left(b_\psi\right) + \frac{1}{SD}\sum_{s=1}^{S}\sum_{d=1}^{D}\left(\Psi\left(\tilde{a}_{\psi,s}\right) - \ln\left(\tilde{b}_{\psi,s,d}\right)\right), \tag{9.39}$$

$$\Psi\left(a_\lambda\right) = \ln\left(b_\lambda\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\Psi\left(\tilde{a}_\lambda\right) - \ln\left(\tilde{b}_{\lambda,k}\right)\right), \tag{9.40}$$

$$b_\alpha^{-1} = \frac{1}{a_\alpha K}\sum_{k=1}^{K}\frac{\tilde{a}_\alpha}{\tilde{b}_{\alpha,k}}, \tag{9.41}$$

$$b_\psi^{-1} = \frac{1}{a_\psi SD}\sum_{s=1}^{S}\sum_{d=1}^{D}\frac{\tilde{a}_{\psi,s}}{\tilde{b}_{\psi,s,d}}, \tag{9.42}$$

$$b_\alpha^{-1} = \frac{1}{a_\lambda K}\sum_{k=1}^{K}\frac{\tilde{a}_\lambda}{\tilde{b}_{\lambda,k}}. \tag{9.43}$$

To speed up the training phase, the hyperparamters are updated every 10 iterations.

Initialization of the training phase is done by setting the hyperparameters to a set of values that yields non-informative prior distributions and “lets the data speak for itself” [Beal 2003]. Thus, the initial values of $\left\{\beta_w, \beta_\mu, a_\alpha, b_\alpha, a_\psi, b_\psi, a_\lambda, b_\lambda\right\}$ are set to 10^{-6} and $\mathbf{m}_\mu$ is set to zero. The empirical mean of each subject’s data is used to initialize $\tilde{\mathbf{m}}_{\mu,s}$. Furthermore, a PCA is performed on the concatenated demeaned data of all training subjects, i.e., the data of each subject is demeaned with respect to its own empirical mean, and the first $K = D - 1$ columns of the PCA loading matrix are used to initialize $\tilde{\mathbf{M}}_w$. Similarly, a PCA is applied to the individual subject’s data. Then, the first K columns of the loading matrix of the PCA is used to initiate subject-level loading matrix $\tilde{\mathbf{W}}_s$. The rest of the variational parameters are initialized to the values of their respective hyperparameters.

The training algorithm is then continued by applying the VBE step to update the approximate distribution of the latent variables and the VBM to update the variational distribution of the parameters. For every 10 iterations, the hyperparameters are updated to increase the lower bound of the evidence log-likelihood. In addition, the ARD-motivated method is used to regularize the columns of the loading matrices, both at individual and group levels, which reduces the value of the redundant components to zero. To remove the redundant components, the lower bound of the evidence log-likelihood with and without the component corresponding to the largest value of $\langle\alpha\rangle$ is calculated at every iteration. An improvement of $\mathcal{L}$ after removing a component confirms the appropriateness of removing the component from the model. Finally, the relative improvement of the lower bound of the log-likelihood of the data is calculated at the end of each iteration using

$$Rel.\ Imprv. = \frac{\mathcal{L}\left(iter\right) - \mathcal{L}\left(iter - 1\right)}{\left|\mathcal{L}\left(iter\right)\right|}, \tag{9.44}$$

and the training algorithm is stopped if the relative improvement is less than 10^{-6}. The pseudo code of the training phase of VBHMSRFA is presented in Algorithm 9.1.

Algorithm 9.1 The training algorithm of variational Bayesian hierarchical multi-subject robust factor analysis (VBHMSRFA).

```
procedure INITIALIZING
    $K = D - 1$, $\mathbf{m}_\mu = \mathbf{0}$, $a_\alpha = b_\alpha = a_\psi = b_\psi = a_\lambda = b_\lambda = \beta_\mu = \beta_w = 10^{-6}$.
    $\tilde{a}_{\psi,s} = a_\psi$, $\tilde{\mathbf{b}}_{\psi,s} = b_\psi$, $\tilde{a}_\lambda = a_\lambda$, $\tilde{\mathbf{b}}_\lambda = b_\lambda$, $\tilde{a}_\alpha = a_\alpha$, $\tilde{\mathbf{b}}_\alpha = b_\alpha$, $\tilde{\beta}_{\mu,s} = \beta_\mu$, $\tilde{\beta}_w = \beta_w$.
    $RelTol = 10^{-6}$, $MaxIter = 1000$.
    for $s = 1$ to $S$ do
        Set $\tilde{\mathbf{m}}_{\mu,s}$ to the empirical mean of the subject $s$ data.
        Set $\tilde{\mathbf{W}}_s$ to the first $K$ columns of the PCA loading matrix of the data of subject $s$.
        $\tilde{\Sigma}_{w,s,d} = \mathbf{I}, \forall d \in \left\{1, \ldots, D\right\}$.
    Set $\tilde{\mathbf{M}}_w$ to the first $K$ components of the PCA coefficients of the concatenated demeaned datasets.
for $iter = 1$ to $MaxIter$ do
    procedure VBE
        for $s = 1$ to $S$ do
            for $n = 1$ to $N_s$ do
                Update the approximate distribution of the latent variables using Equations (9.14) and (9.15).
    procedure VBM
        for $s = 1$ to $S$ do
            Update the subject specific variational parameters using Equations (9.17)–(9.22).
        Update the shared variational parameters using Equations (9.23)–(9.28).
    procedure UPDATE HYPERPARAMETERS
        if Reminder($iter$, 10) is 0 then
            Update $\mathbf{m}_\mu$, $\beta_\mu$, and $\beta_w$ using Equations (9.35)–(9.37).
            Update Gamma hyperparameters by fixed point iterating over Equations (9.38)–(9.43).
    procedure STOPPING CRITERIA
        Calculate the lower bound of the marginal log-likelihood.
        if $\frac{\mathcal{L}\left(iter\right) - \mathcal{L}\left(iter-1\right)}{\left|\mathcal{L}\left(iter\right)\right|} < RelTol$ then
            Stop.                                                    ▹ Converged
    procedure PRUNING LATENT VARIABLES
        Temporarily remove the component corresponding to the highest $\langle\alpha\rangle$.
        Calculate the lower bound of the marginal log-likelihood.
        if $\mathcal{L}$ (after pruning) > $\mathcal{L}$ (before pruning) then     ▹ The component can be removed.
            $\mathcal{L}\left(iter\right) = \mathcal{L}$ (after pruning) and remove the component.
        else                                          ▹ The component can not be removed.
            $\mathcal{L}\left(iter\right) = \mathcal{L}$ (before pruning) and keep the component.
```

9.3.2 Testing phase

The testing phase aims to find the posterior probability of the latent variables given new data, where the MAPs of the latent variables can be used as meta-features for proceeding classification tasks. Analogous to VBMSRFA, the distributions of the mean and noise terms of the test subject have to be estimated from its own data, while the data arrives sequentially. Therefore, an incremental algorithm is used to infer the posterior of the mean and noise incrementally, as well as maximize the evidence likelihood. The latter can be written as

$$\begin{aligned} p\left(\mathbf{x}_n \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathcal{D}_{\text{train}}\right) &= \int p\left(\mathbf{x}_n \,\middle|\, \mathbf{z}_n, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) p\left(\mathbf{W} \,\middle|\, \mathbf{M}_w, \boldsymbol{\alpha}\right) p\left(\mathbf{M}_w, \boldsymbol{\alpha} \,\middle|\, \mathcal{D}_{\text{train}}\right) \\ &\qquad p\left(\mathbf{z}_n \,\middle|\, \boldsymbol{\Lambda}\right) p\left(\boldsymbol{\Lambda} \,\middle|\, \mathcal{D}_{\text{train}}\right) p\left(\boldsymbol{\mu}, \boldsymbol{\Psi} \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right) d\Theta \\ &\approx \int p\left(\mathbf{x}_n \,\middle|\, \mathbf{z}_n, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) p\left(\mathbf{W} \,\middle|\, \mathbf{M}_w, \boldsymbol{\alpha}\right) q\left(\mathbf{M}_w, \boldsymbol{\alpha}\right) \\ &\qquad p\left(\mathbf{z}_n \,\middle|\, \boldsymbol{\Lambda}\right) q\left(\boldsymbol{\Lambda}\right) p\left(\boldsymbol{\mu}, \boldsymbol{\Psi} \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right) d\Theta, \end{aligned} \tag{9.45}$$

where $\Theta = \left\{\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}, \boldsymbol{\Lambda}, \mathbf{M}_w, \boldsymbol{\alpha}\right\}$ is the set of all model parameters and the probability distribution of $\mathbf{M}_w$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Lambda}$ are approximated by their respective training variational distributions. To further simplify Equation (9.45), $\mathbf{M}_w$ can be integrated out, that is

$$\begin{aligned} p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right) &= \int p\left(\mathbf{W} \,\middle|\, \mathbf{M}_w, \boldsymbol{\alpha}\right) q\left(\mathbf{M}_w \,\middle|\, \boldsymbol{\alpha}\right) d\mathbf{M}_w \\ &= \int \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k \,\middle|\, \mathbf{m}_{w,k}, \alpha_k^{-1}\mathbf{I}\right) \mathcal{N}\left(\mathbf{m}_{w,k} \,\middle|\, \tilde{\mathbf{m}}_{w,k}, \left(\tilde{\beta}_w \alpha_k\right)^{-1} \mathbf{I}\right) d\mathbf{m}_{w,k} \\ &= \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k \,\middle|\, \tilde{\mathbf{m}}_{w,k}, \left(\delta \alpha_k\right)^{-1} \mathbf{I}\right), \end{aligned} \tag{9.46}$$

where $\delta = {}^{\tilde{\beta}_w}/\left(1+\tilde{\beta}_w\right)$ is a constant. Furthermore, we added a forgetting factor to the mean and noise distributions to make the system more adaptive. The evidence likelihood can be simplified using Equation (9.46), as

$$\begin{aligned} p\left(\mathbf{x}_n \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathcal{D}_{\text{train}}\right) &\approx \int p\left(\mathbf{x}_n \,\middle|\, \mathbf{z}_n, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right) q\left(\boldsymbol{\alpha}\right) \\ &\qquad p\left(\mathbf{z}_n \,\middle|\, \boldsymbol{\Lambda}\right) q\left(\boldsymbol{\Lambda}\right) p\left(\boldsymbol{\mu}, \boldsymbol{\Psi} \,\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right)^{\kappa} d\Theta, \end{aligned} \tag{9.47}$$

where κ is an exponential forgetting factor constant and can take values between 0 and 1. Using a value of $\kappa = 1$ makes the system remember all the earlier data to estimate the posterior of the mean and noise terms, while $\kappa < 1$ causes the model to assign exponentially smaller weights to the older data. This essentially lets the mean and noise terms adapt to more recent data. A variational EM is employed to approximate the posterior distribution of all of the parameters with a factorized variational distribution as

$$q^*\left(\Theta^{(n)}\right) = q^*\left(\mathbf{W}\right) q^*\left(\boldsymbol{\alpha}\right) q^*\left(\mathbf{z}_n\right) q^*\left(\boldsymbol{\Lambda}\right) q^*\left(\boldsymbol{\mu}^{(n)}, \boldsymbol{\Psi}^{(n)}\right), \tag{9.48}$$

where $\Theta^{(n)}$ corresponds to the variational parameters after observing the n^{th} data. After observing $\mathbf{x}_n$, the approximate distribution of the latent variables is updated by the VBE step

$$\hat{\Sigma}_z^{(n)} = \Big(\langle \boldsymbol{\Lambda} \rangle + \langle \mathbf{W}^\top \boldsymbol{\Psi} \mathbf{W} \rangle\Big)^{-1}, \tag{9.49}$$

$$\hat{\mathbf{m}}_z^{(n)} = \hat{\Sigma}_z^{(n)} \langle \mathbf{W} \rangle^\top \langle \boldsymbol{\Psi} \rangle \left(\mathbf{x}_n - \langle \boldsymbol{\mu} \rangle\right). \tag{9.50}$$

The VBM step then updates the variational parameters to maximize the marginal log-likelihood of the newly observed data by

$$\hat{a}_\alpha^{(n)} = \tilde{a}_\alpha + \frac{D}{2}, \tag{9.51}$$

$$\hat{b}_{\alpha,k}^{(n)} = \frac{\delta}{2} \Big\langle \left(\mathbf{w}_k - \tilde{\mathbf{m}}_{\mathbf{w},k}\right)^\top \left(\mathbf{w}_k - \tilde{\mathbf{m}}_{\mathbf{w},k}\right) \Big\rangle + \tilde{b}_{\alpha,k}, \tag{9.52}$$

$$\hat{\Sigma}_{w,d}^{(n)} = \Big(\delta \big\langle \operatorname{diag}(\boldsymbol{\alpha}) \big\rangle + \langle \psi_d \rangle \langle \mathbf{z}_n \mathbf{z}_n^\top \rangle \Big)^{-1}, \tag{9.53}$$

$$\hat{\mathbf{w}}_{d,.}^{(n)} = \hat{\Sigma}_{w,d}^{(n)} \Big(\delta \big\langle \operatorname{diag}(\boldsymbol{\alpha}) \big\rangle \tilde{\mathbf{m}}_{\mathbf{w},d,.}^\top + \langle \mathbf{z}_n \rangle \langle \psi_d \rangle \left(x_{n,d} - \langle \mu_d \rangle \right) \Big), \tag{9.54}$$

$$\hat{a}_\lambda^{(n)} = \tilde{a}_\lambda + \frac{1}{2}, \tag{9.55}$$

$$\hat{b}_{\lambda,k}^{(n)} = \tilde{b}_{\lambda,k} + \frac{1}{2} \big\langle z_{n,k}^2 \big\rangle, \tag{9.56}$$

$$\hat{\beta}_\mu^{(n)} = 1 + \kappa \hat{\beta}_\mu^{(n-1)}, \tag{9.57}$$

$$\hat{m}_{\mu,d}^{(n)} = \frac{1}{\hat{\beta}_\mu^{(n)}} \Big(\kappa \hat{\beta}_\mu^{(n-1)} \hat{m}_{\mu,d}^{(n-1)} + x_{n,d} - \hat{\mathbf{w}}_{d,.}^{(n)} \hat{\mathbf{z}}^{(n)} \Big), \tag{9.58}$$

$$\hat{a}_\psi^{(n)} = \kappa \left(\hat{a}_\psi^{(n-1)} - \frac{1}{2} \right) + 1, \tag{9.59}$$

$$\begin{aligned} \hat{b}_{\psi,d}^{(n)} = {} & \kappa \hat{b}_{\psi,d}^{(n-1)} + \frac{\kappa}{2} \hat{\beta}_\mu^{(n-1)} \left(\hat{m}_{\mu,d}^{(n-1)} \right)^2 - \frac{1}{2} \hat{\beta}_\mu^{(n)} \left(\hat{m}_{\mu,d}^{(n)} \right)^2 \\ & + \frac{1}{2} \Big(x_{n,d}^2 - 2 x_{n,d} \hat{\mathbf{w}}_{d,.}^{(n)} \hat{\mathbf{z}}^{(n)} + \operatorname{tr}\Big(\langle \mathbf{z}_n \mathbf{z}_n^\top \rangle \big\langle \mathbf{w}_{d,.}^\top \mathbf{w}_{d,.} \big\rangle \Big) \Big). \end{aligned} \tag{9.60}$$

The variational EM converges after a few iterations of the VBE and VBM steps. The lower bound of the log-likelihood of the recently observed data can be used to monitor the convergence. The MAP of the latent variables $\hat{\mathbf{M}}_z = \left\{ \hat{\mathbf{m}}_z^{(1)}, \hat{\mathbf{m}}_z^{(2)}, \ldots, \hat{\mathbf{m}}_z^{(N)} \right\}$ is then used as meta features for further processing and classification.

9.4 RESULTS AND DISCUSSION

9.4.1 Detection and prediction of microsleep states

VBHMSRFA was first applied to various feature sets extracted from 2, 5, and 10 s EEG windows. For all of the feature sets, an arbitrary value of $\kappa = 0.9999$ was used to avoid forgetting the previous observations too quickly, which assigns more weights on the recent ~30 min of data. However, no experiment was done to fine tune the value of κ. Results of the microsleep state detection ($\tau = 0\,\mathrm{s}$) using VBHMSRFA meta-features of a single EEG window and an LDA classifier, are presented in Table 9.1. Significant improvements of AUC-ROC and AUC-PR were observed with the meta-features of MDF. AUC-ROC and AUC-PR of VBHMSRFA meta-features of MDF extracted from 2-s EEG windows were 0.91 (cf. 0.86 for baseline) and 0.37 (cf. 0.31 for baseline), respectively. Similarly, performances of the meta-features of MDF extracted from 5-s EEG windows were AUC-ROC = 0.91 (cf. 0.88 for baseline) and 0.41 (cf. 0.36 for baseline). There were also slight improvements with meta-features of WMSF and WEPF. However, there were no consistent improvements for meta-features of PSF, PSF-IAF, and WLMSF.

Table 9.1 Performance (mean ± SE) of microsleep state detection with VBHMSRFA meta-features and an LDA classifier. A bold value indicates the highest performance of each feature set and italics indicate the highest overall. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements relative to the baselines.

Feature set	Feature type	2-s EEG window		5-s EEG window		10-s EEG window	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.86±0.03	0.31±0.12	0.88±0.03	0.36±0.12	0.86±0.04	0.36±0.12
	VBHMSRFA	**0.91±0.02***	0.37±0.12*	**0.91±0.02***	**0.41±0.13***	0.87±0.03	0.36±0.12
PSF	Baseline	0.89±0.02	0.34±0.12	**0.92±0.01**	**0.41±0.12**	**0.92±0.02**	0.37±0.12
	VBHMSRFA	0.89±0.01	0.33±0.11	**0.92±0.01**	0.38±0.12	0.91±0.02	**0.41±0.13**
PSF-IAF	Baseline	0.88±0.02	0.34±0.12	0.90±0.02	0.37±0.13	**0.91±0.02**	0.35±0.12
	VBHMSRFA	0.88±0.02	0.34±0.12	**0.91±0.02**	**0.40±0.12**	0.89±0.02	0.32±0.13
WMSF	Baseline	0.87±0.03	0.35±0.12	0.91±0.02	0.38±0.14	0.91±0.02	0.37±0.13
	VBHMSRFA	0.90±0.02	0.38±0.13	***0.93±0.01***	***0.43±0.13***$^\sim$	0.91±0.02	0.38±0.13
WLMSF	Baseline	0.88±0.03	0.36±0.13	**0.91±0.02**	0.41±0.13	**0.91±0.02**	0.37±0.13
	VBHMSRFA	0.89±0.02	0.36±0.12	**0.91±0.03**	**0.42±0.12**	0.90±0.03	0.40±0.12
WEPF	Baseline	0.74±0.03	0.21±0.10	0.78±0.03	0.22±0.11	0.81±0.02	0.23±0.11
	VBHMSRFA	0.80±0.04	**0.27±0.10**	**0.82±0.03**	**0.27±0.10**	0.79±0.04	0.23±0.10

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table 9.2 shows the dimension of VBHMSRFA meta-features of individual feature sets. It was observed that the number of meta-features of VBHMSRFA was even lower than VBRFA-2 and VBRFA-3 meta-features, while the best performances were comparable. This indicates that allowing a slightly variable loading matrix to model inter-subject variability, as in VBHMSRFA, can reduce the number of features without sacrificing performance.

Table 9.2 Average number of VBHMSRFA meta-features.

Feature set	Number of features			
	2-s EEG window	5-s EEG window	10-s EEG window	Baseline
MDF	52.3	59.4	64.5	176
PSF	41.8	53	63.1	192
PSF-IAF	42.3	46	57.8	192
WMSF	29.8	33.6	34.4	80
WLMSF	20.9	24.3	27.1	80
WEPF	33.4	35.8	37.3	80

The highest performance of the baseline features was achieved with aggregated features extracted from multiple EEG windows (see Section 7.5). Therefore, aggregated VBHMSRFA meta-features were also created. This was done by concatenating the latent variables of the three independent VBHMSRFA models corresponding to the three EEG segments of feature extraction, i.e., 2, 5, and 10 s. The advantage of three parallel models lies in the lower computational costs due to matrix inversions, i.e., Equations (9.14) and (9.21). The performances of microsleep state detection using the aggregated meta-features and LDA classifiers are shown in Table 9.3. AUC-PR values improved across all of the feature sets. Most of the highest detection performances were achieved using VBHMSRFA meta-features of PSF with an AUC-PR of 0.48 (cf. 0.44 for baseline with PSF-IAF and WLMSF), AUC-ROC of 0.94 (cf. 0.94 for baseline with PSF, PSF-IAF, and WLMSF), GM of 0.81 (cf. 0.76 for baseline with PSF-IAF and WLMSF), and phi of 0.45 (cf. 0.37 for baseline with PSF-IAF). Interestingly, using VBHMSRFA meta-features led to higher values of phi and, in some cases GM, compared to the baseline (with the exception of WEPF), although the values of AUC-ROC were similar or slightly less. This indicates that the threshold of classification with baseline features has inter-subject variability, where the identified threshold from the training data might not lead to the highest GM and phi, whereas VBHMSRFA has a less subject-variable classification-threshold. This is a result of allowing subjects to have individual characteristics as well as sharing information with other subjects.

Table 9.3 Performance (mean ± SE) of microsleep state detection using aggregated VBHMSRFA meta-features of multiple EEG windows with an LDA classifier. A bold value indicates the highest performance in an individual feature set, while italics indicate the highest among all feature sets. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements of VBHMSRFA meta-features compared to the baseline.

Feature set	Feature type	Microsleep state prediction performance with $\tau = 0$					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	0.90±0.02	0.40±0.13	**0.71±0.08**	0.34±0.10	**0.62±0.10**	0.33±0.12
	VBHMSRFA	**0.92±0.01**$^{\sim}$	**0.41±0.12**	0.68±0.08	**0.37±0.09**	0.55±0.10	**0.39±0.12**
PSF	Baseline	***0.94±0.01***	0.43±0.12	0.74±0.06	0.36±0.08	0.68±0.11	0.36±0.12
	VBHMSRFA	***0.94±0.01***	***0.48±0.12***	***0.81±0.04***	***0.45±0.09****	**0.71±0.06**	**0.37±0.12**
PSF-IAF	Baseline	***0.94±0.01***	0.44±0.12	**0.76±0.05**	0.37±0.08	**0.70±0.10**	0.36±0.12
	VBHMSRFA	0.92±0.02	**0.45±0.12**	**0.76±0.04**	**0.42±0.09**	0.62±0.06	**0.39±0.13**
WMSF	Baseline	0.92±0.02	0.40±0.14	0.66±0.10	0.33±0.09	**0.57±0.12**	0.31±0.11
	VBHMSRFA	**0.93±0.02**	**0.43±0.14**	**0.73±0.05**	**0.41±0.10***	0.56±0.06	***0.40±0.13***
WLMSF	Baseline	***0.94±0.01***	0.44±0.13	0.76±0.07	0.36±0.10	0.70±0.11	0.31±0.12
	VBHMSRFA	0.92±0.02	**0.45±0.12**	**0.80±0.06**	**0.42±0.10**	***0.72±0.08***	**0.34±0.11**
WEPF	Baseline	**0.84±0.02**	0.27±0.12	**0.70±0.03**	**0.25±0.07**	**0.67±0.08**	**0.25±0.13**
	VBHMSRFA	**0.84±0.03**	**0.29±0.11**	0.66±0.05	0.23±0.07	0.59±0.10	0.24±0.12

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

The performances of different classifiers on the aggregated VBHMSRFA meta-features for detection of microsleep states are shown in Table 9.4. The highest overall performances were achieved with an LDA classifier and aggregated VBHMSRFA meta-features of PSF

(AUC-ROC = 0.94; AUC-PR = 0.48), whereas for the baselines, the highest AUC-ROC was 0.95 with PSF-IAF and the highest AUC-PR was 0.49 with PSF. The performance of the three classifiers, LDA, linear SVM, and VBLR, were close but the TAN performed the worst of the classifiers. Notwithstanding, the performances of TAN with VBHMSRFA meta-features were substantially higher than those of the same classifier with the baseline features.

Table 9.4 Performance of different classifiers for microsleep state detection with aggregated VBHMSRFA meta-features. A bold value indicates the highest performance among selected classifiers and an italic value is the highest overall.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.90	0.40	**0.93**	**0.48**	0.90	0.42	0.77	0.22
	VBHMSRFA	0.92$^{\sim}$	0.41	**0.93**	0.43	**0.93**	0.40	0.83$^{\sim}$	0.27$^{\sim}$
PSF	Baseline	**0.94**	0.43	**0.94**	***0.49***	**0.94**	0.44	0.70	0.21
	VBHMSRFA	**0.94**	0.48	**0.94**	0.46	**0.94**	0.47	0.87*	0.30$^{\sim}$
PSF-IAF	Baseline	0.94	0.44	***0.95***	**0.47**	0.93	0.45	0.73	0.26
	VBHMSRFA	0.92	0.45	0.92	0.43	0.92	0.44	0.86$^{\sim}$	0.29
WMSF	Baseline	0.92	0.40	**0.93**	0.42	0.92	0.39	0.77	0.22
	VBHMSRFA	**0.93**	0.43	**0.93**	**0.45**	**0.93**	0.44$^{\sim}$	0.82	0.24
WLMSF	Baseline	**0.94**	0.44	0.93	0.45	0.92	0.45	0.76	0.25
	VBHMSRFA	0.92	0.45	0.92	**0.46**	0.92	0.45	0.86$^{\sim}$	0.28
WEPF	Baseline	0.84	0.27	**0.86**	0.29	0.85	0.28	0.82	0.23
	VBHMSRFA	0.84	0.29	**0.86**	**0.30**	0.85	**0.30**	0.84	0.29

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Figure 9.2 depicts the AUC-PR of microsleep state prediction using different feature sets and prediction times of 0 to 1 s with steps of 0.25 s. Slight improvements were observed with VBHMSRFA meta-features of WMSF and WEPF compared to of the baseline. Meta-features of PSF-IAF and WLMSF had similar AUC-PR values relative to the baseline. With $\tau = 0$ s, the highest AUC-PR of VBHMSRFA meta-features was 0.48, achieved with an LDA classifier and meta-features of PSF (cf. 0.49 for baseline with linear SVM and PSF). Using the same settings with a prediction of time of $\tau = 1.0$ s, the AUC-PR dropped to 0.42 (cf. 0.42 for baseline with linear SVM and MDF). The lowest AUC-PR values were consistently achieved with WEPF.

Table 9.5 shows the performance of microsleep state prediction in terms of phi. Two-tailed Wilcoxon singed-rank tests were performed to find significant improvements with VBHMSRFA meta-features relative to of the baselines. Interestingly, although the AUC-PR values of most feature sets were not superior to the baselines, the highest phi values of most feature sets were achieved with VBHMSRFA meta-features. Applying VBHMSRFA to PSF improved phi by an average 17.7% (12.5–24.2%). Similarly, the average improvement of phi with VBHMSRFA meta-features of PSF-IAF and WLMSF were 10.6% (8.5–13.5%) and 15.9% (14.3–17.1%), respectively. Notwithstanding, a slight decline in phi was observed with VBHMSRFA meta-features of WEPF and MDF. Furthermore, the values of phi achieved with an LDA classifier were superior among classifiers for PSF, PSF-IAF, and WLMSF. Linear SVM and VBLR classifiers had slightly higher phi values than other classifiers with VBHMSRFA meta-features of WMSF. The TAN classifier, however, had substantially lower values of phi among classifiers

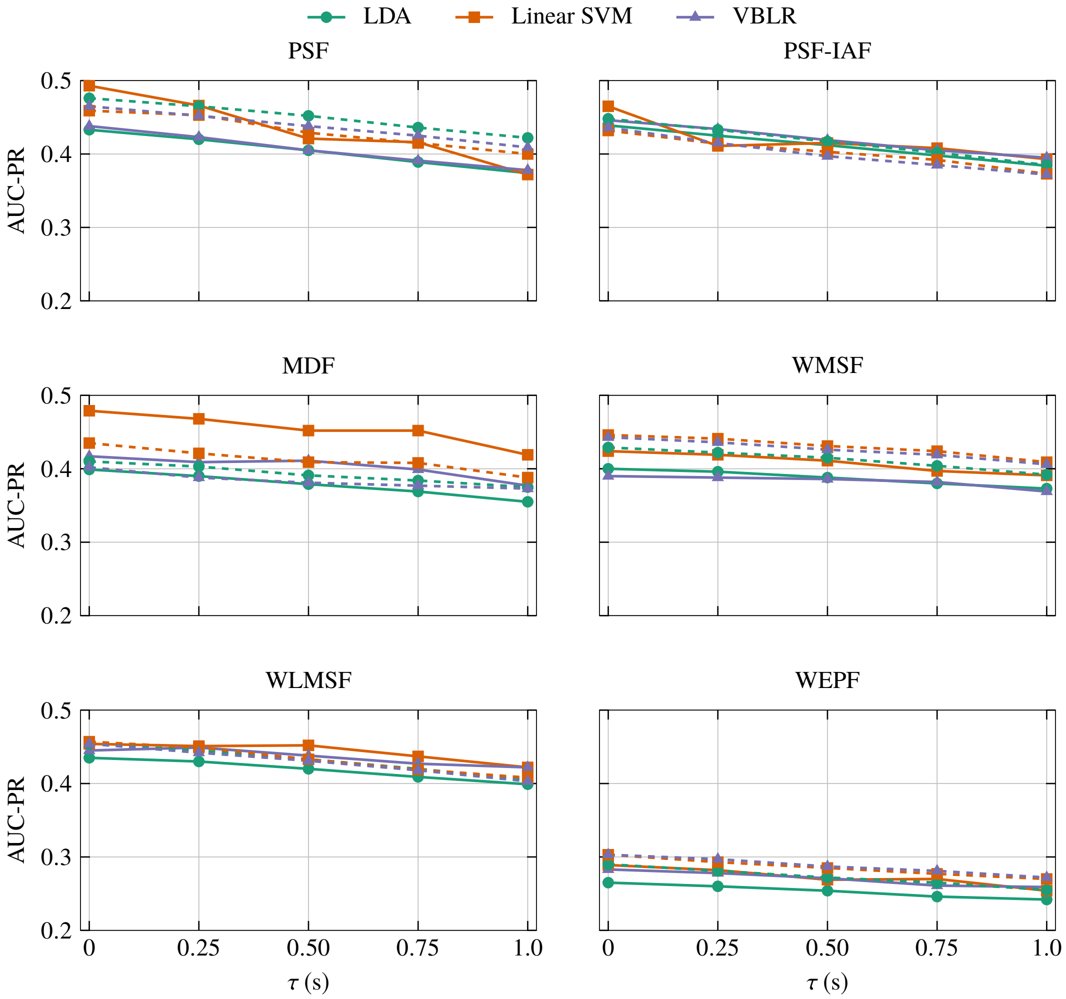


Figure 9.2 Performance of microsleep state prediction in terms of AUC-PR using different classifiers versus prediction time for τ = 0–1 s. Solid lines correspond to the performance of a classifier with baseline features, whereas dashed lines correspond to their respective performance with VBHMSRFA meta-features.

across all feature sets.

The LDA classifier had higher performances in terms of phi, whereas the linear SVM classifier had better AUC-ROC and AUC-PR. The performances of LDA and linear SVM classifiers in terms of sensitivity, precision, and GM for various feature sets across all prediction times are shown in Figures 9.3 and 9.4, respectively. Improvements of GM were achieved by using VBHMSRFA meta-features of PSF, WMSF, and WLMSF. The average improvement of GM across all prediction times using VBHMSRFA meta-features of PSF was 9.9% (9.5–11.4%) with LDA and 8.9% (5.0–12.3%) with linear SVM classifiers. Similarly, using VBHMSRFA meta-features of WLMSF increased GM by an average of 7.0% (5.3–8.2%) with an LDA and 4.8% (2.6–6.8%) with a linear SVM.

The linear SVM classifier had higher GM values with WMSF and PSF compared to LDA, whereas LDA had slightly higher values of GM with WLMSF. For detection, the highest GM of 0.83 was achieved with a linear SVM and VBHMSRFA meta-features of PSF (φ = 0.42;

Table 9.5 Performance of different classifiers in terms of phi for microsleep state prediction with aggregated VBHMSRFA meta-features for τ = 0–1 s. A bold value indicates the highest performance among selected classifiers. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements by using VBHMSRFA meta-features relative to their respective baseline.

Feature set	τ (s)	LDA		Linear SVM		VBLR		TAN	
		Baseline	VBHMSRFA	Baseline	VBHMSRFA	Baseline	VBHMSRFA	Baseline	VBHMSRFA
MDF	0.00	0.34	0.37	**0.38**	0.37	0.36	0.36	0.28	0.27
	0.25	0.33	0.36	**0.37**	0.36	0.35	0.36	0.28	0.27
	0.50	0.33	0.35	**0.37**	0.36	0.36	0.35	0.27	0.27
	0.75	0.32	0.34	**0.37**	0.35	0.36	0.35	0.27	0.26
	1.00	0.31	0.33	**0.35**	0.34	0.34	0.34	0.26	0.25
PSF	0.00	0.36	**0.45***	0.40	0.42	0.37	0.43	0.21	0.30$^{\sim}$
	0.25	0.35	**0.44***	0.37	0.43	0.36	0.42*	0.20	0.30$^{\sim}$
	0.50	0.34	**0.43***	0.36	0.41	0.35	0.42$^{\sim}$	0.20	0.29$^{\sim}$
	0.75	0.34	**0.42***	0.37	0.40	0.36	0.41$^{\sim}$	0.20	0.29$^{\sim}$
	1.00	0.33	**0.41***	0.33	0.39$^{\sim}$	0.34	0.40$^{\sim}$	0.19	0.28*
PSF-IAF	0.00	0.37	**0.42**	0.36	0.40	0.36	0.40	0.26	0.28
	0.25	0.37	**0.41**	0.34	0.39	0.36	0.39	0.25	0.27
	0.50	0.36	**0.40**	0.36	0.38	0.36	0.38	0.25	0.27
	0.75	0.35	**0.38**	0.35	**0.38**	0.34	0.37	0.24	0.27
	1.00	0.34	**0.37**	0.33	**0.37**	0.34	0.36	0.23	0.26
WMSF	0.00	0.33	**0.41***	0.34	**0.41***	0.34	**0.41**$^{\sim}$	0.27	0.27
	0.25	0.33	0.40$^{\sim}$	0.35	**0.41**$^{\sim}$	0.34	**0.41****	0.27	0.27
	0.50	0.33	0.39	0.34	**0.40**	0.33	**0.40****	0.27	0.27
	0.75	0.32	0.39$^{\sim}$	0.34	**0.40**$^{\sim}$	0.34	**0.40***	0.27	0.26
	1.00	0.32	0.37	0.33	**0.39***	0.33	**0.39***	0.26	0.26
WLMSF	0.00	0.36	**0.42**	0.35	**0.42**	0.36	**0.42**	0.24	0.27
	0.25	0.36	**0.42**	0.34	0.41	0.35	0.41	0.24	0.28
	0.50	0.35	**0.41**	0.35	0.40	0.34	0.40	0.24	0.27
	0.75	0.35	**0.40**	0.35	0.39	0.34	0.39	0.24	0.27
	1.00	0.34	**0.39**	0.33	**0.39**	0.34	0.38	0.23	0.26
WEPF	0.00	0.25	0.23	**0.26**	0.24	**0.26**	0.24	0.21	0.23
	0.25	0.24	0.23	**0.26**	0.24	**0.26**	0.24	0.21	0.22
	0.50	0.24	0.22	**0.25**	0.23	**0.25**	0.23	0.21	0.22
	0.75	0.23	0.21	**0.25**	0.22	**0.25**	0.22	0.20	0.21
	1.00	0.23	0.21	0.23	0.22	**0.24**	0.22	0.19	0.21

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Pr = 0.32; Sn = 0.75). Moreover, the lowest GM for microsleep state detection was 0.66 with an LDA classifier and meta-features of WEPF (φ = 0.23; Pr = 0.24; Sn = 0.59). Increasing the prediction time to τ = 1.0 s, the highest GM was 0.80 with a linear SVM and meta-features of WMSF (φ = 0.39; Pr = 0.33; Sn = 0.70).

Conversely, a decline in GM was observed when VBHMSRFA was applied to MDF and WEPF for microsleep state prediction. The average drop of GM with meta-features of MDF was 5.5% (4.6–6.7%) with an LDA and 4.3% (1.2–6.7%) with a linear SVM.

The sensitivity of linear SVM was slightly higher than that of LDA. The highest detection sensitivity was 0.75 (Pr = 0.36) and was achieved with meta-features of WMSF and linear SVM (cf. Sn = 0.74 and Pr = 0.38 for baseline with PSF). Increasing the prediction time to τ = 1.0 s, the highest sensitivity of 0.71 (Pr = 0.34) was achieved with VBHMSRFA meta-features of WLMSF and a linear SVM (cf. Sn = 0.71 and Pr = 0.29 for baseline with PSF-IAF). This shows that although the highest sensitivity of the baseline and VBHMSRFA meta-features were similar, the latter had higher precision and thus a lower number of false positives.

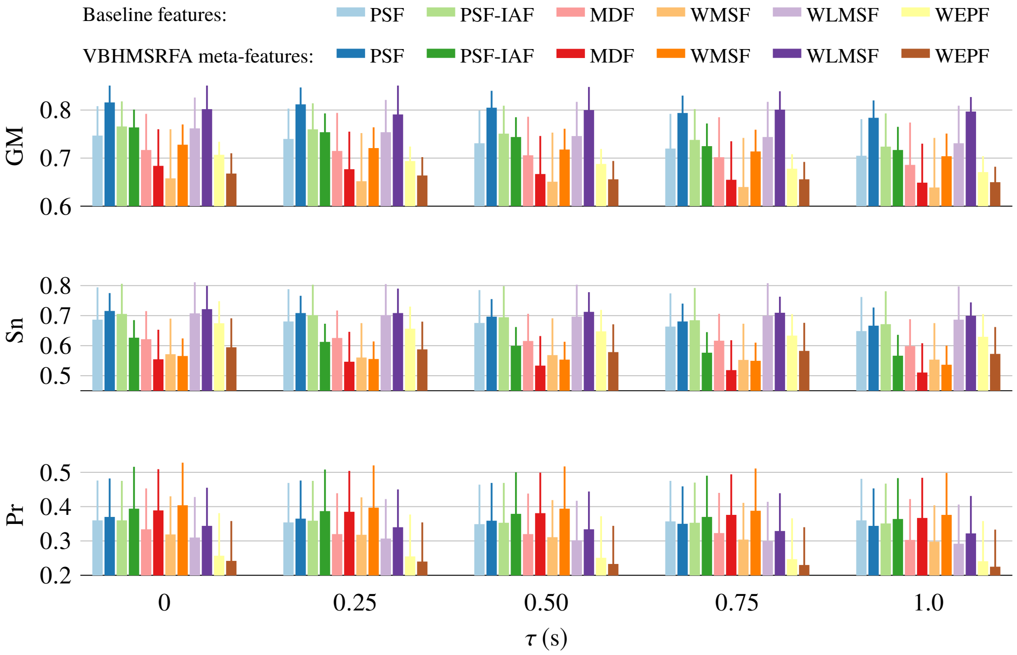


Figure 9.3 Performance (mean ± SE) of microsleep state prediction using VBHMSRFA meta-features with an LDA classifier for τ = 0–1 s.

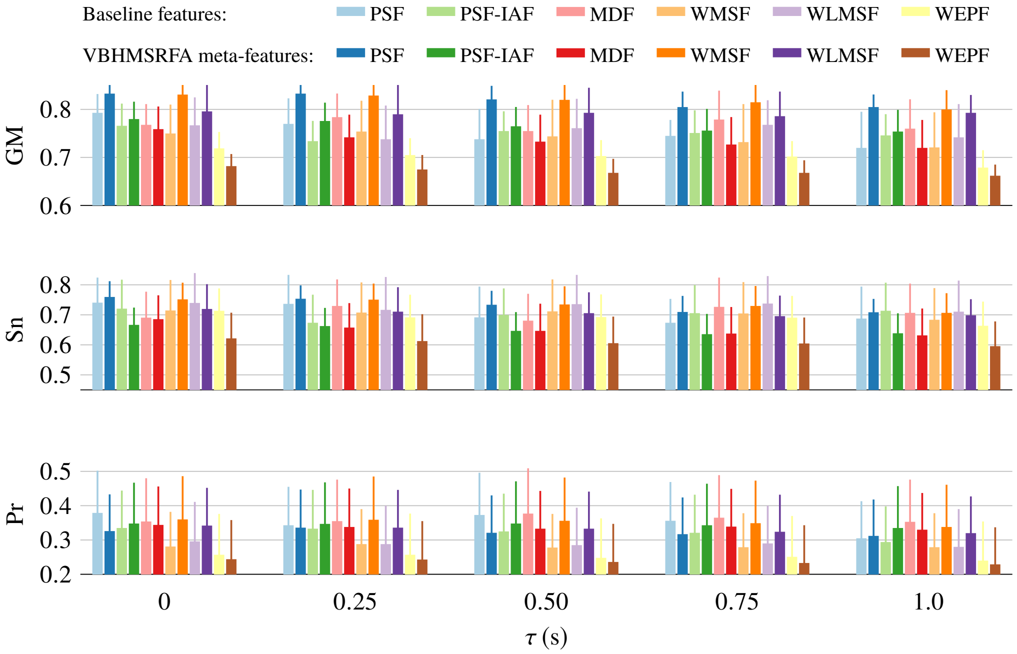


Figure 9.4 Performance measures (mean ± SE) of microsleep state prediction using baseline features and VBHMSRFA meta-features with a linear SVM classifier for τ = 0–1 s.

9.4.2 Detection and prediction of microsleep onsets

The results of microsleep onset detection and prediction using aggregated VBHMSRFA meta-features are presented in this section. These results are limited to aggregated meta-features, due to higher performance compared to single-window features for microsleep state detection.

Table 9.6 presents the performances of microsleep onset detection ($\tau = 0$ s) using VBHMSRFA meta-features and an LDA classifier. Detection performance of the VBHMSRFA meta-features of all feature sets were relatively similar, with an exception of WEPF, with the highest AUC-ROC of 0.90 (cf. 0.88 for baseline with WMSF and WLMSF). The highest AUC-PR was 0.06 with VBHMSRFA meta-features of PSF and PSF-IAF (cf. 0.05 for baseline with WLMSF). Although the highest values of AUC-ROC and AUC-PR were relatively similar to the baselines, improvements of GM were observed in four feature sets. The highest GM of 0.78 was achieved with meta-features of WLMSF (cf. 0.73 for baseline with PSF-IAF and WLMSF).

Table 9.6 Performance (mean ± SE) of microsleep onset detection using VBHMSRFA aggregated meta-features and an LDA classifier. A bold value indicates the highest performance of each feature set, whereas an italic value represents the highest overall.

Feature set	Feature type	Microsleep onset prediction performance with $\tau = 0$					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	0.87±0.02	**0.03±0.01**	**0.69±0.05**	***0.09±0.02***	**0.55±0.08**	0.02±0.01
	VBHMSRFA	***0.90±0.01***	**0.03±0.01**	0.68±0.06	***0.09±0.02***	0.54±0.08	***0.03±0.01***
PSF	Baseline	0.87±0.01	0.04±0.03	0.71±0.04	0.07±0.01	0.63±0.09	**0.02±0.01**
	VBHMSRFA	**0.89±0.02**	***0.06±0.03***	**0.77±0.03**	***0.09±0.01***	***0.70±0.07***	**0.02±0.01**
PSF-IAF	Baseline	0.87±0.01	0.04±0.03	0.73±0.04	0.08±0.01	0.66±0.09	**0.02±0.01**
	VBHMSRFA	**0.89±0.02**	***0.06±0.02***	**0.77±0.04**	***0.09±0.01***	**0.68±0.07**	**0.02±0.01**
WMSF	Baseline	0.88±0.02	0.03±0.02	0.69±0.07	0.08±0.02	0.60±0.11	**0.02±0.01**
	VBHMSRFA	**0.89±0.02**	**0.04±0.02**	**0.72±0.08**	***0.09±0.02***	**0.62±0.10**	**0.02±0.01**
WLMSF	Baseline	0.88±0.02	**0.05±0.03**	0.73±0.05	0.07±0.01	***0.70±0.10***	0.01±0.00
	VBHMSRFA	***0.90±0.01***	**0.05±0.02**	***0.78±0.03***	***0.09±0.02***	0.69±0.05	**0.02±0.01**
WEPF	Baseline	**0.73±0.04**	**0.01±0.01**	**0.63±0.05**	**0.04±0.01**	**0.60±0.11**	**0.01±0.00**
	VBHMSRFA	**0.73±0.04**	**0.01±0.00**	0.59±0.06	0.03±0.01	0.49±0.11	**0.01±0.00**

Table 9.7 shows the AUC-ROC and AUC-PR of microsleep onset detection using different classifiers. An LDA classifier with meta-features of PSF and PSF-IAF had the highest AUC-PR of 0.06 (cf. 0.09 for baseline with a linear SVM and PSF). The highest AUC-ROC of 0.91 was achieved with a linear SVM and meta-features of WLMSF (cf. 0.91 for baseline with MDF, PSF, and PSF-IAF). It is evident that although slight improvements of microsleep onset detection using LDA and VBLR were achieved with VBHMSRFA meta-features, the highest performance in terms of AUC-ROC and AUC-PR was not superior compared to that of the baseline.

Figure 9.5 shows the AUC-ROC of microsleep onset prediction versus prediction times up to 10 s. In most cases, AUC-ROC of VBHMSRFA meta-features were slightly lower than the best baseline. Comparing the performance of various VBHMSRFA meta-feature sets, however, revealed that meta-features of PSF had the highest AUC-ROC of 0.89 with $\tau = 0$ s

Table 9.7 Performance of different classifiers for microsleep onset detection with VBHMSRFA aggregated meta-features. A bold value indicates the highest performance among classifiers, whereas an italic value indicates the highest performance overall.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.87	0.03	***0.91***	**0.06**	0.82	0.05	0.72	0.01
	VBHMSRFA	0.90	0.03	0.82	0.03	0.86	**0.03**	0.72	0.01
PSF	Baseline	0.87	0.04	***0.91***	***0.09***	0.83	0.04	0.70	0.01
	VBHMSRFA	0.89	0.06	0.89	0.06	0.84	0.03	0.75	0.01
PSF-IAF	Baseline	0.87	0.04	***0.91***	**0.07**	0.81	0.04	0.70	0.01
	VBHMSRFA	0.89	0.06	0.85	0.05	0.85	0.04	0.72	0.01
WMSF	Baseline	0.88	0.03	**0.90**	**0.05**	0.80	0.02	0.74	0.01
	VBHMSRFA	**0.89**	**0.04**	0.88	0.03	0.83	0.02	0.69	0.01
WLMSF	Baseline	0.88	0.05	0.90	**0.06**	0.83	0.04	0.79	0.02
	VBHMSRFA	0.90	0.05	***0.91***	0.05	0.85	0.04	0.78	0.02
WEPF	Baseline	0.73	0.01	**0.78**	**0.02**	0.74	0.01	0.76	0.01
	VBHMSRFA	0.73	0.01	0.75	0.01	0.75	0.01	0.70	0.01

(cf. 0.91 for baseline with PSF), whereas meta-features of MDF had the highest AUC-ROC of 0.76 with $\tau = 10\,\mathrm{s}$ (cf. 0.74 for baseline with WLMSF). On the other hand, the worst AUC-ROC of VBHMSRFA meta-features at $\tau = 0\,\mathrm{s}$ was achieved with WEPF and a TAN, i.e., AUC-ROC = 0.74 (cf. 0.70 for baseline with PSF and similar classifier), which dropped to 0.58 with a prediction time of $\tau = 10\,\mathrm{s}$ (cf. 0.52 for baseline with WEPF and a VBLR). Nevertheless, the AUC-ROC performance of VBHMSRFA meta-features of WLMSF and MDF had a lower dropping trend than the other meta-feature sets. Moreover, the VBHMSRFA meta-features of WEPF generally had the lowest performance of the meta-feature sets.

Figure 9.6 depicts the phi performances of microsleep onset prediction. Similar to the VBMSRFA results, the phi measures of VBHMSRFA meta-features were mostly superior to the baseline. Moreover, linear SVM with VBHMSRFA meta-features, in most cases, had superior or similar performances in terms of phi among classifiers. The lowest phi values noted were with WEPF, which was consistent with the baseline performances. The highest phi with $\tau = 0\,\mathrm{s}$ was 0.10 which was achieved with a VBLR and VBHMSRFA meta-features of WLMSF (cf. 0.09 for baseline with an LDA and MDF). As expected, a dropping trend of phi with prediction time τ was observed. The highest phi with a prediction time of $\tau = 10\,\mathrm{s}$ was 0.04 (AUC-ROC = 0.76) which was achieved with a linear SVM and the meta-features of MDF (cf. $\varphi = 0.04$; AUC-ROC = 0.74 for baseline with WMSF).

Figure 9.7 depicts the rest of the performance measures – GM, sensitivity, and precision – of microsleep onset prediction using a linear SVM classifier. The remainder of the results of microsleep onset prediction with VBHMSRFA are limited to a linear SVM due to its better performance relative to the rest of the classifiers. Applying VBHMSRFA to PSF, PSF-IAF, MDF, and WLMSF improved GM performances across all prediction times. However, the improvements achieved with VBHMSRFA meta-features of WMSF and WEPF were not consistent. The highest GM for detection was 0.79 (Sn = 0.73; Sp = 0.87) which was achieved

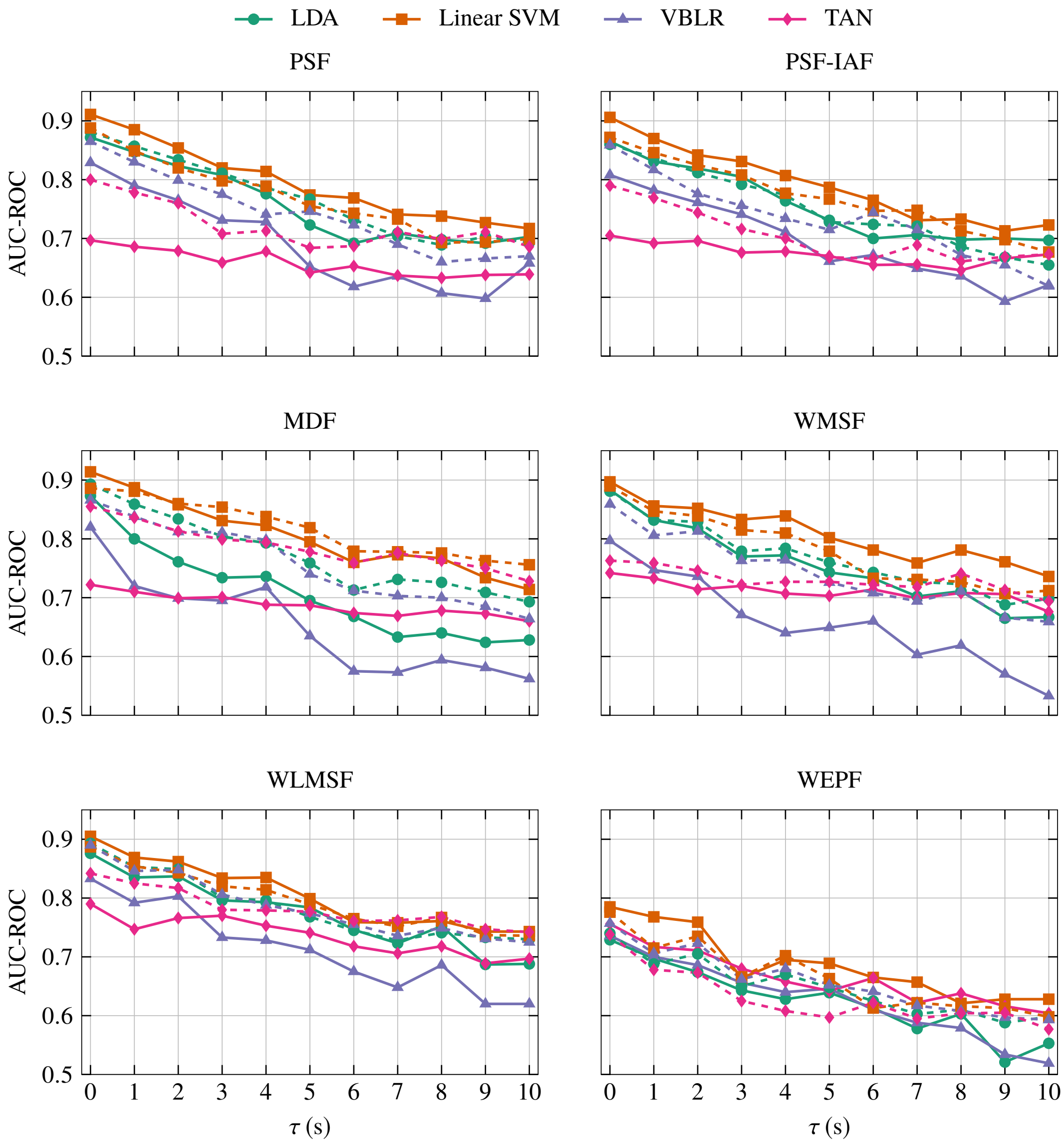


Figure 9.5 AUC-ROC of microsleep onset prediction using different classifiers versus prediction time 0–10 s. Solid lines correspond to the baseline features, whereas dashed lines correspond to the VBHMSRFA meta-features.

with VBHMSRFA meta-features of WMSF (cf. GM = 0.78; Sn = 0.78; Sp = 0.80 for baseline with MDF). Increasing the prediction time to $\tau = 10$ s, the highest GM was 0.66 (Sn = 0.66; Sp = 0.71) which was achieved with VBHMSRFA meta-features of MDF (cf. GM = 0.55; Sn = 0.57; Sp = 0.71 for baseline with MDF). On the other hand, the lowest GM with $\tau = 0$ s was 0.76 (Sn = 0.49; Sp = 0.85), achieved with VBHMSRFA meta-features of WMSF and a TAN classifier (cf. GM = 0.48; Sn = 0.54; Sp = 0.71 for baseline with PSF-IAF).

In terms of sensitivity, no consistent improvement was found. Notwithstanding, using VBHMSRFA meta-features of MDF for microsleep onset prediction achieved higher sensitivities than those of the baselines in most cases. Moreover, using VBHMSRFA meta-features of PSF-IAF, WMSF, and WLMSF led to higher precisions compared to their respective baselines, although all precisions were too low. This indicates that the ratio of false positives to true

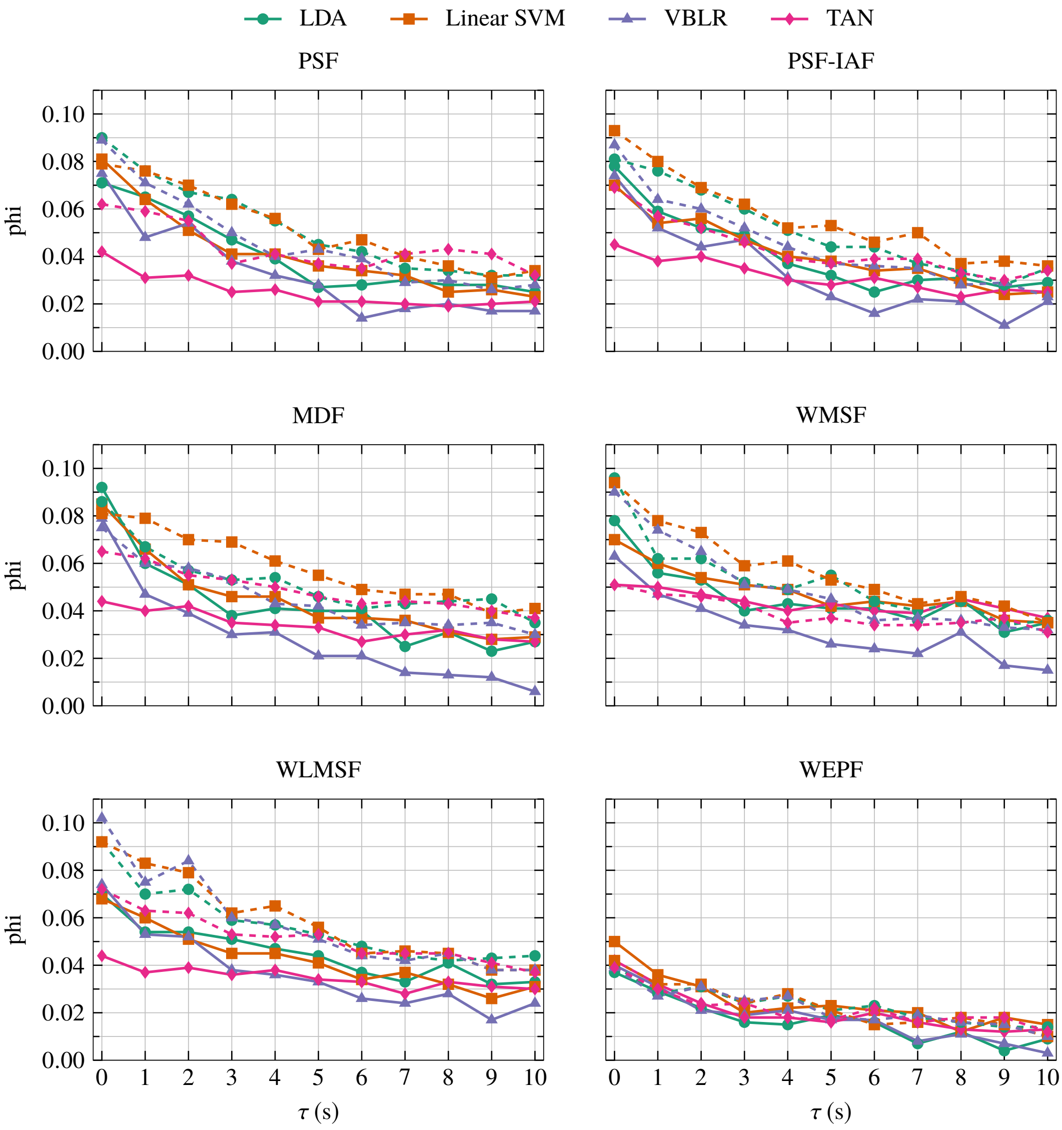


Figure 9.6 Performance of microsleep onset prediction in terms of phi using different classifiers versus prediction time 0–10 s. Solid lines correspond to the baseline features, whereas dashed lines correspond to the VBHMSRFA meta-features.

positives was too high.

9.5 SUMMARY

In this chapter, VBHMSRFA was presented and its results for microsleep detection and prediction were reported. VBHMSRFA is a multi-subject feature reduction method that finds a robust and less inter-subject variable latent space. This is done by allowing subjects to share information via a group-level loading matrix while each subject can vary slightly from the group. Moreover, each subject has its own mean and noise parameters, similar to VBMSRFA described in Chapter 8. VBHMSRFA finds the optimum number of components by employing ARD motivated prior

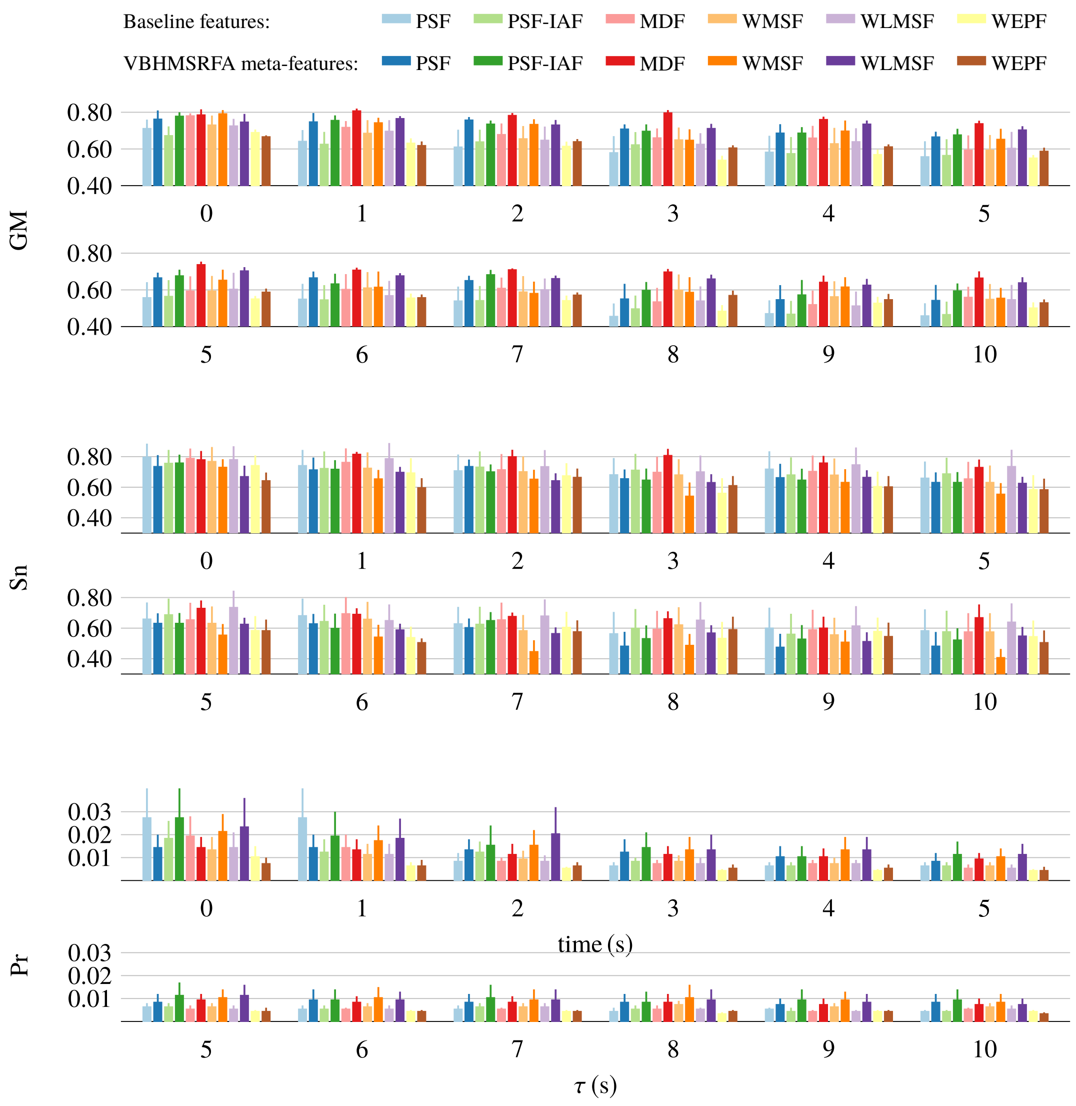


Figure 9.7 Performance (mean ± SE) of microsleep onset prediction using VBHMSRFA meta-features and a single linear SVM classifier for 0–10 s.

distributions over both subject-level and group-level loading matrices. The variational inference for the training and testing phases were derived and presented.

Using VBHMSRFA for detection and prediction of microsleep states led to significant improvements of phi with PSF and WMSF relative to their respective baselines. Substantial improvements of phi were also achieved with PSF-IAF and WLMSF. The highest phi for detection was 0.45 with meta-features of PSF and an LDA classifier (cf. 0.40 for baseline with PSF). The same setup led to the highest phi of 0.41 with a prediction time of τ = 1.0 s (cf. 0.35 for baseline with MDF). Notwithstanding, the performances of microsleep state detection in terms of AUC-ROC with the baseline and VBHMSRFA were similar.

For detection and prediction of microsleep onsets, applying VBHMSRFA slightly lowered the AUC-ROC and AUC-PR compared to the baselines. However, there were improvements in terms of phi. For detection, the highest phi of 0.10 was achieved with meta-features of WLMSF and a VBLR classifier (cf. 0.09 for baseline with an LDA and MDF). Increasing the prediction time to $\tau = 10\,\mathrm{s}$, the highest phi was 0.04 which was achieved with a linear SVM and VBHMSRFA meta-features of MDF (cf. 0.04 for baseline with WMSF). Although the values of phi were low, AUC-ROC and GM were moderate. The highest detection AUC-ROC was 0.91 with meta-features of WLMSF and a linear SVM (cf. 0.91 for baseline with PSF). Moreover, the highest GM was 0.79 with meta-features of WMSF (cf. 0.78 for baseline with MDF).

Chapter 10

VARIATIONAL BAYESIAN HIERARCHICAL MULTI-SUBJECT ROBUST JOINT MATRIX FACTORIZATION

10.1 INTRODUCTION

Three Bayesian feature-reduction methods were presented in Chapters 7–9. However, all of these methods are unsupervised and discard the class information of the gold-standard. Therefore, exploiting the class labels might lead to better performances of microsleep detection and prediction.

In this chapter, VBHMSRFA is extended to use the gold-standard labels in addition to the features, i.e., variational Bayesian hierarchical multi-subject robust joint matrix factorization (VBHMSRJMF). It was expected that by exploiting information of the gold-standard and jointly factorizing all classes, the performance of microsleep detection/prediction would improve. Section 10.2 introduces the proposed Bayesian model and its underlying assumptions and probability distributions. Since inferring the proposed model is intractable, variational inference is used to approximate posterior probabilities. The variational formulation of our proposed method is given in Section 10.3. Finally, the results of detection and prediction of microsleeps using VBHMSRJMF meta-features are presented in Section 10.4.

10.2 BAYESIAN HIERARCHICAL MULTI-SUBJECT ROBUST JOINT MATRIX FACTORIZATION

In the previous chapters, various models of Bayesian FA were discussed. However, all of those methods are unsupervised and do not take the gold-standard into account. In a highly-imbalanced dataset, such as microsleeps, an unsupervised technique might result in a model that only explains the majority class while critical information on the minority class is ignored and discarded. To overcome this issue, we have developed Bayesian hierarchical multi-subject robust joint matrix factorization (BHMSRJMF), an extension of BHMSRFA, that jointly factorizes the data of all the classes, automatically selects the required number of latent variables, and takes subject variability into account.

BHMSRJMF assumes that the data is generated by

$$\mathbf{x}_{s,c,n} = \mathbf{W}_s \mathbf{z}_{s,c,n} + \boldsymbol{\mu}_s + \boldsymbol{\varepsilon}_{s,c,n}, \tag{10.1}$$

where $c \in \left\{1, \ldots, C\right\}$ is the class index, $s \in \left\{1, \ldots, S\right\}$ is the subject index, and $n \in \left\{1, \ldots, N_{s,c}\right\}$ is the observation index. The noise $\boldsymbol{\varepsilon}_{s,c,n}$ is assumed to be drawn from a zero-mean normal distribution,

$$\boldsymbol{\varepsilon}_{s,c,n} \sim \mathcal{N}\left(\boldsymbol{\varepsilon}_{s,c,n} \,\middle|\, \mathbf{0}, \boldsymbol{\Psi}_s^{-1}\right), \tag{10.2}$$

which leads to a linear Gaussian model as

$$\mathbf{x}_{s,c,n} \sim \mathcal{N}\left(\mathbf{x}_{s,c,n} \,\middle|\, \mathbf{W}_s \mathbf{z}_{s,c,n} + \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s^{-1}\right). \tag{10.3}$$

It is apparent that for any subject, the data of each class is being generated by its corresponding latent variables, while the loading matrix is shared between all the classes. In addition, a subject's independent $\boldsymbol{\mu}_s$ and $\boldsymbol{\Psi}_s$ lets the system to adapt to new subjects at the time of testing.

The latent variables are assumed to have zero-mean independent Student-t distributions to make them less sensitive to noise. This is achieved by representing a Student-t distribution as a zero-mean normal distribution with a prior probability of Gamma distribution over its precision, leading to

$$\mathbf{z}_{s,c,n} \sim \mathcal{N}\left(\mathbf{z}_{s,c,n} \,\middle|\, \mathbf{0}, \boldsymbol{\Lambda}_c^{-1}\right), \tag{10.4}$$

$$\boldsymbol{\Lambda}_c \sim \prod_{k=1}^{K} \mathcal{G}\left(\lambda_{c,k} \,\middle|\, a_\lambda, b_\lambda\right), \tag{10.5}$$

where K is the dimension of the latent space.

To reduce the subject variability of the latent spaces, individual subjects are allowed to have different loading matrices. Sharing information between subjects is done via a group level loading matrix $\mathbf{M}_w$ to capture the similar patterns among all subjects. The individual subject loading matrices are then calculated by

$$\mathbf{w}_{s,k} = \mathbf{m}_{w,k} + \boldsymbol{\zeta}_{s,k}, \tag{10.6}$$

$$\boldsymbol{\zeta}_{s,k} \sim \mathcal{N}\left(\boldsymbol{\zeta}_{s,k} \,\middle|\, \mathbf{0}, \left(\alpha_k \mathbf{I}\right)^{-1}\right), \tag{10.7}$$

where $\mathbf{m}_{w,k}$ is the k^{th} column of the group level loading matrix. Equivalently, BHMSRJMF assumes that each column of the subject level loading matrix is drawn from a Gaussian distribution,

$$\mathbf{W}_s \sim \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_{s,k} \,\middle|\, \mathbf{m}_{w,k}, \left(\alpha_k \mathbf{I}\right)^{-1}\right). \tag{10.8}$$

Using a conjugate prior, each column of $\mathbf{M}_w$ is assumed to be a zero-mean normal distribution,

$$\mathbf{M}_w \sim \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{m}_{w,k} \,\middle|\, \mathbf{0}, \beta_w {\alpha_k}^{-1}\mathbf{I}\right). \tag{10.9}$$

The prior distribution of $\boldsymbol{\alpha}$ is assumed to be Gamma distribution,

$$\boldsymbol{\alpha} \sim \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,\middle|\, a_\alpha, b_\alpha\right). \tag{10.10}$$

The subject and group level loading matrices are regularized with the same parameter $\boldsymbol{\alpha}$, motivated by ARD. If the posterior distribution of α_k is concentrated on large values, the k^{th} column of the group level loading matrix $\mathbf{M}_w$ will drop to zero. As a result, the k^{th} component of the Equation (10.8) simplifies to a zero-mean normal distribution with a high precision, and effectively drops to zero. Therefore, the column sparsity of the subject and group level loading matrices are consistent.

To make the BHMSRJMF fully Bayesian, the prior probability over the mean of each class $\boldsymbol{\mu}_s$ is assumed to be a normal distribution,

$$\boldsymbol{\mu}_s \sim \prod_{d=1}^{D} \mathcal{N}\left(\mu_{s,d} \,\middle|\, m_{\mu,d}, \left(\beta_\mu \psi_{s,d}\right)^{-1}\right), \tag{10.11}$$

where $\psi_{s,d}$ is the d^{th} diagonal element of $\boldsymbol{\Psi}_s$ and has a Gamma prior distribution,

$$\boldsymbol{\Psi}_s \sim \prod_{d=1}^{D} \mathcal{G}\left(\psi_{s,d} \,\middle|\, a_\psi, b_\psi\right). \tag{10.12}$$

Using a hyperparameter $\mathbf{m}$ for mean vectors $\boldsymbol{\mu}$ and updating it at the training phase improves the ability to adapt to a test subject.

The graphical model representation of BHMSRJMF is shown in Figure 10.1.

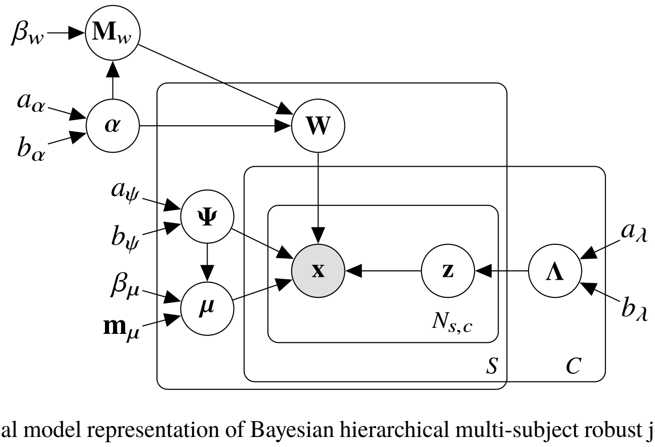


Figure 10.1 Graphical model representation of Bayesian hierarchical multi-subject robust joint matrix factorization (BHMSRJMF) model.

10.3 VARIATIONAL INFERENCE

10.3.1 Training phase

Using an exact Bayesian inference for BHMSRJMF is analytically intractable. To have a fully Bayesian inference, the variational inference was used to approximate the posterior probabilities. The first step is to assume a variational distribution. We assumed that the variational distribution has a factorized structure,

$$q\left(\Theta\right) = q\left(\mathbf{M}_w \mid \alpha\right) q\left(\alpha\right) \prod_{s=1}^{S} \Big(q\left(\mathbf{W}_s\right) q\left(\boldsymbol{\mu}_s, \boldsymbol{\Psi}_s\right) \Big) \prod_{c=1}^{C} q\left(\boldsymbol{\Lambda}_c\right) \prod_{s=1}^{S} \prod_{c=1}^{C} \prod_{n=1}^{N_{s,c}} q\left(\mathbf{z}_{s,c,n}\right), \tag{10.13}$$

where $\Theta = \left\{\mathbf{M}_w, \mathbf{W}, \alpha, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}\right\}$ is a set of all model parameters. Using Equations (3.7) and (10.13), the lower bound of the marginal log-likelihood of data is given by

$$\begin{aligned}
\mathcal{L} = & \left\langle \ln\left(\frac{p\left(\alpha\right)}{q\left(\alpha\right)}\right)\right\rangle + \left\langle \ln\left(\frac{p\left(\mathbf{M}_w \mid \alpha\right)}{q\left(\mathbf{M}_w \mid \alpha\right)}\right)\right\rangle + \sum_{c=1}^{C}\left\langle \ln\left(\frac{p\left(\boldsymbol{\Lambda}_c\right)}{q\left(\boldsymbol{\Lambda}_c\right)}\right)\right\rangle \\
& + \sum_{s=1}^{S}\left(\left\langle \ln\left(\frac{p\left(\boldsymbol{\Psi}_s\right)}{q\left(\boldsymbol{\Psi}_s\right)}\right)\right\rangle + \left\langle \ln\left(\frac{p\left(\boldsymbol{\mu}_s \mid \boldsymbol{\Psi}_s\right)}{q\left(\boldsymbol{\mu}_s \mid \boldsymbol{\Psi}_s\right)}\right)\right\rangle + \left\langle \ln\left(\frac{p\left(\mathbf{W}_s \mid \mathbf{M}_w, \alpha\right)}{q\left(\mathbf{W}_s\right)}\right)\right\rangle\right) \\
& + \sum_{s=1}^{S}\sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}}\left(\left\langle \ln\left(\frac{p\left(\mathbf{z}_{s,c,n} \mid \boldsymbol{\Lambda}_c\right)}{q\left(\mathbf{z}_{s,c,n}\right)}\right)\right\rangle + \left\langle \ln\left(p\left(\mathbf{x}_{s,c,n} \mid \mathbf{W}_s, \boldsymbol{\mu}_s, \mathbf{z}_{s,c,n}, \boldsymbol{\Psi}_s\right)\right)\right\rangle\right),
\end{aligned} \tag{10.14}$$

where the last term plays the role of data fitting objective while the rest of the terms are penalties for model complexities.

To approximate the posterior probabilities, the VBE step updates the variational parameters of the latent variables using

$$\tilde{\Sigma}_{z,s,c} = \left(\left\langle \mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\mathbf{W}_s\right\rangle + \left\langle \boldsymbol{\Lambda}_c\right\rangle\right)^{-1}, \tag{10.15}$$

$$\tilde{\mathbf{m}}_{z,s,c,n} = \tilde{\Sigma}_{z,s,c}\left\langle \mathbf{W}_s^{\top}\right\rangle\left\langle \boldsymbol{\Psi}_s\right\rangle\left(\mathbf{x}_{s,c,n} - \left\langle \boldsymbol{\mu}_s\right\rangle\right), \tag{10.16}$$

where the variational distribution of the latent variables is

$$q\left(\mathbf{Z}_{s,c}\right) = \prod_{n=1}^{N_{s,c}} \mathcal{N}\left(\mathbf{z}_{s,c,n} \mid \tilde{\mathbf{m}}_{z,s,c,n}, \tilde{\Sigma}_{z,s,c}\right). \tag{10.17}$$

After that, the VBM step updates the variational distributions of the model parameters by

$$\tilde{\beta}_w = \beta_w + S, \tag{10.18}$$

$$\tilde{\mathbf{m}}_{\mathbf{w},k} = \tilde{\beta}_w^{-1}\left(\sum_{s=1}^{S}\left\langle \mathbf{w}_{s,k}\right\rangle\right), \tag{10.19}$$

$$\tilde{a}_{\alpha} = \frac{SD}{2} + a_{\alpha}, \tag{10.20}$$

$$\tilde{b}_{\alpha,k} = b_{\alpha} + \frac{1}{2}\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k}^{\top}\mathbf{w}_{s,k} \right\rangle - \frac{1}{2}\tilde{\mathbf{m}}_{w,k}^{\top}\tilde{\mathbf{m}}_{w,k}, \tag{10.21}$$

$$\tilde{a}_{\lambda,c} = a_{\lambda} + \sum_{s=1}^{S} \frac{N_{s,c}}{2}, \tag{10.22}$$

$$\tilde{b}_{\lambda,c,k} = b_{\lambda} + \frac{1}{2}\sum_{s=1}^{S}\sum_{n=1}^{N_{s,c}} \left\langle z_{s,c,n,k}^{2} \right\rangle, \tag{10.23}$$

$$\tilde{\beta}_{\mu,s} = \beta_{\mu} + \sum_{c=1}^{C} N_{s,c}, \tag{10.24}$$

$$\tilde{\mathbf{m}}_{\mu,s} = \frac{1}{\tilde{\beta}_{\mu,s}}\left(\beta_{\mu}\mathbf{m}_{\mu} + \sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left(\mathbf{x}_{s,c,n} - \left\langle \mathbf{w}_{s} \right\rangle \left\langle \mathbf{z}_{s,c,n} \right\rangle\right)\right), \tag{10.25}$$

$$\tilde{a}_{\psi,s} = a_{\psi} + \frac{\sum_{c=1}^{C} N_{s,c}}{2}, \tag{10.26}$$

$$\begin{aligned}\tilde{b}_{\psi,s,d} = b_{\psi,d} + \frac{\beta_{\mu}}{2} m_{\mu,d}^{2} - \frac{\tilde{\beta}_{\mu,s}}{2}\tilde{m}_{\mu,s,d}^{2} + \frac{1}{2}\sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \Bigg(& x_{s,n,c,d}^{2} \\ & - 2x_{s,c,n,d}\left\langle \mathbf{w}_{s,d,.} \right\rangle \left\langle \mathbf{z}_{s,c,n} \right\rangle + \operatorname{tr}\left(\left\langle \mathbf{w}_{s,d,.}^{\top}\mathbf{w}_{s,d,.} \right\rangle \left\langle \mathbf{z}_{s,c,n}\mathbf{z}_{s,c,n}^{\top} \right\rangle\right)\Bigg),\end{aligned} \tag{10.27}$$

$$\tilde{\Sigma}_{w,s,d} = \left(\left\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \right\rangle + \left\langle \psi_{s,d} \right\rangle \sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left\langle \mathbf{z}_{s,c,n}\mathbf{z}_{s,c,n}^{\top} \right\rangle\right)^{-1}, \tag{10.28}$$

$$\tilde{\mathbf{w}}_{s,d} = \tilde{\Sigma}_{w,s,d}\left(\left\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \right\rangle \mathbf{m}_{\mathbf{w},d,.}^{\top} + \left\langle \psi_{s,d} \right\rangle \sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left(\left\langle \mathbf{z}_{s,c,n} \right\rangle \left(x_{s,c,n,d} - \left\langle \mu_{s,d} \right\rangle\right)\right)\right), \tag{10.29}$$

where the variational distributions are

$$q\left(\mathbf{M}_{w} \mid \boldsymbol{\alpha}\right) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{m}_{w,k} \,\middle|\, \tilde{\mathbf{m}}_{w,k}, \left(\tilde{\beta}_{w}\alpha_{k}\right)^{-1}\mathbf{I}\right), \tag{10.30}$$

$$q\left(\boldsymbol{\alpha}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_{k} \mid \tilde{a}_{\alpha}, \tilde{b}_{\alpha,k}\right), \tag{10.31}$$

$$q\left(\boldsymbol{\Lambda}_{c}\right) = \prod_{k=1}^{K} \mathcal{G}\left(\lambda_{c,k} \mid \tilde{a}_{\lambda,c}, \tilde{b}_{\lambda,c,k}\right), \tag{10.32}$$

$$q\left(\boldsymbol{\mu}_{s} \mid \boldsymbol{\Psi}_{s}\right) = \prod_{d=1}^{D} \mathcal{N}\left(\mu_{s,d} \,\middle|\, \tilde{m}_{\mu,s,d}, \left(\tilde{\beta}_{\mu,s}\psi_{s,d}\right)^{-1}\right), \tag{10.33}$$

$$q\left(\boldsymbol{\Psi}_{s}\right) = \prod_{d=1}^{D} \mathcal{G}\left(\psi_{s,d} \mid \tilde{a}_{\psi,s}, \tilde{b}_{\psi,s,d}\right), \tag{10.34}$$

$$q\left(\mathbf{W}_{s}\right) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{w}_{s,d,.}^{\top} \,\middle|\, \tilde{\mathbf{w}}_{s,d}, \tilde{\Sigma}_{w,s,d}\right). \tag{10.35}$$

Initialization of the training phase is done by setting the hyperparameters to correspond to uninformative prior distributions over the model parameters. To this end, the hyperparameters $\{a_\alpha, b_\alpha, a_\lambda, b_\lambda, a_\psi, b_\psi, \beta_\mu, \beta_w\}$ are initialized to a small value, i.e., 10^{-6}, and the hyperparameter $\mathbf{m}_\mu$ is set to a vector of zeros. The empirical mean of each subject is used to set the variational mean parameters $\tilde{\mathbf{m}}_{\mu,s}$. A PCA is performed on the concatenated demeaned data of all subjects and the first $K = D - 1$ columns of the PCA loading matrix is used to initialize $\tilde{\mathbf{M}}_w$ and $\tilde{\mathbf{W}}_s$ for all subjects. All other variational parameters are set to their corresponding prior hyperparameters.

Each iteration of training consists of a VBE step to update the latent variable distributions, followed by a VBM to update the variational distribution of the model parameters. The lower bound of the evidence log-likelihood is calculated at each iteration to monitor the convergence of the training phase. The training phase is stopped when the relative improvement of the lower bound of the evidence log-likelihood drops below a threshold, i.e., 10^{-6}, in adjacent iterations.

Updating the variational distributions over iterations would cause the redundant latent variables to have very small values and effectively be turned off. But this does not remove the extra components which are important for improving the computational speed. Therefore, an extra step is necessary to remove the extra components without reducing the likelihood of data. This is achieved by first calculating the lower bound of the evidence log-likelihood with all the components. It is then followed by temporarily removing the component corresponding to the largest value of $\langle\alpha\rangle$ and calculating the lower bound again. Examining the lower bounds of data with and without the selected component indicates if the component can be removed or not, i.e., the component is retained if the lower bound did not improve. In our implementation, the process of pruning extra components is performed at every iteration.

Initialization of the training phase includes setting the hyperparameters to correspond to uninformative prior distributions. Notwithstanding, those values can be updated to increase the lower bound of the evidence log-likelihood. Taking the derivative of the lower bound with respect to the β_μ, β_w, and $m_{\mu,d}$ leads to the analytical updating formulas,

$$m_{\mu,d} = \frac{\sum_{s=1}^{S} \langle\psi_{s,d}\rangle\langle\mu_{s,d}\rangle}{\sum_{s=1}^{S} \langle\psi_{s,d}\rangle}, \tag{10.36}$$

$$\beta_\mu^{-1} = \frac{\sum_{s=1}^{S} \left(D\tilde{\beta}_{\mu,s}^{-1} + \sum_{d=1}^{D} \langle\psi_{s,d}\rangle \left(\langle\mu_{s,d}\rangle - m_{\mu,d}\right)^2 \right)}{SD}, \tag{10.37}$$

$$\beta_w^{-1} = \tilde{\beta}_w^{-1} + \frac{1}{KD} \sum_{k=1}^{K} \langle\alpha_k\rangle\langle\mathbf{m}_{\mathrm{w},k}^\top\rangle\langle\mathbf{m}_{\mathrm{w},k}\rangle. \tag{10.38}$$

However, updating Gamma priors, i.e., $\{a_\alpha, b_\alpha, a_\psi, b_\psi, a_\lambda, b_\lambda\}$, does not have an analytic solution. We use a coordinate-descent algorithm which updates one parameter at a time and alternates over the parameters. This is achieved by fixed point iteration and solving over the following equations,

$$\mathit{\Psi}\left(a_{\alpha}\right) = \ln\left(b_{\alpha}\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\mathit{\Psi}\left(\tilde{a}_{\alpha}\right) - \ln\left(\tilde{b}_{\alpha,k}\right)\right), \tag{10.39}$$

$$\mathit{\Psi}\left(a_{\psi}\right) = \ln\left(b_{\psi}\right) + \frac{1}{SD}\sum_{s=1}^{S}\sum_{d=1}^{D}\left(\mathit{\Psi}\left(\tilde{a}_{\psi,s}\right) - \ln\left(\tilde{b}_{\psi,s,d}\right)\right), \tag{10.40}$$

$$\mathit{\Psi}\left(a_{\lambda}\right) = \ln\left(b_{\lambda}\right) + \frac{1}{CK}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(\mathit{\Psi}\left(\tilde{a}_{\lambda,c}\right) - \ln\left(\tilde{b}_{\lambda,c,k}\right)\right), \tag{10.41}$$

$$b_{\alpha}^{-1} = \frac{1}{a_{\alpha}K}\sum_{k=1}^{K}\frac{\tilde{a}_{\alpha}}{\tilde{b}_{\alpha,k}}, \tag{10.42}$$

$$b_{\psi}^{-1} = \frac{1}{a_{\psi}SD}\sum_{s=1}^{S}\sum_{d=1}^{D}\frac{\tilde{a}_{\psi,s}}{\tilde{b}_{\psi,s,d}}, \tag{10.43}$$

$$b_{\alpha}^{-1} = \frac{1}{a_{\lambda}CK}\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{\tilde{a}_{\lambda,c}}{\tilde{b}_{\lambda,c,k}}. \tag{10.44}$$

To speed up computations, the hyperparameters are updated every 10 iterations since updating the Gamma hyperparameters can be slow. Algorithm 10.1 represents the pseudo code of VBHMSRJMF.

10.3.2 Testing phase

The aim of the training phase was to find the approximate posterior distributions of the model parameters. At the test step, however, the MAP estimate of the latent variables are desired for a given test data. Moreover, due to individuality of the mean and noise terms, these parameters must be inferred and updated incrementally as test data arrives. Lets assume that the data of a test subject is presented sequentially, the posterior predictive of a new data $\mathbf{x}_n$ given all the past data is

$$\begin{aligned}
p\left(\mathbf{x}_n \mid \mathcal{D}_{\text{train}}, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right) = \int \prod_{c=1}^{C} & \Big(p\left(\mathbf{x}_n \mid \mathbf{W}, \mathbf{z}_{c,n}, \boldsymbol{\mu}, \mathbf{\Psi}\right) p\left(\mathbf{z}_{c,n} \mid \mathbf{\Lambda}_c\right) p\left(\mathbf{\Lambda}_c \mid \mathcal{D}_{\text{train}}\right)\Big) \\
& p\left(\boldsymbol{\mu}, \mathbf{\Psi} \mid \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right)^{\kappa} p\left(\mathbf{W} \mid \mathbf{M}_w, \alpha\right) \\
& p\left(\mathbf{M}_w \mid \alpha, \mathcal{D}_{\text{train}}\right) p\left(\alpha \mid \mathcal{D}_{\text{train}}\right) d\Theta \\
\approx \int \prod_{c=1}^{C} & \Big(p\left(\mathbf{x}_n \mid \mathbf{W}, \mathbf{z}_{c,n}, \boldsymbol{\mu}, \mathbf{\Psi}\right) p\left(\mathbf{z}_{c,n} \mid \mathbf{\Lambda}_c\right) q\left(\mathbf{\Lambda}_c\right)\Big) \\
& p\left(\boldsymbol{\mu}, \mathbf{\Psi} \mid \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right)^{\kappa} p\left(\mathbf{W} \mid \mathbf{M}_w, \alpha\right) \\
& q\left(\mathbf{M}_w \mid \alpha\right) q\left(\alpha\right) d\Theta,
\end{aligned} \tag{10.45}$$

where $\Theta = \left\{\mathbf{W}, \mathbf{M}_w, \alpha, \mathbf{z}, \boldsymbol{\mu}, \mathbf{\Psi}, \mathbf{\Lambda}\right\}$ is the set of all model parameters, $\left\{\mathbf{x}_1, \cdots, \mathbf{x}_{n-1}\right\}$ are the past data, and κ is forgetting factor. The latter can take a value between 0 to 1 where 1 does not

Algorithm 10.1 The training algorithm of variational Bayesian hierarchical multi-subject robust joint matrix factorization (VBHMSRJMF).

```
procedure Initializing
    $K = D - 1$, $a_\alpha = b_\alpha = a_\psi = b_\psi = a_\lambda = b_\lambda = \beta_\mu = \beta_w = 10^{-6}$.
    $\tilde{a}_{\psi,s} = a_\psi$, $\tilde{\mathbf{b}}_{\psi,s} = b_\psi$, $\tilde{a}_{\lambda,c} = a_\lambda$, $\tilde{\mathbf{b}}_{\lambda,c} = b_\lambda$, $\tilde{a}_\alpha = a_\alpha$, $\tilde{\mathbf{b}}_\alpha = b_\alpha$, $\tilde{\beta}_{\mu,s} = \beta_\mu$, $\tilde{\beta}_w = \beta_w$.
    $RelTol = 10^{-6}$, $MaxIter = 1000$.
    for $s = 1$ to $S$ do
        Set $\tilde{\mathbf{m}}_{\mu,s}$ to the empirical mean of the subject $s$ data.
        $\tilde{\Sigma}_{w,s,d} = \mathbf{I}, \forall d \in \{1, \ldots, D\}$.
        Set $\tilde{\mathbf{w}}_s$ to the first $K$ components of the PCA coefficients of the concatenated data of all classes of subject $s$.
        Set $\tilde{\mathbf{m}}_w$ to the first $K$ components of the PCA coefficients of the concatenated demeaned data of all classes and subjects.
for $iter = 1$ to $MaxIter$ do
    procedure VBE
        for $s = 1$ to $S$ do
            for $c = 1$ to $C$ do
                Update $\tilde{\Sigma}_{z,s,c}$ using Equation (10.15).
                for $n = 1$ to $N_{s,c}$ do
                    Update $\tilde{\mathbf{m}}_{z,s,c,n}$ using Equation (10.16).
    procedure VBM
        for $c = 1$ to $C$ do
            Update the class specific variational parameters using Equations (10.22) and (10.23).
        for $s = 1$ to $S$ do
            Update the subject specific variational paramters using Equations (10.24)–(10.29).
        Update the shared variational parameters using Equations (10.18)–(10.21).
    procedure Update hyperparameters
        if Reminder($iter$, 10) is 0 then
            Update $\beta_\mu$, $\beta_w$, and $\mathbf{m}_\mu$ using Equations (10.36)–(10.38).
            Update Gamma hyperparameters by iterating over Equations (10.39)–(10.44).
    procedure Stopping criteria
        Calculate the lower bound of the marginal log-likelihood.
        if $\frac{\mathcal{L}(iter) - \mathcal{L}(iter-1)}{|\mathcal{L}(iter)|} < RelTol$ then
            Stop.                                                    ▷ Converged
    procedure Pruning latent variables
        Temporarily remove the component corresponding to the highest $\langle \alpha \rangle$.
        Calculate the lower bound of the marginal log-likelihood.
        if $\mathcal{L}$ (after pruning) $>$ $\mathcal{L}$ (before pruning) then   ▷ The component can be removed.
            $\mathcal{L}(iter) = \mathcal{L}$ (after pruning) and remove the component.
        else                                          ▷ The component can not be removed.
            $\mathcal{L}(iter) = \mathcal{L}$ (before pruning) and keep the component.
```

forget anything. With a $0 < \kappa < 1$, the model exponentially lowers the weight of observations over time. Therefore, the model adapts itself over time to the more recently observed data.

Furthermore, Equation (10.45) uses the approximate posterior distributions of $\mathbf{M}_w$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Lambda}$ from the training phase. We can simplify the distribution of $\mathbf{W}$ by integrating out $\mathbf{M}_w$, resulting in a normal distribution,

$$p\left(\mathbf{W} \,\middle|\, \alpha\right) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k \,\middle|\, \tilde{\mathbf{m}}_{w,k}, \left(\delta\alpha_k\right)^{-1} \mathbf{I}\right), \tag{10.46}$$

where $\delta = {}^{\tilde{\beta}_w}\big/_{\left(\tilde{\beta}_w+1\right)}$.

Similar to the training phase, exact Bayesian inference is not a viable solution due to the intractable integral. A variational Bayesian is utilized to approximate the posterior distributions that maximize the lower bound of the posterior predictive distribution of a new data, which incrementally updates the mean and noise approximate distributions. Moreover, since the class of the test data point is not available, the latent variables $\mathbf{z}_c$ where $c \in \left\{1, \ldots, C\right\}$ are required to be jointly approximated. To approximate the posterior distributions, a variational EM was used where the variational distribution is assumed to be factorized as

$$q^*\left(\Theta^{(n)}\right) = \prod_{c=1}^{C} \left(q^*\left(\mathbf{z}_{c,n}\right) q^*\left(\boldsymbol{\Lambda}_c\right)\right) q^*\left(\boldsymbol{\mu}^{(n)}, \boldsymbol{\Psi}^{(n)}\right) q^*\left(\mathbf{W}\right) q^*\left(\alpha\right). \tag{10.47}$$

where $\Theta^{(n)} = \left\{\mathbf{W}, \alpha, \mathbf{z}_n, \boldsymbol{\Lambda}, \boldsymbol{\mu}^{(n)}, \boldsymbol{\Psi}^{(n)}\right\}$ is set of all parameters after observing n^{th} test data. Substituting the approximate posterior of the mean and noise terms as the prior probability for the next step leads to

$$\begin{aligned} p\left(\mathbf{x}_n \,\middle|\, \mathcal{D}_{\text{train}}, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\right) \approx \int \prod_{c=1}^{C} &\left(p\left(\mathbf{x}_n \,\middle|\, \mathbf{W}, \mathbf{z}_{c,n}, \boldsymbol{\mu}, \boldsymbol{\Psi}\right) p\left(\mathbf{z}_{c,n} \,\middle|\, \boldsymbol{\Lambda}_c\right) q\left(\boldsymbol{\Lambda}_c\right)\right) \\ & q^*\left(\boldsymbol{\mu}^{(n-1)}, \boldsymbol{\Psi}^{(n-1)}\right)^{\kappa} p\left(\mathbf{W} \,\middle|\, \alpha\right) q\left(\alpha\right) d\Theta. \end{aligned} \tag{10.48}$$

To approximate the posterior probabilities using variational inference, the lower bound of the predictive log-likelihood is

$$\begin{aligned} \mathcal{L} = &\left\langle \ln\left(\frac{p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right)}{q^*\left(\mathbf{W}\right)}\right)\right\rangle + \left\langle \ln\left(\frac{q\left(\boldsymbol{\alpha}\right)}{q^*\left(\boldsymbol{\alpha}\right)}\right)\right\rangle + \left\langle \ln\left(\frac{q^*\left(\boldsymbol{\Psi}^{(n-1)}\right)}{q^*\left(\boldsymbol{\Psi}^{(n)}\right)}\right)\right\rangle \\ &+ \left\langle \ln\left(\frac{q^*\left(\boldsymbol{\mu}^{(n-1)} \,\middle|\, \boldsymbol{\Psi}^{(n-1)}\right)^{\kappa}}{q^*\left(\boldsymbol{\mu}^{(n)} \,\middle|\, \boldsymbol{\Psi}^{(n)}\right)}\right)\right\rangle + \sum_{c=1}^{C}\left(\left\langle \ln\left(\frac{p\left(\mathbf{z}_{c,n} \,\middle|\, \boldsymbol{\Lambda}_c\right)}{q^*\left(\mathbf{z}_{c,n}\right)}\right)\right\rangle + \left\langle \ln\left(\frac{q\left(\boldsymbol{\Lambda}_c\right)}{q^*\left(\boldsymbol{\Lambda}_c\right)}\right)\right\rangle\right) \\ &+ \sum_{c=1}^{C}\left(\left\langle \ln\left(p\left(\mathbf{x}_n \,\middle|\, \mathbf{W}, \boldsymbol{\mu}, \mathbf{z}_{c,n}, \boldsymbol{\Psi}\right)\right)\right\rangle\right), \end{aligned} \tag{10.49}$$

and the objective is to maximize the lower bound. Alternating between the VBE and VBM steps would maximize the lower bound within a few iterations. The VBE step updates the variational distribution of the latent variables by

$$\hat{\Sigma}_{z,c}^{(n)} = \left(\langle \mathbf{W}^\top \boldsymbol{\Psi} \mathbf{W} \rangle + \langle \boldsymbol{\Lambda}_c \rangle\right)^{-1}, \tag{10.50}$$

$$\hat{\mathbf{m}}_{z,c}^{(n)} = \hat{\Sigma}_{z,c}^{(n)} \langle \mathbf{W}^\top \rangle \langle \boldsymbol{\Psi} \rangle \left(\mathbf{x}_n - \langle \boldsymbol{\mu} \rangle\right), \tag{10.51}$$

where the variational distribution of the latent variables is

$$q^* \left(\mathbf{z}_n\right) = \prod_{c=1}^{C} \mathcal{N}\left(\mathbf{z}_{c,n} \,\middle|\, \hat{\mathbf{m}}_{z,c}^{(n)}, \hat{\Sigma}_{z,c}^{(n)}\right). \tag{10.52}$$

The VBM step updates the approximate distribution of the parameters by iterating over

$$\hat{a}_{\alpha}^{(n)} = \tilde{a}_{\alpha} + \frac{D}{2}, \tag{10.53}$$

$$\hat{b}_{\alpha,k}^{(n)} = \tilde{b}_{\alpha,k} + \frac{\delta}{2}\left(\langle \mathbf{w}_k^\top \mathbf{w}_k \rangle - 2\langle \mathbf{w}_k^\top \rangle \tilde{\mathbf{m}}_{\mathbf{w},k} + \tilde{\mathbf{m}}_{\mathbf{w},k}^\top \tilde{\mathbf{m}}_{\mathbf{w},k}\right), \tag{10.54}$$

$$\hat{\Sigma}_{\mathbf{w},d}^{(n)} = \left(\delta \operatorname{diag}\left(\langle \boldsymbol{\alpha} \rangle\right) + \langle \psi_d \rangle \sum_{c=1}^{C} \langle \mathbf{z}_{c,n} \mathbf{z}_{c,n}^\top \rangle\right)^{-1}, \tag{10.55}$$

$$\hat{\mathbf{w}}_{d,.}^{(n)\top} = \hat{\Sigma}_{\mathbf{w},d}^{(n)}\left(\delta \operatorname{diag}\left(\langle \boldsymbol{\alpha} \rangle\right) \tilde{\mathbf{m}}_{w,d,.}^\top + \langle \psi_d \rangle \sum_{c=1}^{C} \langle \mathbf{z}_{c,n} \rangle \left(x_{n,d} - \langle \mu_d \rangle\right)\right), \tag{10.56}$$

$$\hat{a}_{\lambda,c}^{(n)} = \tilde{a}_{\lambda,c} + \frac{1}{2}, \tag{10.57}$$

$$\hat{b}_{\lambda,c,k}^{(n)} = \tilde{b}_{\lambda,c,k} + \frac{1}{2}\langle z_{c,k}^2 \rangle, \tag{10.58}$$

$$\hat{\beta}_{\mu}^{(n)} = \kappa \hat{\beta}_{\mu}^{(n-1)} + C, \tag{10.59}$$

$$\hat{m}_{\mu,d}^{(n)} = \frac{1}{\hat{\beta}_{\mu}^{(n)}}\left(\kappa \hat{\beta}_{\mu}^{(n-1)} \hat{m}_{\mu,d}^{(n-1)} + \sum_{c=1}^{C}\left(x_{n,d} - \langle \mathbf{w}_{d,.} \rangle \langle \mathbf{z}_{c,n} \rangle\right)\right), \tag{10.60}$$

$$\hat{a}_{\psi}^{(n)} = \kappa\left(\hat{a}_{\psi}^{(n-1)} - \frac{1}{2}\right) + \frac{C+1}{2}, \tag{10.61}$$

$$\begin{aligned}\hat{b}_{\psi,d}^{(n)} = {} & \kappa \hat{b}_{\psi,d}^{(n-1)} - \frac{\hat{\beta}_{\mu}^{(n)}}{2}\left(\hat{m}_{\mu,d}^{(n)}\right)^2 + \kappa \frac{\hat{\beta}_{\mu}^{(n-1)}}{2}\left(\hat{m}_{\mu,d}^{(n-1)}\right)^2 \\ & + \frac{1}{2}\sum_{c=1}^{C}\left(x_{n,d}^2 - 2x_{n,d}\langle \mathbf{w}_{d,.} \rangle \langle \mathbf{z}_{c,n} \rangle + \operatorname{tr}\left(\langle \mathbf{w}_{d,.}^\top \mathbf{w}_{d,.} \rangle \langle \mathbf{z}_{c,n} \mathbf{z}_{c,n}^\top \rangle\right)\right),\end{aligned} \tag{10.62}$$

$$\hat{\Sigma}_{w,d}^{(n)} = \left(\tilde{\Sigma}_{w,d}^{-1} + \langle \psi_d \rangle \sum_{c=1}^{C} \langle \mathbf{z}_{c,n} \mathbf{z}_{c,n}^\top \rangle\right)^{-1}, \tag{10.63}$$

$$\hat{\mathbf{m}}_{w,d}^{(n)} = \hat{\Sigma}_{w,d}^{(n)}\left(\tilde{\Sigma}_{w,d}^{-1} \tilde{\mathbf{m}}_{w,d} + \langle \psi_d \rangle \sum_{c=1}^{C} \langle \mathbf{z}_{c,n} \rangle \left(x_{n,d} - \langle \mu_d \rangle\right)\right), \tag{10.64}$$

where all the expectations are taken with respect to the variational distributions of the n^{th} test data.

10.4 RESULTS AND DISCUSSION

10.4.1 Detection and prediction of microsleep states

Similar to VBHMSRFA in Section 9.4, an arbitrary value of $\kappa = 0.9999$ was used to prevent our proposed method forgetting the past too quickly. This value was chosen arbitrarily and no experiments were performed to fine-tune it. Table 10.1 shows the results of microsleep state detection with an LDA classifier and meta-features of various feature sets extracted from single EEG windows (2, 5, and 10 s). Improvements of AUC-ROC and AUC-PR with VBHMSRJMF meta-features of MDF with 2-s windows relative to the baselines were 5.8% and 19.4%, respectively. Similarly, slight improvements were observed with meta-features of WMSF, WLMSF, and WEPF of 2-s EEG windows. However, contrary to our expectation, the performances dropped with VBHMSRJMF meta-features of extracted features from longer EEG windows. Nevertheless, the highest AUC-ROC of 0.93 was achieved with VBHMSRJMF meta-features of WLMSF using 5-s EEG windows, whereas the baseline of the same feature set and PSF had the highest AUC-PR of 0.41.

Table 10.1 Performance (mean ± SE) of microsleep state detection with VBHMSRJMF meta-features and an LDA classifier. A bold value indicates the highest performances of each feature set and italics indicate the highest overall. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements relative to the baselines.

Feature set	Feature type	2-s EEG window		5-s EEG window		10-s EEG window	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.86±0.03	0.31±0.12	0.88±0.03	0.36±0.12	0.86±0.04	0.36±0.12
	VBHMSRJMF	**0.91±0.02***	**0.37±0.11**	0.89±0.02	0.36±0.13	0.84±0.04	0.33±0.12
PSF	Baseline	0.89±0.02	0.34±0.12	**0.92±0.01**	**0.41±0.12**	**0.92±0.02**	0.37±0.12
	VBHMSRJMF	0.89±0.02	0.33±0.12	0.90±0.03	0.37±0.13	0.87±0.03	0.33±0.13
PSF-IAF	Baseline	0.88±0.02	0.34±0.12	0.90±0.02	**0.37±0.13**	**0.91±0.02**	0.35±0.12
	VBHMSRJMF	0.84±0.03	0.28±0.12	0.84±0.03	0.26±0.11	0.79±0.06	0.28±0.13
WMSF	Baseline	0.87±0.03	0.35±0.12	**0.91±0.02**	0.38±0.14	**0.91±0.02**	0.37±0.13
	VBHMSRJMF	0.90±0.02	0.36±0.12	**0.91±0.02**	**0.39±0.13**	0.89±0.02	0.37±0.12
WLMSF	Baseline	0.88±0.03	0.36±0.13	0.91±0.02	***0.41±0.13***	0.91±0.02	0.37±0.13
	VBHMSRJMF	0.89±0.02	0.34±0.12	***0.93±0.01***	0.40±0.13	0.89±0.03	0.38±0.12
WEPF	Baseline	0.74±0.03	0.21±0.10	0.78±0.03	0.22±0.11	**0.81±0.02**	0.23±0.11
	VBHMSRJMF	0.80±0.04	**0.26±0.10**	0.79±0.05	0.25±0.11	0.79±0.03	0.23±0.11

Wilcoxon signed-rank test: * $p < 0.05$

The number of VBHMSRJMF meta-features of various feature sets are shown in Table 10.2. Although the number of VBHMSRJMF meta-features were higher than those of the VBHMSRFA, performances of the latter were mostly superior. Since VBHMSRJMF uses all of the classes and our microsleep datasets are highly imbalanced, the lower-dimension representation of data might have been affected. This might explain the lower performances with VBHMSRJMF meta-features.

Table 10.2 Average number of VBHMSRJMF meta-features.

Feature set	Number of features			
	2-s EEG window	5-s EEG window	10-s EEG window	Baseline
MDF	62.6	69.8	72.8	176
PSF	99.6	124.1	148.9	192
PSF-IAF	50.9	67.9	95.3	192
WMSF	37.1	42.8	43.8	80
WLMSF	26.1	30.9	36.9	80
WEPF	36.1	41.3	46.6	80

To exploit both tonic and transient characteristics of EEG, VBHMSRJMF meta-features of multiple EEG windows, i.e., 2, 5, and 10 s, were aggregated. This was done by applying VBHMSRJMF to features extracted from each window length independently, computing meta-features, and then concatenating meta-features of all windows. VBHMSRJMF was applied independently to the features of each EEG window length to reduce the computation time, which arises from matrix inversions in Equations (10.15) and (10.28). Table 10.3 presents the results of microsleep state detection ($\tau = 0$ s) using aggregated VBHMSRJMF meta-features and an LDA classifier. It is evident that although the AUC-ROC and AUC-PR performances were improved by aggregating meta-features of multiple EEG windows, the performances of aggregated baselines were similar or superior to the VBHMSRJMF meta-features. Notwithstanding, the highest GM and phi performance metrics were achieved with VBHMSRJMF meta-features. The highest GM of 0.81 was achieved with meta-features of WLMSF (cf. 0.76 for baselines with PSF-IAF and WLMSF). The highest phi of 0.43 was achieved with meta-features of PSF (cf. 0.37 for baseline with PSF-IAF). Higher values of GM and phi with relatively similar AUC-ROC and AUC-PR indicate that applying VBHMSRJMF to the features of multiple subjects reduces the inter-subject variability of classification threshold. To our surprise, the performances of microsleep state detection with the baseline PSF-IAF was substantially higher than that with VBHMSRJMF meta-features.

Table 10.4 shows the performance of microsleep state detection with different classifiers and VBHMSRJMF meta-features. The highest AUC-ROC of 0.95 and AUC-PR of 0.49 were achieved with a linear SVM and the baseline features of PSF-IAF and PSF, respectively. With VBHMSRJMF meta-features, the highest AUC-ROC of 0.94 was with PSF and the highest AUC-PR of 0.43 was with WMSF and WLMSF. Although the overall performance with VBHMSRJMF meta-features was slightly lower than that of the baseline, TAN performed substantially better with VBHMSRJMF meta-features. Using VBHMSRJMF meta-features and a TAN classifier, the average improvement of AUC-ROC and AUC-PR over all feature sets compared to the baseline were 10.5% (0.0–22.9%) and 33.6% (3.8–61.9%), respectively.

The performance of microsleep state prediction in terms of AUC-PR with various prediction times (τ = 0–1 s), different feature sets, and different classifiers is shown in Figure 10.2. It is evident that the AUC-PRs with baseline features of PSF, PSF-IAF, MDF, and WLMSF were

Table 10.3 Performance (mean ± SE) of microsleep state detection using aggregated VBHMSRJMF meta-features of multiple EEG windows and an LDA classifier. A bold value indicates the highest performances in individual feature sets, while italics indicate the highest among all feature sets.

Feature set	Feature type	Microsleep state prediction performance with $\tau = 0$					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	**0.90±0.02**	**0.40±0.13**	**0.71±0.08**	0.34±0.10	**0.62±0.10**	**0.33±0.12**
	VBHMSRJMF	**0.90±0.03**	0.39±0.12	0.70±0.08	**0.36±0.09**	0.59±0.10	**0.33±0.12**
PSF	Baseline	***0.94±0.01***	**0.43±0.12**	0.74±0.06	0.36±0.08	**0.68±0.11**	0.36±0.12
	VBHMSRJMF	***0.94±0.01***	0.42±0.13	**0.77±0.04**	***0.43±0.10***	0.63±0.07	***0.40±0.13***
PSF-IAF	Baseline	***0.94±0.01***	***0.44±0.12***	**0.76±0.05**	**0.37±0.08**	**0.70±0.10**	**0.36±0.12**
	VBHMSRJMF	0.86±0.04	0.34±0.13	0.67±0.06	0.31±0.09	0.53±0.09	0.30±0.13
WMSF	Baseline	**0.92±0.02**	**0.40±0.14**	0.66±0.10	0.33±0.09	0.57±0.12	0.31±0.11
	VBHMSRJMF	**0.92±0.02**	**0.40±0.12**	**0.74±0.05**	**0.40±0.09**	**0.61±0.07**	**0.39±0.13**
WLMSF	Baseline	***0.94±0.01***	***0.44±0.13***	0.76±0.07	0.36±0.10	0.70±0.11	0.31±0.12
	VBHMSRJMF	0.93±0.02	0.43±0.13	***0.81±0.03***	**0.41±0.10**	***0.73±0.05***	**0.35±0.13**
WEPF	Baseline	**0.84±0.02**	0.27±0.12	**0.70±0.03**	**0.25±0.07**	**0.67±0.08**	**0.25±0.13**
	VBHMSRJMF	0.83±0.03	**0.29±0.13**	0.65±0.07	0.24±0.09	0.55±0.10	**0.25±0.13**

Table 10.4 Performance of different classifiers for microsleep state detection with aggregated VBHMSRJMF meta-features. A bold value indicates the highest performance among selected classifiers and an italic value is the highest overall. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements of VBHMSRJMF meta-features comparing to the baselines.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.90	0.40	**0.93**	**0.48**	0.90	0.42	0.77	0.22
	VBHMSRJMF	0.90	0.39	0.88	0.40	0.89	0.39	0.85*	0.28
PSF	Baseline	**0.94**	0.43	**0.94**	***0.49***	**0.94**	0.44	0.70	0.21
	VBHMSRJMF	**0.94**	0.42	0.93	0.39	0.92	0.39	0.86*	0.34$^\sim$
PSF-IAF	Baseline	0.94	0.44	***0.95***	**0.47**	0.93	0.45	0.73	0.26
	VBHMSRJMF	0.86	0.34	0.91	0.37	0.88	0.34	0.73	0.27
WMSF	Baseline	0.92	0.40	**0.93**	0.42	0.92	0.39	0.77	0.22
	VBHMSRJMF	0.92	0.40	**0.93**	**0.43**	0.92	0.41	0.82	0.26
WLMSF	Baseline	**0.94**	0.44	0.93	**0.45**	0.92	**0.45**	0.76	0.25
	VBHMSRJMF	0.93	0.43	0.92	0.43	0.92	0.43	0.90*	0.40
WEPF	Baseline	0.84	0.27	**0.86**	0.29	0.85	0.28	0.82	0.23
	VBHMSRJMF	0.83	0.29	0.85	**0.30**	0.84	0.29	0.86	**0.30***

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$

superior to the AUC-PRs with VBHMSRJMF meta-features. Comparative performances were observed with baseline features and VBHMSRJMF meta-features of WMSF and WEPF. With baseline features, the linear SVM classifier performed better or similar to other classifiers. However, with VBHMSRJMF meta-features, the three classifiers of LDA, linear SVM, and VBLR performed relatively similar while TAN had the lowest performance with the exception of WEPF. Notwithstanding, the AUC-PRs of TAN with VBHMSRJMF meta-features were superior to those with baseline features. Nevertheless, the highest detection AUC-PR ($\tau = 0$ s) with VBHMSRJMF meta-features was 0.43, achieved with VBLR and WLMSF (cf. 0.49 for baseline with a linear SVM and PSF). Increasing the prediction time to $\tau = 1.0$ s, the highest AUC-PR of 0.40 was achieved with VBHMSRJMF meta-features of WLMSF and an LDA (cf. 0.42 for baseline with a linear SVM and MDF). Moreover, the AUC-PRs with WEPF were the lowest among all the feature sets.

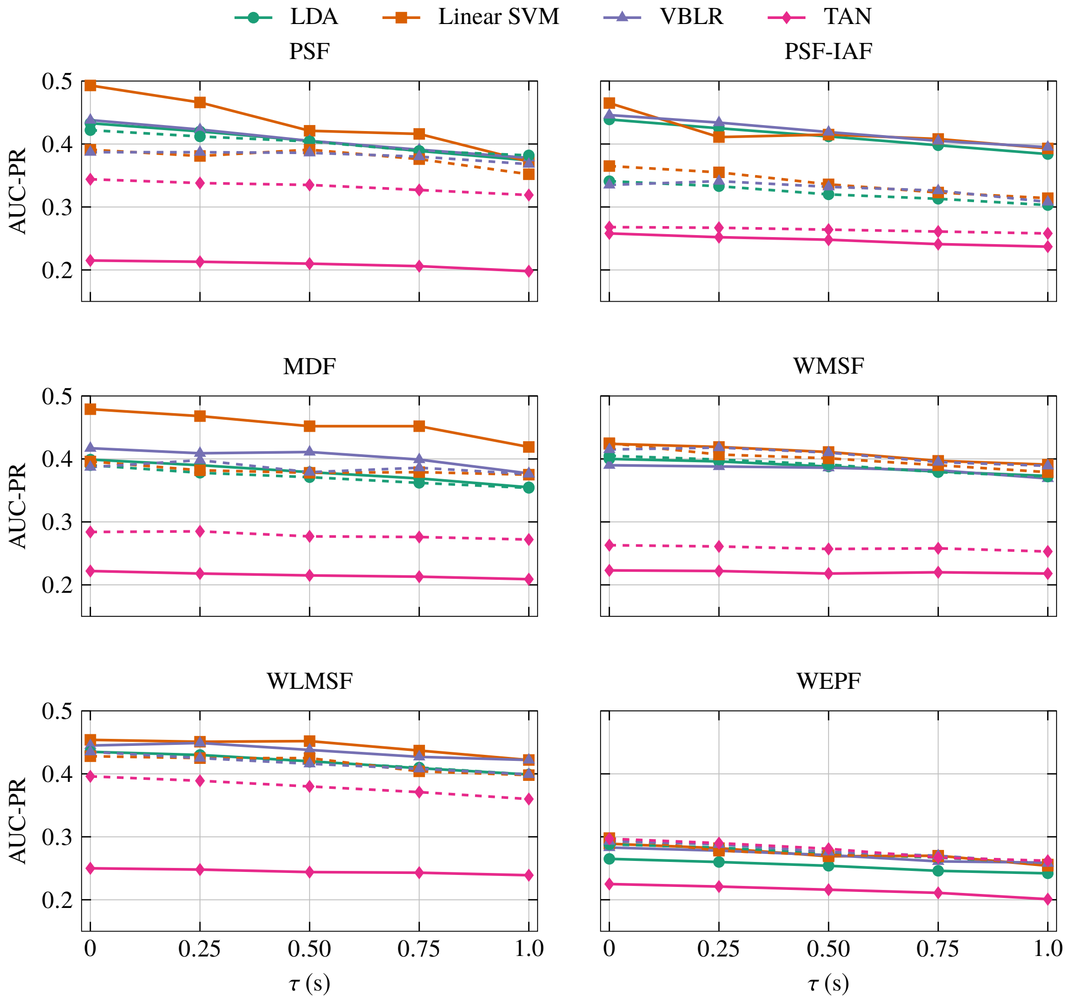


Figure 10.2 Performance of microsleep state prediction in terms of AUC-PR using different classifiers versus prediction time τ = 0–1 s. Solid lines correspond to the performance of a classifier with baseline features, whereas dashed lines correspond to their respective performance with VBHMSRJMF meta-features.

AUC-PR is a threshold-free performance measure and although it provides an overall system performance, it does not capture the operating performance of the system with a specific classification-threshold. Hence, the performances of microsleep state prediction in terms of phi metric are shown in Table 10.5. Surprisingly, phi performances with VBHMSRJMF meta-features of PSF, WMSF, and WLMSF outperformed the baselines. Moreover, the phi measures with meta-features of MDF and a linear SVM were similar to of the baselines, while the AUC-ROCs and AUC-PRs across prediction times with the baseline MDF were on average 4.5% (3.6–5.8%) and 18.8% (11.7–22.5%) higher, respectively. This indicates that although applying VBHMSRJMF did not improve threshold-free performance measures, it reduced inter-subject variability between training and testing subjects. Thus, the threshold found with the training subjects using VBHMSRJMF led to a better phi performance with an independent test subject.

For detection, the highest phi of 0.43 was achieved with an LDA and VBHMSRJMF meta-features of PSF (cf. 0.40 for baseline with a linear SVM and PSF). Increasing the prediction time to $\tau = 1.0$ s, the highest phi of 0.39 was achieved with meta-features of PSF and WLMSF (cf. 0.35 for baseline with a linear SVM and MDF). On average across prediction times, improvements of phi values relative to the best baselines were 11.0% (5.4–14.7%), 14.1% (11.4–20.6%), and 13.7% (11.4–14.7%), with VBHMSRJMF meta-features of PSF, WMSF, and WLMSF, respectively.

Table 10.5 Phi values of different classifiers for microsleep state prediction with aggregated VBHMSRJMF meta-features for τ = 0–1 s. A bold value indicates the highest performance among selected classifiers. Two-tail Wilcoxon signed-rank tests were performed to identify significant improvements by using VBHMSRJMF meta-features relative to their respective baseline.

Feature set	τ (s)	LDA		Linear SVM		VBLR		TAN	
		Baseline	VBHMSRJMF	Baseline	VBHMSRJMF	Baseline	VBHMSRJMF	Baseline	VBHMSRJMF
MDF	0.00	0.34	0.36	**0.38**	**0.38**	0.36	0.37	0.28	0.29
	0.25	0.33	0.36	**0.37**	**0.37**	0.35	**0.37**	0.28	0.30
	0.50	0.33	0.35	**0.37**	0.36	0.36	**0.37**	0.27	0.29
	0.75	0.32	0.35	**0.37**	**0.37**	0.36	**0.37**	0.27	0.29
	1.00	0.31	0.35	0.35	**0.36**	0.34	**0.36**	0.26	0.29
PSF	0.00	0.36	**0.43**	0.40	0.40	0.37	0.39	0.21	0.39*
	0.25	0.35	**0.42**	0.37	0.39	0.36	0.38	0.20	0.38*
	0.50	0.34	**0.41**	0.36	0.40	0.35	0.38	0.20	0.38*
	0.75	0.34	**0.39**$^{\sim}$	0.37	0.37	0.36	0.37	0.20	0.37*
	1.00	0.33	**0.39**$^{\sim}$	0.33	0.36	0.34	0.36	0.19	0.36*
PSF-IAF	0.00	**0.37**	0.31	0.36	0.36	0.36	0.33	0.26	0.26
	0.25	**0.37**	0.31	0.34	0.34	0.36	0.34	0.25	0.26
	0.50	**0.36**	0.30	**0.36**	0.32	**0.36**	0.33	0.25	0.25
	0.75	**0.35**	0.29	**0.35**	0.33	0.34	0.33	0.24	0.25
	1.00	**0.34**	0.28	0.33	0.32	**0.34**	0.31	0.23	0.25
WMSF	0.00	0.33	0.40	0.34	**0.41**	0.34	0.38	0.27	0.27
	0.25	0.33	**0.39**	0.35	**0.39**	0.34	0.38	0.27	0.27
	0.50	0.33	**0.39**	0.34	**0.39**	0.33	0.38	0.27	0.27
	0.75	0.32	**0.38**	0.34	**0.38**	0.34	0.37	0.27	0.27
	1.00	0.32	**0.37**	0.33	**0.37**	0.33	0.36$^{\sim}$	0.26	0.26
WLMSF	0.00	0.36	**0.41**	0.35	0.40	0.36	0.40	0.24	0.40
	0.25	0.36	**0.40**	0.34	0.39	0.35	**0.40**	0.24	0.39
	0.50	0.35	**0.40**	0.35	**0.40**	0.34	**0.40**	0.24	0.39
	0.75	0.35	**0.39**	0.35	**0.39**	0.34	**0.39**	0.24	0.38
	1.00	0.34	**0.39**	0.33	**0.39**	0.34	**0.39**	0.23	0.37$^{\sim}$
WEPF	0.00	0.25	0.24	**0.26**	0.25	**0.26**	0.25	0.21	0.24
	0.25	0.24	0.23	**0.26**	0.25	**0.26**	0.25	0.21	0.24
	0.50	0.24	0.23	**0.25**	0.24	**0.25**	0.24	0.21	0.24
	0.75	0.23	0.22	**0.25**	0.24	**0.25**	0.23	0.20	0.23
	1.00	0.23	0.22	0.23	0.23	**0.24**	0.23	0.19	0.22

Wilcoxon signed-rank test: ~$p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Figure 10.3 shows the performance of microsleep prediction in terms of GM with different classifiers. Interestingly, the highest GMs with meta-features of PSF were achieved with a TAN classifier. However, TAN had substantially lower values of AUC-ROC and AUC-PR compared to the other classifiers. The VBHMSRJMF meta-features of WMSF (with a linear SVM) and WLMSF (with an LDA) had higher values of GM and phi compared to the baseline, but their

AUC-ROC and AUC-PR were relatively similar to the baseline. The average improvement of GM with VBHMSRJMF meta-features across prediction times for PSF, WMSF, and WLMSF were 5.8% (1.4–9.2%), 4.5% (2.3–6.2%), and 6.6% (5.5–7.9%), respectively. For detection, the highest GM of 0.81 ($\varphi = 0.41$; Sn = 0.73; Pr = 0.35) was achieved with VBHMSRJMF meta-features of WLMSF and an LDA (cf. GM = 0.79; $\varphi = 0.40$; Sn = 0.74; Pr = 0.38 for baseline with a linear SVM and PSF). Increasing the prediction time to $\tau = 1.0$ s with a similar setup resulted in the same GM of 0.81, but the other performance metrics slightly dropped, i.e., phi = 0.39; Sn = 0.73; Pr = 0.34 (cf. GM = 0.76; $\varphi = 0.35$; Sn = 0.70; Pr = 0.35 for baseline with a linear SVM and MDF).

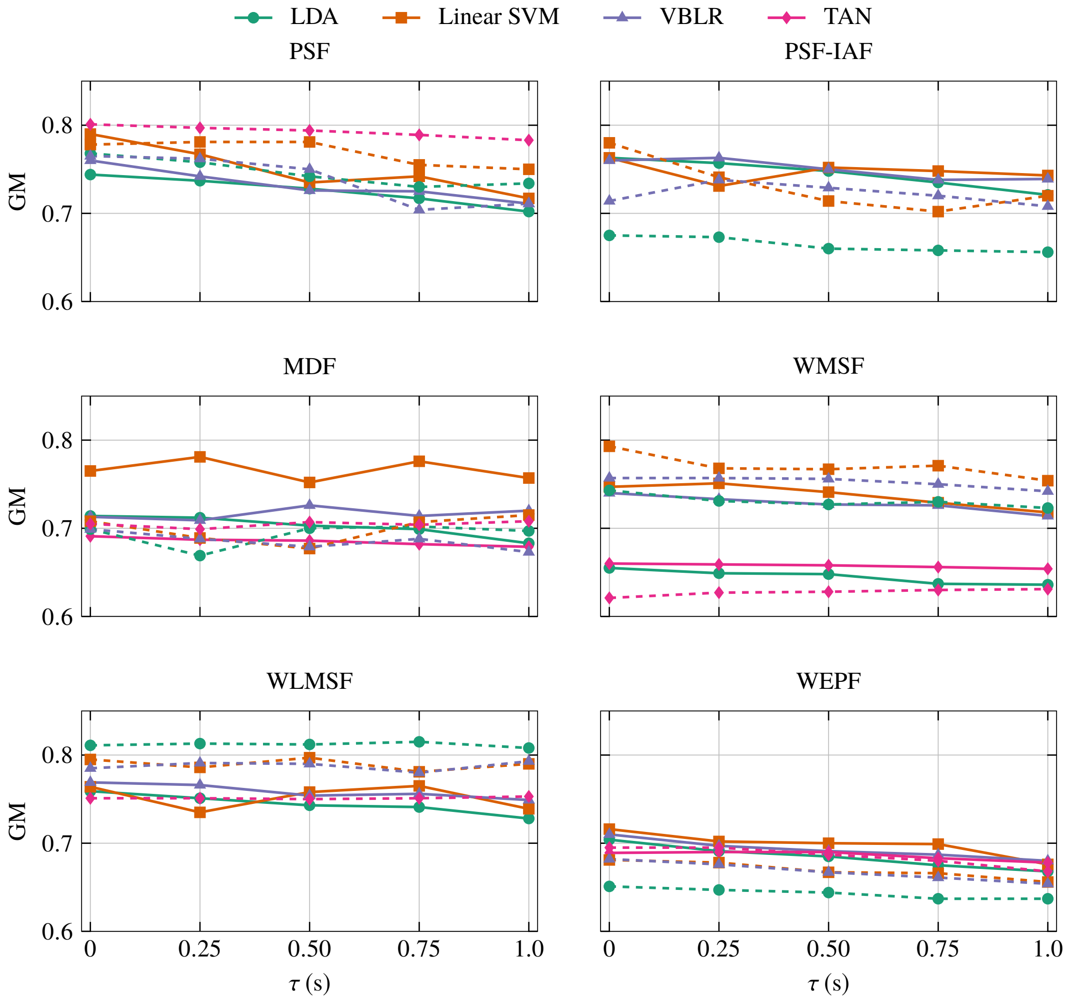


Figure 10.3 Performance of microsleep state prediction in terms of GM using different classifiers versus prediction time τ = 0–1 s. Solid lines correspond to the performance of a classifier with baseline features, whereas dashed lines correspond to their respective performance with VBHMSRJMF meta-features.

Sensitivity and precision values of microsleep state prediction with VBHMSRJMF are shown in Figure 10.4. The highest detection sensitivity of 0.73 (Pr = 0.35) was achieved with VBHMSRJMF meta-features of WLMSF and an LDA (cf. Sn = 0.77; Pr = 0.19 for baseline with a TAN and WEPF). The highest detection precision of 0.40 (Sn = 0.63) was achieved with meta-features of PSF and an LDA (cf. Pr = 0.39; Sn = 0.61 for baseline with WMSF and an LDA). Increasing the prediction time to $\tau = 1.0$ s, the highest sensitivity of 0.73 (Pr = 0.34) was with an LDA and the meta-features of WLMSF (cf. Sn = 0.17; Pr = 0.17 for baseline with a TAN and WEPF). Similarly, the highest precision of 0.37 (Sn = 0.58) was with an LDA and the meta-features of PSF (cf. Pr = 0.36; Sn = 0.64 for baseline with an LDA and PSF). Using VBHMSRJMF meta-features of WLMSF with an LDA or a TAN classifier improved both sensitivity and precision across prediction times. Although using a linear SVM with meta-features of WLMSF and WMSF slightly deteriorated the sensitivities, precisions improved substantially. A similar pattern was observed with VBLR, where there was no consistent improvement of either measure. Using meta-features of PSF and WLMSF with a TAN classifier improved both the sensitivity and precision consistently across prediction times.

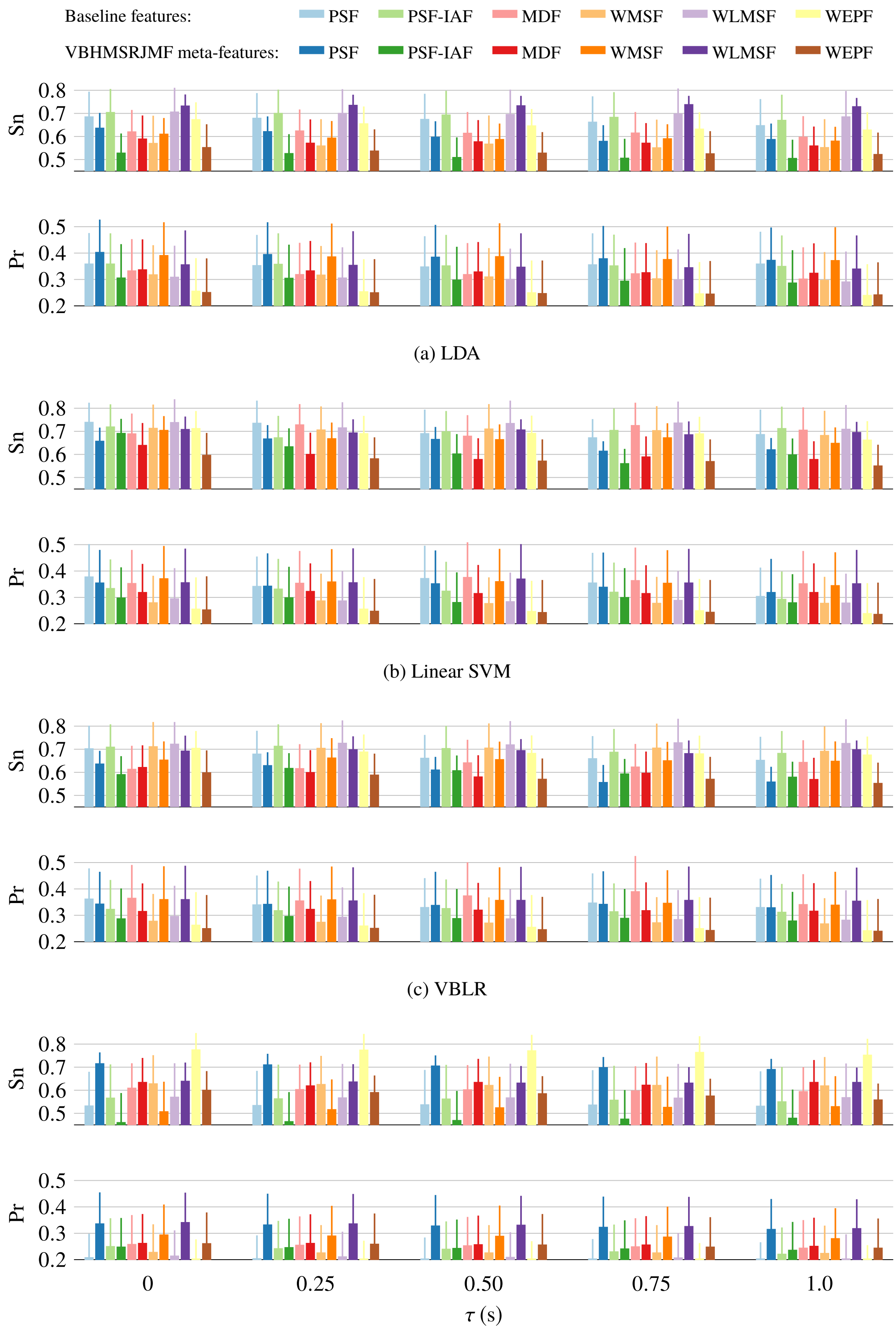
10.4.2 Detection and prediction of microsleep onsets

This section provides the results of microsleep onset detection and prediction using aggregated VBHMSRJMF meta-features with various feature sets and classifiers. The detection performance ($\tau = 0$ s) with an LDA classifier is shown in Table 10.6. Similar to the results of state detection, AUC-ROC and AUC-PR dropped slightly with VBHMSRJMF meta-features, with an average drop of AUC-ROC of 2.0% (0.0–3.6%). The highest detection performances of MDF, PSF, PSF-IAF, and WEPF were achieved with the baseline. Notwithstanding, the highest GM of 0.74 was with meta-features of WLMSF (cf. 0.73 for baseline with PSF-IAF and WLMSF).

Table 10.6 Performance (mean ± SE) of microsleep onset detection using aggregated VBHMSRJMF meta-features and an LDA classifier. Bold values indicate the highest performances of individual feature sets and italic values show the highest overall performances.

Feature set	Feature type	Microsleep onset prediction performance with $\tau = 0$					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	**0.87±0.02**	**0.03±0.01**	**0.69±0.05**	***0.09±0.02***	**0.55±0.08**	**0.02±0.01**
	VBHMSRJMF	0.85±0.02	**0.03±0.01**	0.67±0.04	0.07±0.01	0.54±0.07	0.01±0.01
PSF	Baseline	**0.87±0.01**	**0.04±0.03**	**0.71±0.04**	**0.07±0.01**	**0.63±0.09**	**0.02±0.01**
	VBHMSRJMF	0.84±0.01	0.03±0.01	0.60±0.05	0.06±0.02	0.43±0.07	**0.02±0.01**
PSF-IAF	Baseline	**0.87±0.01**	**0.04±0.03**	**0.73±0.04**	**0.08±0.01**	**0.66±0.09**	**0.02±0.01**
	VBHMSRJMF	0.84±0.03	0.04±0.01	0.68±0.06	0.07±0.01	0.58±0.10	**0.02±0.01**
WMSF	Baseline	***0.88±0.02***	**0.03±0.02**	0.69±0.07	0.08±0.02	**0.60±0.11**	0.02±0.01
	VBHMSRJMF	0.87±0.03	0.02±0.01	**0.70±0.07**	***0.09±0.02***	0.59±0.11	***0.03±0.01***
WLMSF	Baseline	***0.88±0.02***	***0.05±0.03***	0.73±0.05	0.07±0.01	***0.70±0.10***	0.01±0.00
	VBHMSRJMF	0.87±0.02	0.03±0.01	***0.74±0.04***	**0.08±0.02**	0.66±0.08	**0.02±0.01**
WEPF	Baseline	**0.73±0.04**	**0.01±0.01**	**0.63±0.05**	**0.04±0.01**	**0.60±0.11**	**0.01±0.00**
	VBHMSRJMF	**0.73±0.05**	**0.01±0.00**	0.59±0.09	**0.04±0.01**	0.53±0.11	**0.01±0.00**

To compare the effect of different classifiers on microsleep onset detection, the values of AUC-ROC and AUC-PR of different classifiers are shown in Table 10.7. Linear SVM with baseline features had almost the highest performances in individual feature sets. With VBHMSRJMF meta-features of PSF and WMSF, the highest AUC-ROC was 0.89 (cf. 0.91 for



(a) LDA

(b) Linear SVM

(c) VBLR

(d) TAN

Figure 10.4 Precision and sensitivity measures (mean ± SE) of microsleep state prediction using VBHMSRJMF meta-features for τ = 0–1 s.

baseline with MDF, PSF, and PSF-IAF). The highest AUC-PR achieved with VBHMSRJMF meta-features was 0.05 with WMSF and WLMSF (cf. 0.09 for baseline with PSF). An interesting finding was that performances of TAN were improved using VBHMSRJMF meta-features. AUC-ROC measures with a TAN and VBHMSRJMF meta-features were on average 11.0% (3.9–18.6%) higher than those of the baselines across feature sets. Nevertheless, in most cases, performances of the TAN in terms of AUC-ROC and AUC-PR were lower than the LDA and linear SVM classifiers.

Table 10.7 Performance of different classifiers for microsleep onset detection with VBHMSRJMF aggregated meta-features. A bold value indicates the highest performance among classifiers of individual feature sets, whereas an italic value indicates the highest performance overall.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.87	0.03	***0.91***	**0.06**	0.82	0.05	0.72	0.01
	VBHMSRJMF	0.85	0.03	0.87	0.04	0.83	0.03	0.84^{*}	0.02^{*}
PSF	Baseline	0.87	0.04	***0.91***	***0.09***	0.83	0.04	0.70	0.01
	VBHMSRJMF	0.84	0.03	0.89	0.04	0.84	0.03	0.83^{**}	0.02^{*}
PSF-IAF	Baseline	0.87	0.04	***0.91***	**0.07**	0.81	0.04	0.70	0.01
	VBHMSRJMF	0.84	0.04	0.88	0.03	$0.86^{\sim}$	0.03	0.74	0.02
WMSF	Baseline	0.88	0.03	**0.90**	**0.05**	0.80	0.02	0.74	0.01
	VBHMSRJMF	0.87	0.02	0.89	0.05	0.84	0.03	0.83^{**}	0.02^{*}
WLMSF	Baseline	0.88	0.05	**0.90**	**0.06**	0.83	0.04	0.79	0.02
	VBHMSRJMF	0.87	0.03	0.88	0.05	0.86	0.04	$0.86^{\sim}$	$0.04^{\sim}$
WEPF	Baseline	0.73	0.01	0.78	**0.02**	0.74	0.01	0.76	0.01
	VBHMSRJMF	0.73	0.01	0.75	0.01	0.70	0.01	**0.79**	0.01

Wilcoxon signed-rank test: $\sim p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Performance of different classifiers in terms of AUC-ROC for longer prediction times (up to 10 s) are shown in Figure 10.5. This shows that a linear SVM classifier with baseline features has the highest AUC-ROCs in most cases. Notwithstanding, although WEPF had the lowest AUC-ROCs among the feature sets, a TAN classifier with VBHMSRJMF meta-features performed similar or better to that of the baseline features. In addition, using VBHMSRJMF meta-features, linear SVM and TAN classifiers had similar or higher performances comparing to the other classifiers. For detection of microsleep onsets, the highest AUC-ROC of 0.89 (GM = 0.79, $\varphi = 0.09$) was with a TAN and meta-features of WLMSF (cf. AUC-ROC = 0.91, GM = 0.71, and $\varphi = 0.08$ for baseline with a linear SVM and PSF). Increasing the prediction time to $\tau = 10$ s, the highest AUC-ROC of 0.74 (GM = 0.66, $\varphi = 0.04$) was with a TAN and meta-features of WLMSF (cf. AUC-ROC = 0.74, GM = 0.54, and $\varphi = 0.03$ for baseline with a linear SVM and WLMSF).

Since AUC-ROC is a threshold-free performance measure, it does not provide an indication of performance with the selected classification-threshold. Therefore, the phi values of microsleep onset prediction for different feature sets are shown in Figure 10.6. Interestingly, phi measures with VBHMSRJMF were slightly higher than the baselines, with the exception of PSF-IAF. Moreover, linear SVM with VBHMSRJMF meta-features of MDF and WMSF outperformed other classifiers. Similarly, TAN outperformed other classifiers with VBHMSRJMF meta-

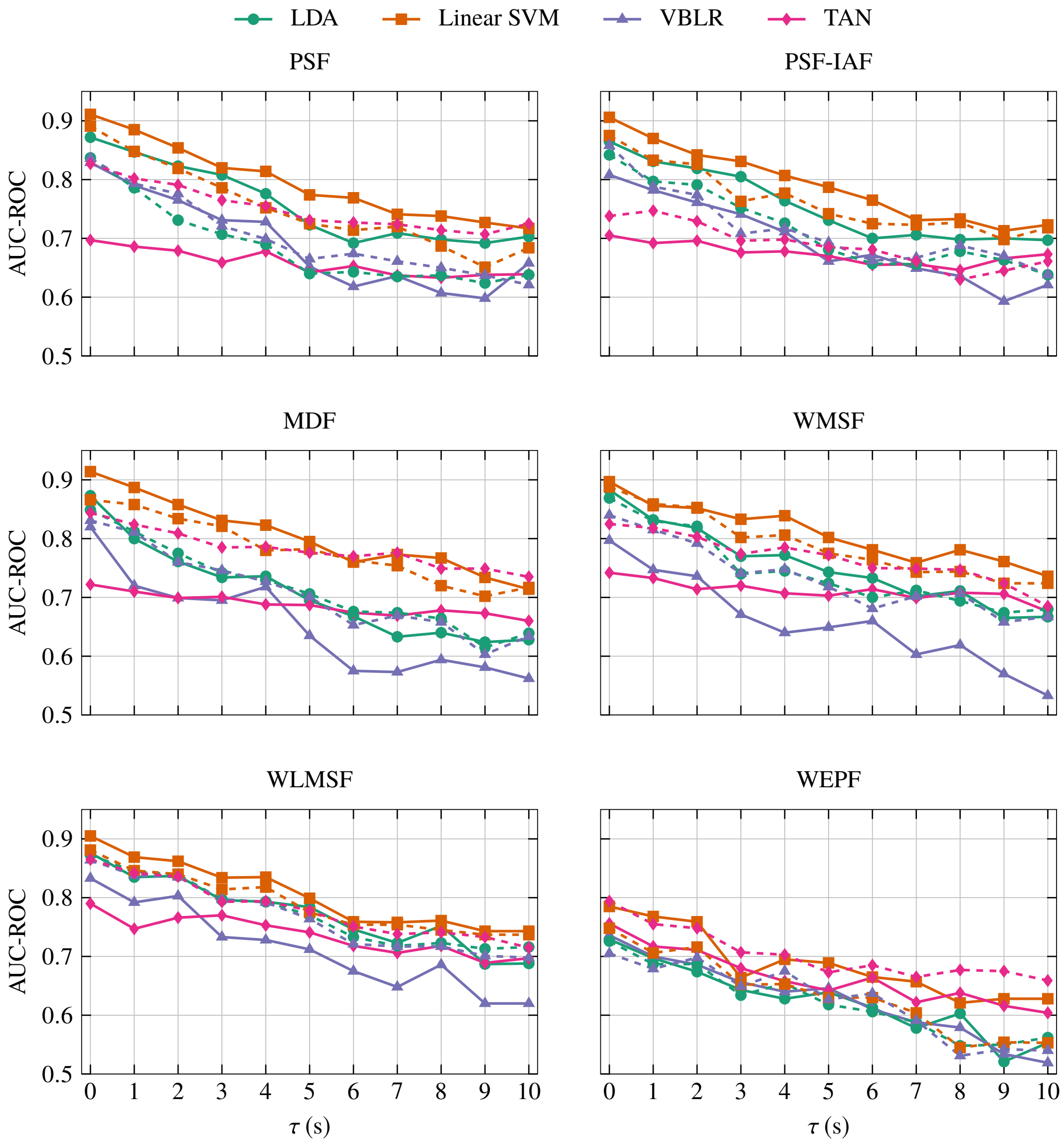


Figure 10.5 Performance of microsleep onset prediction in terms of AUC-ROC using different classifiers versus prediction time τ = 0–10 s. Solid lines correspond to the baseline features, whereas dashed lines correspond to the VBHMSRJMF meta-features.

features of WEPF. For detection, the highest phi of 0.09 (AUC-ROC = 0.89, GM = 0.79) was with the meta-features of WMSF and a linear SVM (cf. $\varphi = 0.09$, AUC-ROC = 0.87, GM = 0.69 for baseline with an LDA and MDF). Increasing the prediction time to $\tau = 10$ s, the highest phi was 0.04 (AUC-ROC = 0.74, GM = 0.66) with the meta-features of WLMSF and a linear SVM (cf. $\varphi = 0.04$, AUC-ROC = 0.74, GM = 0.54 for baseline with a linear SVM and WMSF).

Linear SVM had superior performances with the baseline features and relatively good performances with VBHMSRJMF meta-features. Therefore, we limit the rest of the results of this section to microlseep onset detection/prediction with linear SVM classifiers. The performance of microsleep onset prediction in terms of GM, sensitivity, and precision is shown in Figure 10.7.

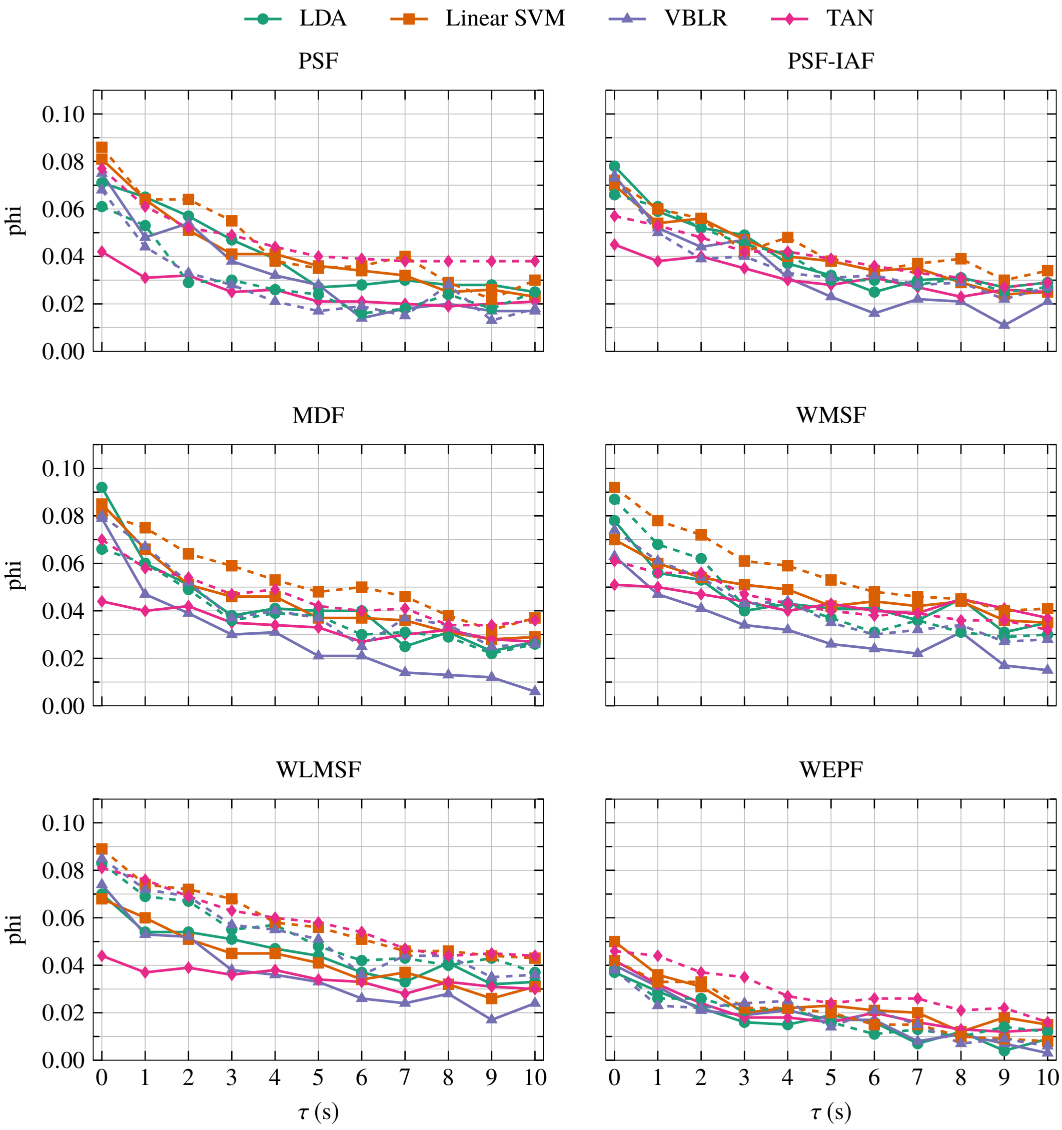


Figure 10.6 Performance of microsleep onset prediction in terms of phi using different classifiers versus prediction time τ = 0–10 s. Solid lines correspond to the baseline features, whereas dashed lines correspond to the VBHMSRJMF meta-features.

It was observed that, in most cases, using VBHMSRJMF meta-features improved GM compared to its respective baseline. On the contrary, no consistent improvement was found in sensitivity and precision. This might be due to the highly imbalanced nature of our dataset and highly variable imbalance ratios among subjects.

For detection of microsleep onsets, the highest GM was 0.79 (Sn = 0.75, Pr = 0.02, Sp = 0.85) with meta-features of WMSF (cf. GM = 0.78, Sn = 0.78, Pr = 0.02, and Sp = 0.80 for baseline with MDF). With the same setup, however, the lowest GM was 0.64 (Sn = 0.57, Pr = 0.01, Sp = 0.77) with meta-features of WEPF (cf. GM = 0.67, Sn = 0.75, Pr = 0.02, and Sp = 0.69 for baseline with PSF-IAF). Increasing the prediction time to τ = 10 s, the highest GM was 0.66 (Sn = 0.59, Pr = 0.01, Sp = 0.75) with meta-features of WLMSF (cf. GM = 0.55, Sn = 0.57, Pr = 0.00, and Sp = 0.71 for baseline with MDF). Similarly, the lowest GM was 0.49 (Sn = 0.38, Pr = 0.00, Sp = 0.68) with meta-features of WEPF (cf. GM = 0.45, Sn = 0.58, Pr = 0.00, and Sp = 0.62 for baseline with PSF).

Although the AUC-ROCs and AUC-PRs with VBHMSRJMF meta-features were not superior to the baselines, phi and GM measures were slightly better. This is a result of finding a lower-dimension representation of data with a lower inter-subject variability. However, the performances were not substantially better which might be due to the highly imbalanced nature of microsleep dataset. The highest precision among all feature sets and classifiers was 0.03 which was too low.

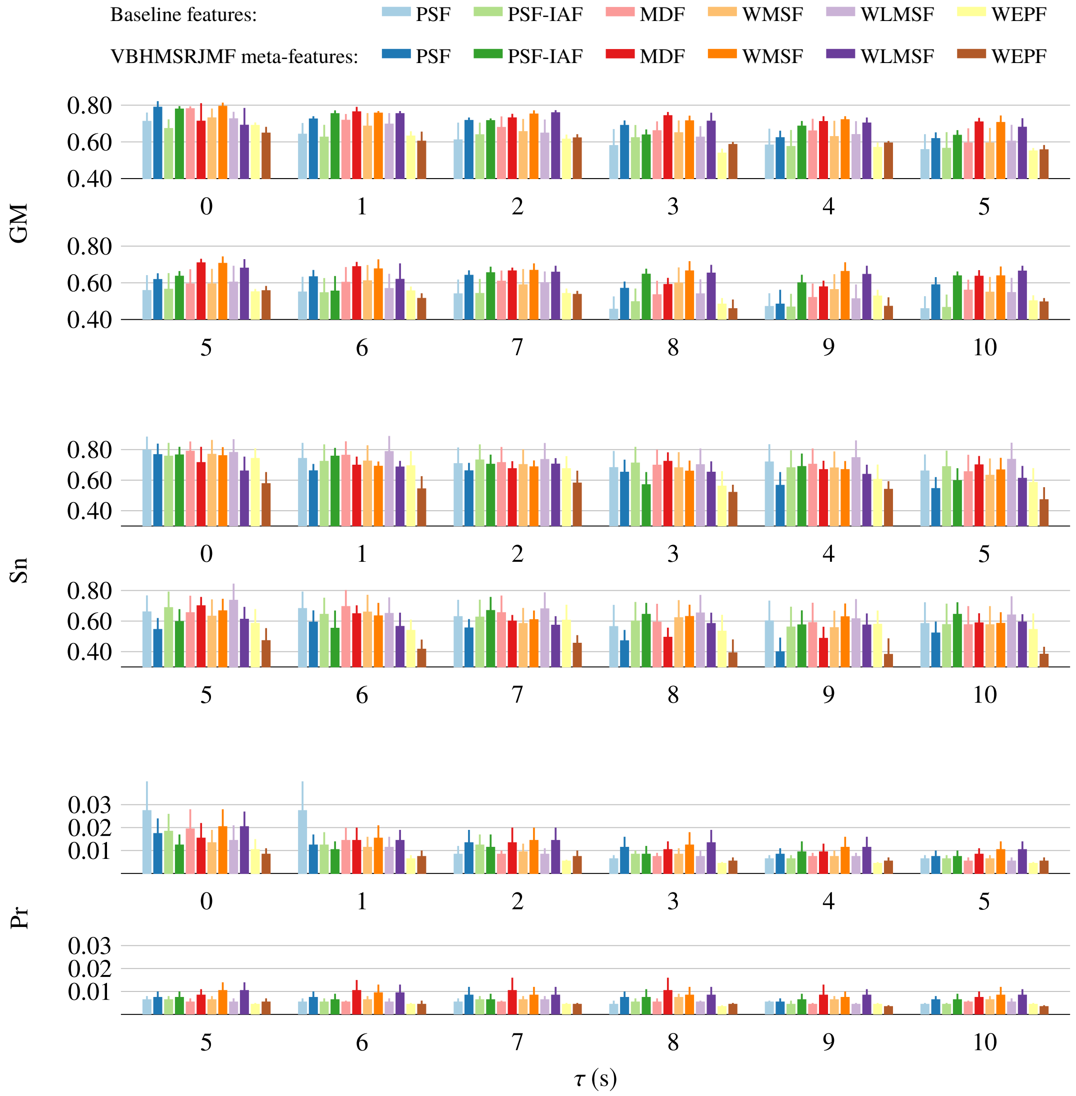


Figure 10.7 Performance (mean ± SE) of microsleep onset prediction in terms of GM, sensitivity, and precision using VBHMSRJMF meta-features and a linear SVM classifier for τ = 0–10 s.

10.5 SUMMARY

VBHMSRJMF, its underlying assumptions, and the results of detection and prediction of microsleeps were presented in this chapter. VBHMSRJMF is a Bayesian feature reduction method that jointly factorizes all classes of data and finds a less inter-subject variable robust latent space. Similar to VBHMSRFA, subjects share information through a group-level loading matrix, while each subject has its own slightly different loading matrix. Moreover, individual subjects have independent mean and noise terms. The latent space is shared among all subjects and optimized to maximize the log-likelihood of data. The optimum number of components is selected automatically using ARD-motivated prior distributions over the loading matrices. Variational formulation of VBHMSRJMF for both the training and testing phases were also derived.

For microsleep state detection and prediction, substantial improvements of GM and phi were achieved with VBHMSRJMF meta-features. However, performances in terms of AUC-ROC and AUC-PR did not improve and even slightly dropped. This indicates that reducing inter-subject variability could potentially lead to a better detection/prediction generalization for test subjects, where the classification threshold is selected based on the training subjects and is applied to an unseen test subject. The highest detection AUC-ROC of 0.94 was with an LDA and the VBHMSRJMF meta-features of PSF (cf. 0.95 for baseline with a linear SVM and PSF-IAF). The highest detection AUC-PR was 0.43 with meta-features of WMSF and WLMSF (cf. 0.49 for baseline with a linear SVM and PSF). In terms of phi, however, the highest value for detection was 0.43 with an LDA and meta-features of PSF (cf. 0.40 for baseline with a linear SVM and PSF).

Performances of microsleep state prediction with a prediction time of $\tau = 1.0\,\mathrm{s}$ dropped slightly. The highest AUC-ROC of 0.92 and AUC-PR of 0.40 were achieved with an LDA and meta-features of WLMSF (GM = 0.81; $\varphi = 0.39$; Sn = 0.73; Pr = 0.34), whereas the highest baselines were AUC-ROC of 0.92 and AUC-PR of 0.42 with a linear SVM and MDF (GM = 0.76; $\varphi = 0.35$; Sn = 0.70; Pr = 0.35). This shows that higher values of GM and phi were achieved with VBHMSRJMF meta-features, although the AUC-ROC and AUC-PR were relatively similar.

Results of microsleep onset detection and prediction showed a similar pattern: AUC-ROC and AUC-PR dropped slightly while GM and phi improved slightly. With $\tau = 0\,\mathrm{s}$, the highest AUC-ROC of 0.89 and AUC-PR of 0.05 were achieved with a linear SVM and VBHMSRJMF meta-features of WMSF (GM = 0.79; $\varphi = 0.09$; Sn = 0.75; Pr = 0.02; Sp = 0.85). The highest values of the baseline were AUC-ROC of 0.91 and AUC-PR of 0.09 with a linear SVM and PSF (GM = 0.71; $\varphi = 0.08$; Sn = 0.79; Pr = 0.03; Sp = 0.71).

Although these results provide evidence that microsleeps can be predicted, the overall performance is too low and thus not practical for real-life application.

Chapter 11

DISCUSSION

Proposed Bayesian methods for feature reduction have been described and evaluated, as presented in Chapters 7–10. Each Bayesian method was investigated for detection and prediction of microsleeps, both in terms of states and onsets, and performances were reported in the respective chapters. The results of each method were also compared to baseline approaches. The aim of this chapter is to compare results of the proposed methods and provide a discussion on the methods and results.

11.1 MICROSLEEP STATE DETECTION AND PREDICTION

11.1.1 Window length for feature extraction

For detection of microsleep states, features extracted from 2-s EEG windows performed inferior to those extracted from longer EEG windows, as shown in Table 11.1. These results suggest that although microsleeps are transient phenomena, a 2-s EEG window does not capture all of the information prior to a microsleep. On the other hand, longer EEG-windows ranging from 5 to 10 s have been used in the literature for drowsiness detection [Akin et al. 2008, Chen et al. 2015, Qian et al. 2017, Subasi et al. 2005, Yeo et al. 2009]. Since microsleeps are usually accompanied with drowsy behavioural cues such as slow eye-closure [Peiris et al. 2006, Poudel et al. 2014], using longer EEG windows might better identify microsleeps through their relationship with drowsiness. Notwithstanding, increasing the EEG-window length from 5 s to 10 s caused a slight drop in the performance of microsleep state detection in most cases, which is in line with the transient nature of microsleeps.

As shown in Table 11.1, aggregated features of multiple EEG windows outperformed those from a single EEG window. Using the aggregated features (baseline) and an LDA classifier for microsleep state detection led to average improvements of 2.6% (1.1–3.7%) and 10.8% (4.9–18.9%) relative to the best of single-window features for AUC-ROC and AUC-PR, respectively. These results highlight that exploiting both the tonic and phasic characteristics of EEG improves the accuracy of microsleep detection. A similar pattern was observed for the proposed Bayesian methods, in which aggregated meta-features of multiple EEG-windows performed similar or superior to those with single-window meta-features. However, performance improvements for

Table 11.1 Performance of microsleep state detection with an LDA classifier versus EEG window for feature extraction. A bold value indicates the highest performances of each feature set and italic indicates the highest of overall.

Feature set	Feature type	2-s EEG window		5-s EEG window		10-s EEG window		Aggregated windows	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.86	0.31	0.88	0.36	0.86	0.36	0.90	0.40
	VBRFA-3	0.85	0.28	0.89	0.31	0.89	0.33	0.92	0.38
	VBMSRFA	0.91	0.36	0.91	0.41	0.88	0.38	**0.93**	**0.44**
	VBHMSRFA	0.91	0.37	0.91	0.41	0.87	0.36	0.92	0.41
	VBHMSRJMF	0.91	0.37	0.89	0.36	0.84	0.33	0.90	0.39
PSF	Baseline	0.89	0.34	0.92	0.41	0.92	0.37	0.94	0.43
	VBRFA-3	0.88	0.33	0.92	0.41	0.92	0.39	0.94	0.46
	VBMSRFA	0.92	0.37	0.93	0.42	0.92	0.39	***0.95***	0.46
	VBHMSRFA	0.89	0.33	0.92	0.38	0.91	0.41	0.94	**0.48**
	VBHMSRJMF	0.89	0.33	0.90	0.37	0.87	0.33	0.94	0.42
PSF-IAF	Baseline	0.88	0.34	0.90	0.37	0.91	0.35	0.94	0.44
	VBRFA-3	0.88	0.34	0.91	0.38	0.92	0.36	***0.95***	**0.47**
	VBMSRFA	0.90	0.37	0.86	0.37	0.90	0.37	0.94	**0.47**
	VBHMSRFA	0.88	0.34	0.91	0.40	0.89	0.32	0.92	0.45
	VBHMSRJMF	0.84	0.28	0.84	0.26	0.79	0.28	0.86	0.34
WMSF	Baseline	0.87	0.35	0.91	0.38	0.91	0.37	0.92	0.40
	VBRFA-3	0.87	0.32	0.90	0.35	0.91	0.36	0.92	0.39
	VBMSRFA	0.88	0.36	0.92	0.40	0.90	0.37	0.92	0.41
	VBHMSRFA	0.90	0.38	**0.93**	**0.43**	0.91	0.38	**0.93**	**0.43**
	VBHMSRJMF	0.90	0.36	0.91	0.39	0.89	0.37	0.92	0.40
WLMSF	Baseline	0.88	0.36	0.91	0.41	0.91	0.37	0.94	0.44
	VBRFA-3	0.87	0.32	0.90	0.35	0.91	0.36	0.92	0.39
	VBMSRFA	0.93	0.42	0.94	0.45	0.91	0.41	***0.95***	***0.49***
	VBHMSRFA	0.89	0.36	0.91	0.42	0.90	0.40	0.92	0.45
	VBHMSRJMF	0.89	0.34	0.93	0.40	0.89	0.38	0.93	0.43
WEPF	Baseline	0.74	0.21	0.78	0.22	0.81	0.23	0.84	0.27
	VBRFA-3	0.87	0.32	0.90	0.35	0.91	0.36	**0.92**	**0.39**
	VBMSRFA	0.81	0.26	0.81	0.26	0.81	0.25	0.86	0.32
	VBHMSRFA	0.80	0.27	0.82	0.27	0.79	0.23	0.84	0.29
	VBHMSRJMF	0.80	0.26	0.79	0.25	0.79	0.23	0.83	0.29

our proposed Bayesian methods were smaller compared to the baseline. Thus, for the remainder of this chapter, we limit the discussion to the performances with aggregated-features of multiple EEG-windows, i.e., 2, 5, and 10 s, and will use the term *features* to refer to the aggregated features.

11.1.2 Classification methods

As shown in Table 11.2, the linear SVM outperformed other classifiers for microsleep state detection with baseline features in terms of AUC-ROC and AUC-PR in most of the cases. The highest AUC-ROC for microsleep state detection with baseline features was 0.95, achieved with PSF-IAF, and the highest AUC-PR of 0.49 was achieved with PSF. In terms of classifiers, performances with LDA was slightly lower than those of linear SVM. These small differences were likely to happen because LDA uses the whole training data to estimate its parameters, whereas linear SVM uses a few support vectors (data points) of the training data [Murphy 2012]. That is, there are likely to have been outliers in the training data that affected LDA parameter estimation, while these not being selected as support vectors of the linear SVM.

Table 11.2 Performance of different classifiers for microsleep state detection. A bold value indicates the highest performance among selected classifiers and an italic value indicates the highest overall.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.90	0.40	0.93	**0.48**	0.90	0.42	0.77	0.22
	VBRFA-3	0.92	0.38	**0.94**	0.46	0.91	0.40	0.77	0.22
	VBMSRFA	0.93	0.44	**0.94**	0.46	0.92	0.42	0.72	0.20
	VBHMSRFA	0.92	0.41	0.93	0.43	0.93	0.40	0.83	0.27
	VBHMSRJMF	0.90	0.39	0.88	0.40	0.89	0.39	0.85	0.28
PSF	Baseline	0.94	0.43	0.94	***0.49***	0.94	0.44	0.70	0.21
	VBRFA-3	0.94	0.46	***0.95***	0.48	0.94	0.47	0.81	0.30
	VBMSRFA	***0.95***	0.46	0.93	0.40	0.94	0.41	0.67	0.17
	VBHMSRFA	0.94	0.48	0.94	0.46	0.94	0.47	0.87	0.30
	VBHMSRJMF	0.94	0.42	0.93	0.39	0.92	0.39	0.86	0.34
PSF-IAF	Baseline	0.94	0.44	***0.95***	**0.47**	0.93	0.45	0.73	0.26
	VBRFA-3	***0.95***	**0.47**	***0.95***	0.46	0.94	0.44	0.80	0.29
	VBMSRFA	0.94	**0.47**	0.92	0.42	0.92	0.40	0.62	0.21
	VBHMSRFA	0.92	0.45	0.92	0.43	0.92	0.44	0.86	0.29
	VBHMSRJMF	0.86	0.34	0.91	0.37	0.88	0.34	0.73	0.27
WMSF	Baseline	0.92	0.40	**0.93**	0.42	0.92	0.39	0.77	0.22
	VBRFA-3	0.92	0.39	**0.93**	0.43	0.91	0.39	0.77	0.23
	VBMSRFA	0.92	0.41	**0.93**	0.43	0.92	0.43	0.74	0.20
	VBHMSRFA	**0.93**	0.43	**0.93**	**0.45**	**0.93**	0.44	0.82	0.24
	VBHMSRJMF	0.92	0.40	**0.93**	0.43	0.92	0.41	0.82	0.26
WLMSF	Baseline	0.94	0.44	0.93	0.45	0.92	0.45	0.76	0.25
	VBRFA-3	0.92	0.39	0.93	0.43	0.91	0.39	0.77	0.23
	VBMSRFA	***0.95***	***0.49***	***0.95***	***0.49***	0.94	0.48	0.79	0.29
	VBHMSRFA	0.92	0.45	0.92	0.46	0.92	0.45	0.86	0.28
	VBHMSRJMF	0.93	0.43	0.92	0.43	0.92	0.43	0.90	0.40
WEPF	Baseline	0.84	0.27	0.86	0.29	0.85	0.28	0.82	0.23
	VBRFA-3	0.92	0.39	**0.93**	**0.43**	0.91	0.39	0.77	0.23
	VBMSRFA	0.86	0.32	0.88	0.33	0.88	0.34	0.68	0.20
	VBHMSRFA	0.84	0.29	0.86	0.30	0.85	0.30	0.84	0.29
	VBHMSRJMF	0.83	0.29	0.85	0.30	0.84	0.29	0.86	0.30

In contrast, the lowest baseline performances were achieved with a TAN classifier which were substantially lower than the other classifiers. With a TAN classifier for detection of microsleep states, the highest AUC-ROC and AUC-PR were 0.90 and 0.40, respectively, achieved with VBHMSRJMF meta-features of WLMSF (cf. AUC-ROC = 0.95 and AUC-PR = 0.49 for LDA and linear SVM with VBMSRFA meta-features of WLMSF). A reason for such poor performance compared to other classifiers could be the restriction of TAN on the classifier's structure. As mentioned in Section 6.3.4, the TAN classifier restricts the conditional dependency of a variable to only one other variable, apart from the class node. Thus, it fits a generative model to the training data based on an approximated dependence structure, which might not accurately represent the data. In addition, although TAN uses the Bayes rule to find the posterior probability of each class-label given data, it does not incorporate a full Bayesian treatment to find parameters of the generative model. It uses MLE of the training data to find the model parameters and does not perform sparse learning and hence is prone to overfitting [Pernkopf and Bilmes 2010].

Although performance of the VBLR classifier was, in most cases, similar to those of LDA and linear SVM, it had slightly inferior performances and, contrary to expectation, did not outperform the other two linear classifiers. This is likely to be due to two reasons. First, VBLR uses a local approximation of the sigmoid function which leads to a quadratic form for the posterior probability [Bishop 2006, Drugowitsch 2013]. Such approximation is required since using the sigmoid function does not lead to a conjugate distribution [Bishop 2006, Drugowitsch 2013]. Second, VBLR uses a variational inference to find posterior approximations, which is prone to find a local optimum. With these approximations, it is possible to find a non-optimum solution.

11.1.3 Feature sets

Of all the feature sets, WEPF led to the lowest performances for microsleep detection, as shown in Table 11.2. With the baseline WEPF, the highest achieved AUC-ROC and AUC-PR for microsleep state detection were 0.86 and 0.29, respectively, which were achieved with a linear SVM (cf. AUC-ROC of 0.95 with PSF-IAF and AUC-PR of 0.49 with PSF). Of our proposed Bayesian feature reduction methods, applying VBRFA-3 to WEPF with a linear SVM classifier improved the detection performances to AUC-ROC = 0.93 and AUC-ROC = 0.43 but WEPF still performed inferior to the other feature sets. This is in line with the findings of Peiris et al. [2011], in which performance of microsleep detection using normalized spectral power features was lower than non-normalized spectral power features. A reason for lower performances of WEPF can be the fact that WEPF represents the distribution of signal energy among different frequency bands, hence it does not have any information regarding the absolute energy. The other feature sets, on the other hand, contain information with respect to the absolute energy or absolute power of EEG.

The original PSF and PSF-IAF achieved similar AUC-ROC and AUC-PR for microsleep states detection. The highest baseline phi of 0.40 and GM of 0.79 were achieved with a linear SVM and PSF (AUC-ROC = 0.94; AUC-PR = 0.49, Sn = 0.74; Pr = 0.38). However, although a linear SVM and PSF-IAF had similar performances in terms of AUC-ROC and AUC-PR (AUC-ROC = 0.95; AUC-PR = 0.47), phi and GM with PSF-IAF was lower (φ = 0.36; GM = 0.76). The decline in phi and GM for PSF-IAF is potentially attributed to the calculation of IAF. In the current study, as mentioned in Section 6.2.2, an 8-s EEG-segment was used to approximate IAF but is likely to be inaccurate, especially when an individual is performing a visuomotor task with eyes open leading to alpha-blockage [Freeman and Quiroga 2013].

Performance with baseline WLMSF for microsleep state detection was inferior to the baseline PSF. The highest detection performances with WLMSF were AUC-ROC = 0.94, AUC-PR = 0.44, GM = 0.76, φ = 0.36, Sn = 0.70, and Pr = 0.31 and were achieved with an LDA classifier. Notwithstanding, these results are comparable with those of baseline PSF-IAF, while WLMSF has 80 features (cf. 192 for baseline PSF and PSF-IAF). Both baseline PSF and baseline PSF-IAF used sub-bands of standard EEG frequency bands (e.g., alpha1 and alpha2), whereas wavelet features were only extracted from the wavelet decompositions corresponding to the standard EEG frequency bands. These results suggest that splitting frequency-bands into sub-bands has the potential to improve performance of microsleep state detection (and prediction). Therefore, a future study can investigate the effect of sub-bands using wavelet packets [Stéphane 2009].

Comparing the performance of microsleep detection with the original WLMSF and WMSF confirmed that transforming data to logarithm-scale slightly improves the performance. The highest detection performance for baseline with WMSF were AUC-ROC = 0.93, AUC-PR = 0.42, GM = 0.75, φ = 0.34, Sn = 0.71, and Pr = 0.28, achieved with a linear SVM classifier (cf. AUC-ROC = 0.94, AUC-PR = 0.44, GM = 0.76, φ = 0.36, Sn = 0.70, and Pr = 0.31 for baseline WLMSF and an LDA). Although the improvement of performances were marginal, it shows that a preprocessing step to map data into a normally-distributed space has the potential to achieve higher performances with linear classifiers.

11.1.4 Proposed Bayesian feature-reduction methods

All the proposed Bayesian models of this project were implemented in Matlab R2016a[1]. In addition, Boost[2] and Armadillo[3] C++ libraries were integrated with Matlab's MEX interface. A personal computer with 48 GB of RAM, an Intel Xeon X5670 Processor[4], and Ubuntu operating system[5] was used to evaluate runtime of the proposed algorithm for both training and testing. Table 11.3 shows the runtime of our proposed methods to extracted PSF from 5-s EEG windows. It is evident that multi-subject algorithms take much longer for training and testing comparing to VBRFA. Additionally, VBMSRFA showed the slowest testing time per iteration. This can be a direct result of retaining larger number of meta-features. Notwithstanding, all of these algorithms are capable of running in real-time with a temporal resolution of 0.25 s.

Table 11.3 Runtime of the proposed Bayesian algorithms applied to extracted PSF from 5-s EEG windows.

Method	Average training time (min)	Average test time (ms/iter)	Average number of meta-features
VBRFA-1	1.2 (0.9–1.8)	0.26 (0.17–0.47)	19.9
VBRFA-2	4.0 (3.8–4.1)	1.85 (1.65–1.94)	72.0
VBRFA-3	5.2 (4.8–5.9)	2.12 (2.10–2.13)	91.9
VBMSRFA	34.2 (30.6–36.3)	114.25 (93.78–129.51)	154.4
VBHMSRFA	45.7 (38.7–47.9)	26.30 (12.39–30.09)	53.0
VBHMSRJMF	46.7 (34.6–52.3)	48.50 (33.16–53.24)	124.1

For detection of microsleep states, VBMSRFA achieved the highest AUC-ROC of 0.95 and AUC-PR of 0.49 of all the proposed methods. Notwithstanding, the highest AUC-ROC values of all of the proposed methods were 0.94–0.95. Moreover, the highest values of AUC-PR

[1] https://mathworks.com/products/matlab.html

[2] http://www.boost.org/

[3] http://arma.sourceforge.net

[4] https://ark.intel.com/products/47920/Intel-Xeon-Processor-X5670

[5] https://www.ubuntu.com

of three of the proposed methods, i.e., VBMSRFA, VBRFA, and VBHMSRFA, were similar 0.48–0.49, whereas the highest AUC-PR with VBHMSRJMF was 0.43. A key concept of VBHMSRJMF is to make use of class labels by incorporating this information in finding a less subject-variable latent space. However, the number of microsleep instances in our dataset is limited and thus there is a large uncertainty on the data of microsleeps. In addition, VBHMSRJMF uses variational inference which is prone to find a local optimum and hence is sensitive to the initialization of the model parameters [Bishop 2006]. Therefore, using PCA to initialize parameters of VBHMSRJMF can result in poor representation of data. Changing the initialization of VBHMSRJMF is likely to improve its performance.

It is evident that in terms of highest AUC-ROC and AUC-PR, our proposed methods did not improve the performance of microsleep state detection, as shown in Table 11.2. However, finding a less subject-variable latent space improved performance in terms of phi and GM, as shown in Table 11.4 for microsleep detection with an LDA classifier. The highest GM for microsleep state detection was 0.84, achieved with VBMSRFA meta-features of WLMSF and a linear SVM classifier (φ = 0.45; Sn = 0.76; Sp = 0.94; Pr = 0.36), whereas the highest GM achieved with original features was 0.79 with PSF (φ = 0.40; Sn = 0.74; Sp = 0.89; Pr = 0.38). In addition, the former had 237 meta-features whereas the latter had 576 features. Interestingly, the highest value of phi was also achieved with VBMSRFA meta-features of WLMSF, but with an LDA classifier (φ = 0.47; GM = 0.83; Sn = 0.74; Sp = 0.95; Pr = 0.38). The highest value of phi with baseline features was 0.40 with PSF.

Phi and GM are calculated after a classification threshold has been applied, as opposed to AUC-ROC and AUC-PR which are threshold-free performance measures. Improvements of phi and GM without any change in AUC-ROC and AUC-PR indicates that our proposed multi-subject methods reduced the variability of classification threshold between the training and test subjects. If the classification threshold is highly variable between training and test subjects, AUC-ROC and AUC-PR do not demonstrate it, since these measures integrate over all possible values of the classification threshold. However, phi, GM, and F-measure are calculated based on the contingency table, which is computed after fixing the classification threshold. Our proposed multi-subject Bayesian models were able to reduce the variability of the classification threshold between training and test subjects and hence achieve higher performances in terms of phi and GM.

As shown in Table 7.1, applying VBRFA to concatenated data of multiple subjects, i.e., VBRFA-1 meta-features, deteriorated performance in most cases. Such decline in performance is potentially attributed to inter-subject variability in the EEG data. A VBRFA model assumes that the distribution of data is consistent but this assumption does not necessarily hold for the concatenated data of multiple subjects if there is substantial inter-subject variability. The VBRFA-2 meta-features solved this issue by applying VBRFA to the data of individual subject independently and combining all of the models into one. Although the results were comparable to the baseline, no improvements were achieved. A shortcoming of applying multiple VBRFA models to individual subjects' data is that if there are shared components between subjects, they

Table 11.4 Performance of microsleep state detection with an LDA classifier and aggregated features of multiple EEG windows. A bold value indicates the highest performances in individual feature sets, while italic indicates the highest overall.

Feature set	Feature type	Microsleep detection performance					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	0.90	0.40	**0.71**	0.34	**0.62**	0.33
	VBRFA-3	0.92	0.38	0.69	0.32	0.59	0.34
	VBMSRFA	**0.93**	**0.44**	0.66	0.36	0.51	***0.42***
	VBHMSRFA	0.92	0.41	0.68	**0.37**	0.55	0.39
	VBHMSRJMF	0.90	0.39	0.70	0.36	0.59	0.33
PSF	Baseline	0.94	0.43	0.74	0.36	0.68	0.36
	VBRFA-3	0.94	0.46	0.77	0.38	**0.71**	0.35
	VBMSRFA	***0.95***	0.46	**0.81**	**0.46**	0.70	**0.41**
	VBHMSRFA	0.94	**0.48**	**0.81**	0.45	**0.71**	0.37
	VBHMSRJMF	0.94	0.42	0.77	0.43	0.63	0.40
PSF-IAF	Baseline	0.94	0.44	0.76	0.37	0.70	0.36
	VBRFA-3	***0.95***	**0.47**	**0.80**	0.39	***0.74***	0.34
	VBMSRFA	0.94	**0.47**	0.76	**0.43**	0.61	**0.41**
	VBHMSRFA	0.92	0.45	0.76	0.42	0.62	0.39
	VBHMSRJMF	0.86	0.34	0.67	0.31	0.53	0.30
WMSF	Baseline	0.92	0.40	0.66	0.33	0.57	0.31
	VBRFA-3	0.92	0.39	0.67	0.33	0.60	0.30
	VBMSRFA	0.92	0.41	0.70	0.36	0.59	0.35
	VBHMSRFA	**0.93**	**0.43**	0.73	**0.41**	0.56	**0.40**
	VBHMSRJMF	0.92	0.40	**0.74**	0.40	**0.61**	0.39
WLMSF	Baseline	0.94	0.44	0.76	0.36	0.70	0.31
	VBRFA-3	0.92	0.39	0.67	0.33	0.60	0.30
	VBMSRFA	***0.95***	***0.49***	***0.83***	***0.47***	***0.74***	**0.38**
	VBHMSRFA	0.92	0.45	0.80	0.42	0.72	0.34
	VBHMSRJMF	0.93	0.43	0.81	0.41	0.73	0.35
WEPF	Baseline	0.84	0.27	**0.70**	0.25	**0.67**	0.25
	VBRFA-3	**0.92**	**0.39**	0.67	**0.33**	0.60	**0.30**
	VBMSRFA	0.86	0.32	0.63	0.26	0.47	0.29
	VBHMSRFA	0.84	0.29	0.66	0.23	0.59	0.24
	VBHMSRJMF	0.83	0.29	0.65	0.24	0.55	0.25

will create redundancy in the VBRFA-2 meta-features. As a result, performance improvement may not be achieved. Moreover, using multiple VBRFA models for individual training subjects finds a rather less subject-variable model for the training data, but it does not adapt to a new test subject. Hence, if the classification-threshold of a test subject is different from the one found from training subjects, the test performance will not be maximal.

The VBRFA-3 meta-features combined both the VBRFA-1 and VBRFA-2 meta-features and showed similar results to those of the baseline. For microsleep state detection, the highest AUC-ROC with VBRFA-3 meta-features was 0.95 with PSF and PSF-IAF (cf. 0.95 for baseline with PSF-IAF) and the highest AUC-PR was 0.48 with PSF (cf. 0.49 for baseline with PSF). Although these performances are very similar, VBRFA-3 used a substantially lower number of meta-features. The average number of VBRFA-3 meta-features, shown in Table 7.2, of PSF and PSF-IAF were 272.8 and 281.8, respectively (cf. 576 for the baseline of both feature sets). In terms of phi and GM values, the highest values with VBRFA-3 meta-features were GM = 0.80 and $\varphi = 0.39$ and were achieved with an LDA classifier and PSF-IAF (cf. GM = 0.79 and $\varphi = 0.40$ for baseline with a linear SVM and PSF). Since VBRFA-3 combines the VBRFA-1 and VBRFA-2 meta-features, it suffers from the shortcomings of both meta-features and hence it was expected that its results would not be superior to those of the baseline. Nevertheless, although our proposed VBRFA method, presented in Chapter 7, did not improve the performance of microsleep detection, it was able to find a compact representation of the data with performance similar to the baseline.

The VBMSRFA method, described in Chapter 8, was proposed to take inter-subject variability into account. The VBMSRFA meta-features of WLMSF led to the highest AUC-ROC and AUC-PR of 0.95 and 0.49, respectively, for microsleep state detection which were similar to those of the baseline with PSF. However, the number of VBMSRFA meta-features were lower, i.e., 237 for the VBMSRFA meta-features of WLMSF compared to 576 for baseline PSF. Although using VBMSRFA did not improve the performance in terms of the highest AUC-ROC and AUC-PR, it substantially improved the performance in terms of phi and GM. The highest phi and GM values with VBMSRFA were 0.47 (with an LDA) and 0.84 (with a linear SVM) and were achieved with WLMSF (cf. GM = 0.79 and $\varphi = 0.40$ for baseline with a linear SVM and PSF). As shown in Table 8.5 and Figure 8.3, the improvements of phi and GM values were consistent across prediction times, especially for PSF, PSF-IAF, and WLMSF. With a prediction time of $\tau = 1$ s, the highest performances of VBMSRFA in terms of AUC-ROC, AUC-PR, and GM were 0.94, 0.45, and 0.81, respectively, and these results were achieved with a linear SVM and WLMSF (cf. AUC-ROC = 0.92, AUC-PR = 0.42, and GM = 0.76 for baseline with an LDA and MDF). The highest performance in terms of phi, however, was 0.44 with the same feature set but an LDA classifier (cf. 0.35 for baseline).

Chapter 9 described the VBHMSRFA to further extend our proposed VBMSRFA method to incorporate inter-subject variability into the loading matrix, in addition to the mean and noise terms. Using VBHMSRFA meta-features for microsleep detection led to an AUC-ROC and AUC-PR of 0.94 and 0.48, respectively, with an LDA and PSF. These results were slightly lower than those of baseline (AUC-ROC = 0.95 with PSF-IAF; AUC-PR = 0.49 with PSF). However, the performance of VBHMSRFA in terms of phi and GM was superior to the baseline. The highest GM of VBHMSRFA was 0.83 with a linear SVM and PSF and the highest phi was 0.45 with the same meta-features and an LDA. Although these performances were superior to those of baseline, they were slightly lower than those of VBMSRFA, i.e., $\varphi = 0.47$ and GM = 0.84. Such small differences are trivial, especially with an imbalanced dataset such as microsleeps. Notwithstanding, a reason for those small differences is likely to be due to usage of MAP of the latent variables as meta-features. The MAP of a probability distribution represents the most likely value and removes the surrounding uncertainty which is prone to eliminate some information and hence is likely to result in a lower performance. On the other hand, VBHMSRFA found a more compact representation of data (shown in Table 9.2) compared to the VBMSRFA (shown in Table 8.2), albeit with a slight performance drop. With a prediction time of $\tau = 1$ s, the highest achieved phi and GM with VBHMSRFA were 0.41 (with an LDA and PSF) and 0.80 (with a VBLR and PSF), respectively (cf. GM = 0.81 and $\varphi = 0.44$ for VBMSRFA with WLMSF).

It is evident that VBMSRFA outperformed the other methods in terms of phi and GM. Notwithstanding, in terms of GM and phi, three of our proposed methods, i.e., VBMSRFA, VBHMSRFA, and VBHMSRJMF, had higher performances compared to those of the baseline. Surprisingly, the performances of VBHMSRJMF in terms of AUC-ROC and AUC-PR were inferior to the baselines, but outperformed the baselines with respect to GM and phi. These results indicate that incorporating inter-subject variability while finding a lower-dimension latent-space and adapting to a new subject's data can improve the performance of microsleep state detection and prediction in terms of GM and phi by reducing the variability of classification threshold between training and test subjects, even if the threshold-free measures, i.e., AUC-ROC and AUC-PR, have declined.

11.2 MICROSLEEP ONSET DETECTION AND PREDICTION

Prediction of the microsleep state does not guarantee prediction of a microsleep *event* before its occurrence. In contrast, it aims to predict the state of an individual's responsiveness. This is equivalent to predicting an imminent microsleep and if it was missed, keep trying to detect it as soon as possible. On the other hand, prediction of microsleep onset is only concerned about prediction of an imminent microsleep event. If the algorithm fails to predict a microsleep onset, then it has failed to predict the complete event of microsleep. The rationale for microsleep onset prediction is to ensure prediction of microsleep events. The downside is that the number of microsleep events are considerably lower than responsives. As a result, the microsleep onset dataset is severely imbalanced. Based on the data presented in Table 5.3, the average imbalance ratio of microsleep onsets to responsive labels (with a 4-Hz temporal resolution) is 0.002 (0.0002–0.006). Therefore, the results of onset prediction are not comparable to those of the states prediction.

As shown in Table 11.5, using baseline features the highest microsleep onset detection performances in terms of AUC-ROC and AUC-PR were achieved with a linear SVM. The highest AUC-ROC and AUC-PR were 0.91 (with MDF, PSF, and PSF-IAF) and 0.09 (with PSF), respectively ($\varphi = 0.08$; GM = 0.71; Sn = 0.79; Pr = 0.03). It is evident that performance in terms of AUC-ROC was relatively good but the AUC-PR was very low. These low values of AUC-PR are a result of highly imbalanced dataset [Saito and Rehmsmeier 2015]. Furthermore, with the same settings as shown in Table 11.6, the highest value of GM and phi were 0.78 (with a linear SVM and MDF) and 0.09 (with an LDA and MDF), respectively.

Table 11.5 Performance of different classifiers for microsleep onset detection. A bold value indicates the highest performance among selected classifiers and an italic value indicates the highest overall.

Feature set	Feature type	LDA		Linear SVM		VBLR		TAN	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
MDF	Baseline	0.87	0.03	***0.91***	**0.06**	0.82	0.05	0.72	0.01
	VBRFA-3	0.86	0.02	0.86	**0.06**	0.83	0.03	0.76	0.01
	VBMSRFA	0.90	0.03	0.82	0.03	0.86	0.03	0.72	0.01
	VBHMSRFA	0.90	0.03	0.82	0.03	0.86	0.03	0.72	0.01
	VBHMSRJMF	0.85	0.03	0.87	0.04	0.83	0.03	0.84	0.02
PSF	Baseline	0.87	0.04	***0.91***	***0.09***	0.83	0.04	0.70	0.01
	VBRFA-3	0.89	0.04	0.90	0.06	0.85	0.03	0.84	0.02
	VBMSRFA	0.89	0.06	0.89	0.06	0.84	0.03	0.75	0.01
	VBHMSRFA	0.89	0.06	0.89	0.06	0.84	0.03	0.75	0.01
	VBHMSRJMF	0.84	0.03	0.89	0.04	0.84	0.03	0.83	0.02
PSF-IAF	Baseline	0.87	0.04	***0.91***	**0.07**	0.81	0.04	0.70	0.01
	VBRFA-3	0.88	0.05	0.90	0.06	0.85	0.03	0.83	0.02
	VBMSRFA	0.89	0.06	0.85	0.05	0.85	0.04	0.72	0.01
	VBHMSRFA	0.89	0.06	0.85	0.05	0.85	0.04	0.72	0.01
	VBHMSRJMF	0.84	0.04	0.88	0.03	0.86	0.03	0.74	0.02
WMSF	Baseline	0.88	0.03	**0.90**	0.05	0.80	0.02	0.74	0.01
	VBRFA-3	0.89	0.04	0.89	**0.06**	0.85	0.04	0.74	0.01
	VBMSRFA	0.89	0.04	0.88	0.03	0.83	0.02	0.69	0.01
	VBHMSRFA	0.89	0.04	0.88	0.03	0.83	0.02	0.69	0.01
	VBHMSRJMF	0.87	0.02	0.89	0.05	0.84	0.03	0.83	0.02
WLMSF	Baseline	0.88	0.05	0.90	**0.06**	0.83	0.04	0.79	0.02
	VBRFA-3	0.89	0.04	0.89	**0.06**	0.85	0.04	0.74	0.01
	VBMSRFA	0.90	0.05	***0.91***	0.05	0.85	0.04	0.78	0.02
	VBHMSRFA	0.90	0.05	***0.91***	0.05	0.85	0.04	0.78	0.02
	VBHMSRJMF	0.87	0.03	0.88	0.05	0.86	0.04	0.86	0.04
WEPF	Baseline	0.73	0.01	0.78	0.02	0.74	0.01	0.76	0.01
	VBRFA-3	**0.89**	0.04	**0.89**	**0.06**	0.85	0.04	0.74	0.01
	VBMSRFA	0.73	0.01	0.75	0.01	0.75	0.01	0.70	0.01
	VBHMSRFA	0.73	0.01	0.75	0.01	0.75	0.01	0.70	0.01
	VBHMSRJMF	0.73	0.01	0.75	0.01	0.70	0.01	0.79	0.01

A value of 0.5 for AUC-ROC represents a random classifier, irrespective of the class imbalance-ratio [He and Garcia 2009, Saito and Rehmsmeier 2015]. Whereas a value of AUC-PR representing a random classifier is not fixed and depends on the class imbalance ratio. Saito and Rehmsmeier [2015] suggested that the random-classifier threshold for AUC-PR is $^{P}/_{(P+N)}$, where P is the number of positive class instances, i.e., microsleep onsets, and N is the number of negative class instances, i.e., responsive labels. Since the number of microsleep onsets are very much smaller than the responsive labels, a random-classifier threshold for AUC-PR is almost the same as class imbalance ratio, which is 0.002 (0.0002–0.006) for our microsleep onset dataset. This shows that although the AUC-PR of microsleep onset prediction is low, it is substantially higher than a random classifier. A similar difference exists between GM and phi performance measures. The GM measures a point on the ROC curve and hence is not sensitive to imbalance ratio. However, the computation of phi measure involves the number of false positives, which can easily outnumber true positives in highly imbalanced data (e.g., microsleep onsets). As a result, the phi measure is highly sensitive to the class imbalance ratio.

Table 11.6 Performance of microsleep onset detection with an LDA classifier and aggregated features of multiple EEG windows. A bold value indicates the highest performances in individual feature sets, while italic indicates the highest overall.

Feature set	Feature type	Microsleep detection performance					
		AUC-ROC	AUC-PR	GM	phi	Sn	Pr
MDF	Baseline	0.87	**0.03**	**0.69**	***0.09***	**0.55**	0.02
	VBRFA-3	0.86	0.02	0.64	0.07	0.49	0.02
	VBMSRFA	***0.90***	**0.03**	0.68	***0.09***	0.54	***0.03***
	VBHMSRFA	***0.90***	**0.03**	0.68	***0.09***	0.54	***0.03***
	VBHMSRJMF	0.85	**0.03**	0.67	0.07	0.54	0.01
PSF	Baseline	0.87	0.04	0.71	0.07	0.63	**0.02**
	VBRFA-3	**0.89**	0.04	0.71	0.07	0.66	0.01
	VBMSRFA	**0.89**	***0.06***	**0.77**	***0.09***	***0.70***	**0.02**
	VBHMSRFA	**0.89**	***0.06***	**0.77**	***0.09***	***0.70***	**0.02**
	VBHMSRJMF	0.84	0.03	0.60	0.06	0.43	**0.02**
PSF-IAF	Baseline	0.87	0.04	0.73	0.08	0.66	**0.02**
	VBRFA-3	0.88	0.05	0.73	0.07	**0.68**	0.01
	VBMSRFA	**0.89**	***0.06***	**0.77**	***0.09***	**0.68**	**0.02**
	VBHMSRFA	**0.89**	***0.06***	**0.77**	***0.09***	**0.68**	**0.02**
	VBHMSRJMF	0.84	0.04	0.68	0.07	0.58	**0.02**
WMSF	Baseline	0.88	0.03	0.69	0.08	0.60	0.02
	VBRFA-3	**0.89**	**0.04**	0.63	0.07	0.55	0.02
	VBMSRFA	**0.89**	**0.04**	**0.72**	***0.09***	**0.62**	0.02
	VBHMSRFA	**0.89**	**0.04**	**0.72**	***0.09***	**0.62**	0.02
	VBHMSRJMF	0.87	0.02	0.70	***0.09***	0.59	***0.03***
WLMSF	Baseline	0.88	**0.05**	0.73	0.07	***0.70***	0.01
	VBRFA-3	0.89	0.04	0.63	0.07	0.55	**0.02**
	VBMSRFA	***0.90***	**0.05**	***0.78***	***0.09***	0.69	**0.02**
	VBHMSRFA	***0.90***	**0.05**	***0.78***	***0.09***	0.69	**0.02**
	VBHMSRJMF	0.87	0.03	0.74	0.08	0.66	**0.02**
WEPF	Baseline	0.73	0.01	**0.63**	0.04	**0.60**	0.01
	VBRFA-3	**0.89**	**0.04**	**0.63**	**0.07**	0.55	**0.02**
	VBMSRFA	0.73	0.01	0.59	0.03	0.49	0.01
	VBHMSRFA	0.73	0.01	0.59	0.03	0.49	0.01
	VBHMSRJMF	0.73	0.01	0.59	0.04	0.53	0.01

Three of our proposed multi-subject methods, i.e., VBMSRFA, VBHMSRFA, and VBHMSRJMF, outperformed baseline in terms of GM, especially with longer prediction times. The highest GM for onset detection was 0.81 which was achieved with a linear SVM and VBMSRFA meta-features of WLMSF (φ = 0.09, Sn = 0.77, Sp = 0.85, Pr = 0.02). Similarly, the highest value of GM with VBHMSRFA meta-features was 0.79 with an LDA and WLMSF (φ = 0.09, Sn = 0.71, Sp = 0.87, Pr = 0.02). Baseline MDF and a linear SVM classifier had a marginally lower GM of 0.78 (φ = 0.08, Sn = 0.78, Sp = 0.80, Pr = 0.02).

With a prediction time of τ = 1 s, a linear SVM and VBHMSRFA meta-features of MDF achieved the highest phi and GM for microsleep onset prediction (Sn = 0.81; Pr = 0.01; Sp = 0.80). Similarly, the highest performance of baseline for microsleep onset prediction was achieved a linear SVM and MDF (GM = 0.71; φ = 0.07; Sn = 0.76; Pr = 0.01; Sp = 0.74). When the prediction time was increased to τ = 10 s, the highest GM was 0.66, achieved with VBHMSRFA with MDF (φ = 0.04; Sn = 0.66; Pr = 0.01; Sp = 0.71). The highest GM of the baseline was 0.54 with a linear SVM and WMSF (φ = 0.04; GM = 0.54; Sn = 0.57; Pr = 0.01; Sp = 0.73). Notwithstanding, the performance in terms of phi was very low across all the methods due to the highly imbalanced property of the dataset.

Although our proposed methods were able to achieve slightly higher GM values compared to those of the baseline, the improvement in performance was lower than expected. The low number of microsleeps and subject variability make it difficult for the proposed methods to identify a set of consistent patterns. Since the data is variable between subjects, multi-subject methods are required to reduce the variability between subjects. However, the number of microsleep onsets is limited and it appears that, as a consequence, our proposed Bayesian algorithms were unable to maximally capture the novelty of microsleep onsets. As a result, the patterns of microsleep onsets could have been partially removed from the data and incorrectly recognized as noise.

11.3 COMPARISON WITH PREVIOUS STUDIES

Baseer et al. [2017] used the data of Study A and our refined gold-standard to predict microsleep states at 0.25 s ahead. They filtered EEG data into four standard EEG frequency-bands, i.e., delta, theta, alpha, and beta. For each of the frequency bands, two sets of inter-channel features, joint entropy and mutual-information, were extracted from 5-s EEG windows, which led to a total of 600 features in each feature set. A feature selection method was applied by excluding redundant features, i.e., correlation of more than 0.9, applying a ranking method, and forward selection of features based on the AUC-ROC of an LDA (3-fold cross-validation). Their highest performance of microsleep state prediction with τ = 0.25 s was achieved with joint entropy features (AUC-ROC = 0.93; GM = 0.75; φ = 0.38; Sn = 0.68; Pr = 0.33). In the current study, our highest performance with a prediction time of τ = 0.25 s for microsleep state prediction was achieved with VBMSRFA meta-features of WLMSF and an LDA classifier (AUC-ROC = 0.95; AUC-PR = 0.48; GM = 0.83; φ = 0.46; Sn = 0.73; Pr = 0.38). This shows that our proposed methods achieved substantially higher performances compared to those of Baseer et al.

The findings of this research are not directly comparable to the remaining of the literature of Study A due to the change of gold-standard but the following is an attempt to provide an unbiased comparison between this study and the literature. Important to note that the current project aimed to identify *behavioural microsleeps*, whereas *lapses of responsiveness* were the main focus of the previous works on Study A [Ayyagari et al. 2015, Ayyagari 2017, Davidson et al. 2007, LaRocco 2015, Peiris et al. 2011]. As mentioned in Section 2.2, behavioural microsleeps are episodes of *sleep-related* suspension of performance, whereas lapses of responsiveness include all short episodes of failure to respond in goal-directed tasks. As described in Section 5.3, previous works on Study A defined lapses as occurrence of tracking flat-spots and/or lapse video-ratings. Hence, a lapse could have been due to a tracking flat-spot, while the subject was awake. However, we defined microsleeps as non-tracking episodes accompanied with video ratings of deep-drowsy or lapse.

Davidson et al. [2007] performed EEG-based lapse detection with a 1-s temporal resolution using Study A. They used 2-s EEG-windows to extract log-power spectral features and utilized an LSTM recurrent neural network as classifier. Their highest reported performances were AUC-ROC = 0.84, AUC-PR = 0.41, and $\varphi = 0.38$. In a similar study, Ayyagari [2017] used the same features with a stack generalization of leaky echo-state neural networks to detect lapses using Study A. His highest performances were AUC-ROC = 0.88, AUC-PR = 0.45, and $\varphi = 0.44$. Both of these studies used the whole dataset without any data pruning. Evidently, the highest performances of the current study have higher values to those of Davidson et al. and Ayyagari. However, a true comparison is not possible as they detected lapses, whereas we performed microsleep detection and prediction. In addition, as mentioned in Section 6.4.3, phi and AUC-PR are sensitive to the class distribution (i.e., class-imbalance ratio) of data. Comparing the imbalance ratio, i.e., ratio of microsleeps (or lapses) to responsive labels, of refined gold-standard (Table 5.3) and the earlier lapse-index gold-standard (Table 5.4) shows that the two gold-standards have different imbalance ratios, where the average imbalance ratio of the refined gold-standard and lapse-index were 0.10 (0.001–0.44) and 0.11 (0.01–0.36), respectively. Although these average class-imbalance ratios are similar, the lowest extremes are very different. As a result of the difference between the imbalance ratios, a fair comparison between performances in terms of phi and AUC-PR. Nevertheless, the highest AUC-ROC of the current study for microsleep state detection with a temporal resolution of 0.25 s was 0.95 (cf. 0.88 from Ayyagari [2017] and 0.84 from Davidson et al. [2007] for lapse detection).

Peiris et al. [2011], Ayyagari et al. [2015], and LaRocco [2015] also performed lapse detection with a 1-s temporal resolution. However, they cleansed the EEG-data by rejecting the noisy epochs and performing the lapse detection task on the cleaned EEG-epochs only. Peiris et al. [2011] used a stack generalization of LDA classifiers and achieved AUC-ROC = 0.84, AUC-PR = 0.43, and $\varphi = 0.39$. LaRocco [2015] used a stack generalization of LDA and 150 principal components and achieved $\varphi = 0.40$, but he did not report other performance measures. Ayyagari et al. [2015] used a stack generalization of leaky echo-state neural networks and achieved AUC-ROC = 0.91, AUC-PR = 0.47, and $\varphi = 0.51$. To compare the effect of data cleansing on the class-imbalance ratio, the epoch rejection process of these studies was replicated. After rejection of noisy epochs, the average imbalance ratio of the remaining data across all subjects was 0.12 (0.01–0.38). As previously mentioned, due to the change of imbalance ratios, comparison of phi and AUC-PR values can be misleading. Notwithstanding, these studies were able to achieve higher AUC-ROC compared to the unpruned dataset. However, we did not attempt to prune noisy EEG epochs, but reconstructed them instead (as described in Section 5.2).

Our highest microsleep detection performance in terms of AUC-ROC was 0.95, which is higher than all previous works on Study A.

Peiris et al. [2011] investigated detection of lapse events using Study A, in which they did not aim to detect onsets. In their analysis, they continuously, with a 1-s temporal resolution, performed lapse detection. Lapse events were marked as detected if any point within an event was successfully identified as the lapse state. Additionally, lapse events without recognizing any points as lapse, were called missed events. To identify lapses, they used a stack generalization of multiple LDA classifiers and log-power spectral features. They achieved an average sensitivity and specificity of 0.74 and 0.59, respectively. Our results for microsleep onset detection was Sn = 0.77 and Sp = 0.85 with a linear SVM and VBMSRFA meta-features of WLMSF. Moreover, our highest performances in terms of sensitivity and specificity for microsleep onset prediction with $\tau = 1$ s were Sn = 0.81 and Sp = 0.80 with a linear SVM and VBHMSRFA meta-features of MDF, which are substantially higher than those of Peiris et al. for lapse detection.

LaRocco [2015] attempted to predict the onset of lapses up to 1 s prior to each event, using Study A. He achieved a phi of 0.05 with a combination of an LDA, a PCA feature reduction, and spectral features. Notwithstanding, he did not provide other performance measures and hence a fair comparison cannot be made due to the differences in class imbalance ratios and gold-standards. Nevertheless, our highest phi for microsleep onset prediction with $\tau = 1$ s was 0.08 with VBHMSRFA meta-features of MDF and a linear SVM classifier.

Subasi et al. [2005] used wavelet decomposition of 5-s EEG epochs to detect vigilance state. They used an ANN classifier to identify wakefulness, i.e., awake, drowsy, and sleep, using the wavelet approximation and details corresponding to the standard EEG frequency bands. They achieved an accuracy of 95% for vigilance state detection. Akin et al. [2008] acquired an EMG signal from the chin of participant, in addition to the EEG. They improved the performance to 99%. However, these studies used only EEG to define the state of vigilance and therefore discarded behavioural information. Therefore, although they found a high accuracy for vigilance state detection, this gives little indication of detection of behavioural microsleeps.

Yeo et al. [2009] investigated an EEG-based system for drowsiness detection. They extracted frequency-domain features of EEG and fed them to an SVM classifier. An accuracy of 99.3% was achieved for drowsiness detection. However, their experimental task was simulation of a driving session by watching a video clip, which was not active and hence did not require feedback from participants. Since the participants were not required to actively provide feedback, it is difficult to accurately identify behavioural microsleeps with their task. In addition, although they incorporated blink frequency and duration as part of defining drowsiness, an expert also rated EEG activity of alpha and beta bands as part of identification of alert and drowsy. Thus, their gold-standard was not independent of EEG data.

Golz et al. [2005, 2007] and Sommer et al. [2005] investigated microsleep detection using EEG. They used 3-s EEG epochs to extract two sets of features, power spectral and delay vector variances, which were then used for microsleep detection. Their highest accuracy of 88.8% was achieved with an SVM classifier and the fusion of both feature sets. Notwithstanding, they suggested only using power spectral features as they achieved a similar accuracy of 88.0%. However, their reported accuracy is misleading for two reasons. First, they concatenated data of all of the subjects and then performed a cross-validation to evaluate the performance of microsleep detection. This introduces dependency between training and test data, as the data of a subject can be partially in both training data and testing data. Second, only a portion of the non-microsleep epochs were selected to balance the class-ratio. Although this procedure is acceptable for training of a classifier, the testing phase must be done on independent and unselected data. For these reasons, their accuracies cannot be used to estimate performance on new unseen individuals. It is expected that if they evaluate their performance using LOSO-CV with a temporal resolution of 0.25 s, their performances would be substantially lower.

Chapter 12

CONCLUSIONS AND FUTURE RESEARCH

12.1 SUMMARY AND KEY FINDINGS

The motivations for this thesis were (1) to improve the accuracy of microsleep detection and (2) to investigate prediction of microsleeps. A system capable of accurately identifying microsleeps could potentially be used to continuously monitor individual's responsiveness, especially in occupations requiring extended visuomotor performance (e.g., truck driver). The focus of this research was on investigation of Bayesian methods for EEG-based microsleep detection/prediction, which could take noise and uncertainty of EEG-features into account.

The previous study [Peiris 2008] collected data of 15 non-sleep-deprived participants while performing a 1-D continuous tracking task (CTT), i.e., Study A. Participants took part in two 1-h sessions resulting in 30 h worth of data. Notwithstanding, they only found 8 subjects with at least a microsleep during the two 1-hour sessions. In the current study, the data of those 8 subjects of Study A with at least a microsleep were used to investigate microsleeps detection/prediction.

The original gold-standards, i.e., lapse index and BM, were prone to having wrong labels. As mentioned in Section 2.5.2, lapses were previously defined as tracking flat-spots and/or lapse video-ratings. As a result, there could be lapses due to tracking flat-spots while awake, which is different from a microsleep. On the other hand, microsleeps were previously defined as occurrence of both video-lapses *and* tracking flat-spots. However, everything other than microsleeps were labelled as responsive, which is likely to include other types of lapses. In addition, due to limitations of the experimental CTT of Study A, there are low-velocity segments in the task which do not require feedback from the user. Therefore, identification of tracking flat-spots is not possible for those low-velocity segments. Therefore, a conservative analysis was performed in the current project to refine the gold-standard prior to machine learning algorithms. Satisfactory tracking for at least 5 s was labelled as responsive, irrespective of the expert's video ratings. A microsleep was defined as the conjunction of a non-tracking episode and a video rating of deep-drowsy or lapse. The remainder of the refined gold-standard was labelled as uncertain, due to a lack of information to accurately identify the state of responsiveness. For the purpose of our analyses, the refined gold-standard, with a 4-Hz temporal resolution, was used and the uncertain labels were nulled out.

We proposed an extension of variational Bayesian factor analysis (VBFA) in Chapter 7, i.e., variational Bayesian robust FA (VBRFA), which uses a robust latent-space. VBRFA uses independent Student-t distributions to represent latent variables, which was achieved with a hierarchical normal-Gamma distribution. Moreover, VBRFA exploits an automatic relevance determination (ARD) motivated prior distribution on the columns of the loading matrix to automatically identify the optimum number of latent variables.

Variational Baysian multi-subject RFA (VBMSRFA), described in Chapter 8, was proposed to extend VBRFA to account for inter-subject variability. VBMSRFA uses an independent mean and noise terms for individual subjects, while a common loading matrix is used among all subjects. This structure allows individual subjects to share information via a common loading matrix, but the differences between subjects are assumed to be limited to the mean and noise terms. Additionally, VBMSRFA could find the optimum number of latent variables, similar to VBRFA, by exploiting an ARD-motivated prior distribution over the columns of the loading matrix.

An extension of VBMSRFA was presented in Chapter 9, i.e., variational Bayesian hierarchical MSRFA (VBHMSRFA), to allow individual subjects to have different loading matrices, while sharing a group loading matrix, in addition to the individual mean and noise terms. This introduces a hierarchical structure on the loading matrix, where the subject-level loading matrices are drawn from a group-level loading matrix. Our proposed VBHMSRFA method uses an ARD-motivated prior distribution over the subject-level and group-level loading matrices to find the optimum number of latent variables across all the subjects.

Variational Bayesian hierarchical multi-subject robust joint matrix factorization (VBHMSRJMF) was proposed in Chapter 10 to extend VBHMSRFA to exploit class-label information while finding latent spaces. If all data were from one class, VBHMSRJMF would simplify to VBHMSRFA. However, with more than one class, VBHMSRJMF simultaneously finds a latent space per class with shared loading matrices. This structure allows our proposed model to identify differences between classes using multiple latent-spaces. At test time, a data vector is jointly factorized into multiple latent-spaces, which are then used as meta-features. The VBHMSRJMF also exploits an ARD-motivated prior distribution over both group- and subject-level loading matrices to find the optimum number of latent variables.

This study focused on microsleep detection/prediction using two prediction strategies: microsleep state and onset identification. Microsleep state detection/prediction continuously identifies the state of responsiveness. Therefore, prediction of the microsleep state with a prediction time of τ s aims to predict an imminent microsleep τ s prior to its occurrence, however if it misses the prediction, it will try to predict/detect the microsleep as soon as possible. On the other hand, microsleep onset prediction continuously attempts to predict the onset of an imminent microsleep but does not further attempt to detect the microsleep if the onset was missed.

Different features of EEG data were extracted and investigated for detection and prediction of microsleeps, namely power spectral features (PSF), power spectral features using individual alpha frequency (PSF-IAF), multiple domain features (MDF), wavelet mean squared features (WMSF), wavelet log mean squared features (WLMSF), and wavelet energy percentage features (WEPF). Additionally, discriminating microsleeps and responsives was done by four classifiers, namely linear discriminant analysis (LDA), linear support vector machine (SVM), tree augmented naïve Bayes (TAN), and variational Bayesian logistic regression (VBLR). All of the features were extracted from three different EEG window-lengths of 2, 5, and 10 s. Features extracted from shorter EEG windows contain phasic/transient information, whereas features of longer windows correspond to tonic changes of EEG. We found that aggregating features of the three window-lengths performed superior to individual single-window features. The current study focused on microsleep detection and prediction using Bayesian feature-reduction methods. As mentioned in Section 6.5, for the sake of comparison, we performed detection and prediction of microsleeps using the original features. Furthermore, four feature-reduction methods – principal component analysis (PCA), Bayesian PCA, FA, and VBFA – and two feature selection methods – greedy forward feature-selection algorithm based on mutual information [Battiti 1994, Kwak and Choi 2002] and greedy forward feature-selection based on Hellinger distance [Yin et al. 2013] – were applied. We found that the original features without any feature selection/reduction method achieved similar or higher performances compared to those with feature reduction/selection methods. Therefore, performances with the original features were used as a baseline for evaluating our proposed methods. In addition, as mentioned in Section 6.3.5, two oversampling methods – synthetic minority over-sampling technique (SMOTE) and adaptive synthetic sampling (ADASYN) – and a cost-sensitive technique were used to reduce bias of training classifiers on imbalanced data. The results showed little difference but cost-sensitive learning was substantially faster. Therefore, cost-sensitive learning was used for training all classifiers.

The best performance in terms of phi for microsleep state detection, i.e., $\tau = 0$ s, was $\varphi = 0.47$ and was achieved with an LDA and VBMSRFA meta-features of WLMSF (GM = 0.83; Sn = 0.74; Pr = 0.38). The second best performance was $\varphi = 0.45$ which was with an LDA and VBHMSRFA meta-features of PSF (GM = 0.81; Sn = 0.71; Pr = 0.37). The highest performance of VBHMSRJMF was $\varphi = 0.43$ with with an LDA and meta-features of PSF (GM = 0.81; Sn = 0.71; Pr = 0.37). Using these methods improved the performance of microsleep state detection compared to the baseline, especially in terms of phi, where the highest performance was $\varphi = 0.40$ and was achieved with a linear SVM and PSF (GM = 0.79; Sn = 0.74; Pr = 0.38). Notwithstanding, in contrast to our expectation, the VBRFA method achieved similar performances to those of the baseline but with lower-dimension representation of the data.

The highest performances of microsleep state detection in terms of AUC-ROC and AUC-PR using various methods were relatively similar to those of the baseline. The highest values of AUC-ROC for the methods used in this thesis were 0.94–0.95. Similarly, the highest values of the AUC-PR were 0.48–0.49, with an exception of VBHMSRJMF which could only achieve an AUC-PR of 0.43. Notwithstanding, all of our proposed multi-subject methods improved the performance in terms of phi and GM, which indicates that finding a less subject-variable latent space reduced the variability between classification thresholds of training and test subjects.

When the prediction time was increased to $\tau = 1$ s, the highest performance in terms of phi was $\varphi = 0.44$ and was achieved with an LDA and VBMSRFA meta-features of WLMSF (AUC-ROC = 0.94; AUC-PR = 0.44; GM = 0.80; Sn = 0.69; Pr = 0.36). However, the highest performance of the baseline was $\varphi = 0.35$ with a linear SVM and MDF (AUC-ROC = 0.92; AUC-PR = 0.42; GM = 0.76; Sn = 0.70; Pr = 0.35). Overall, the highest performances for both detection and prediction of microsleep states were achieved with VBMSRFA meta-features of WLMSF.

For microsleep onset detection, the baseline features of PSF with a linear SVM achieved an AUC-ROC of 0.91 and AUC-PR of 0.09 ($\varphi = 0.08$; GM = 0.71; Sn = 0.79; Pr = 0.03). However, our proposed Bayesian multi-subject methods were able to achieve higher GM and marginally better phi values, although the values of AUC-ROC and AUC-PR were slightly inferior to those of the baseline. The highest phi of 0.10 was achieved with a combination of VBHMSRFA meta-features of WLMSF and a VBLR (AUC-ROC = 0.89; AUC-PR = 0.05; GM = 0.78; Sn = 0.70; Pr = 0.03). Furthermore, the highest GM was 0.81 with a linear SVM and VBMSRFA meta-features of WLMSF (AUC-ROC = 0.91; AUC-PR = 0.05; $\varphi = 0.09$; Sn = 0.77; Pr = 0.02).

Prediction of microsleep onsets $\tau = 10$ s ahead had a poor performance. The highest performance was achieved with a linear SVM and VBHMSRFA meta-features of MDF (AUC-ROC = 0.76; AUC-PR = 0.01; $\varphi = 0.04$; GM = 0.66; Sn = 0.66; Pr = 0.01). On the other hand, the highest AUC-ROC of baseline was 0.74 and was achieved with a linear SVM and WMSF (AUC-PR = 0.01; $\varphi = 0.04$; GM = 0.54; Sn = 0.57; Pr = 0.01). Due to the highly imbalanced nature of the microsleep onset dataset, i.e., an average ratio of 0.002 (0.0002–0.006) for microsleep onsets relative to responsives, the values of AUC-PR and phi were highly affected by false positives.

Although our proposed Bayesian methods were able to achieve better results, the overall performances are still insufficient for real-life applications. Notwithstanding, we found evidence that EEG contains information regarding ongoing and even imminent microsleeps. The sensitivity of microsleep detection/prediction was moderate but with low precision, i.e., too many false positives.

12.2 REVIEW OF HYPOTHESES

Hypothesis 1 — Explicit modelling of noise and uncertainty will improve performance of the microsleep detection/prediction system.

Minimal evidence was found to support this hypothesis. A Bayesian model was proposed in Chapter 7, i.e., VBRFA, to perform feature reduction with a robust latent space. The highest baseline performances were relatively similar to those of VBRFA meta-features. Although

VBRFA was able to find a lower-dimension representation of data with comparative performance, no performance improvements were found with VBRFA compared to the best of baseline.

Hypothesis 2 — Using a Bayesian model to find a common latent-space among all subjects will improve the performance of a microsleep detection/prediction system.

Given that inter-subject variability exists in the EEG data, we hypothesized that finding a common latent-space among subjects using a Bayesian model could improve the performance. Three multi-subject Bayesian feature reduction models were proposed in Chapters 8–10 to incorporate inter-subject variability. The results of these methods in terms of phi and GM were higher than those of the baseline which supports our hypothesis. These findings suggest that using a less subject-variable latent-space reduces the variability of classification threshold between training and test subjects and result in higher phi and GM values.

Hypothesis 3 — There are specific changes in the EEG before microsleeps which can be identified in real-time and used to predict imminent microsleeps.

Evidence was found to support this hypothesis. The results of microsleep onset prediction, Chapters 7–10, were substantially better than a random guess. We found relatively high values of AUC-ROC with short prediction times but low AUC-PR. Although our performance was poor and not ready for real-life applications, it showed that microsleeps can be predicted with a moderate sensitivity several seconds before their occurrence.

12.3 CRITIQUE

A limitation of the current research was the small number of participants in Study A. Having a small group of 15 participants, of which only 8 had at least one definite microsleep, limits the generalization of our findings. Additionally, finding a common latent-space among 7 training subjects results in a large uncertainty in the latent variables. Furthermore, due to the small number of participants, we reported individual p-values without correcting for multiple comparisons throughout the thesis. As a result, type-II errors (false negatives) were prevented. However, it might have led to some false positives, i.e., type-I errors, in our results. In the future, it is of importance to collect more data so that correcting for multiple comparisons is feasible and hence avoid such shortcomings.

As mentioned in Section 5.3, the velocity of the target of CTT fell to zero 34 times in each 128-s cycle. In addition, certain segments of the target have a low velocity, i.e., target's flat-spots, and hence no input from the user is needed to keep the tracking cursor on the target trace. Estimating the state of responsiveness during these episodes is nearly impossible. Although we refined the gold-standard and marked such regions as uncertain, all the microsleeps might not have been identified. Such limitation could be prevented in future studies by using a 2-D continuous tracking task [Poudel et al. 2014].

In the current study, we extracted EEG features from individual channels, which discards inter-channel relationships. Therefore, any microsleep-related information in the synchrony of different brain regions, i.e., different EEG electrodes, is missed in our system. Baseer et al. [2017] showed that joint entropy as a set of inter-channel-relationship features performs slightly better than PSF for microsleep state prediction. Therefore, a future study can investigate the effect of different inter-channel relationships, together with VBMSRFA feature reduction, for microsleep prediction.

Aggregated features of multiple EEG windows, i.e., 2, 5, and 10 s, outperformed single-window features for microsleep state detection. This is evidence that both the tonic and phasic characteristics of EEG contain microsleep-related information and that incorporating this temporal information can improve performance of microsleep detection and prediction. However, the length of EEG windows were fixed in the current study and therefore our performances may not be maximal. Additionally, our proposed methods do not take dynamic temporal information into account, while reducing the number of features. A future study can focus on extracting and incorporating dynamic temporal information of EEG for a microsleep prediction system.

In the current study, Bayesian methods were used as a means of feature reduction and classification. Although these methods provide a framework to handle uncertainty, they were applied independent of each other. First, features of EEG were fed to a Bayesian feature-reduction method to reduce the dimensionality of the data. Then, the maximum a posteriori (MAP) estimates of latent variables were used for classification. This process removes the uncertainty of meta-features (latent variables) but, at the same time, is prone to loss of some useful information. Integrating the two steps of feature-reduction and classification into a unified Bayesian model is expected to improve performance of microsleep detection and prediction, which can be the focus of a future study.

On a positive note, our proposed multi-subject Bayesian feature-reduction methods reduced subject-variability of classification-threshold, achieving higher performances in terms of GM and phi with similar values of AUC-ROC and AUC-PR. These methods take inter-subject variability of EEG into account and provide a framework for handling inherent noise and uncertainty of EEG features. Moreover, automatic identification of optimum number of meta-features, in the lower-dimension space, minimizes the chance of overfitting without a need for cross-validation.

Our results show that microsleeps can be predicted prior to their occurrence with a moderate sensitivity and specificity, especially with shorter prediction times. However, there were too many false positives relative to true positives, i.e., low precision, which makes our system impractical for real-life applications.

12.4 SUGGESTED FUTURE WORK

As mentioned in Section 12.3, one of the limitations of the current study was the low number of participants. Considering inter-subject variability in the EEG data and an inherently high class-imbalance ratio, a large number of subjects should substantially improve the generalization of microsleep detection/prediction methods to new unseen subjects. Additionally, calculating performance on a larger number of subjects would provide a better generalized estimate of bias–variance trade-off of machine learning methods for EEG-based microsleep prediction systems.

Apart from the shortcomings of the CTT used in Study A, as mentioned in Section 12.3, the CTT was not intended to simulate driving. As a result, our findings may not be directly applicable to real-life driving. Using a sophisticated driving simulator, with traffic, pedestrians, windy roads, intersections, etc., would be much more indicative of real-life driving, although even this would still lack the undoubtedly important effect of 'consequences' on propensity for microsleeps.

A refined gold-standard was presented in the current study to minimize errors in class labels and subsequent deleterious effects on machine learning/classification. However, we used the video ratings of an expert, who used the recording of a camera 1 m from eyes, i.e., not close-up, to rate responsiveness. As a result, the video ratings were subjective, conservative, and likely contain numerous errors. Even in our refined gold-standard, these errors are likely to have led to incorrect labels. Supervised machine learning methods use a labelled dataset to learn about a certain task and explore the relationship between features and classes. The presence of label-noise in such algorithms has a significant impact on performance, especially when the dataset itself is imbalanced. Using noise-robust methods, algorithmically relabelling the data based on spatial location in feature space, or even discarding the labels with high classification uncertainties at the time of training can effectively improve the performance of classifiers [Bootkrajang and Kabán 2014, Frenay and Verleysen 2014, Lausser et al. 2014]. The effect of label-noise on the performance of microsleep prediction system could be investigated in a future study.

Microsleep prediction using various feature sets was reported in the current study. Microsleep detection and prediction using MDF had similar results to the spectral features. However, the effect of fusion of multiple feature sets was not investigated. Therefore, it is reasonable to assume that different feature sets could potentially contain distinct/orthogonal information. Although the performance of individual feature sets might be somewhat similar, fusion of multiple feature-sets could potentially combine information of multiple domains and hence improve the detection/prediction performance.

As mentioned earlier, aggregation of features extracted from multiple EEG windows, i.e., 2, 5, and 10 s, outperformed single-window features. This provides evidence that EEG has temporal information which can improve the detection/prediction performance. In the current study, the lengths of EEG windows were fixed and unlikely to be optimal. Therefore, exploiting methods such as Kalman filter, hidden Markov models, and autoregressive models to extract temporal information of EEG might provide additional information resulting in higher accuracy of microsleep detection and prediction.

In the current study, we extracted features from individual EEG-channels independently. However, these features make no use of synchronization information across EEG channels.

Baseer et al. [2017] used joint-entropy of every two EEG-channels for microsleep state prediction. They achieved slightly higher performances than those of power spectral features. This suggests that incorporating relationships between EEG channels, such as coherence and connectivity features, will likely improve accuracy of microsleep prediction. It is expected that applying our proposed Bayesian methods to those inter-channel features will further improve performance.

In this study, uncertain labels of the refined gold-standard were discarded prior to the training phase of machine learning methods. However, although the features corresponding to uncertain labels do not provide any information for evaluating performance of the system, they can provide spatial information in the feature space to improve the training of supervised methods. Semi-supervised methods use both labelled and unlabelled data to train a machine learning method [Schwenker and Trentin 2014]. Investigation of semi-supervised machine-learning methods for microsleep prediction could be an important focus of future research.

Four classifiers – LDA, linear SVM, VBLR, and TAN – were investigated in the current study to discriminate between the microsleeps and the responsives. In our design, a single classifier was used to perform the prediction or detection task. However, incorporating ensemble classifiers, i.e, aggregating multiple weak classifiers to find a stronger learner – is likely to improve performance [Abbasi et al. 2016, Murphy 2012, Ofek et al. 2017]. Notwithstanding, although ensemble classifiers have higher flexibility to find a better separation between data of different classes, they are prone to overfitting. A future study can investigate different types of ensemble classifiers such as cluster based, bagging, and boosting for microsleep detection and prediction.

In the current study, Bayesian methods were used as two independent tools for feature reduction and classification. However, this is prone to losing information, where uncertainty of meta-features (latent variables) is removed. A unified Bayesian model capable of finding a multi-subject latent-space and performing the classification task is expected to have higher accuracy for microsleep prediction. A future study can concentrate on further developing Bayesian models, especially supervised, for prediction of microsleeps.

The current study used a fixed prediction-time for imminent microsleep prediction, but a fixed prediction-time does not give the system a chance to correct itself. The prediction time can be a range (e.g., 10 s) prior to a microsleep, in which the system is expected to identify the imminent microsleep. Therefore, instead of assuming that a microsleep is going to occur at exactly τ-s ahead, we let the system to give us a time-range such that a microsleep is about to happen within the next τ s. Employing this perspective, the microsleep prediction system can initiate multiple attempts at identifying imminent microsleeps and yet be in the specified prediction time-range, which is expected to improve both of sensitivity and precision. For instance, microsleep onset prediction within the next 10 s with a temporal resolution of 0.25 s allows the system to identify the event in 40 attempts. This concept can be used for future studies to find a better estimate of prediction ability of an EEG-based microsleep prediction system.

REFERENCES

Abbasi, E., Shiri, M.E. and Ghatee, M. (2016), 'Root-quatric mixture of experts for complex classification problems', *Expert Systems with Applications*, Vol. 53, pp. 192–203.

Akin, M., Kurt, M., Sezgin, N. and Bayram, M. (2008), 'Estimating vigilance level by using EEG and EMG signals', *Neural Computing and Applications*, Vol. 17, No. 3, pp. 227–236.

Albera, L., Kachenoura, A., Comon, P., Karfoul, A., Wendling, F., Senhadji, L. and Merlet, I. (2012), 'ICA-based EEG denoising: a comparative analysis of fifteen methods', In *Bulletin of the Polish Academy of Sciences: Technical Sciences*, Vol. 60, pp. 407–418.

Alpert, N.M. and Yuan, F. (2009), 'A general method of Bayesian estimation for parametric imaging of the brain', *NeuroImage*, Vol. 45, No. 4, pp. 1183–1189.

Anund, A. and Åkerstedt, T. (2010), 'Perception of sleepiness before falling asleep', *Sleep Medicine*, Vol. 11, No. 8, pp. 743–744.

Asyali, M.H., Berry, R.B., Khoo, M.C.K. and Altinok, A. (2007), 'Determining a continuous marker for sleep depth', *Computers in Biology and Medicine*, Vol. 37, No. 11, pp. 1600–1609.

Avendaño-Valencia, L., Godino-Llorente, J., Blanco-Velasco, M. and Castellanos-Dominguez, G. (2010), 'Feature extraction from parametric time–frequency representations for heart murmur detection', *Annals of Biomedical Engineering*, Vol. 38, No. 8, pp. 2716–2732.

Ayyagari, S., Jones, R. and Weddell, S. (2015), 'Optimized echo state networks with leaky integrator neurons for EEG-based microsleep detection', In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3775–3778.

Ayyagari, S. (2017), *Reservoir computing approaches to EEG-based detection of microsleeps*, PhD thesis, University of Canterbury.

Baas, P.H., Charlton, S.G. and Bastin, G.T. (2000), 'Survey of New Zealand truck driver fatigue and fitness for duty', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 3, No. 4, pp. 185–193.

Badillo, S., Desmidt, S., Ginisty, C. and Ciuciu, P. (2014), 'Multi-subject Bayesian joint detection and estimation in fMRI', In *International Workshop on Pattern Recognition in Neuroimaging*, pp. 1–4.

Bagley, S.C., White, H. and Golomb, B.A. (2001), 'Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain', *Journal of Clinical Epidemiology*, Vol. 54, No. 10, pp. 979–985.

Baillet, S. and Garnero, L. (1997), 'A Bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem', *IEEE Transactions on Biomedical Engineering*, Vol. 44, No. 5, pp. 374–385.

Bajaj, V. and Pachori, R.B. (2013), 'Automatic classification of sleep stages based on the time-frequency image of EEG signals', *Computer Methods and Programs in Biomedicine*, Vol. 112, No. 3, pp. 320–328.

Barber, D. (2012), *Bayesian Reasoning and Machine Learning*, Cambridge University Press.

Baseer, A., Weddell, S.J. and Jones, R.D. (2017), 'Prediction of microsleeps using pairwise joint entropy and mutual information between EEG channels', In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4495–4498.

Battiti, R. (1994), 'Using mutual information for selecting features in supervised neural net learning', *IEEE Transactions on Neural Networks*, Vol. 5, No. 4, pp. 537–550.

Baulk, S.D., Kandelaars, K.J., Lamond, N., Roach, G.D., Dawson, D. and Fletcher, A. (2007), 'Does variation in workload affect fatigue in a regular 12-hour shift system?', *Sleep & Biological Rhythms*, Vol. 5, No. 1, pp. 74–77.

Beal, M.J. (2003), *Variational algorithms for approximate Bayesian inference*, PhD thesis, University College London.

Belardinelli, P., Ortiz, E., Barnes, G., Noppeney, U. and Preissl, H. (2012), 'Source reconstruction accuracy of MEG and EEG Bayesian inversion approaches', *PLoS ONE*, Vol. 7, No. 12, pp. 1–16.

Bielza, C. and Larrañaga, P. (2014), 'Discrete Bayesian network classifiers: A survey', *ACM Computing Surveys*, Vol. 47, No. 1, pp. 60:1–60:43.

Bishop, C.M. (2012), 'Model-based machine learning', *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 371, No. 1984, pp. 1–17.

Bishop, C. (1999), 'Variational principal components', In *Proceedings of 9th International Conference on Artificial Neural Networks*, pp. 509–514.

Bishop, C. (2006), *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer-Verlag, New York.

Blanco, R., Inza, I., Merino, M., Quiroga, J. and Larrañaga, P. (2005), 'Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS', *Journal of Biomedical Informatics*, Vol. 38, No. 5, pp. 376–388.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.R. and Curio, G. (2007), 'The non-invasive Berlin brain–computer interface: Fast acquisition of effective performance in untrained subjects', *NeuroImage*, Vol. 37, No. 2, pp. 539–550.

Bojić, T., Vuckovic, A. and Kalauzi, A. (2010), 'Modeling EEG fractal dimension changes in wake and drowsy states in humans–a preliminary study', *Journal of Theoretical Biology*, Vol. 262, No. 2, pp. 214–222.

Bootkrajang, J. and Kabán, A. (2014), 'Learning kernel logistic regression in the presence of class label noise', *Pattern Recognition*, Vol. 47, No. 11, pp. 3641–3655.

Botvinick, M.M. and Bylsma, L.M. (2005), 'Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context', *Psychonomic Bulletin & Review*, Vol. 12, No. 6, pp. 1011–1017.

Boyle, L.N., Tippin, J., Paul, A. and Rizzo, M. (2008), 'Driver performance in the moments surrounding a microsleep', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 11, No. 2, pp. 126–136.

Buckley, R.J., Helton, W.S., Innes, C.R., Dalrymple-Alford, J.C. and Jones, R.D. (2016), 'Attention lapses and behavioural microsleeps during tracking, psychomotor vigilance, and dual tasks', *Consciousness and Cognition*, Vol. 45, pp. 174–183.

Centre for Road Safety and Transport for NSW (2014), 'Road traffic crashes in New South Wales: Statistical statement for the year ended 31 December 2014', `http://roadsafety.transport.nsw.gov.au/downloads/crashstats2014.pdf`. Accessed February 2017.

Chai, R., Tran, Y., Naik, G.R., Nguyen, T.N., Ling, S.H., Craig, A. and Nguyen, H.T. (2016), 'Classification of EEG based-mental fatigue using principal component analysis and Bayesian neural network', In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4654–4657.

Chapotot, F. and Becq, G. (2010), 'Automated sleep–wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules', *International Journal of Adaptive Control and Signal Processing*, Vol. 24, No. 5, pp. 409–423.

Chawla, N.V. (2010), 'Data mining for imbalanced datasets: An overview', In O. Maimon and L. Rokach (editors), *Data Mining and Knowledge Discovery Handbook*, pp. 875–886, Springer, Boston, MA.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. (2002), 'SMOTE: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357.

Chen, L., Zhao, Y., Zhang, J. and Zou, J. (2015), 'Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning', *Expert Systems with Applications*, Vol. 42, No. 21, pp. 7344–7355.

Chien, J.T. and Chen, J.C. (2009), 'Recursive Bayesian linear regression for adaptive classification', *IEEE Transactions on Signal Processing*, Vol. 57, No. 2, pp. 565–575.

Chow, C. and Liu, C. (1968), 'Approximating discrete probability distributions with dependence trees', *IEEE Transactions on Information Theory*, Vol. 14, No. 3, pp. 462–467.

Clercq, W.D., Vergult, A., Vanrumste, B., Van Paesschen, W. and Van Huffel, S. (2006), 'Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram', *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 12, pp. 2583–2587.

Cortes, J.M., Lopez, A., Molina, R. and Katsaggelos, A.K. (2012), 'Variational Bayesian localization of EEG sources with generalized Gaussian priors', *The European Physical Journal Plus*, Vol. 127, No. 11 (140), pp. 1–12.

Costa, F., Batatia, H., Chaari, L. and Tourneret, J.Y. (2015), 'Sparse EEG source localization using Bernoulli Laplacian priors', *IEEE Transactions on Biomedical Engineering*, Vol. 62, No. 12, pp. 2888–2898.

Croce, P., Basti, A., Marzetti, L., Zappasodi, F. and Gratta, C.D. (2016), 'EEG–fMRI Bayesian framework for neural activity estimation: a simulation study', *Journal of Neural Engineering*, Vol. 13, No. 6 (066017), pp. 1–13.

Daly, I., Scherer, R., Billinger, M. and Muller-Putz, G. (2015), 'FORCe: Fully online and automated artifact removal for brain-computer interfacing', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 23, No. 5, pp. 725–736.

Daly, I., Nicolaou, N., Nasuto, S.J. and Warwick, K. (2013), 'Automated artifact removal from the electroencephalogram: A comparative study', *Clinical EEG and Neuroscience*, Vol. 44, No. 4, pp. 291–306.

Das, B., Krishnan, N.C. and Cook, D.J. (2015), 'RACOG and wRACOG: Two probabilistic oversampling techniques', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 1, pp. 222–234.

Daunizeau, J., Mattout, J., Clonda, D., Goulard, B., Benali, H. and Lina, J.M. (2006), 'Bayesian spatio-temporal approach for EEG source reconstruction: conciliating ECD and distributed models', *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 3, pp. 503–516.

Davidson, P., Jones, R. and Peiris, M. (2005), 'Detecting behavioral microsleeps using EEG and LSTM recurrent neural networks', In *Annual International Conference of the Engineering in Medicine and Biology Society*, pp. 5754–5757.

Davidson, P., Jones, R. and Peiris, M. (2007), 'EEG-based lapse detection with high temporal resolution', *IEEE Transactions on Biomedical Engineering*, Vol. 54, No. 5, pp. 832–839.

de Rochefort, L., Liu, T., Kressler, B., Liu, J., Spincemaille, P., Lebon, V., Wu, J. and Wang, Y. (2010), 'Quantitative susceptibility map reconstruction from MR phase data using Bayesian regularization: Validation and application to brain imaging', *Magnetic Resonance in Medicine*, Vol. 63, No. 1, pp. 194–206.

Delorme, A. and Makeig, S. (2004), 'EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis', *Journal of Neuroscience Methods*, Vol. 134, No. 1, pp. 9–21.

Delorme, A., Sejnowski, T. and Makeig, S. (2007), 'Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis', *NeuroImage*, Vol. 34, No. 4, pp. 1443–1449.

DelPozo-Banos, M., Travieso, C.M., Weidemann, C.T. and Alonso, J.B. (2015), 'EEG biometric identification: a thorough exploration of the time-frequency domain', *Journal of Neural Engineering*, Vol. 12, No. 5 (056019), pp. 1–23.

Diez, P.F., Laciar, E., Mut, V., Avila, E. and Torres, A. (2008), 'A comparative study of the performance of different spectral estimation methods for classification of mental tasks', In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1155–1158.

Ding, Y., Xie, R., Zou, Y. and Guo, J. (2016), 'NMR data compression method based on principal component analysis', *Applied Magnetic Resonance*, Vol. 47, No. 3, pp. 297–307.

dos Santos, E.B., Hruschka Jr., E.R., Hruschka, E.R. and Ebecken, N.F.F. (2011), 'Bayesian network classifiers: Beyond classification accuracy', *Intelligent Data Analysis*, Vol. 15, No. 3, pp. 279–298.

Dreiseitl, S. and Ohno-Machado, L. (2002), 'Logistic regression and artificial neural network classification models: a methodology review', *Journal of Biomedical Informatics*, Vol. 35, No. 5, pp. 352–359.

Drugowitsch, J. (2013), 'Variational Bayesian inference for linear and logistic regression', *ArXiv e-prints*, pp. 1–28.

Duffy, F., Iyer, V. and Surwillo, W. (1989), 'Brain electrical activity: An introduction to EEG recording', In *Clinical Electroencephalography and Topographic Brain Mapping*, Springer, New York.

Eddy, S.R. (2004), 'What is bayesian statistics?', *Nature Biotechnology*, Vol. 22, No. 9, pp. 1177–1178.

Faust, O., Acharya, U.R., Adeli, H. and Adeli, A. (2015), 'Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis', *Seizure*, Vol. 26, pp. 56–64.

Fell, J., Röschke, J., Mann, K. and Schäffner, C. (1996), 'Discrimination of sleep stages: a comparison between spectral and nonlinear EEG measures', *Electroencephalography and Clinical Neurophysiology*, Vol. 98, No. 5, pp. 401–410.

Finkbeiner, K.M., Wilson, K.M., Russell, P.N. and Helton, W.S. (2015), 'The effects of warning cues and attention-capturing stimuli on the sustained attention to response task', *Experimental Brain Research*, Vol. 233, No. 4, pp. 1061–1068.

Folleco, A.A., Khoshgoftaar, T.M., Van Hulse, J. and Napolitano, A. (2009), 'Identifying learners robust to low quality data.', *Informatica*, Vol. 33, No. 3, pp. 245–259.

Forsman, P.M., Vila, B.J., Short, R.A., Mott, C.G. and Dongen, H.P.V. (2013), 'Efficient driver drowsiness detection at moderate levels of drowsiness', *Accident Analysis & Prevention*, Vol. 50, pp. 341–350.

Freeman, W. and Quiroga, R. (2013), *Imaging Brain Function With EEG: Advanced Temporal and Spatial Analysis of Electroencephalographic Signals*, Springer, New York.

Frenay, B. and Verleysen, M. (2014), 'Classification in the presence of label noise: A survey', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 25, No. 5, pp. 845–869.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J. and Penny, W. (2007), 'Variational free energy and the Laplace approximation', *NeuroImage*, Vol. 34, No. 1, pp. 220–234.

Gander, P.H., Mulrine, H.M., van den Berg, M.J., Smith, A.A.T., Signal, T.L., Wu, L.J. and Belenky, G. (2014), 'Pilot fatigue: Relationships with departure and arrival times, flight duration, and direction', *Aviation, Space, and Environmental Medicine*, Vol. 85, No. 8, pp. 833–840.

Geiger-Brown, J., Rogers, V.E., Trinkoff, A.M., Kane, R.L., Bausell, R.B. and Scharf, S.M. (2012), ‘Sleep, sleepiness, fatigue, and performance of 12-hour-shift nurses’, *Chronobiology International*, Vol. 29, No. 2, pp. 211–219.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013), *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Geng, S., Zhou, W., Yuan, Q., Cai, D. and Zeng, Y. (2011), ‘EEG non-linear feature extraction using correlation dimension and hurst exponent’, *Neurological Research*, Vol. 33, No. 9, pp. 908–912.

Georgieva, P., Bouaynaya, N., Silva, F., Mihaylova, L. and Jain, L.C. (2016), ‘A beamformer-particle filter framework for localization of correlated EEG sources’, *IEEE Journal of Biomedical and Health Informatics*, Vol. 20, No. 3, pp. 880–892.

Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., Ramsey, N., Haselager, P., Vuurpijl, L., Gielen, S. and Desain, P. (2009), ‘The brain–computer interface cycle’, *Journal of Neural Engineering*, Vol. 6, No. 4 (041001), pp. 1–10.

Ghahramani, Z. (2015), ‘Probabilistic machine learning and artificial intelligence’, *Nature*, Vol. 521, No. 7553, pp. 452–459.

Ghahramani, Z. and Beal, M. (2000), ‘Variational inference for Bayesian mixtures of factor analysers’, In *Advances in Neural Information Processing Systems*, pp. 449–455.

Ghahramani, Z. (2012), ‘Bayesian non-parametrics and the probabilistic approach to modelling’, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 371, No. 1984, pp. 1–20.

Goljahani, A., D’Avanzo, C., Schiff, S., Amodio, P., Bisiacchi, P. and Sparacino, G. (2012), ‘A novel method for the determination of the EEG individual alpha frequency’, *NeuroImage*, Vol. 60, No. 1, pp. 774–786.

Golz, M., Sommer, D. and Krajewski, J. (2016), ‘Prediction of immediately occurring microsleep events from brain electric signals’, *Current Directions in Biomedical Engineering*, Vol. 2, No. 1, pp. 149–153.

Golz, M., Sommer, D. and Mandic, D. (2005), ‘Microsleep detection in electrophysiological signals’, In *Proceedings of 1st International Workshop Biosignal Processing and Classification (BPC)*, pp. 102–109.

Golz, M., Sommer, D., Chen, M., Trutschel, U. and Mandic, D. (2007), ‘Feature fusion for the detection of microsleep events’, *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, Vol. 49, No. 2, pp. 329–342.

Gong, J. and Kim, H. (2017), ‘RHSBoost: Improving classification performance in imbalance data’, *Computational Statistics & Data Analysis*, Vol. 111, pp. 1–13.

Greene, B., Faul, S., Marnane, W., Lightbody, G., Korotchikova, I. and Boylan, G. (2008), ‘A comparison of quantitative EEG features for neonatal seizure detection’, *Clinical Neurophysiology*, Vol. 119, No. 6, pp. 1248–1261.

Güneş, S., Polat, K. and Şebnem Yosunkaya (2010), ‘Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting’, *Expert Systems with Applications*, Vol. 37, No. 12, pp. 7922–7928.

Haegens, S., Cousijn, H., Wallis, G., Harrison, P.J. and Nobre, A.C. (2014), ‘Inter- and intra-individual variability in alpha peak frequency’, *NeuroImage*, Vol. 92, pp. 46–55.

Häkkänen, H. and Summala, H. (2001), ‘Fatal traffic accidents among trailer truck drivers and accident causes as viewed by other truck drivers’, *Accident Analysis & Prevention*, Vol. 33, No. 2, pp. 187–196.

Härmä, M., Sallinen, M., Ranta, R., Mutanen, P. and Müller, K. (2002), ‘The effect of an irregular shift system on sleepiness at work in train drivers and railway traffic controllers’, *Journal of Sleep Research*, Vol. 11, No. 2, pp. 141–151.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics, Springer, New York.

He, H. and Garcia, E. (2009), ‘Learning from imbalanced data’, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, pp. 1263–1284.

He, H., Bai, Y., Garcia, E. and Li, S. (2008), ‘ADASYN: Adaptive synthetic sampling approach for imbalanced learning’, In *IEEE International Joint Conference on Neural Networks*, pp. 1322–1328.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B. (1998), ‘Support vector machines’, *IEEE Intelligent Systems and their Applications*, Vol. 13, No. 4, pp. 18–28.

Hensman, J., Matthews, A. and Ghahramani, Z. (2015), ‘Scalable variational Gaussian process classification’, *Journal of Machine Learning Research*, Vol. 38, pp. 351–360.

Higgins, J.S., Michael, J., Austin, R., Åkerstedt, T., Van Dongen, H.P.A., Watson, N., Czeisler, C., Pack, A.I. and Rosekind, M.R. (2017), ‘Asleep at the wheel–the road to addressing drowsy driving’, *Sleep*, Vol. 40, No. 2 (zsx001), pp. 1–9.

Hinne, M., Heskes, T., Beckmann, C.F. and van Gerven, Marcel, A. (2013), ‘Bayesian inference of structural brain networks’, *NeuroImage*, Vol. 66, pp. 543–52.

Holub, M., Šrutová, M. and Lhotská, L. (2015), ‘Adaptable microsleep detection based on EOG signals: A feasibility study’, In *International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, pp. 1–4.

Hopfinger, J.B., Buonocore, M.H. and Mangun, G.R. (2000), ‘The neural mechanisms of top-down attentional control’, *Nature Neuroscience*, Vol. 3, No. 3, pp. 284–291.

Hor, C.Y., Yang, C.B., Yang, Z.J. and Tseng, C.T. (2013), ‘Prediction of protein essentiality by the support vector machine with statistical tests’, *Evolutionary Bioinformatics*, Vol. 9, pp. 387–416.

Huang, K., Huang, T., Chuang, C., King, J., Wang, Y., Lin, C. and Jung, T. (2016), ‘An EEG-based fatigue detection and mitigation system’, *International Journal of Neural Systems*, Vol. 26, No. 4 (1650018), pp. 1–14.

Huang, Q., Yang, J. and Zhou, Y. (2007), 'Variational Bayesian method for speech enhancement', *Neurocomputing*, Vol. 70, No. 16-18, pp. 3063–3067.

Huang, R.S., Jung, T.P., Delorme, A. and Makeig, S. (2008), 'Tonic and phasic electroencephalographic dynamics during continuous compensatory tracking', *NeuroImage*, Vol. 39, No. 4, pp. 1896–1909.

Huang, Y., Zhang, Y., Li, N., Wu, Z. and Chambers, J.A. (2017), 'A novel robust student's t-based kalman filter', *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 53, No. 3, pp. 1545–1554.

Imtiaz, S.A., Saremi-Yarahmadi, S. and Rodriguez-Villegas, E. (2013), 'Automatic detection of sleep spindles using Teager energy and spectral edge frequency', In *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 262–265.

Innes, C.R.H., Poudel, G.R. and Jones, R.D. (2013), 'Efficient and regular patterns of nighttime sleep are related to increased vulnerability to microsleeps following a single night of sleep restriction', *Chronobiology International*, Vol. 30, No. 9, pp. 1187–1196.

Jap, B.T., Lal, S., Fischer, P. and Bekiaris, E. (2009), 'Using EEG spectral components to assess algorithms for detecting fatigue', *Expert Systems with Applications*, Vol. 36, No. 2, Part 1, pp. 2352–2359.

Jatupaiboon, N., Pan-ngum, S., and Israsena, P. (2013), 'Real-time EEG-based happiness detection system', *The Scientific World Journal*, Vol. 2013, pp. 1–12.

Jay, S.M., Dawson, D., Ferguson, S.A. and Lamond, N. (2008), 'Driver fatigue during extended rail operations', *Applied Ergonomics*, Vol. 39, No. 5, pp. 623–629.

Jebara, T. (2003), *Machine Learning: Discriminative and Generative*, The Springer International Series in Engineering and Computer Science, Springer, US.

Jensen, F.V. and Nielsen, T.D. (2007), *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York, 2nd ed.

Jenssen, R. (2013), 'Mean vector component analysis for visualization and clustering of nonnegative data', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 24, No. 10, pp. 1553–1564.

Jiang, L., Cai, Z., Wang, D. and Zhang, H. (2012), 'Improving tree augmented naive Bayes for class probability estimation', *Knowledge-Based Systems*, Vol. 26, pp. 239–245.

Jones, R., Poudel, G., Innes, C., Davidson, P.R., Peiris, M., Malla, A., Signal, T., Carroll, G., Watts, R. and Bones, P. (2010), 'Lapses of responsiveness: Characteristics, detection, and underlying mechanisms', In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1788–1791.

Jung, T.P., Makeig, S., Stensmo, M. and Sejnowski, T.J. (1997), 'Estimating alertness from the EEG power spectrum', *IEEE Transactions on Biomedical Engineering*, Vol. 44, No. 1, pp. 60–69.

Kanemura, H., Mizorogi, S., Aoyagi, K., Sugita, K. and Aihara, M. (2012), 'EEG characteristics predict subsequent epilepsy in children with febrile seizure', *Brain and Development*, Vol. 34, No. 4, pp. 302–307.

Kang, H. and Choi, S. (2014), 'Bayesian common spatial patterns for multi-subject EEG classification', *Neural Networks*, Vol. 57, pp. 39–50.

Khalighi, S., Sousa, T. and Nunes, U. (2012), 'Adaptive automatic sleep stage classification under covariate shift', In *34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2259–2262.

Kiebel, S.J., Daunizeau, J., Phillips, C. and Friston, K.J. (2008), 'Variational Bayesian inversion of the equivalent current dipole model in EEG/MEG', *NeuroImage*, Vol. 39, No. 2, pp. 728–741.

Killeen, P. (2013), 'Absent without leave; a neuroenergetic theory of mind wandering', *Frontiers in Psychology*, Vol. 4, No. 373, pp. 1–8.

Kim, H.C. and Ghahramani, Z. (2006), 'Bayesian Gaussian process classification with the EM-EP algorithm', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 12, pp. 1948–1959.

Klados, M.A., Papadelis, C., Braun, C. and Bamidis, P.D. (2011), 'Reg-ICA: A hybrid methodology combining blind source separation and regression techniques for the rejection of ocular artifacts', *Biomedical Signal Processing and Control*, Vol. 6, No. 3, pp. 291–300.

Klami, A., Virtanen, S. and Kaski, S.b. (2013), 'Bayesian canonical correlation analysis', *Journal of Machine Learning Research*, Vol. 14, No. 1, pp. 965–1003.

Klimesch, W. (1999), 'EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis', *Brain Research Reviews*, Vol. 29, No. 2–3, pp. 169–195.

Klonovs, J., Petersen, C.K., Olesen, H. and Hammershoj, A. (2013), 'ID proof on the go: Development of a mobile EEG-based biometric authentication system', *IEEE Vehicular Technology Magazine*, Vol. 8, No. 1, pp. 81–89.

Knopp, S.J. (2015), *A multi-modal device for application in microsleep detection*, PhD thesis, University of Canterbury.

Ko, K., Yang, H. and Sim, K. (2009), 'Emotion recognition using EEG signals with relative power values and Bayesian network', *International Journal of Control, Automation and Systems*, Vol. 7, No. 5, pp. 865–870.

Krajewski, J., Batliner, A. and Wieland, R. (2008), 'Multiple classifier applied on predicting microsleep from speech', In *19th International Conference on Pattern Recognition (ICPR)*, pp. 1–4.

Krajewski, J., Golz, M., Sommer, D. and Wieland, R. (2009), 'Genetic algorithm based feature selection applied on predicting microsleep from speech', In J. Sloten, P. Verdonck, M. Nyssen and J. Haueisen (editors), *4th European Conference of the International Federation for Medical and Biological Engineering*, Springer, Berlin, Heidelberg, pp. 184–187.

Kumar, Y., Dewal, M. and Anand, R. (2014), 'Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine', *Neurocomputing*, Vol. 133, pp. 271–279.

Kurt, M.B., Sezgin, N., Akin, M., Kirbas, G. and Bayram, M. (2009), 'The ANN-based computing of drowsy level', *Expert Systems with Applications*, Vol. 36, No. 2, Part 1, pp. 2534–2542.

Kwak, N. and Choi, C.H. (2002), 'Input feature selection by mutual information based on parzen window', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 12, pp. 1667–1671.

Lal, S.K. and Craig, A. (2005), 'Reproducibility of the spectral components of the electroencephalogram during driver fatigue', *International Journal of Psychophysiology*, Vol. 55, No. 2, pp. 137–143.

LaRocco, J. (2015), *Detection of microsleeps from the EEG via optimized classification techniques*, PhD thesis, University of Canterbury.

Larue, G.S., Rakotonirainy, A. and Pettitt, A.N. (2015), 'Predicting reduced driver alertness on monotonous highways', *IEEE Pervasive Computing*, Vol. 14, No. 2, pp. 78–85.

Larue, G.S., Rakotonirainy, A. and Pettitt, A.N. (2010), 'Real-time performance modelling of a sustained attention to response task', *Ergonomics*, Vol. 53, No. 10, pp. 1205–1216.

Larue, G.S., Rakotonirainy, A. and Pettitt, A.N. (2011), 'Driving performance impairments due to hypovigilance on monotonous roads', *Accident Analysis & Prevention*, Vol. 43, No. 6, pp. 2037–2046.

Lausser, L., Schmid, F., Schmid, M. and Kestler, H.A. (2014), 'Unlabeling data can improve classification accuracy', *Pattern Recognition Letters*, Vol. 37, pp. 15–23.

Lawrence, N.D. and Schölkopf, B. (2001), 'Estimating a kernel fisher discriminant in the presence of label noise', In *Proceedings of the 18th International Conference on Machine Learning*, ICML '01, Citeseer, pp. 306–313.

Lemm, S., Blankertz, B., Dickhaus, T. and Müller, K.R. (2011), 'Introduction to machine learning for brain imaging', *NeuroImage*, Vol. 56, No. 2, pp. 387–399.

Li, M., Luo, X. and Yang, J. (2016), 'Extracting the nonlinear features of motor imagery EEG using parametric t-SNE', *Neurocomputing*, Vol. 218, pp. 371–381.

Lic, H. and Summala, H. (2000), 'Sleepiness at work among commercial truck drivers', *Sleep*, Vol. 23, No. 1, pp. 1–9.

Lin, C.T., Huang, K.C., Chao, C.F., Chen, J.A., Chiu, T.W., Ko, L.W. and Jung, T.P. (2010), 'Tonic and phasic EEG and behavioral changes induced by arousing feedback', *NeuroImage*, Vol. 52, No. 2, pp. 633–642.

Lin, C.T., Huang, K.C., Chuang, C.H., Ko, L.W. and Jung, T.P. (2013), 'Can arousing feedback rectify lapses in driving? prediction from EEG power spectra', *Journal of Neural Engineering*, Vol. 10, No. 5 (056024), pp. 1–10.

Liu, H., Shah, S. and Jiang, W. (2004), 'On-line outlier detection and data cleaning', *Computers & Chemical Engineering*, Vol. 28, No. 9, pp. 1635–1647.

López, J., Litvak, V., Espinosa, J., Friston, K. and Barnes, G. (2014), 'Algorithmic procedures for Bayesian MEG/EEG source reconstruction in SPM', *NeuroImage*, Vol. 84, pp. 476–487.

López, V., Fernández, A., García, S., Palade, V. and Herrera, F. (2013), 'An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics', *Information Sciences*, Vol. 250, pp. 113–141.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F. and Arnaldi, B. (2007), 'A review of classification algorithms for EEG-based brain–computer interfaces', *Journal of Neural Engineering*, Vol. 4, No. 2, pp. R1–R13.

Lucka, F., Pursiainen, S., Burger, M. and Wolters, C.H. (2012), 'Hierarchical Bayesian inference for the EEG inverse problem using realistic FE head models: Depth localization and source separation for focal primary currents', *NeuroImage*, Vol. 61, No. 4, pp. 1364–1382.

Mackay, D.J.C. (1995), 'Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks', *Network: Computation in Neural Systems*, Vol. 6, No. 3, pp. 469–505.

Madden, M.G. (2009), 'On the classification performance of TAN and general Bayesian networks', *Knowledge-Based Systems*, Vol. 22, No. 7, pp. 489–495.

Makeig, S., Jung, T. and Sejnowski, T.J. (2000), 'Awareness during drowsiness: Dynamics and electrophysiological correlates', *Canadian Journal of Experimental Psychology*, Vol. 54, No. 4, pp. 266–273.

Malla, A., Davidson, P.R., Bones, P., Green, R. and Jones, R. (2010), 'Automated video-based measurement of eye closure for detecting behavioral microsleep', In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6741–6744.

Martinez-Leon, J.A., Cano-Izquierdo, J.M. and Ibarrola, J. (2016), 'Are low cost brain computer interface headsets ready for motor imagery applications?', *Expert Systems with Applications*, Vol. 49, pp. 136–144.

Matiko, J.W., Wei, Y., Torah, R., Grabham, N., Paul, G., Beeby, S. and Tudor, J. (2015), 'Wearable EEG headband using printed electrodes and powered by energy harvesting for emotion monitoring in ambient assisted living', *Smart Materials and Structures*, Vol. 24, No. 12 (125028), pp. 1–11.

McGrory, C.A. and Titterington, D.M. (2009), 'Variational Bayesian analysis for hidden Markov models', *Australian & New Zealand Journal of Statistics*, Vol. 51, No. 2, pp. 227–244.

Mohammadiha, N., Smaragdis, P. and Leijon, A. (2013), 'Supervised and unsupervised speech enhancement using nonnegative matrix factorization', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 10, pp. 2140–2151.

Moller, H.J., Kayumov, L., Bulmash, E.L., Nhan, J. and Shapiro, C.M. (2006), 'Simulator performance, microsleep episodes, and subjective sleepiness: normative data using convergent methodologies to assess driver drowsiness', *Journal of Psychosomatic Research*, Vol. 61, No. 3, pp. 335–342.

Mukuta, Y. and Harada, T. (2014), 'Probabilistic partial canonical correlation analysis', In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Vol. 4, pp. 3309–3321.

Mullen, T., Kothe, C., Chi, Y.M., Ojeda, A., Kerth, T., Makeig, S., Cauwenberghs, G. and Jung, T.P. (2013), 'Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG', In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2184–2187.

Mullen, T., Kothe, C., Chi, Y., Ojeda, A., Kerth, T., Makeig, S., Jung, T.P. and Cauwenberghs, G. (2015), 'Real-time neuroimaging and cognitive monitoring using wearable dry EEG', *IEEE Transactions on Biomedical Engineering*, Vol. 62, No. 11, pp. 2553–2567.

Müller, P. and Mitra, R. (2013), 'Bayesian nonparametric inference - why and how', *Bayesian analysis*, Vol. 8, No. 2, pp. 1–23.

Murphy, K.P. (2012), *Machine learning: a probabilistic perspective*, MIT Press, Cambridge, MA.

Naderi, M. and Mahdavi-Nasab, H. (2010), 'Analysis and classification of EEG signals using spectral analysis and recurrent neural networks', In *17th Iranian Conference of Biomedical Engineering (ICBME)*, pp. 1–4.

Nakajima, S., Sugiyama, M., Babacan, S.D. and Tomioka, R. (2013), 'Global analytic solution of fully-observed variational Bayesian matrix factorization', *Journal of Machine Learning Research*, Vol. 14, pp. 1–37.

Navascués, M.A. and Sebastián, M.V. (2009), 'Time domain indices and discrete power spectrum in electroencephalographic processing', *International Journal of Computer Mathematics*, Vol. 86, No. 10–11, pp. 1968–1978.

Nguyen, T.M. and Wu, Q.M.J. (2012), 'Robust student's-t mixture model with spatial constraints and its application in medical image segmentation', *IEEE Transactions on Medical Imaging*, Vol. 31, No. 1, pp. 103–116.

NHTSA (2011), 'Traffic safety facts', `www-nrd.nhtsa.dot.gov/pubs/811449.pdf`. Accessed September 2017.

Nicolaou, N. and Georgiou, J. (2011), 'The use of permutation entropy to characterize sleep electroencephalograms', *Clinical EEG and Neuroscience*, Vol. 42, No. 1, pp. 24–28.

Nolan, H., Whelan, R. and Reilly, R. (2010), 'FASTER: Fully automated statistical thresholding for EEG artifact rejection', *Journal of Neuroscience Methods*, Vol. 192, No. 1, pp. 152–162.

NTSB (2016), 'Most wanted transportation safety improvements', `www.ntsb.gov/safety/mwl/Documents/MWL_2016_factsheet01.pdf`. Accessed September 2017.

Ofek, N., Rokach, L., Stern, R. and Shabtai, A. (2017), 'Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem', *Neurocomputing*, Vol. 243, pp. 88–102.

Oostenveld, R. and Praamstra, P. (2001), 'The five percent electrode system for high-resolution EEG and ERP measurements', *Clinical Neurophysiology*, Vol. 112, No. 4, pp. 713–719.

PARAMANATHAN, P. AND UTHAYAKUMAR, R. (2008), ‘Application of fractal theory in analysis of human electroencephalographic signals’, *Computers in Biology and Medicine*, Vol. 38, No. 3, pp. 372–378.

PAVITHRA, M., NIRANJANAKRUPA, B., SASIDHARAN, A., KUTTY, B.M. AND LAKKANNAVAR, M. (2014), ‘Fractal dimension for drowsiness detection in brainwaves’, In *International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 757–761.

PEARSON, R. (2002), ‘Outliers in process modeling and identification’, *IEEE Transactions on Control Systems Technology*, Vol. 10, No. 1, pp. 55–63.

PEIRIS, M. (2008), *Lapses in Responsiveness:Characteristics and Detection from the EEG*, PhD thesis, University of Canterbury.

PEIRIS, M.T.R., JONES, R.D., DAVIDSON, P.R., CARROLL, G.J. AND BONES, P.J. (2006), ‘Frequent lapses of responsiveness during an extended visuomotor tracking task in non-sleep-deprived subjects’, *Journal of Sleep Research*, Vol. 15, No. 3, pp. 291–300.

PEIRIS, M.T.R., DAVIDSON, P.R., BONES, P.J. AND JONES, R.D. (2011), ‘Detection of lapses in responsiveness from the EEG’, *Journal of Neural Engineering*, Vol. 8, No. 1 (016003), pp. 1–15.

PEIRIS, M., JONES, R., DAVIDSON, P., CARROLL, G., SIGNAL, T., PARKIN, P., VAN DEN BERG, M. AND BONES, P. (2005), ‘Identification of vigilance lapses using EEG/EOG by expert human raters’, In *27th Annual International Conference of the Engineering in Medicine and Biology Society*, pp. 5735–5737.

PÉREZ, A., NAGA, P.L. AND NAKI INZA, I. (2006), ‘Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes’, *International Journal of Approximate Reasoning*, Vol. 43, No. 1, pp. 1–25.

PERNKOPF, F. AND BILMES, J.A. (2010), ‘Efficient heuristics for discriminative structure learning of Bayesian network classifiers.’, *Journal of Machine Learning Research*, Vol. 11, No. 8, pp. 2323–2360.

PHILLIPS, C., MATTOUT, J., RUGG, M.D., MAQUET, P. AND FRISTON, K.J. (2005), ‘An empirical Bayesian solution to the source reconstruction problem in EEG’, *NeuroImage*, Vol. 24, No. 4, pp. 997–1011.

POLYCHRONAKI, G.E., KTONAS, P.Y., GATZONIS, S., SIATOUNI, A., ASVESTAS, P.A., TSEKOU, H., SAKAS, D. AND NIKITA, K.S. (2010), ‘Comparison of fractal dimension estimation algorithms for epileptic seizure onset detection’, *Journal of Neural Engineering*, Vol. 7, No. 4 (046007), pp. 1–18.

POSTHUMA, D., NEALE, M.C., BOOMSMA, D.I. AND DE GEUS, E.J.C. (2001), ‘Are smarter brains running faster? heritability of alpha peak frequency, IQ, and their interrelation’, *Behavior Genetics*, Vol. 31, No. 6, pp. 567–579.

POUDEL, G.R., INNES, C.R. AND JONES, R.D. (2013), ‘Distinct neural correlates of time-on-task and transient errors during a visuomotor tracking task after sleep restriction’, *NeuroImage*, Vol. 77, pp. 105–113.

Poudel, G.R., Innes, C.R., Bones, P.J., Watts, R. and Jones, R.D. (2014), 'Losing the struggle to stay awake: Divergent thalamic and cortical activity during microsleeps', *Human Brain Mapping*, Vol. 35, No. 1, pp. 257–269.

Puuronen, J. and Hyvärinen, A. (2014), 'A Bayesian inverse solution using independent component analysis', *Neural Networks*, Vol. 50, pp. 47–59.

Qian, D., Wang, B., Qing, X., Zhang, T., Zhang, Y., Wang, X. and Nakamura, M. (2017), 'Drowsiness detection by Bayesian-Copula discriminant classifier based on EEG signals during daytime short nap', *IEEE Transactions on Biomedical Engineering*, Vol. 64, No. 4, pp. 743–754.

Quitadamo, L.R., Cavrini, F., Sbernini, L., Riillo, F., Bianchi, L., Seri, S. and Saggio, G. (2017), 'Support vector machines to detect physiological patterns for EEG and EMG-based human–computer interaction: a review', *Journal of Neural Engineering*, Vol. 14, No. 1 (011001), pp. 1–27.

Rasmussen, C.E. (2004), 'Gaussian processes in machine learning', In O. Bousquet, U. von Luxburg and G. Rätsch (editors), *Advanced Lectures on Machine Learning*, pp. 63–71, Springer, Berlin, Heidelberg.

Rasmussen, C.E. and Williams, C.K.I. (2006), *Gaussian Processes for Machine Learning*, Adaptive computation and machine learning, The MIT Press.

Rodríguez-Bermúdez, G., García-Laencina, P.J., Roca-González, J. and Roca-Dorda, J. (2013), 'Efficient feature selection and linear discrimination of EEG signals', *Neurocomputing*, Vol. 115, pp. 161–165.

Saito, T. and Rehmsmeier, M. (2015), 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PLoS ONE*, Vol. 10, No. 3 (e0118432), pp. 1–21.

Sanei, S., Chambers, J., Sanei, S. and Chambers, J. (2007), 'Fundamentals of EEG signal processing', In *EEG Signal Processing*, pp. 35–125, John Wiley & Sons Ltd.

Schad, A., Schindler, K., Schelter, B., Maiwald, T., Brandt, A., Timmer, J. and Schulze-Bonhage, A. (2008), 'Application of a multivariate seizure detection and prediction method to non-invasive and intracranial long-term EEG recordings', *Clinical Neurophysiology*, Vol. 119, No. 1, pp. 197–211.

Schwenker, F. and Trentin, E. (2014), 'Pattern classification and clustering: A review of partially supervised learning approaches', *Pattern Recognition Letters*, Vol. 37, pp. 4–14.

Şen, B., Peker, M., Çavuşoğlu, A. and Çelebi, F. (2014), 'A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms', *Journal of Medical Systems*, Vol. 38, No. 3, 18.

Sheng, H., Chen, Y. and Qiu, T. (2012), 'Multifractional property analysis of human sleep electroencephalogram signals', In *Fractional Processes and Fractional-Order Signal Processing: Techniques and Applications*, pp. 243–250, Springer, London.

Smith, A.P. (2016), 'A UK survey of driving behaviour, fatigue, risk taking and road traffic accidents', *BMJ Open*, Vol. 6, No. 8, pp. 1–6.

SOMMER, D., CHEN, M., GOLZ, M., TRUTSCHEL, U. AND MANDIC, D. (2005), 'Fusion of state space and frequency-domain features for improved microsleep detection', In W. Duch, J. Kacprzyk, E. Oja and S. Zadrożny (editors), *Artificial Neural Networks: Formal Models and Their Applications (ICANN)*, pp. 753–759, Springer, Berlin, Heidelberg.

STAHLHUT, C., MØRUP, M., WINTHER, O. AND HANSEN, L.K. (2011), 'Simultaneous EEG source and forward model reconstruction (SOFOMORE) using a hierarchical Bayesian approach', *Journal of Signal Processing Systems*, Vol. 65, No. 3, pp. 431–444.

STANLEY, N. (1996), 'The future of sleep staging', *Human Psychopharmacology: Clinical & Experimental*, Vol. 11, No. 3, pp. 253–256.

STÉPHANE, M. (2009), 'Wavelet packet and local cosine bases', In M. Stéphane (editor), *A Wavelet Tour of Signal Processing (3rd Edition)*, chap. 8, pp. 377–434, Academic Press, Boston, third edition ed.

STROBBE, G., VAN MIERLO, P., VOS, M.D., MIJOVIĆ, B., HALLEZ, H., HUFFEL, S.V., LÓPEZ, J.D. AND VANDENBERGHE, S. (2014), 'Bayesian model selection of template forward models for EEG source reconstruction', *NeuroImage*, Vol. 93, No. Part 1, pp. 11–22.

SUBASI, A., KIYMIK, M., AKIN, M. AND EROGUL, O. (2005), 'Automatic recognition of vigilance state by using a wavelet-based artificial neural network', *Neural Computing & Applications*, Vol. 14, No. 1, pp. 45–55.

SUN, H., LAM, K., CHUNG, S., DONG, W., GU, M. AND SUN, J. (2005), 'Efficient vector quantization using genetic algorithm', *Neural Computing & Applications*, Vol. 14, No. 3, pp. 203–211.

SUN, S. (2013), 'A review of deterministic approximate inference techniques for Bayesian machine learning', *Neural Computing and Applications*, Vol. 23, No. 7, pp. 2039–2050.

SUN, Y., WONG, A.K.C. AND KAMEL, M.S. (2009), 'Classification of imbalanced data: A review', *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 23, No. 04, pp. 687–719.

SUNDAR, H., SEELAMANTULA, C.S. AND SREENIVAS, T.V. (2012), 'A mixture model approach for formant tracking and the robustness of student's-t distribution', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 10, pp. 2626–2636.

TANEJA, N. (2007), 'Fatigue in aviation: A survey of the awareness and attitudes of indian air force pilots.', *International Journal of Aviation Psychology*, Vol. 17, No. 3, pp. 275–284.

TEFFT, B.C. (2010), 'Asleep at the wheel: The prevalence and impact of drowsy driving', `www.aaafoundation.org/sites/default/files/2010DrowsyDrivingReport_1.pdf`. Accessed January 2015.

TEFFT, B.C. (2014), 'Prevalence of motor vehicle crashes involving drowsy drivers, United States, 2009–2013', `https://www.aaafoundation.org/sites/default/files/AAAFoundation-DrowsyDriving-Nov2014.pdf`. Accessed August 2017.

THAI-NGHE, N., GANTNER, Z. AND SCHMIDT-THIEME, L. (2010), 'Cost-sensitive learning methods for imbalanced data', In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

THOMAS, K.P. AND VINOD, A.P. (2016), 'Biometric identification of persons using sample entropy features of EEG during rest state', In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3487–3492.

TIPPING, M.E. AND BISHOP, C.M. (1999), 'Probabilistic principal component analysis', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 61, No. 3, pp. 611–622.

TIPPING, M.E. AND LAWRENCE, N.D. (2005), 'Variational inference for student-t models: Robust Bayesian interpolation and generalised component analysis', *Neurocomputing*, Vol. 69, No. 1, pp. 123–141.

TONG, S. AND THAKOR, N. (2009), *Quantitative EEG analysis methods and clinical applications*, Artech House engineering in medicine & biology series, Artech House.

TRUJILLO-BARRETO, N.J., AUBERT-VÁZQUEZ, E. AND PENNY, W.D. (2008), 'Bayesian M/EEG source reconstruction with spatio-temporal priors', *NeuroImage*, Vol. 39, No. 1, pp. 318–335.

TZIKAS, D.G., LIKAS, A.C. AND GALATSANOS, N.P. (2008), 'The variational approximation for Bayesian inference', *IEEE Signal Processing Magazine*, Vol. 25, No. 6, pp. 131–146.

UNSWORTH, N. AND ROBISON, M.K. (2016), 'Pupillary correlates of lapses of sustained attention', *Cognitive, Affective, & Behavioral Neuroscience*, Vol. 16, No. 4, pp. 601–615.

UPADHYAY, R., MANGLICK, A., REDDY, D., PADHY, P. AND KANKAR, P. (2015), 'Channel optimization and nonlinear feature extraction for electroencephalogram signals classification', *Computers & Electrical Engineering*, Vol. 45, pp. 222–234.

VANLAAR, W., SIMPSON, H., MAYHEW, D. AND ROBERTSON, R. (2008), 'Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors', *Journal of Safety Research*, Vol. 39, No. 3, pp. 303–309.

VIDAURRE, C., KRÄMER, N., BLANKERTZ, B. AND SCHLÖGL, A. (2009), 'Time domain parameters as a feature for EEG-based brain-computer interfaces', *Neural Networks*, Vol. 22, No. 9, pp. 1313–1319.

VIHINEN, M. (2012), 'How to evaluate performance of prediction methods? measures and their interpretation in variation effect analysis', *BMC Genomics*, Vol. 13, No. 4 (S2), pp. 1–10.

VIHINEN, M. (2013), 'Guidelines for reporting and using prediction tools for genetic variation analysis', *Human Mutation*, Vol. 34, No. 2, pp. 275–282.

WANG, C. (2007), 'Variational Bayesian approach to canonical correlation analysis', *IEEE Transactions on Neural Networks*, Vol. 18, No. 3, pp. 905–910.

WANG, S., WANG, J., WANG, Z. AND JI, Q. (2014a), 'Enhancing multi-label classification by modeling dependencies among labels', *Pattern Recognition*, Vol. 47, No. 10, pp. 3405–3413.

WANG, Y.T., HUANG, K.C., WEI, C.S., HUANG, T.Y., KO, L.W., LIN, C.T., CHENG, C.K. AND JUNG, T.P. (2014b), 'Developing an EEG based on-line closed-loop lapse detection and mitigation system', *Frontiers in Neuroscience*, Vol. 8, No. 321, pp. 1–11.

WANG, Z., LEUNG, C., ZHU, Y. AND WONG, T. (2004), 'Data compression on the illumination adjustable images by PCA and ICA', *Signal Processing: Image Communication*, Vol. 19, No. 10, pp. 939–954.

Watling, C.N. (2014), 'Sleepy driving and pulling over for a rest: Investigating individual factors that contribute to these driving behaviours', *Personality and Individual Differences*, Vol. 56, pp. 105–110.

Wei, C.S., Lin, Y.P., Wang, Y.T., Jung, T.P., Bigdely-Shamlo, N. and Lin, C.T. (2015), 'Selective transfer learning for EEG-based drowsiness detection', In *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3229–3232.

Wei, Y., Wu, Y. and Tudor, J. (2017), 'A real-time wearable emotion detection headband based on EEG measurement', *Sensors and Actuators A: Physical*, Vol. 263, pp. 614–621.

Weiss, B., Clemens, Z., Bódizs, R. and Halász, P. (2011), 'Comparison of fractal and power spectral EEG features: Effects of topography and sleep stages', *Brain Research Bulletin*, Vol. 84, No. 6, pp. 359–375.

Weissman, D.H., Roberts, K.C., Visscher, K.M. and Woldorff, M.G. (2006), 'The neural bases of momentary lapses in attention', *Nature neuroscience*, Vol. 9, No. 7, pp. 971–978.

Welch, P. (1967), 'The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms', *IEEE Transactions on Audio and Electroacoustics*, Vol. 15, No. 2, pp. 70–73.

Westhall, E., Rosén, I., Rossetti, A.O., van Rootselaar, A.F., Kjaer, T.W., Horn, J., Ullén, S., Friberg, H., Nielsen, N. and Cronberg, T. (2014), 'Electroencephalography (EEG) for neurological prognostication after cardiac arrest and targeted temperature management; rationale and study design', *BMC Neurology*, Vol. 14, pp. 159–166.

Williams, H.L., Granda, A.M., Jones, R.C., Lubin, A. and Armington, J.C. (1962), 'EEG frequency and finger pulse volume as predictors of reaction time during sleep loss', *Electroencephalography and Clinical Neurophysiology*, Vol. 14, No. 1, pp. 64–70.

Wipf, D. and Nagarajan, S. (2009), 'A unified Bayesian framework for MEG/EEG source imaging', *NeuroImage*, Vol. 44, No. 3, pp. 947–966.

Woldorff, M.G., Hazlett, C.J., Fichtenholtz, H.M., Weissman, D.H., Dale, A.M. and Song, A.W. (2004), 'Functional parcellation of attentional control regions of the brain', *Journal of Cognitive Neuroscience*, Vol. 16, No. 1, pp. 149–165.

World Health Organization (2015), 'Global status report on road safety', `http://apps.who.int/iris/bitstream/10665/189242/1/9789241565066_eng.pdf`. Accessed October 2017.

Wu, W., Chen, Z., Gao, X., Li, Y., Brown, E. and Gao, S. (2015), 'Probabilistic common spatial patterns for multichannel EEG analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 3, pp. 639–653.

Wu, W., Nagarajan, S. and Chen, Z. (2016), 'Bayesian machine learning: EEG/MEG signal processing measurements', *IEEE Signal Processing Magazine*, Vol. 33, No. 1, pp. 14–36.

Wu, W., Chen, Z., Gao, S. and Brown, E. (2009), 'A probabilistic framework for learning robust common spatial patterns', In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4658–4661.

Wu, W., Chen, Z., Gao, S. and Brown, E.N. (2011), 'A hierarchical Bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG', *NeuroImage*, Vol. 56, No. 4, pp. 1929–1945.

Wu, W., Wu, C., Gao, S., Liu, B., Li, Y. and Gao, X. (2014), 'Bayesian estimation of ERP components from multicondition and multichannel EEG', *NeuroImage*, Vol. 88, pp. 319–339.

Xanthopoulos, P., Pardalos, P. and Trafalis, T. (2012), *Robust Data Mining*, Springer, New York.

Xu, L., Johnson, T.D., Nichols, T.E. and Nee, D.E. (2009), 'Modeling inter-subject variability in fMRI activation location: A Bayesian hierarchical spatial model', *Biometrics*, Vol. 65, No. 4, pp. 1041–1051.

Yeo, M.V., Li, X., Shen, K. and Wilder-Smith, E.P. (2009), 'Can SVM be used for automatic EEG detection of drowsiness during car driving?', *Safety Science*, Vol. 47, No. 1, pp. 115–124.

Yin, L., Ge, Y., Xiao, K., Wang, X. and Quan, X. (2013), 'Feature selection for high-dimensional imbalanced data', *Neurocomputing*, Vol. 105, pp. 3–11.

Yoon, H.J. and Chung, S.Y. (2013), 'EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm', *Computers in Biology and Medicine*, Vol. 43, No. 12, pp. 2230–2237.

Yu, K. and Gales, M.J.F. (2007), 'Bayesian adaptive inference and adaptive training', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 6, pp. 1932–1943.

Yu, X., Chum, P. and Sim, K.B. (2014), 'Analysis the effect of PCA for feature reduction in non-stationary EEG based motor imagery of BCI system', *Optik - International Journal for Light and Electron Optics*, Vol. 125, No. 3, pp. 1498–1502.

Yuan, Q., Zhou, W., Li, S. and Cai, D. (2011), 'Epileptic EEG classification based on extreme learning machine and nonlinear features', *Epilepsy Research*, Vol. 96, No. 1–2, pp. 29–38.

Zhang, G., Yau, K.K., Zhang, X. and Li, Y. (2016a), 'Traffic accidents involving fatigue driving and their extent of casualties', *Accident Analysis & Prevention*, Vol. 87, pp. 34–42.

Zhang, J. and Yang, Y. (2003), 'Robustness of regularized linear classification methods in text categorization', In *Proceedings of The 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 190–197.

Zhang, L., Guindani, M. and Vannucci, M. (2015), 'Bayesian models for functional magnetic resonance imaging data analysis', *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 7, No. 1, pp. 21–41.

Zhang, T. and Chan, A.H. (2014), 'Sleepiness and the risk of road accidents for professional drivers: A systematic review and meta-analysis of retrospective studies', *Safety Science*, Vol. 70, pp. 180–188.

Zhang, Y., Zhou, G., Jin, J., Zhao, Q., Wang, X. and Cichocki, A. (2016b), 'Sparse Bayesian classification of EEG for brain-computer interface', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 27, No. 11, pp. 2256–2267.

Zhang, Y. and Hua, C. (2015), 'Driver fatigue recognition based on facial expression analysis using local binary patterns', *Optik - International Journal for Light and Electron Optics*, Vol. 126, No. 23, pp. 4501–4505.

Zhao, J. and Jiang, Q. (2006), 'Probabilistic PCA for t distributions', *Neurocomputing*, Vol. 69, No. 16, pp. 2217–2226.

Zhao, J. and Yu, P.L. (2009), 'A note on variational Bayesian factor analysis', *Neural Networks*, Vol. 22, No. 7, pp. 988–997.

Zhao, Q., Meng, D., Xu, Z., Zuo, W. and Yan, Y. (2015a), 'L_1-norm low-rank matrix factorization by variational Bayesian method', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 4, pp. 825–839.

Zhao, Q., Zhang, L. and Cichocki, A. (2015b), 'Bayesian CP factorization of incomplete tensors with automatic rank determination', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 9, pp. 1751–1763.

Zhao, Q., Zhou, G., Zhang, L., Cichocki, A. and Amari, S.I. (2016), 'Bayesian robust tensor factorization for incomplete multiway data', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 27, No. 4, pp. 736–748.

Zhao, Q., Peng, H., Hu, B., Liu, Q., Liu, L., Qi, Y. and Li, L. (2010), 'Improving individual identification in security check with an EEG based biometric solution', In Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong and J. Huang (editors), *Brain Informatics: International Conference Proceedings*, Springer, Berlin, Heidelberg, pp. 145–155.

Zhu, H., Leung, H. and He, Z. (2013), 'A variational Bayesian approach to robust sensor fusion based on student-t distribution', *Information Sciences*, Vol. 221, pp. 201–214.

Zou, Q., Xie, S., Lin, Z., Wu, M. and Ju, Y. (2016), 'Finding the best classification threshold in imbalanced classification', *Big Data Research*, Vol. 5, pp. 2–8.

Zumer, J.M., Attias, H.T., Sekihara, K. and Nagarajan, S.S. (2007), 'A probabilistic algorithm integrating source localization and noise suppression for MEG and EEG data', *NeuroImage*, Vol. 37, No. 1, pp. 102–115.

Appendix A

BAYESIAN ROBUST FACTOR ANALYSIS

A.1 VARIATIONAL BAYESIAN DERIVATION

Approximate posterior distribution of α (Equations (7.15), (7.23) and (7.24))

$$
\begin{aligned}
\ln\left(q\left(\boldsymbol{\alpha}\right)\right) &= \left\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle + \ln\left(p\left(\boldsymbol{\alpha}\right)\right) + \text{const} \\
&= \sum_{k=1}^{K} \left\{ \left(\frac{D}{2} + a_{\alpha} - 1\right) \ln\left(\alpha_k\right) - \left(b_{\alpha} + \frac{\left\langle \mathbf{w}_k^{\top} \mathbf{w}_k \right\rangle}{2}\right) \alpha_k \right\} + \text{const}, \\
q\left(\boldsymbol{\alpha}\right) &= \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,\middle|\, \tilde{a}_{\alpha}, \tilde{b}_{\alpha,k}\right), \\
\tilde{a}_{\alpha} &= \frac{D}{2} + a_{\alpha}, \\
\tilde{b}_{\alpha,k} &= b_{\alpha} + \frac{\left\langle \mathbf{w}_k^{\top} \mathbf{w}_k \right\rangle}{2}.
\end{aligned}
$$

Approximate posterior distribution of W (Equations (7.19), (7.29) and (7.30))

$$
\begin{aligned}
\ln\left(q\left(\mathbf{W}\right)\right) &= \sum_{s=1}^{S} \left\langle \ln\left(p\left(\mathbf{X}_s \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}_s, \mathbf{W}\right)\right) + \ln\left(p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle + \text{const} \\
&= -\frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\langle \mathbf{z}_{s,n}^{\top} \mathbf{W}^{\top} \boldsymbol{\Psi} \mathbf{W} \mathbf{z}_{s,n} - 2\mathbf{z}_{s,n}^{\top} \mathbf{W}^{\top} \boldsymbol{\Psi}\left(\mathbf{x}_{s,n} - \boldsymbol{\mu}\right) \right\rangle \\
&\quad - \frac{1}{2} \operatorname{tr}\left(\mathbf{W} \left\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \right\rangle \mathbf{W}^{\top}\right) + \text{const} \\
&= -\frac{1}{2} \sum_{d=1}^{D} \left\{ \mathbf{w}_{d,\cdot} \left(\left\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \right\rangle + \left\langle \psi_d \right\rangle \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^{\top} \right\rangle \right) \mathbf{w}_{d,\cdot}^{\top} \right. \\
&\quad \left. - 2\mathbf{w}_{d,\cdot} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left(\left\langle \mathbf{z}_{s,n} \right\rangle \left(\left\langle \psi_d \right\rangle x_{s,n,d} - \left\langle \psi_d \mu_d \right\rangle \right) \right) \right\} + \text{const}, \\
q\left(\mathbf{W}\right) &= \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{w}_{d,\cdot}^{\top} \,\middle|\, \tilde{\mathbf{w}}_{w,d}, \tilde{\Sigma}_{w,d}\right),
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\Sigma}_{w,d} &= \left(\left\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \right\rangle + \left\langle \psi_d \right\rangle \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^{\top} \right\rangle \right)^{-1}, \\
\tilde{\mathbf{w}}_{w,d} &= \tilde{\Sigma}_{w,d} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \Big(\left\langle \mathbf{z}_{s,n} \right\rangle \Big(\left\langle \psi_d \right\rangle x_{s,n,d} - \left\langle \psi_d \mu_d \right\rangle \Big) \Big).
\end{aligned}
$$

Approximate posterior distributions of μ and Ψ (Equations (7.16), (7.17) and (7.25)–(7.28))

$$
\begin{aligned}
\ln\left(q\left(\boldsymbol{\mu}, \boldsymbol{\Psi}\right)\right) &= \sum_{s=1}^{S} \left\langle \ln\left(p\left(\mathbf{X}_s \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}_s, \mathbf{W}\right)\right) \right\rangle + \ln\left(p\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) + \ln\left(p\left(\boldsymbol{\Psi}\right)\right) + \text{const} \\
&= \sum_{d=1}^{D} \left\{ \left(a_{\psi} + \frac{\sum_{s=1}^{S} N_s + 1}{2} - 1 \right) \ln\left(\psi_d\right) - b_{\psi,d}\psi_d - \frac{\beta_0 \psi_d}{2} \mu_d^2 \right\} \\
&\qquad - \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\{ \left\langle \Big(\mathbf{x}_{s,n} - \left(\mathbf{W}\mathbf{z}_{s,n} + \boldsymbol{\mu}\right) \Big)^{\top} \boldsymbol{\Psi} \Big(\mathbf{x}_{s,n} - \left(\mathbf{W}\mathbf{z}_{s,n} + \boldsymbol{\mu}\right) \Big) \right\rangle \right\} + \text{const} \\
&= \sum_{d=1}^{D} \Bigg\{ \left(a_{\psi} + \frac{\sum_{s=1}^{S} N_s + 1}{2} - 1 \right) \ln\left(\psi_d\right) - \Bigg(\sum_{s=1}^{S} \sum_{n=1}^{N_s} \Bigg(\frac{x_{s,n,d}^2}{2} \\
&\qquad + \frac{1}{2} \operatorname{tr}\Big(\left\langle \mathbf{w}_{d,.}^{\top} \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^{\top} \right\rangle \Big) - x_{s,n,d} \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle \Bigg) + b_{\psi,d} \Bigg) \psi_d \\
&\qquad - \frac{\psi_d}{2} \Bigg(\left(\beta_0 + \sum_{s=1}^{S} N_s \right) \mu_d^2 - 2\mu_d \Bigg(\sum_{s=1}^{S} \sum_{n=1}^{N_s} \Big(x_{s,n,d} - \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle \Big) \Bigg) \Bigg) \Bigg\} + \text{const} \\
&= \sum_{d=1}^{D} \Bigg\{ \ln\left(q\left(\mu_d \,\middle|\, \psi_d\right)\right) + \left(a_{\psi} + \frac{\sum_{s=1}^{S} N_s}{2} - 1 \right) \ln\left(\psi_d\right) - \psi_d \Bigg(b_{\psi} - \frac{\beta_{\mu}}{2} \tilde{m}_{\mu,d}^2 \\
&\qquad + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \Big(x_{s,n,d}^2 - 2x_{s,n,d} \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle + \operatorname{tr}\Big(\left\langle \mathbf{w}_{d,.}^{\top} \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^{\top} \right\rangle \Big) \Big) \Bigg) \Bigg\} + \text{const} \\
&= \ln\left(q\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) + \ln\left(q\left(\boldsymbol{\Psi}\right)\right), \\
q\left(\boldsymbol{\mu}, \boldsymbol{\Psi}\right) &= \prod_{d=1}^{D} \left(\mathcal{N}\left(\mu_d \,\middle|\, \tilde{\mu}_d, \left(\beta_{\mu}\psi_d\right)^{-1}\right) \mathcal{G}\left(\psi_d \,\middle|\, \tilde{a}_{\psi}, \tilde{b}_{\psi,d}\right) \right), \\
\beta_{\mu} &= \sum_{s=1}^{S} N_s + \beta_0, \\
\tilde{\mu}_d &= \beta_{\mu}^{-1} \left(\sum_{s=1}^{S} \sum_{n=1}^{N_s} \Big(x_{s,n,d} - \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle \Big) \right), \\
\tilde{a}_{\psi} &= a_{\psi} + \frac{1}{2} \sum_{s=1}^{S} N_s, \\
\tilde{b}_{\psi,d} &= b_{\psi} - \frac{\beta_{\mu}}{2} \tilde{m}_{\mu,d}^2 + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \Big(x_{s,n,d}^2 - 2x_{s,n,d} \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle + \operatorname{tr}\Big(\left\langle \mathbf{w}_{d,.}^{\top} \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^{\top} \right\rangle \Big) \Big).
\end{aligned}
$$

Approximate posterior distribution of Z (Equations (7.20)–(7.22))

$$\begin{aligned}
\ln\left(q\left(\mathbf{Z}_s\right)\right) &= \left\langle \ln\left(p\left(\mathbf{X}_s \mid \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}_s, \mathbf{W}\right)\right) + \ln\left(p\left(\mathbf{Z}_s \mid \boldsymbol{\Lambda}\right)\right) \right\rangle + \text{const} \\
&= \sum_{n=1}^{N_s} \left\{ -\frac{1}{2} \left(\left\langle \left(\mathbf{x}_{s,n} - \mathbf{W}\mathbf{z}_{s,n} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Psi} \left(\mathbf{x}_{s,n} - \mathbf{W}\mathbf{z}_{s,n} - \boldsymbol{\mu}\right) + \mathbf{z}_{s,n}^\top \boldsymbol{\Lambda} \mathbf{z}_{s,n} \right\rangle \right) \right\} + \text{const} \\
&= \sum_{n=1}^{N_s} \left\{ -\frac{1}{2} \left(\mathbf{z}_{s,n}^\top \left(\left\langle \mathbf{W}^\top \boldsymbol{\Psi} \mathbf{W} \right\rangle + \left\langle \boldsymbol{\Lambda} \right\rangle \right) \mathbf{z}_{s,n} - 2\mathbf{z}_{s,n}^\top \left\langle \mathbf{W}^\top \right\rangle \left(\left\langle \boldsymbol{\Psi} \right\rangle \mathbf{x}_{s,n} - \left\langle \boldsymbol{\Psi}\boldsymbol{\mu} \right\rangle \right) \right) \right\} + \text{const}, \\
q\left(\mathbf{Z}\right) &= \prod_{s=1}^{S} \prod_{n=1}^{N_s} \mathcal{N}\left(\mathbf{z}_{s,n} \mid \tilde{\mathbf{m}}_{z,s,n}, \tilde{\Sigma}_z\right), \\
\tilde{\Sigma}_z &= \left(\left\langle \mathbf{W}^\top \boldsymbol{\Psi} \mathbf{W} \right\rangle + \left\langle \boldsymbol{\Lambda} \right\rangle \right)^{-1}, \\
\tilde{\mathbf{m}}_{z,s,n} &= \tilde{\Sigma}_z \left\langle \mathbf{W}^\top \right\rangle \left(\left\langle \boldsymbol{\Psi} \right\rangle \mathbf{x}_{s,n} - \left\langle \boldsymbol{\Psi}\boldsymbol{\mu} \right\rangle \right).
\end{aligned}$$

Approximate posterior distribution of Λ (Equations (7.18), (7.31) and (7.32))

$$\begin{aligned}
\ln\left(q\left(\boldsymbol{\Lambda}\right)\right) &= \sum_{s=1}^{S} \left\langle \ln\left(p\left(\mathbf{Z}_s \mid \boldsymbol{\Lambda}\right)\right) \right\rangle + \ln\left(p\left(\boldsymbol{\Lambda}\right)\right) + \text{const} \\
&= \sum_{k=1}^{K} \left\{ \left(a_\lambda + \sum_{s=1}^{S} \frac{N_s}{2} - 1 \right) \ln\left(\lambda_k\right) + \left(b_\lambda + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\langle z_{s,n,k}^2 \right\rangle \right) \lambda_k \right\} + \text{const}, \\
q\left(\boldsymbol{\Lambda}\right) &= \prod_{k=1}^{K} \mathcal{G}\left(\lambda_k \mid \tilde{a}_\lambda, \tilde{b}_{\lambda,k}\right), \\
\tilde{a}_\lambda &= a_\lambda + \frac{1}{2} \sum_{s=1}^{S} N_s, \\
\tilde{b}_{\lambda,k} &= b_\lambda + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\langle z_{s,n,k}^2 \right\rangle.
\end{aligned}$$

A.2 LOWER BOUND

$$\begin{aligned}
\mathcal{L} = &\left\langle \ln\left(p\left(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}, \mathbf{W}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\mathbf{W} \mid \boldsymbol{\alpha}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\boldsymbol{\alpha}\right)\right) \right\rangle \\
&+ \left\langle \ln\left(p\left(\mathbf{Z} \mid \boldsymbol{\Lambda}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\boldsymbol{\Lambda}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\boldsymbol{\mu} \mid \boldsymbol{\Psi}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\boldsymbol{\Psi}\right)\right) \right\rangle \\
&- \left\langle \ln\left(q\left(\mathbf{W}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\alpha}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\mathbf{Z}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\Lambda}\right)\right) \right\rangle \\
&- \left\langle \ln\left(q\left(\boldsymbol{\mu} \mid \boldsymbol{\Psi}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\Psi}\right)\right) \right\rangle,
\end{aligned}$$

$$\begin{aligned}\big\langle \ln\left(p\left(\mathbf{X}\,\middle|\,\boldsymbol{\mu},\boldsymbol{\Psi},\mathbf{Z},\mathbf{W}\right)\right)\big\rangle =\ & \frac{1}{2}\sum_{s=1}^{S} N_s\Big(\ln\left(|\boldsymbol{\Psi}|\right) - D\ln\left(2\pi\right)\Big) - \frac{1}{2}\sum_{s=1}^{S}\sum_{n=1}^{N_s}\Big(\mathbf{x}_{s,n}^{\top}\langle\boldsymbol{\Psi}\rangle\mathbf{x}_{s,n} \\ & + \langle\boldsymbol{\mu}^{\top}\boldsymbol{\Psi}\boldsymbol{\mu}\rangle + \operatorname{tr}\left(\langle\mathbf{W}^{\top}\boldsymbol{\Psi}\mathbf{W}\rangle\langle\mathbf{z}_{s,n}\mathbf{z}_{s,n}^{\top}\rangle\right) \\ & - 2\mathbf{x}_{s,n}^{\top}\langle\boldsymbol{\Psi}\rangle\Big(\langle\mathbf{W}\rangle\langle\mathbf{z}_{s,n}\rangle + \langle\boldsymbol{\mu}\rangle\Big) + 2\langle\mathbf{z}_{s,n}\rangle^{\top}\langle\mathbf{W}^{\top}\rangle\langle\boldsymbol{\Psi}\rangle\langle\boldsymbol{\mu}\rangle\Big),\end{aligned}$$

$$\big\langle \ln\left(p\left(\mathbf{W}\,\middle|\,\boldsymbol{\alpha}\right)\right)\big\rangle = \sum_{k=1}^{K}\left(\frac{D}{2}\Big(\big\langle\ln\left(\alpha_k\right)\big\rangle - \ln\left(2\pi\right)\Big) - \frac{\langle\alpha_k\rangle}{2}\langle\mathbf{w}_k^{\top}\mathbf{w}_k\rangle\right),$$

$$\big\langle \ln\left(q\left(\mathbf{W}\right)\right)\big\rangle = -\frac{KD}{2}\ln\left(2\pi\right) - \frac{KD}{2} - \frac{1}{2}\sum_{d=1}^{D}\Big(\ln\left(\big|\tilde{\Sigma}_{w,d}\right|)\Big),$$

$$\big\langle \ln\left(p\left(\boldsymbol{\alpha}\right)\right)\big\rangle = K\Big(-\ln\left(\Gamma\left(a_{\alpha}\right)\right) + a_{\alpha}\ln\left(b_{\alpha}\right)\Big) + \sum_{k=1}^{K}\Big(\left(a_{\alpha}-1\right)\big\langle\ln\left(\alpha_k\right)\big\rangle - b_{\alpha}\langle\alpha_k\rangle\Big),$$

$$\big\langle \ln\left(q\left(\boldsymbol{\alpha}\right)\right)\big\rangle = -K\ln\left(\Gamma\left(\tilde{a}_{\alpha}\right)\right) + \sum_{k=1}^{K}\Big(\tilde{a}_{\alpha}\ln\left(\tilde{b}_{\alpha,k}\right) + \left(\tilde{a}_{\alpha}-1\right)\big\langle\ln\left(\alpha_k\right)\big\rangle - \tilde{b}_{\alpha,k}\langle\alpha_k\rangle\Big),$$

$$\big\langle \ln\left(p\left(\mathbf{Z}\,\middle|\,\boldsymbol{\Lambda}\right)\right)\big\rangle = \sum_{s=1}^{S}\frac{N_s}{2}\Big(\ln\left(|\boldsymbol{\Lambda}|\right) - K\ln\left(2\pi\right)\Big) - \frac{1}{2}\sum_{s=1}^{S}\sum_{n=1}^{N_s}\operatorname{tr}\left(\langle\mathbf{z}_{s,n}\mathbf{z}_{s,n}^{\top}\rangle\langle\boldsymbol{\Lambda}\rangle\right),$$

$$\big\langle \ln\left(q\left(\mathbf{Z}\right)\right)\big\rangle = -\sum_{s=1}^{S}\frac{N_s}{2}\Big(\ln\left(\big|\tilde{\Sigma}_z\right|) + K\ln\left(2\pi\right)\Big) - \frac{1}{2}\sum_{s=1}^{S}N_sK,$$

$$\big\langle \ln\left(p\left(\boldsymbol{\Lambda}\right)\right)\big\rangle = K\Big(-\ln\left(\Gamma\left(a_{\lambda}\right)\right) + a_{\lambda}\ln\left(b_{\lambda}\right)\Big) + \sum_{k=1}^{K}\Big(\left(a_{\lambda}-1\right)\big\langle\ln\left(\lambda_k\right)\big\rangle - b_{\lambda}\langle\lambda_k\rangle\Big),$$

$$\big\langle \ln\left(q\left(\boldsymbol{\Lambda}\right)\right)\big\rangle = -K\ln\left(\Gamma\left(\tilde{a}_{\lambda}\right)\right) + \sum_{k=1}^{K}\Big(\tilde{a}_{\lambda}\ln\left(\tilde{b}_{\lambda,k}\right) + \left(\tilde{a}_{\lambda}-1\right)\big\langle\ln\left(\lambda_k\right)\big\rangle - \tilde{b}_{\lambda,k}\langle\lambda_k\rangle\Big),$$

$$\big\langle \ln\left(p\left(\boldsymbol{\mu}\,\middle|\,\boldsymbol{\Psi}\right)\right)\big\rangle = \frac{D}{2}\Big(\ln\left(\beta_0\right) - \ln\left(2\pi\right)\Big) + \sum_{d=1}^{D}\left(\frac{\big\langle\ln\left(\psi_d\right)\big\rangle}{2} - \frac{\beta_0}{2}\langle\psi_d\mu_d^2\rangle\right),$$

$$\big\langle \ln\left(q\left(\boldsymbol{\mu}\,\middle|\,\boldsymbol{\Psi}\right)\right)\big\rangle = \frac{D}{2}\Big(\ln\left(\beta_{\mu}\right) - \ln\left(2\pi\right)\Big) + \sum_{d=1}^{D}\left(\frac{\big\langle\ln\left(\psi_d\right)\big\rangle}{2} - \frac{\beta_{\mu}}{2}\right),$$

$$\big\langle \ln\left(p\left(\boldsymbol{\Psi}\right)\right)\big\rangle = D\Big(-\ln\left(\Gamma\left(a_{\psi}\right)\right) + a_{\psi}\ln\left(b_{\psi}\right)\Big) + \sum_{d=1}^{D}\Big(\left(a_{\psi}-1\right)\big\langle\ln\left(\psi_d\right)\big\rangle - b_{\psi}\langle\psi_d\rangle\Big),$$

$$\big\langle \ln\left(q\left(\boldsymbol{\Psi}\right)\right)\big\rangle = -D\ln\left(\Gamma\left(\tilde{a}_{\psi}\right)\right) + \sum_{d=1}^{D}\Big(\tilde{a}_{\psi}\ln\left(\tilde{b}_{\psi,d}\right) + \left(\tilde{a}_{\psi}-1\right)\big\langle\ln\left(\psi_d\right)\big\rangle - \tilde{b}_{\psi,d}\langle\psi_d\rangle\Big).$$

Appendix B

BAYESIAN MULTI-SUBJECT ROBUST FACTOR ANALYSIS

B.1 VARIATIONAL BAYESIAN DERIVATION

Approximate posterior distribution of α (Equations (8.8), (8.21) and (8.22))

$$\begin{aligned}
\ln\left(q\left(\boldsymbol{\alpha}\right)\right) &= \left\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle + \ln\left(p\left(\boldsymbol{\alpha}\right)\right) + \text{const} \\
&= \sum_{k=1}^{K} \left\{ \left(\frac{D}{2} + a_{\alpha} - 1\right) \ln\left(\alpha_k\right) - \left(b_{\alpha} + \frac{\left\langle \mathbf{w}_k^{\top}\mathbf{w}_k \right\rangle}{2}\right) \alpha_k \right\} + \text{const}, \\
q\left(\boldsymbol{\alpha}\right) &= \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,\middle|\, \tilde{a}_{\alpha}, \tilde{b}_{\alpha,k}\right), \\
\tilde{a}_{\alpha} &= a_{\alpha} + \frac{D}{2}, \\
\tilde{b}_{\alpha,k} &= b_{\alpha} + \frac{\left\langle \mathbf{w}_k^{\top}\mathbf{w}_k \right\rangle}{2}.
\end{aligned}$$

Approximate posterior distribution of W (Equations (8.11), (8.17) and (8.18))

$$\begin{aligned}
\ln\left(q\left(\mathbf{W}\right)\right) &= \left\langle \sum_{s=1}^{S} \ln\left(p\left(\mathbf{X}_s \,\middle|\, \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_s, \mathbf{W}\right)\right) + \ln\left(p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle + \text{const} \\
&= -\frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\langle \mathbf{z}_{s,n}^{\top}\mathbf{W}^{\top}\boldsymbol{\Psi}_s\mathbf{W}\mathbf{z}_{s,n} - 2\mathbf{z}_{s,n}^{\top}\mathbf{W}^{\top}\boldsymbol{\Psi}_s\left(\mathbf{x}_{s,n} - \boldsymbol{\mu}_s\right) \right\rangle \\
&\quad - \frac{1}{2}\operatorname{tr}\left(\mathbf{W}\left\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \right\rangle \mathbf{W}^{\top}\right) + \text{const} \\
&= -\frac{1}{2} \sum_{d=1}^{D} \left\{ \mathbf{w}_{d,.} \left(\left\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \right\rangle + \sum_{s=1}^{S} \left\langle \psi_{s,d} \right\rangle \sum_{n=1}^{N_s} \left\langle \mathbf{z}_{s,n}\mathbf{z}_{s,n}^{\top} \right\rangle \right) \mathbf{w}_{d,.}^{\top} \right. \\
&\quad \left. - 2\mathbf{w}_{d,.} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left(\left\langle \mathbf{z}_{s,n} \right\rangle \left(\left\langle \psi_{s,d} \right\rangle x_{s,n,d} - \left\langle \psi_{s,d}\mu_{s,d} \right\rangle \right) \right) \right\} + \text{const}, \\
q\left(\mathbf{W}\right) &= \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{w}_{d,.}^{\top} \,\middle|\, \tilde{\mathbf{m}}_{w,d}, \tilde{\Sigma}_{w,d}\right),
\end{aligned}$$

$$
\begin{aligned}
\tilde{\Sigma}_{w,d} &= \left(\left\langle \operatorname{diag}\left(\boldsymbol{\alpha}\right) \right\rangle + \sum_{s=1}^{S} \left\langle \psi_{s,d} \right\rangle \sum_{n=1}^{N_s} \left\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^{\top} \right\rangle \right)^{-1}, \\
\tilde{\mathbf{m}}_{w,d} &= \tilde{\Sigma}_{w,d} \sum_{s=1}^{S} \left\langle \psi_{s,d} \right\rangle \sum_{n=1}^{N_s} \Big(\left\langle \mathbf{z}_{s,n} \right\rangle \left(x_{s,n,d} - \left\langle \mu_{s,d} \right\rangle \right) \Big).
\end{aligned}
$$

Approximate posterior distributions of μ and Ψ (Equations (8.9), (8.10) and (8.13)–(8.16))

$$
\begin{aligned}
\ln\left(q\left(\boldsymbol{\mu}_s, \boldsymbol{\Psi}_s\right)\right) &= \left\langle \ln\left(p\left(\mathbf{X}_s \,\middle|\, \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_s, \mathbf{W}\right)\right) \right\rangle + \ln\left(p\left(\boldsymbol{\mu}_s \,\middle|\, \boldsymbol{\Psi}_s\right)\right) + \ln\left(p\left(\boldsymbol{\Psi}_s\right)\right) + \text{const} \\
&= \sum_{d=1}^{D} \left\{ \left(a_{\psi} + \frac{N_s + 1}{2} - 1 \right) \ln\left(\psi_{s,d}\right) - b_{\psi,s,d} - \frac{\beta_0 \psi_{s,d}}{2} \left(\mu_{s,d} - m_{\mu,d}\right)^2 \right\} \\
&\quad - \frac{1}{2} \sum_{n=1}^{N_s} \left\{ \left\langle \Big(\mathbf{x}_{s,n} - \left(\mathbf{W}\mathbf{z}_{s,n} + \boldsymbol{\mu}_s\right) \Big)^{\top} \boldsymbol{\Psi}_s \Big(\mathbf{x}_{s,n} - \left(\mathbf{W}\mathbf{z}_{s,n} + \boldsymbol{\mu}_s\right) \Big) \right\rangle \right\} + \text{const} \\
&= \sum_{d=1}^{D} \Bigg\{ \left(a_{\psi} + \frac{N_s + 1}{2} - 1 \right) \ln\left(\psi_{s,d}\right) - \Bigg(\sum_{n=1}^{N_s} \Bigg(\frac{x_{s,n,d}^2}{2} + \frac{1}{2} \operatorname{tr}\left(\left\langle \mathbf{w}_{d,.}^{\top} \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^{\top} \right\rangle \right) \\
&\quad - x_{s,n,d} \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle \Bigg) + \frac{\beta_0}{2} m_{\mu,d}^2 + b_{\psi} \Bigg) \psi_{s,d} - \frac{\psi_{s,d}}{2} \Bigg(\left(\beta_0 + N_s\right) \mu_{s,d}^2 \\
&\quad - 2\mu_{s,d} \Bigg(\beta_0 m_{\mu,d} + \sum_{n=1}^{N_s} \left(x_{s,n,d} - \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle \right) \Bigg) \Bigg) \Bigg\} + \text{const} \\
&= \sum_{d=1}^{D} \Bigg\{ \ln\left(q\left(\mu_{s,d} \,\middle|\, \psi_{s,d}\right)\right) + \left(a_{\psi} + \frac{N_s}{2} - 1 \right) \ln\left(\psi_{s,d}\right) - \psi_{s,d} \Bigg(b_{\psi} + \frac{\beta_0}{2} m_{\mu,d}^2 - \frac{\beta_{\mu,s}}{2} \tilde{m}_{\mu,s,d}^2 \\
&\quad + \frac{1}{2} \sum_{n=1}^{N_s} \Big(x_{s,n,d}^2 - 2x_{s,n,d} \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle + \operatorname{tr}\left(\left\langle \mathbf{w}_{d,.}^{\top} \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^{\top} \right\rangle \right) \Big) \Bigg) \Bigg\} + \text{const} \\
&= \ln\left(q\left(\boldsymbol{\mu}_s \,\middle|\, \boldsymbol{\Psi}_s\right)\right) + \ln\left(q\left(\boldsymbol{\Psi}_s\right)\right), \\
q\left(\boldsymbol{\mu}_s \,\middle|\, \boldsymbol{\Psi}_s\right) &= \prod_{d=1}^{D} \mathcal{N}\left(\mu_{s,d} \,\middle|\, \tilde{m}_{\mu,s,d}, \left(\beta_{\mu,s} \psi_{s,d}\right)^{-1} \right), \\
q\left(\boldsymbol{\Psi}_s\right) &= \prod_{d=1}^{D} \mathcal{G}\left(\psi_{s,d} \,\middle|\, \tilde{a}_{\psi,s}, \tilde{b}_{\psi,s,d} \right), \\
\beta_{\mu,s} &= N_s + \beta_0, \\
\tilde{m}_{\mu,s,d} &= \frac{1}{\beta_{\mu,s}} \left(\beta_0 m_{\mu,d} + \sum_{n=1}^{N_s} \left(x_{s,n,d} - \left\langle \mathbf{w}_{d,.} \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle \right) \right), \\
\tilde{a}_{\psi,s} &= a_{\psi} + \frac{N_s}{2},
\end{aligned}
$$

$$\begin{aligned}\tilde{b}_{\psi,s,d} &= b_\psi + \frac{\beta_0}{2} m_{\mu,d}^2 - \frac{\beta_{\mu,s}}{2} \tilde{m}_{\mu,s,d}^2 \\ &\quad + \frac{1}{2} \sum_{n=1}^{N_s} \left(x_{s,n,d}^2 - 2 x_{s,n,d} \langle \mathbf{w}_{d,.} \rangle \langle \mathbf{z}_{s,n} \rangle + \mathrm{tr}\left(\langle \mathbf{w}_{d,.}^\top \mathbf{w}_{d,.} \rangle \langle \mathbf{z}_{s,n} \mathbf{z}_{s,n}^\top \rangle \right) \right).\end{aligned}$$

Approximate posterior distribution of Z (Equations (8.5)–(8.7))

$$\begin{aligned}\ln\left(q\left(\mathbf{Z}_s\right)\right) &= \left\langle \ln\left(p\left(\mathbf{X}_s \mid \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_s, \mathbf{W}\right)\right) + \ln\left(p\left(\mathbf{Z}_s \mid \boldsymbol{\Lambda}\right)\right) \right\rangle + \mathrm{const} \\ &= \sum_{n=1}^{N_s} \left\{ -\frac{1}{2} \left(\left\langle \left(\mathbf{x}_{s,n} - \mathbf{W}\mathbf{z}_{s,n} - \boldsymbol{\mu}_s\right)^\top \boldsymbol{\Psi}_s \left(\mathbf{x}_{s,n} - \mathbf{W}\mathbf{z}_{s,n} - \boldsymbol{\mu}_s\right) + \mathbf{z}_{s,n}^\top \boldsymbol{\Lambda} \mathbf{z}_{s,n} \right\rangle \right) \right\} + \mathrm{const} \\ &= \sum_{n=1}^{N_s} \left\{ -\frac{1}{2} \left(\mathbf{z}_{s,n}^\top \left(\langle \mathbf{W}^\top \boldsymbol{\Psi}_s \mathbf{W} \rangle + \langle \boldsymbol{\Lambda} \rangle \right) \mathbf{z}_{s,n} - 2 \mathbf{z}_{s,n}^\top \langle \mathbf{W}^\top \rangle \left(\langle \boldsymbol{\Psi}_s \rangle \mathbf{x}_{s,n} - \langle \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \rangle \right) \right) \right\} + \mathrm{const}\end{aligned}$$

$$\begin{aligned}q\left(\mathbf{Z}_s\right) &= \prod_{n=1}^{N_s} \mathcal{N}\left(\mathbf{z}_{s,n} \mid \tilde{\mathbf{m}}_{z,s,n}, \tilde{\Sigma}_{z,s}\right), \\ \tilde{\Sigma}_{z,s} &= \left(\langle \mathbf{W}^\top \boldsymbol{\Psi}_s \mathbf{W} \rangle + \langle \boldsymbol{\Lambda} \rangle \right)^{-1}, \\ \tilde{\mathbf{m}}_{z,s,n} &= \tilde{\Sigma}_{z,s} \langle \mathbf{W}^\top \rangle \left(\langle \boldsymbol{\Psi}_s \rangle \mathbf{x}_{s,n} - \langle \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \rangle \right)\end{aligned}$$

Approximate posterior distribution of Λ (Equations (8.12), (8.19) and (8.20))

$$\begin{aligned}\ln\left(q\left(\boldsymbol{\Lambda}\right)\right) &= \sum_{s=1}^{S} \left\langle \ln\left(p\left(\mathbf{Z}_s \mid \boldsymbol{\Lambda}\right)\right) \right\rangle + \ln\left(p\left(\boldsymbol{\Lambda}\right)\right) + \mathrm{const} \\ &= \sum_{k=1}^{K} \left\{ \left(a_\lambda + \frac{1}{2} \sum_{s=1}^{S} N_s - 1 \right) \ln\left(\lambda_k\right) + \left(b_\lambda + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \langle z_{s,n,k}^2 \rangle \right) \lambda_k \right\} + \mathrm{const},\end{aligned}$$

$$\begin{aligned}q\left(\boldsymbol{\Lambda}\right) &= \prod_{k=1}^{K} \mathcal{G}\left(\lambda_k \mid \tilde{a}_\lambda, \tilde{b}_{\lambda,k}\right), \\ \tilde{a}_\lambda &= a_\lambda + \frac{1}{2} \sum_{s=1}^{S} N_s, \\ \tilde{b}_{\lambda,k} &= b_\lambda + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \langle z_{s,n,k}^2 \rangle.\end{aligned}$$

B.2 LOWER BOUND

$$
\begin{aligned}
\mathcal{L} = {} & \big\langle \ln\left(p\left(\mathbf{X} \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}, \mathbf{W}\right)\right) \big\rangle + \big\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \boldsymbol{\alpha}\right)\right) \big\rangle + \big\langle \ln\left(p\left(\boldsymbol{\alpha}\right)\right) \big\rangle \\
& + \big\langle \ln\left(p\left(\mathbf{Z} \,\middle|\, \boldsymbol{\Lambda}\right)\right) \big\rangle + \big\langle \ln\left(p\left(\boldsymbol{\Lambda}\right)\right) \big\rangle + \big\langle \ln\left(p\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \big\rangle + \big\langle \ln\left(p\left(\boldsymbol{\Psi}\right)\right) \big\rangle \\
& - \big\langle \ln\left(q\left(\mathbf{W}\right)\right) \big\rangle - \big\langle \ln\left(q\left(\boldsymbol{\alpha}\right)\right) \big\rangle - \big\langle \ln\left(q\left(\mathbf{Z}\right)\right) \big\rangle - \big\langle \ln\left(q\left(\boldsymbol{\Lambda}\right)\right) \big\rangle \\
& - \big\langle \ln\left(q\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \big\rangle - \big\langle \ln\left(q\left(\boldsymbol{\Psi}\right)\right) \big\rangle, \\
\big\langle \ln\left(p\left(\mathbf{X} \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}, \mathbf{W}\right)\right) \big\rangle = {} & \sum_{s=1}^{S} \sum_{n=1}^{N_s} \Bigg\{ \frac{-D}{2} \ln\left(2\pi\right) + \frac{1}{2} \big\langle \ln\left(\left|\boldsymbol{\Psi}_s\right|\right) \big\rangle - \frac{1}{2} \Bigg(\mathbf{x}_{s,n}^{\top} \langle \boldsymbol{\Psi}_s \rangle \mathbf{x}_{s,n} \\
& + \langle \mathbf{z}_{s,n}^{\top} \mathbf{W}^{\top} \boldsymbol{\Psi}_s \mathbf{W} \mathbf{z}_{s,n} \rangle + \langle \boldsymbol{\mu}_s^{\top} \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \rangle - 2\mathbf{x}_{s,n}^{\top} \Big(\langle \boldsymbol{\Psi}_s \rangle \langle \mathbf{W} \rangle \langle \mathbf{z}_{s,n} \rangle \\
& + \langle \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \rangle \Big) + 2 \langle \mathbf{z}_{s,n}^{\top} \rangle \langle \mathbf{W}^{\top} \rangle \langle \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \rangle \Bigg) \Bigg\}, \\
\big\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \alpha\right)\right) \big\rangle = {} & \sum_{k=1}^{K} \left(\frac{D}{2} \Big(\big\langle \ln\left(\alpha_k\right) \big\rangle - \ln\left(2\pi\right) \Big) - \frac{\langle \alpha_k \rangle}{2} \langle \mathbf{w}_k^{\top} \mathbf{w}_k \rangle \right), \\
\big\langle \ln\left(q\left(\mathbf{W}\right)\right) \big\rangle = {} & -\frac{KD}{2} - \frac{KD}{2} \ln\left(2\pi\right) - \frac{1}{2} \sum_{d=1}^{D} \ln\left(\left|\tilde{\Sigma}_{w,d}\right|\right), \\
\big\langle \ln\left(p\left(\boldsymbol{\alpha}\right)\right) \big\rangle = {} & \sum_{k=1}^{K} \Big(-\ln\left(\Gamma\left(a_{\alpha}\right)\right) + a_{\alpha} \ln\left(b_{\alpha}\right) + \left(a_{\alpha} - 1\right) \big\langle \ln\left(\alpha_k\right) \big\rangle - b_{\alpha} \langle \alpha_k \rangle \Big), \\
\big\langle \ln\left(q\left(\boldsymbol{\alpha}\right)\right) \big\rangle = {} & \sum_{k=1}^{K} \Big(-\ln\left(\Gamma\left(\tilde{a}_{\alpha}\right)\right) + \tilde{a}_{\alpha} \ln\left(\tilde{b}_{\alpha,k}\right) + \left(\tilde{a}_{\alpha} - 1\right) \big\langle \ln\left(\alpha_k\right) \big\rangle - \tilde{b}_{\alpha,k} \langle \alpha_k \rangle \Big), \\
\big\langle \ln\left(p\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \big\rangle = {} & \sum_{s=1}^{S} \Bigg(\frac{D}{2} \Big(\ln\left(\beta_0\right) - \ln\left(2\pi\right) \Big) \\
& + \sum_{d=1}^{D} \left(\frac{\big\langle \ln\left(\psi_{s,d}\right) \big\rangle}{2} - \frac{\beta_0}{2} \left\langle \psi_{s,d} \left(\mu_{s,d} - m_{\mu,d}\right)^2 \right\rangle \right) \Bigg), \\
\big\langle \ln\left(q\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \big\rangle = {} & \sum_{s=1}^{S} \left\{ \frac{D}{2} \Big(\ln\left(\beta_{\mu,s}\right) - \ln\left(2\pi\right) - 1 \Big) + \sum_{d=1}^{D} \frac{\big\langle \ln\left(\psi_{s,d}\right) \big\rangle}{2} \right\}, \\
\big\langle \ln\left(p\left(\boldsymbol{\Psi}\right)\right) \big\rangle = {} & \sum_{s=1}^{S} \sum_{d=1}^{D} \Big(-\ln\left(\Gamma\left(a_{\psi}\right)\right) + a_{\psi} \ln\left(b_{\psi,d}\right) + \left(a_{\psi} - 1\right) \big\langle \ln\left(\psi_{s,d}\right) \big\rangle - b_{\psi,d} \langle \psi_{s,d} \rangle \Big), \\
\big\langle \ln\left(q\left(\boldsymbol{\Psi}\right)\right) \big\rangle = {} & \sum_{s=1}^{S} \sum_{d=1}^{D} \Big(-\ln\left(\Gamma\left(\tilde{a}_{\psi,s}\right)\right) + \tilde{a}_{\psi,s} \ln\left(\tilde{b}_{\psi,s,d}\right) \\
& + \left(\tilde{a}_{\psi,s} - 1\right) \big\langle \ln\left(\psi_{s,d}\right) \big\rangle - \tilde{b}_{\psi,s,d} \langle \psi s, d \rangle \Big), \\
\big\langle \ln\left(p\left(\mathbf{Z} \,\middle|\, \boldsymbol{\Lambda}\right)\right) \big\rangle = {} & -\frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \sum_{k=1}^{K} \Big(\ln\left(2\pi\right) - \big\langle \ln\left(\lambda_k\right) \big\rangle + \langle \lambda_k \rangle \langle \mathbf{z}_{s,n,k}^2 \rangle \Big),
\end{aligned}
$$

$$\begin{aligned}
\big\langle \ln\left(q\left(\mathbf{Z}\right)\right)\big\rangle &= -\sum_{s=1}^{S} \frac{N_s}{2}\left(\ln\left(\left|\tilde{\Sigma}_{z,s}\right|\right) + K\ln\left(2\pi\right) + K\right),\\
\big\langle \ln\left(p\left(\mathbf{\Lambda}\right)\right)\big\rangle &= \sum_{k=1}^{K}\Big(-\ln\left(\Gamma\left(a_\lambda\right)\right) + a_\lambda \ln\left(b_\lambda\right) + \left(a_\lambda - 1\right)\big\langle \ln\left(\lambda_k\right)\big\rangle - b_\lambda\big\langle\lambda_k\big\rangle\Big),\\
\big\langle \ln\left(q\left(\mathbf{\Lambda}\right)\right)\big\rangle &= \sum_{k=1}^{K}\left(-\ln\left(\Gamma\left(\tilde{a}_\lambda\right)\right) + \tilde{a}_\lambda \ln\left(\tilde{b}_{\lambda,k}\right) + \left(\tilde{a}_\lambda - 1\right)\big\langle \ln\left(\lambda_k\right)\big\rangle - \tilde{b}_{\lambda,k}\big\langle\lambda_k\big\rangle\right).
\end{aligned}$$

Appendix C

BAYESIAN HIERARCHICAL MULTI-SUBJECT ROBUST FACTOR ANALYSIS

C.1 VARIATIONAL BAYESIAN DERIVATION

Approximate posterior distributions of $\mathbf{M}_{\mathbf{w}}$ and α (Equations (9.23)–(9.26), (9.32) and (9.33))

$$
\begin{aligned}
\ln\left(q\left(\mathbf{M}_{\mathbf{w}}, \alpha\right)\right) &= \left\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \mathbf{M}_{\mathbf{w}}, \alpha\right)\right) \right\rangle + \ln\left(p\left(\mathbf{M}_{\mathbf{w}} \mid \alpha\right)\right) + \ln\left(p\left(\alpha\right)\right) + \text{const} \\
&= \sum_{k=1}^{K} \left\{ \left(\frac{SD}{2} + a_{\alpha} - 1\right) \ln\left(\alpha_k\right) - \left(b_{\alpha} + \frac{\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k}^{\top} \mathbf{w}_{s,k} \right\rangle}{2}\right) \alpha_k \right. \\
&\qquad \left. + \frac{D}{2} \ln\left(\alpha_k\right) - \frac{\alpha_k}{2} \left(\left(S + \beta_w\right) \mathbf{m}_{\mathbf{w},k}^{\top} \mathbf{m}_{\mathbf{w},k} - 2\mathbf{m}_{\mathbf{w},k}^{\top} \left(\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k} \right\rangle \right) \right) \right\} + \text{const} \\
&= \ln\left(q\left(\mathbf{M}_{\mathbf{w}} \mid \alpha\right)\right) + \ln\left(q\left(\alpha\right)\right), \\
q\left(\mathbf{M}_{\mathbf{w}} \mid \alpha\right) &= \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{m}_{w,k} \,\middle|\, \tilde{\mathbf{m}}_{w,k}, \left(\tilde{\beta}_w \alpha_k \mathbf{I}\right)^{-1}\right), \\
q\left(\alpha\right) &= \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \mid \tilde{a}_{\alpha}, \tilde{b}_{\alpha,k}\right), \\
\tilde{\beta}_w &= S + \beta_w, \\
\tilde{\mathbf{m}}_{\mathbf{w},k} &= \tilde{\beta}_w^{-1} \left(\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k} \right\rangle \right), \\
\tilde{a}_{\alpha} &= a_{\alpha} + \frac{SD}{2}, \\
\tilde{b}_{\alpha,k} &= b_{\alpha} + \frac{\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k}^{\top} \mathbf{w}_{s,k} \right\rangle - \tilde{\beta}_w \tilde{\mathbf{m}}_{\mathbf{w},k}^{\top} \tilde{\mathbf{m}}_{\mathbf{w},k}}{2}.
\end{aligned}
$$

Approximate posterior distribution of W (Equations (9.21), (9.22) and (9.31))

$$\begin{aligned}
\ln\left(q\left(\mathbf{W}\right)\right) &= \sum_{s=1}^{S}\left(\left\langle \ln\left(p\left(\mathbf{X}_s \,\middle|\, \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_s, \mathbf{W}_s\right)\right) + \ln\left(p\left(\mathbf{W}_s \,\middle|\, \mathbf{M}_{\mathbf{w}}, \alpha\right)\right)\right\rangle\right) + \text{const} \\
&= -\frac{1}{2}\sum_{s=1}^{S}\left\{ \operatorname{tr}\left(\mathbf{W}_s\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle\mathbf{W}_s^{\top} - 2\mathbf{W}_s\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle\left\langle\mathbf{M}_{\mathbf{w}}^{\top}\right\rangle\right)\right. \\
&\quad \left. + \sum_{n=1}^{N_s}\left\langle \mathbf{z}_{s,n}^{\top}\mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\mathbf{W}_s\mathbf{z}_{s,n} - 2\mathbf{z}_{s,n}^{\top}\mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\left(\mathbf{x}_{s,n} - \boldsymbol{\mu}_s\right)\right\rangle\right\} + \text{const} \\
&= -\frac{1}{2}\sum_{s=1}^{S}\sum_{d=1}^{D}\left\{ \mathbf{w}_{s,d,.}\left(\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle + \left\langle\psi_{s,d}\right\rangle\sum_{n=1}^{N_s}\left\langle \mathbf{z}_{s,n}\mathbf{z}_{s,n}^{\top}\right\rangle\right)\mathbf{w}_{s,d,.}^{\top}\right. \\
&\quad \left. - 2\mathbf{w}_{s,d,.}\left(\sum_{n=1}^{N_s}\left\langle\mathbf{z}_{s,n}\right\rangle\left(\left\langle\psi_{s,d}\right\rangle x_{s,n,d} - \left\langle\psi_{s,d}\mu_{s,d}\right\rangle\right) + \left\langle \operatorname{diag}\left(\alpha\right)\right\rangle\left\langle\mathbf{m}_{\mathbf{w},d,.}^{\top}\right\rangle\right)\right\} + \text{const}, \\
q\left(\mathbf{W}\right) &= \prod_{s=1}^{S}\prod_{d=1}^{D}\mathcal{N}\left(\mathbf{w}_{s,d,.}^{\top}\,\middle|\,\tilde{\mathbf{w}}_{s,d}, \tilde{\Sigma}_{w,s,d}\right), \\
\tilde{\Sigma}_{w,s,d} &= \left(\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle + \left\langle\psi_{s,d}\right\rangle\sum_{n=1}^{N_s}\left\langle\mathbf{z}_{s,n}\mathbf{z}_{s,n}^{\top}\right\rangle\right)^{-1}, \\
\tilde{\mathbf{w}}_{s,d} &= \tilde{\Sigma}_{w,s,d}\left(\sum_{n=1}^{N_s}\left\langle\mathbf{z}_{s,n}\right\rangle\left(\left\langle\psi_{s,d}\right\rangle x_{s,n,d} - \left\langle\psi_{s,d}\mu_{s,d}\right\rangle\right) + \left\langle \operatorname{diag}\left(\alpha\right)\right\rangle\left\langle\mathbf{m}_{\mathbf{w},d,.}^{\top}\right\rangle\right).
\end{aligned}$$

Approximate posterior distribution of Z (Equations (9.14)–(9.16))

$$\begin{aligned}
\ln\left(q\left(\mathbf{Z}_s\right)\right) &= \left\langle \ln\left(p\left(\mathbf{X}_s \,\middle|\, \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_s, \mathbf{W}_s\right)\right) + \ln\left(p\left(\mathbf{Z}_s \,\middle|\, \boldsymbol{\Lambda}\right)\right)\right\rangle + \text{const} \\
&= \sum_{n=1}^{N_s}\left\{ -\frac{1}{2}\left(\left\langle \left(\mathbf{x}_{s,n} - \mathbf{W}_s\mathbf{z}_{s,n} - \boldsymbol{\mu}_s\right)^{\top}\boldsymbol{\Psi}_s\left(\mathbf{x}_{s,n} - \mathbf{W}\mathbf{z}_{s,n} - \boldsymbol{\mu}_s\right) + \mathbf{z}_{s,n}{}^{\top}\boldsymbol{\Lambda}\mathbf{z}_{s,n}\right\rangle\right)\right\} + \text{const} \\
&= \sum_{n=1}^{N_s}\left\{ -\frac{1}{2}\left(\mathbf{z}_{s,n}{}^{\top}\left(\left\langle\mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\mathbf{W}_s\right\rangle + \left\langle\boldsymbol{\Lambda}\right\rangle\right)\mathbf{z}_{s,n} - 2\mathbf{z}_{s,n}{}^{\top}\left\langle\mathbf{W}_s^{\top}\right\rangle\left(\left\langle\boldsymbol{\Psi}_s\right\rangle\mathbf{x}_{s,n} - \left\langle\boldsymbol{\Psi}_s\boldsymbol{\mu}_s\right\rangle\right)\right)\right\} + \text{const}, \\
q\left(\mathbf{Z}\right) &= \prod_{s=1}^{S}\prod_{n=1}^{N_s}\mathcal{N}\left(\mathbf{z}_{s,n}\,\middle|\,\tilde{\mathbf{m}}_{z,s,n}, \tilde{\Sigma}_{z,s}\right), \\
\tilde{\Sigma}_{z,s} &= \left(\left\langle\mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\mathbf{W}_s\right\rangle + \left\langle\boldsymbol{\Lambda}\right\rangle\right)^{-1}, \\
\tilde{\mathbf{m}}_{z,s,n} &= \tilde{\Sigma}_{z,s}\left\langle\mathbf{W}_s^{\top}\right\rangle\left(\left\langle\boldsymbol{\Psi}_s\right\rangle\mathbf{x}_{s,n} - \left\langle\boldsymbol{\Psi}_s\boldsymbol{\mu}_s\right\rangle\right).
\end{aligned}$$

Approximate posterior distributions of μ and Ψ (Equations (9.17)–(9.20), (9.29) and (9.30))

$$
\begin{aligned}
\ln\left(q\left(\boldsymbol{\mu}_s, \boldsymbol{\Psi}_s\right)\right) &= \left\langle \ln\left(p\left(\mathbf{X}_s \,\middle|\, \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_s, \mathbf{W}_s\right)\right)\right\rangle + \ln\left(p\left(\boldsymbol{\mu}_s \,\middle|\, \boldsymbol{\Psi}_s\right)\right) + \ln\left(p\left(\boldsymbol{\Psi}_s\right)\right) + \text{const} \\
&= \sum_{d=1}^{D}\left\{\left(a_\psi + \frac{N_s+1}{2} - 1\right)\ln\left(\psi_{s,d}\right) - b_{\psi,d}\psi_{s,d} - \frac{\beta_\mu \psi_{s,d}}{2}\left(\mu_{s,d} - m_{\mu,d}\right)^2\right\} \\
&\quad - \frac{1}{2}\sum_{n=1}^{N_s}\left\{\left\langle\left(\mathbf{x}_{s,n} - \left(\mathbf{W}_s\mathbf{z}_{s,n} + \boldsymbol{\mu}_s\right)\right)^\top \boldsymbol{\Psi}_s\left(\mathbf{x}_{s,n} - \left(\mathbf{W}_s\mathbf{z}_{s,n} + \boldsymbol{\mu}_s\right)\right)\right\rangle\right\} + \text{const} \\
&= \sum_{d=1}^{D}\Bigg\{\left(a_\psi + \frac{N_s+1}{2} - 1\right)\ln\left(\psi_{s,d}\right) - \Bigg(\sum_{n=1}^{N_s}\Bigg(\frac{x_{s,n,d}^2}{2} + \frac{1}{2}\operatorname{tr}\left(\left\langle\mathbf{w}_{s,d,.}^\top\mathbf{w}_{s,d,.}\right\rangle\left\langle\mathbf{z}_{s,n}\mathbf{z}_{s,n}^\top\right\rangle\right) \\
&\quad - x_{s,n,d}\left\langle\mathbf{w}_{s,d,.}\right\rangle\left\langle\mathbf{z}_{s,n}\right\rangle + \frac{\beta_\mu}{2}m_{\mu,d}^2\Bigg) + b_{\psi,d}\Bigg)\psi_{s,d} \\
&\quad - \frac{\psi_{s,d}}{2}\left(\left(\beta_\mu + N_s\right)\mu_{s,d}^2 - 2\mu_{s,d}\left(\beta_\mu m_{\mu,d} + \sum_{n=1}^{N_s}\left(x_{s,n,d} - \left\langle\mathbf{w}_{s,d,.}\right\rangle\left\langle\mathbf{z}_{s,n}\right\rangle\right)\right)\right)\Bigg\} + \text{const} \\
&= \sum_{d=1}^{D}\Bigg\{\ln\left(q\left(\mu_{s,d} \,\middle|\, \psi_{s,d}\right)\right) + \left(a_\psi + \frac{N_s}{2} - 1\right)\ln\left(\psi_{s,d}\right) - \psi_{s,d}\Bigg(b_{\psi,d} + \frac{\beta_\mu}{2}m_{\mu,d}^2 - \frac{\tilde{\beta}_{s,\mu}}{2}\tilde{m}_{\mu,s,d}^2 \\
&\quad + \frac{1}{2}\sum_{n=1}^{N_s}\left(x_{s,n,d}^2 - 2x_{s,n,d}\left\langle\mathbf{w}_{s,d,.}\right\rangle\left\langle\mathbf{z}_{s,n}\right\rangle + \operatorname{tr}\left(\left\langle\mathbf{w}_{s,d,.}^\top\mathbf{w}_{s,d,.}\right\rangle\left\langle\mathbf{z}_{s,n}\mathbf{z}_{s,n}^\top\right\rangle\right)\right)\Bigg)\Bigg\} + \text{const} \\
&= \ln\left(q\left(\boldsymbol{\mu}_s \,\middle|\, \boldsymbol{\Psi}_s\right)\right) + \ln\left(q\left(\boldsymbol{\Psi}_s\right)\right) \\
q\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right) &= \prod_{s=1}^{S}\prod_{d=1}^{D}\mathcal{N}\left(\mu_{s,d} \,\middle|\, \tilde{m}_{\mu,s,d}, \left(\tilde{\beta}_{\mu,s}\psi_{s,d}\right)^{-1}\right), \\
q\left(\boldsymbol{\Psi}\right) &= \prod_{s=1}^{S}\prod_{d=1}^{D}\mathcal{G}\left(\psi_{s,d} \,\middle|\, \tilde{a}_{\psi,d}, \tilde{b}_{\psi,s,d}\right), \\
\tilde{\beta}_{\mu,s} &= N_s + \beta_\mu, \\
\tilde{m}_{\mu,s,d} &= \frac{1}{\tilde{\beta}_{\mu,s}}\left(\beta_\mu m_{\mu,d} + \sum_{n=1}^{N_s}\left(x_{s,n,d} - \left\langle\mathbf{w}_{s,d,.}\right\rangle\left\langle\mathbf{z}_{s,n}\right\rangle\right)\right), \\
\tilde{a}_{\psi,s} &= a_\psi + \frac{N_s}{2}, \\
\tilde{b}_{\psi,s,d} &= b_{\psi,d} + \frac{\beta_\mu}{2}m_{\mu,d}^2 - \frac{\tilde{\beta}_{s,\mu}}{2}\tilde{m}_{\mu,s,d}^2 + \frac{1}{2}\sum_{n=1}^{N_s}\Bigg(x_{s,n,d}^2 - 2x_{s,n,d}\left\langle\mathbf{w}_{s,d,.}\right\rangle\left\langle\mathbf{z}_{s,n}\right\rangle \\
&\quad + \operatorname{tr}\left(\left\langle\mathbf{w}_{s,d,.}^\top\mathbf{w}_{s,d,.}\right\rangle\left\langle\mathbf{z}_{s,n}\mathbf{z}_{s,n}^\top\right\rangle\right)\Bigg).
\end{aligned}
$$

Approximate posterior distribution of Λ (Equations (9.27), (9.28) and (9.34))

$$
\ln\left(q\left(\boldsymbol{\Lambda}\right)\right) = \sum_{s=1}^{S}\left\langle\ln\left(p\left(\mathbf{Z}_s \,\middle|\, \boldsymbol{\Lambda}\right)\right)\right\rangle + \ln\left(p\left(\boldsymbol{\Lambda}\right)\right) + \text{const}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} \left\{ \left(a_\lambda + \frac{1}{2} \sum_{s=1}^{S} N_s - 1 \right) \ln\left(\lambda_k\right) + \left(b_\lambda + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\langle z_{s,n,k}^2 \right\rangle \right) \lambda_k \right\} + \text{const}, \\
q\left(\boldsymbol{\Lambda}\right) &= \prod_{k=1}^{K} \mathcal{G}\left(\lambda_k \,\middle|\, \tilde{a}_\lambda, \tilde{b}_{\lambda,k}\right), \\
\tilde{a}_\lambda &= a_\lambda + \frac{1}{2} \sum_{s=1}^{S} N_s, \\
\tilde{b}_{\lambda,k} &= b_\lambda + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \left\langle z_{s,n,k}^2 \right\rangle.
\end{aligned}
$$

C.2 LOWER BOUND

$$
\begin{aligned}
\mathcal{L} = &\left\langle \ln\left(p\left(\mathbf{X} \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}, \mathbf{W}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \mathbf{M}_{\mathbf{w}}, \boldsymbol{\alpha}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\mathbf{M}_{\mathbf{w}} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle \\
&+ \left\langle \ln\left(p\left(\boldsymbol{\alpha}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\mathbf{Z} \,\middle|\, \boldsymbol{\Lambda}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\boldsymbol{\Lambda}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \right\rangle \\
&+ \left\langle \ln\left(p\left(\boldsymbol{\Psi}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\mathbf{W}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\mathbf{M}_{\mathbf{w}} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\alpha}\right)\right) \right\rangle \\
&- \left\langle \ln\left(q\left(\mathbf{Z}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\Lambda}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\Psi}\right)\right) \right\rangle,
\end{aligned}
$$

$$
\begin{aligned}
\left\langle \ln\left(p\left(\mathbf{X} \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}, \mathbf{W}\right)\right) \right\rangle = &\sum_{s=1}^{S} \sum_{n=1}^{N_s} \Bigg\{ \frac{-D}{2} \ln\left(2\pi\right) + \frac{1}{2} \left\langle \ln\left(\left|\boldsymbol{\Psi}_s\right|\right) \right\rangle - \frac{1}{2} \Big(\mathbf{x}_{s,n}^\top \left\langle \boldsymbol{\Psi}_s \right\rangle \mathbf{x}_{s,n} \\
&+ \left\langle \mathbf{z}_{s,n}^\top \mathbf{W}_s^\top \boldsymbol{\Psi}_s \mathbf{W}_s \mathbf{z}_{s,n} \right\rangle + \left\langle \boldsymbol{\mu}_s^\top \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \right\rangle + 2 \left\langle \mathbf{z}_{s,n}^\top \right\rangle \left\langle \mathbf{W}_s^\top \right\rangle \left\langle \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \right\rangle \\
&- 2 \mathbf{x}_{s,n}^\top \Big(\left\langle \boldsymbol{\Psi}_s \right\rangle \left\langle \mathbf{W}_s \right\rangle \left\langle \mathbf{z}_{s,n} \right\rangle + \left\langle \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \right\rangle \Big) \Big) \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\left\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \mathbf{M}_{\mathbf{w}}, \boldsymbol{\alpha}\right)\right) \right\rangle = &\sum_{s=1}^{S} \sum_{k=1}^{K} \Bigg(\frac{D}{2} \Big(\left\langle \ln\left(\alpha_k\right) \right\rangle - \ln\left(2\pi\right) \Big) - \frac{1}{2} \Big(\left\langle \alpha_k \right\rangle \left\langle \mathbf{w}_{s,k}^\top \mathbf{w}_{s,k} \right\rangle \\
&- 2 \left\langle \alpha_k \right\rangle \left\langle \mathbf{w}_{s,k} \right\rangle \left\langle \mathbf{m}_{\mathbf{w},k} \right\rangle + \left\langle \alpha_k \mathbf{m}_{\mathbf{w},k}^\top \mathbf{m}_{\mathbf{w},k} \right\rangle \Big) \Bigg),
\end{aligned}
$$

$$
\left\langle \ln\left(q\left(\mathbf{W}\right)\right) \right\rangle = \sum_{s=1}^{S} \sum_{d=1}^{D} \left(-\frac{K}{2} - \frac{K}{2} \ln\left(2\pi\right) - \frac{1}{2} \ln\left(\left|\tilde{\Sigma}_{w,s,d}\right|\right) \right),
$$

$$
\left\langle \ln\left(p\left(\mathbf{M}_{\mathbf{w}} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle = \sum_{k=1}^{K} \left(\frac{D}{2} \Big(\ln\left(\beta_w\right) + \left\langle \ln\left(\alpha_k\right) \right\rangle - \ln\left(2\pi\right) \Big) - \frac{\beta_w}{2} \left\langle \alpha_k \mathbf{m}_{\mathbf{w},k}^\top \mathbf{m}_{\mathbf{w},k} \right\rangle \right),
$$

$$
\left\langle \ln\left(q\left(\mathbf{M}_{\mathbf{w}} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle = \sum_{k=1}^{K} \left(\frac{D}{2} \Big(\ln\left(\tilde{\beta}_w\right) + \left\langle \ln\left(\alpha_k\right) \right\rangle - \ln\left(2\pi\right) \Big) - \frac{D}{2} \right),
$$

$$
\left\langle \ln\left(p\left(\boldsymbol{\alpha}\right)\right) \right\rangle = \sum_{k=1}^{K} \Big(-\ln\left(\Gamma\left(a_\alpha\right)\right) + a_\alpha \ln\left(b_\alpha\right) + \left(a_\alpha - 1\right) \left\langle \ln\left(\alpha_k\right) \right\rangle - b_\alpha \left\langle \alpha_k \right\rangle \Big),
$$

$$
\left\langle \ln\left(q\left(\boldsymbol{\alpha}\right)\right) \right\rangle = \sum_{k=1}^{K} \Big(-\ln\left(\Gamma\left(\tilde{a}_\alpha\right)\right) + \tilde{a}_\alpha \ln\left(\tilde{b}_{\alpha,k}\right) + \left(\tilde{a}_\alpha - 1\right) \left\langle \ln\left(\alpha_k\right) \right\rangle - \tilde{b}_{\alpha,k} \left\langle \alpha_k \right\rangle \Big),
$$

$$
\begin{aligned}
\big\langle \ln\left(p\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \big\rangle &= \sum_{s=1}^{S} \left(\frac{D}{2}\Big(\ln\left(\beta_{\mu}\right) - \ln\left(2\pi\right)\Big) + \sum_{d=1}^{D} \left(\frac{\big\langle \ln\left(\psi_{s,d}\right) \big\rangle}{2} - \frac{\beta_{\mu}}{2}\Big(\big\langle \psi_{s,d}\mu_{s,d} \big\rangle^{2} \right.\right. \\
&\qquad \left.\left. - 2\big\langle \psi_{s,d} \big\rangle \big\langle \mu_{s,d} \big\rangle m_{0,d} + \big\langle \psi_{s,d} \big\rangle m_{0,d}^{2} \Big) \right) \right) \\
&= \sum_{s=1}^{S} \left(\frac{D}{2}\Big(\ln\left(\beta_{\mu}\right) - \ln\left(2\pi\right) - \frac{\beta_{\mu}}{\tilde{\beta}_{\mu,s}} \Big) \right. \\
&\qquad \left. + \sum_{d=1}^{D} \frac{1}{2}\Big(\big\langle \ln\left(\psi_{s,d}\right) \big\rangle - \beta_{\mu} \big\langle \psi_{s,d} \big\rangle \left(\tilde{m}_{\mu,s,d} - m_{0,d}\right)^{2} \Big) \right), \\
\big\langle \ln\left(q\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \big\rangle &= \sum_{s=1}^{S} \left(\frac{D}{2}\Big(\ln\left(\tilde{\beta}_{\mu,s}\right) - \ln\left(2\pi\right) - 1 \Big) + \sum_{d=1}^{D} \frac{\big\langle \ln\left(\psi_{s,d}\right) \big\rangle}{2} \right), \\
\big\langle \ln\left(p\left(\boldsymbol{\Psi}\right)\right) \big\rangle &= \sum_{s=1}^{S} \sum_{d=1}^{D} \Big(- \ln\left(\Gamma\left(a_{\psi}\right)\right) + a_{\psi} \ln\left(b_{\psi,d}\right) + \left(a_{\psi} - 1\right) \big\langle \ln\left(\psi_{s,d}\right) \big\rangle - b_{\psi,d} \big\langle \psi_{s,d} \big\rangle \Big), \\
\big\langle \ln\left(q\left(\boldsymbol{\Psi}\right)\right) \big\rangle &= \sum_{s=1}^{S} \sum_{d=1}^{D} \Big(- \ln\left(\Gamma\left(\tilde{a}_{\psi,s}\right)\right) + \tilde{a}_{\psi,s} \ln\left(\tilde{b}_{\psi,s,d}\right) \\
&\qquad + \left(\tilde{a}_{\psi,s} - 1\right) \big\langle \ln\left(\psi_{s,d}\right) \big\rangle - \tilde{b}_{\psi,s,d} \big\langle \psi_{s,d} \big\rangle \Big), \\
\big\langle \ln\left(p\left(\mathbf{Z} \,\middle|\, \boldsymbol{\Lambda}\right)\right) \big\rangle &= -\frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_s} \sum_{k=1}^{K} \Big(\ln\left(2\pi\right) - \big\langle \ln\left(\lambda_k\right) \big\rangle + \big\langle \lambda_k \big\rangle \big\langle \mathbf{z}_{s,n,k}^{2} \big\rangle \Big), \\
\big\langle \ln\left(q\left(\mathbf{Z}\right)\right) \big\rangle &= - \sum_{s=1}^{S} \frac{N_s}{2} \Big(\ln\left(\left|\tilde{\Sigma}_{z,s}\right|\right) + K \ln\left(2\pi\right) + K \Big), \\
\big\langle \ln\left(p\left(\boldsymbol{\Lambda}\right)\right) \big\rangle &= \sum_{k=1}^{K} \Big(- \ln\left(\Gamma\left(a_{\lambda}\right)\right) + a_{\lambda} \ln\left(b_{\lambda}\right) + \left(a_{\lambda} - 1\right) \big\langle \ln\left(\lambda_k\right) \big\rangle - b_{\lambda} \big\langle \lambda_k \big\rangle \Big), \\
\big\langle \ln\left(q\left(\boldsymbol{\Lambda}\right)\right) \big\rangle &= \sum_{k=1}^{K} \Big(- \ln\left(\Gamma\left(\tilde{a}_{\lambda}\right)\right) + \tilde{a}_{\lambda} \ln\left(\tilde{b}_{\lambda,k}\right) + \left(\tilde{a}_{\lambda} - 1\right) \big\langle \ln\left(\lambda_k\right) \big\rangle - \tilde{b}_{\lambda,k} \big\langle \lambda_k \big\rangle \Big).
\end{aligned}
$$

Appendix D

BAYESIAN HIERARCHICAL MULTI-SUBJECT ROBUST JOINT MATRIX FACTORIZATION

D.1 VARIATIONAL BAYESIAN DERIVATION

Approximate posterior distributions of $\mathbf{M}_\mathbf{w}$ and α (Equations (10.18)–(10.21), (10.30) and (10.31))

$$\begin{aligned}
\ln\left(q\left(\mathbf{M}_\mathbf{w}, \boldsymbol{\alpha}\right)\right) &= \left\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \mathbf{M}_\mathbf{w}, \boldsymbol{\alpha}\right)\right) \right\rangle + \ln\left(p\left(\mathbf{M}_\mathbf{w} \,\middle|\, \boldsymbol{\alpha}\right)\right) + \ln\left(p\left(\boldsymbol{\alpha}\right)\right) + \text{const} \\
&= \sum_{k=1}^{K} \left\{ \left(\frac{SD}{2} + a_\alpha - 1\right) \ln\left(\alpha_k\right) - \left(b_\alpha + \frac{1}{2}\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k}^\top \mathbf{w}_{s,k} \right\rangle\right) \alpha_k \right\} \\
&\quad + \sum_{k=1}^{K} \left\{ \frac{D}{2}\ln\left(\alpha_k\right) - \frac{\alpha_k}{2}\left(\left(\beta_w + S\right) \mathbf{m}_{\mathbf{w},k}^\top \mathbf{m}_{\mathbf{w},k} - 2\mathbf{m}_{\mathbf{w},k}^\top \left(\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k} \right\rangle \right)\right)\right\} + \text{const} \\
&= \sum_{k=1}^{K} \left\{ \left(\frac{SD}{2} + a_\alpha - 1\right) \ln\left(\alpha_k\right) - \left(b_\alpha + \frac{1}{2}\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k}^\top \mathbf{w}_{s,k} \right\rangle - \frac{1}{2}\tilde{\mathbf{m}}_{w,k}^\top \tilde{\mathbf{m}}_{w,k}\right) \alpha_k \right\} \\
&\quad + \ln\left(q\left(\mathbf{M}_\mathbf{w} \,\middle|\, \boldsymbol{\alpha}\right)\right) + \text{const}, \\
q\left(\mathbf{M}_\mathbf{w} \,\middle|\, \boldsymbol{\alpha}\right) &= \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{m}_{w,k} \,\middle|\, \tilde{\mathbf{m}}_{w,k}, \left(\tilde{\beta}_w \alpha_k\right)^{-1} \mathbf{I}\right), \\
q\left(\boldsymbol{\alpha}\right) &= \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,\middle|\, \tilde{a}_\alpha, \tilde{b}_{\alpha,k}\right), \\
\tilde{\beta}_w &= \beta_w + S, \\
\tilde{\mathbf{m}}_{\mathbf{w},k} &= \tilde{\beta}_w^{-1} \left(\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k} \right\rangle \right), \\
\tilde{a}_\alpha &= \frac{SD}{2} + a_\alpha, \\
\tilde{b}_{\alpha,k} &= b_\alpha + \frac{1}{2}\sum_{s=1}^{S} \left\langle \mathbf{w}_{s,k}^\top \mathbf{w}_{s,k} \right\rangle - \frac{1}{2}\tilde{\mathbf{m}}_{w,k}^\top \tilde{\mathbf{m}}_{w,k}.
\end{aligned}$$

Approximate posterior distribution of W (Equations (10.28), (10.29) and (10.35))

$$
\begin{aligned}
\ln\left(q\left(\mathbf{W}_s\right)\right) &= \left\langle \sum_{c=1}^{C} \ln\left(p\left(\mathbf{X}_{s,c} \,\middle|\, \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_{s,c}, \mathbf{W}_s\right)\right) + \ln\left(p\left(\mathbf{W}_s \,\middle|\, \mathbf{M}_{\mathbf{w}}, \alpha\right)\right) \right\rangle + \text{const} \\
&= -\frac{1}{2}\operatorname{tr}\left(\mathbf{W}_s\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle \mathbf{W}_s^{\top} - 2\mathbf{W}_s\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle \mathbf{M}_{\mathbf{w}}^{\top}\right) \\
&\quad - \frac{1}{2}\sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left\langle \mathbf{z}_{s,c,n}^{\top}\mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\mathbf{W}_s\mathbf{z}_{s,c,n} - 2\mathbf{z}_{s,c,n}^{\top}\mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\left(\mathbf{x}_{s,c,n} - \boldsymbol{\mu}_s\right)\right\rangle + \text{const} \\
&= -\frac{1}{2}\sum_{d=1}^{D}\left\{ \mathbf{w}_{s,d,\cdot}\left(\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle + \left\langle \psi_{s,d}\right\rangle \sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left\langle \mathbf{z}_{s,c,n}\mathbf{z}_{s,c,n}^{\top}\right\rangle\right)\mathbf{w}_{s,d,\cdot}^{\top} \right. \\
&\quad \left. - 2\mathbf{w}_{s,d,\cdot}\left(\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle \mathbf{m}_{\mathbf{w},d,\cdot}^{\top} + \left\langle \psi_{s,d}\right\rangle \sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left(\left\langle \mathbf{z}_{s,c,n}\right\rangle \left(x_{s,c,n,d} - \left\langle \mu_{s,d}\right\rangle\right)\right)\right)\right\} + \text{const},
\end{aligned}
$$

$$
q\left(\mathbf{W}\right) = \prod_{s=1}^{S}\prod_{d=1}^{D} \mathcal{N}\left(\mathbf{w}_{s,d,\cdot}^{\top} \,\middle|\, \tilde{\mathbf{w}}_{s,d}, \tilde{\Sigma}_{w,s,d}\right),
$$

$$
\tilde{\Sigma}_{w,s,d} = \left(\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle + \left\langle \psi_{s,d}\right\rangle \sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left\langle \mathbf{z}_{s,c,n}\mathbf{z}_{s,c,n}^{\top}\right\rangle\right)^{-1},
$$

$$
\tilde{\mathbf{w}}_{s,d} = \tilde{\Sigma}_{w,s,d}\left(\left\langle \operatorname{diag}\left(\alpha\right)\right\rangle \mathbf{m}_{\mathbf{w},d,\cdot}^{\top} + \left\langle \psi_{s,d}\right\rangle \sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left(\left\langle \mathbf{z}_{s,c,n}\right\rangle \left(x_{s,c,n,d} - \left\langle \mu_{s,d}\right\rangle\right)\right)\right).
$$

Approximate posterior distribution of Z (Equations (10.15)–(10.17))

$$
\begin{aligned}
\ln\left(q\left(\mathbf{Z}\right)\right) &= \sum_{s=1}^{S}\sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left\langle \ln\left(p\left(\mathbf{X}_{s,c} \,\middle|\, \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_{s,c}, \mathbf{W}_s\right)\right) + \ln\left(p\left(\mathbf{Z}_{s,c} \,\middle|\, \boldsymbol{\Lambda}_c\right)\right)\right\rangle + \text{const} \\
&= \sum_{s=1}^{S}\sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left\{ -\frac{1}{2}\left(\left\langle \left(\mathbf{x}_{s,c,n} - \mathbf{W}_s\mathbf{z}_{s,c,n} - \boldsymbol{\mu}_s\right)^{\top}\boldsymbol{\Psi}_s\left(\mathbf{x}_{s,c,n} - \mathbf{W}_s\mathbf{z}_{s,c,n} - \boldsymbol{\mu}_s\right)\right\rangle \right.\right. \\
&\quad \left.\left. + \left\langle \mathbf{z}_{s,c,n}{}^{\top}\boldsymbol{\Lambda}_c\mathbf{z}_{s,c,n}\right\rangle\right)\right\} + \text{const} \\
&= \sum_{s=1}^{S}\sum_{c=1}^{C}\sum_{n=1}^{N_{s,c}} \left\{ -\frac{1}{2}\left(\mathbf{z}_{s,c,n}^{\top}\left(\left\langle \mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\mathbf{W}_s\right\rangle + \left\langle \boldsymbol{\Lambda}_c\right\rangle\right)\mathbf{z}_{s,c,n} \right.\right. \\
&\quad \left.\left. - 2\mathbf{z}_{s,c,n}{}^{\top}\left\langle \mathbf{W}_s^{\top}\right\rangle\left\langle \boldsymbol{\Psi}_s\right\rangle\left(\mathbf{x}_{s,c,n} - \left\langle \boldsymbol{\mu}_s\right\rangle\right)\right)\right\} + \text{const},
\end{aligned}
$$

$$
q\left(\mathbf{Z}\right) = \prod_{s=1}^{S}\prod_{c=1}^{C}\prod_{n=1}^{N_{s,c}} \mathcal{N}\left(\mathbf{z}_{s,c,n} \,\middle|\, \tilde{\mathbf{m}}_{z,s,c,n}, \Sigma_{z,s,c}\right),
$$

$$
\tilde{\Sigma}_{z,s,c} = \left(\left\langle \mathbf{W}_s^{\top}\boldsymbol{\Psi}_s\mathbf{W}_s\right\rangle + \left\langle \boldsymbol{\Lambda}_c\right\rangle\right)^{-1},
$$

$$
\tilde{\mathbf{m}}_{z,s,c,n} = \tilde{\Sigma}_{z,s,c}\left\langle \mathbf{W}_s^{\top}\right\rangle\left\langle \boldsymbol{\Psi}_s\right\rangle\left(\mathbf{x}_{s,c,n} - \left\langle \boldsymbol{\mu}_s\right\rangle\right).
$$

Approximate posterior distributions of μ and Ψ (Equations (10.24)–(10.27), (10.33) and (10.34))

$$
\begin{aligned}
\ln\left(q\left(\boldsymbol{\mu}_s, \boldsymbol{\Psi}_s\right)\right) &= \sum_{c=1}^{C} \left\langle \ln\left(p\left(\mathbf{X}_{s,c} \mid \boldsymbol{\mu}_s, \boldsymbol{\Psi}_s, \mathbf{Z}_{s,c}, \mathbf{W}_s\right)\right)\right\rangle + \ln\left(p\left(\boldsymbol{\mu}_s \mid \boldsymbol{\Psi}_s\right)\right) + \ln\left(p\left(\boldsymbol{\Psi}_s\right)\right) + \text{const} \\
&= \sum_{d=1}^{D} \left\{ \left(a_\psi + \frac{1}{2} - 1\right) \ln\left(\psi_{s,d}\right) - b_\psi \psi_{s,d} - \frac{\beta_\mu \psi_{s,d}}{2} \left(\mu_{s,d} - m_{\mu,d}\right)^2 \right\} + \sum_{c=1}^{C} \sum_{n=1}^{N_{s,c}} \left\{ \frac{1}{2} \ln\left(|\boldsymbol{\Psi}_s|\right) \right. \\
&\quad \left. - \frac{1}{2} \left\langle \left(\mathbf{x}_{s,c,n} - \left(\mathbf{W}_s \mathbf{z}_{s,c,n} + \boldsymbol{\mu}_s\right)\right)^\top \boldsymbol{\Psi}_s \left(\mathbf{x}_{s,c,n} - \left(\mathbf{W}_s \mathbf{z}_{s,c,n} + \boldsymbol{\mu}_s\right)\right)\right\rangle \right\} + \text{const} \\
&= \sum_{d=1}^{D} \left\{ \left(a_\psi + \frac{\sum_{c=1}^{C} N_{s,c} + 1}{2} - 1\right) \ln\left(\psi_{s,d}\right) - b_\psi \psi_{s,d} - \frac{\beta_\mu}{2} m_{\mu,d}^2 \right. \\
&\quad - \frac{\psi_{s,d}}{2} \left(\left(\beta_\mu + \sum_{c=1}^{C} N_{s,c}\right) \mu_{s,d}^2 - 2\mu_{s,d} \left(\beta_\mu m_{\mu,d} + \sum_{c=1}^{C} \sum_{n=1}^{N_{s,c}} \left(x_{s,c,n,d} - \left\langle \mathbf{w}_{s,d,.}\right\rangle \left\langle \mathbf{z}_{s,c,n}\right\rangle\right)\right) \right. \\
&\quad \left.\left. + \sum_{c=1}^{C} \sum_{n=1}^{N_{s,c}} \left(x_{s,n,c,d}^2 - 2x_{s,c,n,d} \left\langle \mathbf{w}_{s,d,.}\right\rangle \left\langle \mathbf{z}_{s,c,n}\right\rangle + \operatorname{tr}\left(\left\langle \mathbf{w}_{s,d,.}^\top \mathbf{w}_{s,d,.}\right\rangle \left\langle \mathbf{z}_{s,c,n} \mathbf{z}_{s,c,n}^\top\right\rangle\right)\right)\right)\right\} + \text{const} \\
&= \ln\left(q\left(\boldsymbol{\mu}_s \mid \boldsymbol{\Psi}_s\right)\right) + \sum_{d=1}^{D} \left\{ \left(a_\psi + \frac{\sum_{c=1}^{C} N_{s,c}}{2} - 1\right) \ln\left(\psi_{s,d}\right) - \frac{\psi_{s,d}}{2} \left(2b_\psi - \tilde{\beta}_{\mu,s} \tilde{m}_{\mu,s,d}^2 + \beta_\mu m_{\mu,d}^2 \right.\right. \\
&\quad \left.\left. + \sum_{c=1}^{C} \sum_{n=1}^{N_{s,c}} \left(x_{s,c,n,d}^2 - 2x_{s,c,n,d} \left\langle \mathbf{w}_{s,d,.}\right\rangle \left\langle \mathbf{z}_{s,c,n}\right\rangle + \operatorname{tr}\left(\left\langle \mathbf{w}_{s,d,.}^\top \mathbf{w}_{s,d,.}\right\rangle \left\langle \mathbf{z}_{s,c,n} \mathbf{z}_{s,c,n}^\top\right\rangle\right)\right)\right)\right\} + \text{const} \\
&= \ln\left(q\left(\boldsymbol{\mu}_s \mid \boldsymbol{\Psi}_s\right)\right) + \ln\left(q\left(\boldsymbol{\Psi}_s\right)\right), \\
q\left(\boldsymbol{\mu} \mid \boldsymbol{\Psi}\right) &= \prod_{s=1}^{S} \prod_{d=1}^{D} \mathcal{N}\left(\mu_{s,d} \,\middle|\, \tilde{m}_{\mu,s,d}, \left(\tilde{\beta}_{\mu,s} \psi_{s,d}\right)^{-1}\right), \\
q\left(\boldsymbol{\Psi}\right) &= \prod_{s=1}^{S} \prod_{d=1}^{D} \mathcal{G}\left(\psi_{s,d} \mid \tilde{a}_{\psi,s}, \tilde{b}_{\psi,s,d}\right), \\
\tilde{\beta}_{\mu,s} &= \beta_\mu + \sum_{c=1}^{C} N_{s,c}, \\
\tilde{m}_{\mu,s,d} &= \frac{1}{\tilde{\beta}_{\mu,s}} \left(\beta_\mu m_{\mu,d} + \sum_{c=1}^{C} \sum_{n=1}^{N_{s,c}} \left(x_{s,c,n,d} - \left\langle \mathbf{w}_{s,d,.}\right\rangle \left\langle \mathbf{z}_{s,c,n}\right\rangle\right)\right), \\
\tilde{a}_{\psi,s} &= a_\psi + \frac{\sum_{c=1}^{C} N_{s,c}}{2}, \\
\tilde{b}_{\psi,s,d} &= b_\psi + \frac{\beta_\mu}{2} m_{\mu,d}^2 - \frac{\tilde{\beta}_{\mu,s}}{2} \tilde{m}_{\mu,s,d}{}^2 \\
&\quad + \frac{1}{2} \sum_{c=1}^{C} \sum_{n=1}^{N_{s,c}} \left(x_{s,n,c,d}^2 - 2x_{s,c,n,d} \left\langle \mathbf{w}_{s,d,.}\right\rangle \left\langle \mathbf{z}_{s,c,n}\right\rangle + \operatorname{tr}\left(\left\langle \mathbf{w}_{s,d,.}^\top \mathbf{w}_{s,d,.}\right\rangle \left\langle \mathbf{z}_{s,c,n} \mathbf{z}_{s,c,n}^\top\right\rangle\right)\right).
\end{aligned}
$$

Approximate posterior distribution of $\boldsymbol{\Lambda}$ (Equations (10.22), (10.23) and (10.32))

$$
\begin{aligned}
\ln\left(q\left(\boldsymbol{\Lambda}_c\right)\right) &= \sum_{s=1}^{S} \left\langle \ln\left(p\left(\mathbf{Z}_{s,c} \,\middle|\, \boldsymbol{\Lambda}_c\right)\right) \right\rangle + \ln\left(p\left(\boldsymbol{\Lambda}_c\right)\right) + \text{const} \\
&= \sum_{k=1}^{K} \Bigg\{ \left(a_{\lambda,c} - 1\right) \ln\left(\lambda_{c,k}\right) - b_{\lambda,c}\lambda_{c,k} + \frac{\sum_{s=1}^{S} N_{s,c}}{2} \ln\left(\lambda_{c,k}\right) \\
&\qquad - \frac{\lambda_{c,k}}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_{s,c}} \left\langle z^2_{s,c,n,k} \right\rangle \Bigg\} + \text{const} \\
&= \sum_{k=1}^{K} \left\{ \left(a_{\lambda,c} + \frac{\sum_{s=1}^{S} N_{s,c}}{2} - 1 \right) \ln\left(\lambda_{c,k}\right) - \left(b_{\lambda,c} + \frac{1}{2} \sum_{s=1}^{S} \sum_{n=1}^{N_{s,c}} \left\langle z^2_{s,c,n,k} \right\rangle \right) \lambda_{c,k} \right\} + \text{const}, \\
q\left(\boldsymbol{\Lambda}\right) &= \prod_{c=1}^{C} \prod_{k=1}^{K} \mathcal{G}\left(\lambda_{c,k} \,\middle|\, \tilde{a}_{\lambda,c}, \tilde{b}_{\lambda,c,k}\right), \\
\tilde{a}_{\lambda,c} &= a_{\lambda,c} + \frac{N_c}{2}, \\
\tilde{b}_{\lambda,c,k} &= b_{\lambda,c} + \frac{1}{2} \sum_{n=1}^{N_c} \left\langle z^2_{c,n,k} \right\rangle.
\end{aligned}
$$

D.2 LOWER BOUND

$$
\begin{aligned}
\mathcal{L} = &\left\langle \ln\left(p\left(\mathbf{X} \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}, \mathbf{W}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\mathbf{W} \,\middle|\, \mathbf{M}_{\mathbf{w}}, \boldsymbol{\alpha}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\mathbf{M}_{\mathbf{w}} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle \\
&+ \left\langle \ln\left(p\left(\boldsymbol{\alpha}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\mathbf{Z} \,\middle|\, \boldsymbol{\Lambda}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\boldsymbol{\Lambda}\right)\right) \right\rangle + \left\langle \ln\left(p\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \right\rangle \\
&+ \left\langle \ln\left(p\left(\boldsymbol{\Psi}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\mathbf{W}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\mathbf{M}_{\mathbf{w}} \,\middle|\, \boldsymbol{\alpha}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\alpha}\right)\right) \right\rangle \\
&- \left\langle \ln\left(q\left(\mathbf{Z}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\Lambda}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\mu} \,\middle|\, \boldsymbol{\Psi}\right)\right) \right\rangle - \left\langle \ln\left(q\left(\boldsymbol{\Psi}\right)\right) \right\rangle,
\end{aligned}
$$

$$
\begin{aligned}
\left\langle \ln\left(p(\mathbf{X}, \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Psi}, \mathbf{Z}, \mathbf{W})\right) \right\rangle = &\sum_{s=1}^{S} \sum_{c=1}^{C} \Bigg\{ \frac{N_{s,c}}{2} \Big(\left\langle \ln\left(\left|\boldsymbol{\Psi}_s\right|\right) \right\rangle - D \ln\left(2\pi\right) \Big) - \frac{1}{2} \sum_{n=1}^{N_{s,c}} \Bigg(\mathbf{x}^{\top}_{s,n,c} \left\langle \boldsymbol{\Psi}_s \right\rangle \mathbf{x}_{s,n,c} \\
&+ \operatorname{tr}\left(\left\langle \mathbf{W}^{\top}_{s} \boldsymbol{\Psi}_s \mathbf{W}_s \right\rangle \left\langle \mathbf{z}_{s,n,c} \mathbf{z}^{\top}_{s,n,c} \right\rangle \right) + \left\langle \boldsymbol{\mu}^{\top}_{s} \boldsymbol{\Psi}_s \boldsymbol{\mu}_s \right\rangle \\
&- 2\mathbf{x}^{\top}_{s,n,c} \left\langle \boldsymbol{\Psi}_s \right\rangle \Big(\left\langle \mathbf{W}_s \right\rangle \left\langle \mathbf{z}_{s,n,c} \right\rangle + \left\langle \boldsymbol{\mu}_s \right\rangle \Big) + 2 \left\langle \mathbf{z}_{s,n,c} \right\rangle^{\top} \left\langle \mathbf{W}^{\top}_{s} \right\rangle \left\langle \boldsymbol{\Psi}_s \right\rangle \left\langle \boldsymbol{\mu}_s \right\rangle \Bigg) \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\left\langle \ln\left(p(\mathbf{W} \,\middle|\, \mathbf{M}_{\mathbf{w}}, \boldsymbol{\alpha})\right) \right\rangle = &\sum_{s=1}^{S} \sum_{k=1}^{K} \Bigg(\frac{D}{2} \Big(\left\langle \ln\left(\alpha_k\right) \right\rangle - \ln\left(2\pi\right) - \tilde{\beta}_{w}^{-1} \Big) \\
&- \frac{\left\langle \alpha_k \right\rangle}{2} \Big(\left\langle \mathbf{w}^{\top}_{s,k} \mathbf{w}_{s,k} \right\rangle - 2 \left\langle \mathbf{w}^{\top}_{s,k} \right\rangle \left\langle \mathbf{m}_{\mathbf{w},k} \right\rangle + \left\langle \mathbf{m}^{\top}_{\mathbf{w},k} \right\rangle \left\langle \mathbf{m}_{\mathbf{w},k} \right\rangle \Big) \Bigg),
\end{aligned}
$$

$$
\left\langle \ln\left(q(\mathbf{W})\right) \right\rangle = \sum_{s=1}^{S} \left\{ - \frac{KD}{2} \ln\left(2\pi\right) - \frac{1}{2} \sum_{d=1}^{D} \left(\ln\left(\left|\tilde{\Sigma}_{w,s,d}\right|\right) + K \right) \right\},
$$

$$\left\langle \ln\left(p(\mathbf{M}_{\mathbf{w}} \mid \boldsymbol{\alpha})\right) \right\rangle = \sum_{k=1}^{K} \left(\frac{D}{2}\Big(\left\langle \ln\left(\alpha_k\right) \right\rangle + \ln\left(\beta_w\right) - \ln\left(2\pi\right) - \frac{\beta_w}{\tilde{\beta}_w} \Big) - \frac{\beta_w}{2} \left\langle \alpha_k \right\rangle \left\langle \mathbf{m}_{\mathbf{w},k}^{\top} \right\rangle \left\langle \mathbf{m}_{\mathbf{w},k} \right\rangle \right),$$

$$\left\langle \ln\left(q(\mathbf{M}_{\mathbf{w}} \mid \boldsymbol{\alpha})\right) \right\rangle = \sum_{k=1}^{K} \left(\frac{D}{2}\Big(\left\langle \ln\left(\alpha_k\right) \right\rangle + \ln\left(\tilde{\beta}_w\right) - \ln\left(2\pi\right) - 1 \Big) \right),$$

$$\left\langle \ln\left(p(\boldsymbol{\alpha})\right) \right\rangle = K\Big(- \ln\left(\Gamma\left(a_\alpha\right)\right) + a_\alpha \ln\left(b_\alpha\right) \Big) + \sum_{k=1}^{K} \Big(\left(a_\alpha - 1\right) \left\langle \ln\left(\alpha_k\right) \right\rangle - b_\alpha \left\langle \alpha_k \right\rangle \Big),$$

$$\left\langle \ln\left(q(\boldsymbol{\alpha})\right) \right\rangle = -K \ln\left(\Gamma\left(\tilde{a}_\alpha\right)\right) + \sum_{k=1}^{K} \left(\tilde{a}_\alpha \ln\left(\tilde{b}_{\alpha,k}\right) + \left(\tilde{a}_\alpha - 1\right) \left\langle \ln\left(\alpha_k\right) \right\rangle - \tilde{b}_{\alpha,k} \left\langle \alpha_k \right\rangle \right),$$

$$\begin{aligned} \left\langle \ln\left(p(\boldsymbol{\mu} \mid \boldsymbol{\Psi})\right) \right\rangle = \sum_{s=1}^{S} \Bigg\{ & \frac{D}{2}\Big(\ln\left(\beta_\mu\right) - \ln\left(2\pi\right) - \frac{\beta_\mu}{\tilde{\beta}_{\mu,s}} \Big) \\ & + \sum_{d=1}^{D} \left(\frac{\left\langle \ln\left(\psi_{s,d}\right) \right\rangle}{2} - \frac{\beta_\mu \left\langle \psi_{s,d} \right\rangle}{2} \Big(\left\langle \mu_{s,d} \right\rangle - m_{\mu,d} \Big)^2 \right) \Bigg\}, \end{aligned}$$

$$\left\langle \ln\left(q(\boldsymbol{\mu} \mid \boldsymbol{\Psi})\right) \right\rangle = \sum_{s=1}^{S} \left\{ \frac{D}{2}\Big(\ln\left(\tilde{\beta}_{\mu,s}\right) - \ln\left(2\pi\right) - 1 \Big) + \frac{1}{2} \sum_{d=1}^{D} \left\langle \ln\left(\psi_{s,d}\right) \right\rangle \right\},$$

$$\begin{aligned} \left\langle \ln\left(p(\boldsymbol{\Psi})\right) \right\rangle = \sum_{s=1}^{S} \Bigg\{ & D\Big(- \ln\left(\Gamma\left(a_\psi\right)\right) + a_\psi \ln\left(b_{\psi,d}\right) \Big) \\ & + \sum_{d=1}^{D} \Big(\left(a_\psi - 1\right) \left\langle \ln\left(\psi_{s,d}\right) \right\rangle - b_{\psi,d} \left\langle \psi_{s,d} \right\rangle \Big) \Bigg\}, \end{aligned}$$

$$\begin{aligned} \left\langle \ln\left(q(\boldsymbol{\Psi})\right) \right\rangle = \sum_{s=1}^{S} \Bigg\{ & - D \ln\left(\Gamma\left(\tilde{a}_{\psi,s}\right)\right) \\ & + \sum_{d=1}^{D} \left(\tilde{a}_{\psi,s} \ln\left(\tilde{b}_{\psi,s,d}\right) + \left(\tilde{a}_{\psi,s} - 1\right) \left\langle \ln\left(\psi_{s,d}\right) \right\rangle - \tilde{b}_{\psi,s,d} \left\langle \psi_{s,d} \right\rangle \right) \Bigg\}, \end{aligned}$$

$$\left\langle \ln\left(p(\mathbf{Z} \mid \boldsymbol{\Lambda})\right) \right\rangle = \sum_{s=1}^{S} \sum_{c=1}^{C} \left\{ \frac{N_{s,c}}{2}\Big(\ln\left(\left|\boldsymbol{\Lambda}_c\right|\right) - K \ln\left(2\pi\right) \Big) - \frac{1}{2} \sum_{n=1}^{N_{s,c}} \operatorname{tr}\left(\left\langle \mathbf{z}_{s,n,c} \mathbf{z}_{s,n,c}^{\top} \right\rangle \left\langle \boldsymbol{\Lambda}_c \right\rangle \right) \right\},$$

$$\left\langle \ln\left(q(\mathbf{Z})\right) \right\rangle = \sum_{s=1}^{S} \sum_{c=1}^{C} \left\{ - \frac{N_{s,c}}{2} \left(\ln\left(\left|\tilde{\Sigma}_{z,s,c}\right|\right) + K \ln\left(2\pi\right) + K \right) \right\},$$

$$\begin{aligned} \left\langle \ln\left(p(\boldsymbol{\Lambda})\right) \right\rangle = {} & K \sum_{c=1}^{C} \Big(- \ln\left(\Gamma\left(a_{\lambda,c}\right)\right) + a_\lambda \ln\left(b_{\lambda,c}\right) \Big) \\ & + \sum_{c=1}^{C} \sum_{k=1}^{K} \Big(\left(a_{\lambda,c} - 1\right) \left\langle \ln\left(\lambda_{c,k}\right) \right\rangle - b_{\lambda,c} \left\langle \lambda_{c,k} \right\rangle \Big), \end{aligned}$$

$$\left\langle \ln\left(q(\boldsymbol{\Lambda})\right) \right\rangle = -K \sum_{c=1}^{C} \ln\left(\Gamma\left(\tilde{a}_{\lambda,c}\right)\right)$$

$$+ \sum_{c=1}^{C} \sum_{k=1}^{K} \Big(\tilde{a}_{\lambda,c} \ln \big(\tilde{b}_{\lambda,c,k} \big) + \big(\tilde{a}_{\lambda,c} - 1 \big) \left\langle \ln \big(\lambda_{c,k} \big) \right\rangle - \tilde{b}_{\lambda,c,k} \langle \lambda_{c,k} \rangle \Big).$$