Classification Methods for Multi-Class Multivariate Longitudinal Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting Wiskunde
te verdedigen door

KRISTIEN WOUTERS

Promotor: Prof. dr. Geert Molenberghs
Copromotor: dr. José Cortiñas

universiteit hasselt
CENTRUM
VOOR STATISTIEK

Contents

List of Abbreviations

AIC	Akaike Information Criterion
AW	Active Wake
DHSLA	Doubly Hierarchical Supervised Learning Analysis
EEG	Electro-Encephalogram
EMG	Electro-myogram
IS	Intermediate Stage Sleep
FDA	Flexible Discriminant Analysis
FLDA	Functional Linear Discriminant Analysis
FP	Fractional Polynomial
LDA	Linear Discriminant Analysis
MARS	Multivariate Adaptive Regression Splines
MI	Mutual Information
MDA	Mixture Discriminant Analysis
MFLDA	Multivariate Functional Linear Discriminant Analysis
PW	Passive Wake
RS	Rapid Eye Movement (REM) Sleep

SWS1	Light Sleep
SWS2	Deep Sleep

List of Tables

List of Figures

List of Publications

This thesis has been based on the following scientific papers:

Wouters, K., Ahnaou, A., Cortiñas, J., Molenberghs, G., Geys, H., Bijnens, L., and Drinkenbrug, W.H.I.M. (2007). Psychotropic drug classification based on sleep-wake behaviour of rats. *Journal of the Royal Statistical Society, Series C*, **56**(2), 223–234.

Wouters, K., Cortinas, J., Molenberghs, G., Ahnaou, A., Drinkenburg, W.H.I.M and Bijnens, L (2007). A comparison of doubly hierarchical discriminant analyses for multiple class longitudinal data from EEG experiments. *Journal of Biopharmaceutical Statistics*, **18**, 000–000.

Wouters, K., Cortinas, J., Molenberghs, G., Geys, H., Ahnaou, A., Drinkenburg, W.H.I.M. and Bijnens, L. (2008). Correction for model selection bias using a modified model averaging approach for supervised learning methods applied to EEG experiments. *Manuscript submitted for publication.*

Wouters, K., Cortinas, J., Molenberghs, G., Geys, H., Ahnaou, A., Drinkenburg, W.H.I.M. and Bijnens, L. (2008). Mutual information as a tool to visualize classes in EEG data. *Manuscript submitted for publication.*

Wouters, K., Cortinas, J., Molenberghs, G., Geys, H., Ahnaou, A., Drinkenburg, W.H.I.M. and Bijnens, L. (2008). Multivariate Functional Linear Discriminant Analysis Based on Pairwise Pseudo-Likelihood Modeling Combined with Splines. *Manuscript submitted for publication.*

1

Introduction

1.1 Brief Introduction to Machine Learning

Machine learning is a very interdisciplinary field, which comes as a result of the interaction between three main areas of research, computer science and engineering, applied mathematics and statistics. It is used in a wide range of applications, including speech and handwriting recognition, detection of credit card fraud, game playing and identification of spam-mail. It has been gaining popularity over the last years in medical research. An application in this area can be found for example in the preliminary diagnosis of a patient's disease in view of instantaneous selection of the treatment while awaiting conclusive test results. Machine learning has also been used in the pharmaceutical industry, in the discovery process of new active compounds.

Two major branches can be distinguished in the machine learning framework: supervised learning and unsupervised learning. Supervised learning is the term applied in the machine learning field to techniques used to find a function mapping pairs of inputs and desired outputs based on some training data. Inputs are typically of the discrete, continuous or mixed types. Outputs are in general of two types, continuous, in which it is called a regression problem, or a vector of class membership, in which it is then called classification or discriminant analysis. After training the procedure with a training dataset, new samples can be classified into one of the predefined classes.

In the field of unsupervised learning, the focus of the researcher is different. The idea is now to find or establish the existence of classes or clusters present in the data at hand. This second situation is the topic of cluster analysis (Johnson and Wichern, 1992).

Discriminant analysis is a well-known procedure which dates back to the first half of the twentieth century (Fisher, 1936). Since then, several procedures have been proposed which enhanced the original ideas of Fisher. Flexible discriminant analysis (Hastie, Tibshirani and Buja, 1994), penalized discriminant analysis (Hastie, Buja and Tibshirani, 1995), mixture discriminant analysis (Hastie and Tibshirani, 1996), functional linear discriminant analysis (James and Hastie, 2001), are some of these methods to mention a few. Nowadays, data-mining procedures such as random forests (Breiman, 2001), neural networks (Haykin, 1999) and support vector machines (Vapnik, 1998) are also gaining popularity in the supervised learning field and their good performances have been shown in several applications.

1.2 Longitudinal Data Analysis

In health related research one is often confronted with longitudinal data, where measurements of individuals are taken repeatedly over time. Since observations coming from the same subject tend to be more alike than observations from different subjects, they are said to be correlated. In order to be able to draw valid conclusions, one needs to account for this correlation when analyzing this type of data.

In order to model continuous, normally distributed longitudinal outcomes, the linear mixed effects model (Verbeke and Molenberghs, 2000) has become the most commonly used tool. This model uses both population average (or fixed) and subject-specific (or random) effects. Furthermore, it marginalizes to a multivariate normal model with directly interpretable mean and covariance parameters.

In case of a discrete or categorical outcome variable, extensions of the generalized linear model can be used. Two important representatives are generalized estimating equations or GEE (Liang and Zeger, 1986) and the generalized linear mixed model or GLMM (Breslow and Clayton (1993), Molenberghs and Verbeke (2005)).

1.3 Scope of this Thesis

In this thesis we will focus on the classification of multiple-class longitudinal data. The motivation for this research was found in pre-clinical pharmaco-electroencephalogram

(EEG) studies aiming at characterizing psychotropic drug effects on the basis of spectral EEG analysis.

For thousands of years, psychoactive substances (i.e. pharmacological agents that act on the central nervous system) have been used by humans in all known societies. Despite the wide variety of effects that such substances can exert on the central nervous system, attempts have been made to categorize drugs into psychoactive classes based on therapeutic efficacy, such as antidepressants, antipsychotics, anxiolytics, hypnotics and stimulants. In this thesis, we will base our classification procedure on this categorization. Alternative classifications have been proposed, however so far with limited preclinical predictive value for clinical usability. For a number of psychiatric disorders more or less effective reference drugs exist. The availability of these reference drugs makes it possible to classify novel, putative psychotropic agents in a direct comparison of their (electro-)physiological profiles.

Pharmaco-ElectroEncephaloGraphy (EEG) is extensively used in humans for the discrimination of clinically active, psychotropic drugs, which has fostered the development of corresponding animal pharmaco-EEG models. In our motivating study, rats were given a psychoactive compound and monitored during 16 hours. Six clearly defined, spontaneously occurring sleep-wake stages are separated out: Active Wake, Passive Wake, Light Sleep, Deep Sleep, Intermediate Stage Sleep and REM Sleep. These six sleeping stages will be used for the classification of the psychotropic drugs into the five reference classes.

From a statistical point of view, analyzing EEG data poses a number of important challenges. First, there is the high-dimensionality of raw EEG data. Even after the usual initial reduction of dimensionality involving a spectral analysis, in which the power spectrum is subdivided into several, predefined frequency bands (e.g., *delta, theta*, etc.), there is still a multitude of variables to be analyzed. Moreover, these variables are measured repeatedly over time, hence we are dealing with longitudinal data.

Secondly, there is no generally accepted functional form for the evolution of the EEG activity over time. The longitudinal profiles are usually highly dynamic and unpredictable within a given short time frame, the variability both between and within subjects can be considerably large under influence of a certain dosage of the pharmacological treatment. Thus, finding a suitable statistical model is therefore a non-trivial task.

Lastly, given these complexities it is evident that to find out about the psychoactivity of a novel drug (i.e. the prediction to which psychoactive class it belongs at a certain dose) is not an easy process. While conventional discriminant analysis can be used, a fully satisfactory answer requires appropriately tailored methods.

The aim of this thesis is therefore to propose new procedures to classify psychotropic pharmacological agents into one of the 5 major classes of psycho-activity or placebo, based on the sleep-wake behaviour of rats as defined by EEG, EMG, and locomotor activity.

1.4 Structure of this Thesis

In Chapter 2, the data used throughout this thesis is presented, and some background information regarding EEG-experiments and psychotropic drug classes is provided. A similar study has been conducted previously (Ruigt *et al*, 1993). The authors propose a classification procedure of psychotropic drugs based on sleep-wake behaviour. Here we will also apply this classification procedure in order to be able to compare their results with those obtained using our proposed methodology.

In Chapter 3, we propose an exploratory tool, based on information theory, for the visualization of classes in EEG data, which is applied to our motivating dataset.

Chapter 4 deals with the longitudinal aspect of the data. Two flexible modeling techniques for longitudinal data are described and applied to the EEG dataset.

In Chapter 5 a new two-stage procedure for the classification of multiple-class longitudinal data is proposed, called doubly hierarchical supervised learning analysis (DHSLA). In the first stage, a flexible modeling technique, e.g. fractional polynomial mixed models, is used to obtain a summary of the data, which is further used in the second stage in a stepwise discriminant procedure. In the second step linear, flexible and mixture discriminant analysis will be used. In this chapter we elaborate on the methodology and in Chapter 6 the DHSLA is applied to our motivating study, with different discriminant techniques in the second stage.

The procedure proposed possibly introduces selection bias. In order to deal with this problem, we suggest a novel modification of the general model averaging approach used in regression problems to the particular case of classification problems. This model averaging will be integrated in the second stage of the DHSLA. In Chapter 7, this approach will be outlined and applied to the EEG dataset.

For the classification of univariate longitudinal profiles, James and Hastie (2001) proposed a functional linear discriminant analysis (FLDA). In Chapter 8, we present an extension of this methodology for the case of multivariate longitudinal profiles using a pseudo-likelihood modeling approach to deal with the multivariate characteristics of the data. The performance of the multivariate functional linear discriminant analysis is evaluated on the EEG dataset and through simulations.

Finally, in Chapter 9, concluding remarks are formulated and a perspective for future research is presented.

2

Motivating Study

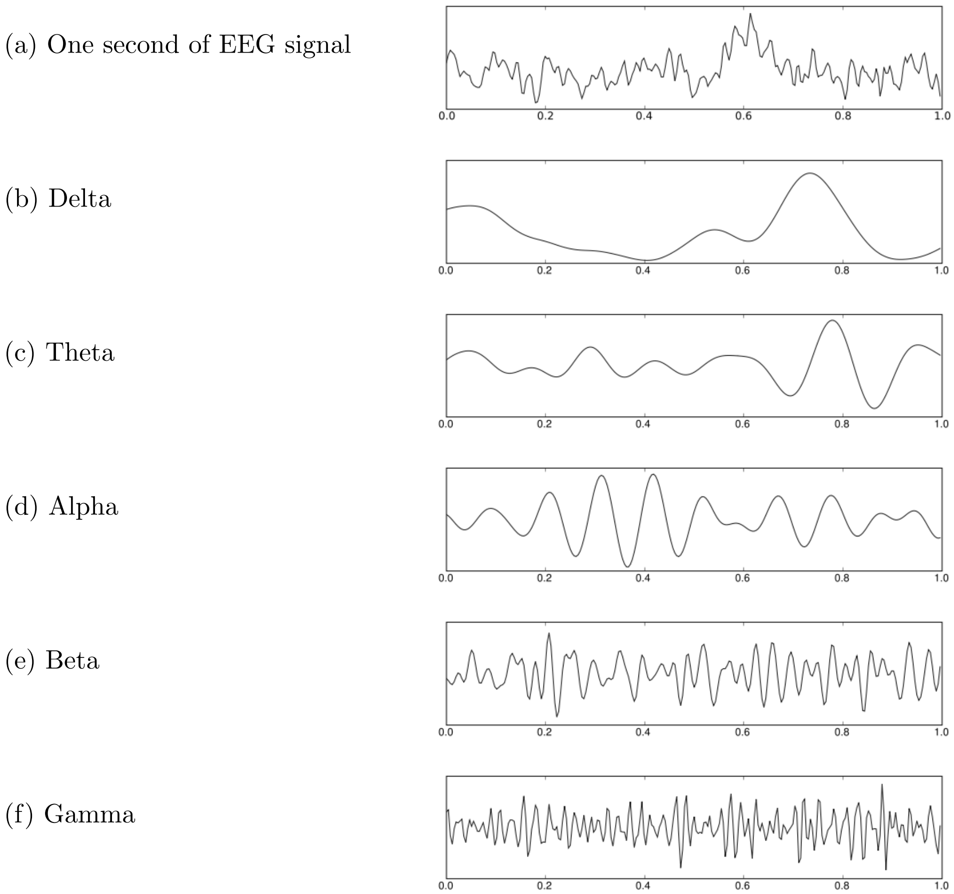
The dataset used throughout the thesis is coming from an electro-encephalogram study conducted at Janssen Pharmaceutica in Beerse (Belgium) aiming at classification of potential new psychotropic drugs into one of five earlier defined psychotropic drug classes and placebo.

We will first introduce the basic concepts of electro-encephalogram (EEG) studies in section 2.1 and we will further expand on the definition of the five drug classes in section 2.2. Thereafter, the setup of the experiment is described and finally we briefly review the classification method for this type of data proposed by Ruigt *et al* (1993).

2.1 Introduction to Electro-Encephalogram Studies

Pharmaco-electro-encephalographical studies aim at characterizing psychotropic drug effects, usually on the basis of spectral electro-encephalograms, which reflect cortical brain activity. Frequency measurements range from below 3.5 Hz per second (so-called delta activity), 4–7.5 Hz/s (theta activity), 8–12 Hz/s (alpha activity), 13–30 Hz/s (beta activity) and above 30 Hz/s (gamma activity). In Figure 2.1, the different activities are illustrated on an EEG sample of one second.

Delta activity tends to have the highest amplitude and the slowest waves. It is normally seen in babies or in adults in slow wave sleep. Theta activity may be seen in children or during drowsiness or arousal in adults. It can also be seen in


(a) One second of EEG signal

(b) Delta

(c) Theta

(d) Alpha

(e) Beta

(f) Gamma

Figure 2.1: *(a) One second sample of an EEG. (b) – (f) delta-, theta-, alpha-, beta- and gamma-waves filtered from sample (a).*

meditation. Alpha waves occur when a person is alert in a relaxed way. Alpha activity decreases with sleepiness and when the eyes are open. Low amplitude beta waves are often associated with active, busy or anxious thinking and active concentration, while rhythmic beta waves are generally linked with pathological or drug-related causes. Gamma waves are associated with strong mental activity like solving problems, fear and awareness.

EEG registrations are reliably carried out in humans and mammals alike. In rodents the EEG can be used to determine sleep-wake architecture, when carried out in conjunction with movement monitoring and a so-called electromyogram (EMG) that records muscle activity. A crucial problem for pharmaco-EEG studies is that the pharmacological effects on the EEG are easily confounded by marked EEG alterations associated with spontaneous changes in behaviour or vigilance. Several approaches have been proposed to overcome this problem, however, in our dataset rats were left undisturbed.

Typically, six sleep-wake stages are distinguished: (1) *active wake (AW)*, characterized by movement, theta activity and high EMG, (2) *passive wake (PW)*, with similar characteristics as the previous one, but without movement, (3) *light sleep* or *slow wave sleep 1 (SWS1)*, characterized by EEG spindles (short lasting burst of phasic brain activity, indicative of transitions in neuronal synchronization), (4) *deep sleep* or *slow wave sleep 2 (SWS2)*, with slow waves and prominent delta activity, (5) *intermediate stage sleep (IS)*, with spindle-like activity against a background of theta activity and low EMG, and (6) *Rapid Eye Movement* or *REM Sleep (RS)*, with theta activity and very low EMG.

2.2 Psychotropic Drugs, Background and Classification

For thousands of years, humans in all known societies have used psychotropic drugs (substances that act on the central nervous system (CNS) and affect mood, thinking, and behaviour). Psychotropic drugs, whatever the substances used, rank second to tenth among the most consumed medicinal products in Western nations (Zarifian (1996)). Although there may be some controversies on classification for clinical purposes (Zarifian (1988), American Psychiatric Association (2000)), psychotropic drugs can be divided into 5 major classes according to their main indication in psychiatry: antidepressants, antipsychotics, anxiolytics, hypnotics and stimulants (Deniker (1982), Oughourlian (1984), Cohen and Cailloux-Cohen (1995)).

Antidepressants are amongst the drugs most commonly prescribed by psychiatrists and general practitioners. They are used for alleviating depression or dysthymia (milder depression). Clinical depression is characterized by a pervasive low mood, loss of interest or pleasure in usual activities and a deep feeling of sadness. Known antidepressants are bupropion, venlafaxine and paroxetine, to list some of them.

Antipsychotic drugs are used to treat psychosis, which is a generic term for a mental state often described as involving a 'loss of contact with reality'. People experiencing psychosis may report hallucinations or delusional beliefs and may exhibit personality changes and disorganized thinking. This may be accompanied with unusual behaviour and difficulty with social interaction. Common conditions with which psychosis might be reported include schizophrenia, bipolar disorder, mania and delirium, which is an acute decline in attention-focus, perception and cognition. Some known antipsychotics are risperidone, haloperidol and chlorpromazine.

The term anxiolytic is applied to a group of drugs used to relieve anxiety or prevent anxiety attacks. The most common anti-anxiety medications include: flurazepam, oxazepam and diazepam.

Hypnotic drugs induce sleep and are used in the treatment of insomnia and in surgical anesthesia. Included in this class are zolpidem and zopiclone among others.

Stimulant drugs enhance the activity of the nervous system. They are used to increase alertness and awareness in patients with narcolepsy, attention-deficit hyperactivity disorder and short-term treatment of obesity. Well known stimulants are cocaine, caffeine, amphetamine, nicotine, etc. The more powerful variants of these drugs are often prescription medicines or illegal drugs.

Classifying drugs only on the basis of chemical structure would create numerous categories, which would not necessarily be indicative of their therapeutic use and is therefore not advisable. A better approach is to classify new chemical entities based on their potential therapeutic activity. This classification ideally should be as early as possible in the drug discovery process. Availability of an advanced classification model or tool that uses a standardized physiological read-out (e.g., the electroencephalogram or EEG) would greatly aid efficient determination of psychoactive properties of newly synthesized chemicals.

The potential of using EEG-derived parameters (pharmaco-EEG) and characteristic fingerprints on rodent sleep-wake architecture for the classification of drugs has been recognized for several decades and is used as a valuable tool in both preclinical drug discovery and clinical drug development (Fink (1959), Krijzer *et al* (1993), Ruigt *et al* (1993), Depoortere *et al* (1995), Edgar (2002), Drinkenburg and Ahnaou (2004), Uchida *et al* (2007)). In addition EEG technology has been used to identify biomarkers that have predictive validity for clinical, pharmacological activity and even for possible efficacy (i.e. as surrogate endpoint) for some diseases such as Major Depressive Disorder (MDD) (Mucci *et al* (2006), Murck *et al* (2003), Staner *et al* (2004)). All these have motivated the research on the classification of psychotropic drugs based on EEG-information and its derivatives.

2.3 Description of the Experimental Study

The motivating study includes 26 psychoactive agents at 4 different doses, including dose 0. To each compound, 32 rats are randomly assigned, i.e., 8 per dose group. After a washout period of 3 weeks, the same rats can be used in another experiment. In total, 342 rats are included in the study. The number of times the rats are used ranges from one up to eight times. Note that the same compound may belong to different classes at different doses.

The brain signals of the rats are monitored during 16 hours, starting with a light period of 10 hours, followed by a period of darkness of 6 hours. The administration of the treatment is done at the beginning of the light period. Every two seconds, the sleep-wake state of the rat is recorded. This information is then summarized by measuring the time spent in each of the six sleep-wake stages per interval of 30 minutes.

It is well known that rats are nocturnal animals. In the conducted study this is reflected in the jump in number of minutes spent in Active Wake and the decrease in Light and Deep Sleep at time period 20. This feature is seen in all the classes. This impact of the presence or absence of light is the reason why the light and dark period were introduced in the first place.

This data is further subdivided into a training and a test dataset. In both cases, there is expert knowledge available regarding the class membership of the compound-by-dose combinations. The final training dataset contains 61 treatments: 23 placebos, 14 antidepressants, 7 antipsychotics, 2 anxiolytics, 5 hypnotics and 10 stimulants. In the test dataset, there are 3 placebos, 4 antidepressants, 2 antipsychotics, 2 hypnotics and 3 stimulants.

The compound-by-dose combinations in the training and the test dataset are given in Tables 2.1 and 2.2. As an illustration the number of minutes spent in the six sleep-wake stages for the eight rats who got clomipramine, which belongs to the antidepressant class, are plotted in Figure 2.2. As we can see, the profiles are very irregular, showing a high variability within and between rats receiving the same drug.

In Figure 2.3 an overview of the profiles in all compound-dose combinations in the six psychotropic drug classes is given. Each line in the plot represents the mean of the eight rats of one compound-dose combination. The plots show high variability within the classes and within the compound-dose combinations. While this variability is less pronounced for placebo and antidepressant, it is highly present in stimulants and anxiolytics. Since we have only 2 compound-dose combinations in the anxiolytic class, this class will be disregarded in further analyses. This leaves us with 59 compound-

Table 2.1: *Compound-by-dose combinations in the training dataset, sorted per class.*

Class	Drug	Dose (mg)	Class	Drug	Dose (mg)
Antidepressant	Clomipramine	22	Antipsychotic	Chlorpromazine	1
	Ritanserin	2.5		Clozapine	1
	Paroxetine	3		Clozapine	3
	Fluvoxamine	7.3		Haloperidol	1
	Fluvoxamine	22		Haloperidol	3
	Fluoxetine	10		Olanzapine	3
	Mirtazapine	3		Risperidone	1
	Desipramine	1	Anxiolytic	Oxazepam	3.2
	Desipramine	3		Buspirone	2.2
	Imipramine	3	Stimulant	Tacrine	10
	Bupropion	10		Apomorphine	1
	Citalopram	3		Amphetamine	1
	Citalopram	10		Amphetamine	3
	Bupropion	10		Amphetamine	10
Hypnotic	Zolpidem	3		Cocaine	10
	Zolpidem	10		Caffeine	10
	Zopiclone	3		Caffeine	22
	Zopiclone	10		Nicotine	0.5
	Flurazepam	3		Nicotine	1

dose combinations or 472 rats in the training dataset.

When comparing the six classes, we see that a slightly different behaviour is observed in different classes. For example, rats who were administered with a stimulant compound spend more time in Active Wake and less in Light and Deep Sleep, while rats who received an antidepressant spent less time in REM Sleep.

Table 2.3 presents an overview of the changes in EEG-defined sleep-wake behaviour that are generally observed to be associated with the compounds belonging to the different drug classes (Ruigt *et al*, 1993).

Table 2.2: *Compound-by-dose combinations in the test dataset, sorted per class.*

Class	Drug	Dose	Class	Drug	Dose
		(mg)			(mg)
Antidepressant	Paroxetine	1	Hypnotic	Gaboxadol	5
	Fluvoxamine	3.7		Zopiclone	1
	Mirtazapine	10	Stimulant	Tacrine	2.5
	Imipramine	10		Apomorphine	3
Antipsychotic	Chlorpromazine	10		Apomorphine	10
	Olanzapine	10			

Table 2.3: *Overview of generally observed changes in sleep-wake behaviour associated with the six psychotropic drug classes.*

	Sleep-Wake Stage					
Class	Active Wake	Passive Wake	Light Sleep	Deep Sleep	Intermediate Stage	REM Sleep
Antidepressant		⇈				⇊
Antipsychotic			⇈		↑	
Anxiolytic			⇈	↓		
Hypnotic			(↑)	⇈		(↑)
Stimulant	⇈					

2.4 Brief Review of Existing Methodology

Ruigt *et al* (1993) propose a method for prediction of psychotropic drug classes based on a discriminant analysis of drug effects on rat sleep. They consider only the first 8 hours after drug injection and regroup the 30-minute periods for every sleep-wake stage in a particular way, which will be described next, in order to cover the general temporal characteristics of the drug effects on the sleep-wake stages.

For Active Wake, the considered periods are 0.5h – 3h, 3h – 5h and 5h – 8h. For Passive Wake, the time periods are regrouped into 0h – 2h, 2h – 5h and 5h – 8h. For Light and Deep Sleep, the periods 0h – 2h, 2h – 4h, 4h – 6h and 6h – 8h are considered. Finally for REM Sleep and Intermediate Stage Sleep the considered time intervals are 0h – 1.5h, 1.5h – 4.5h, and 4.5h – 8h. In every period, the change in percentage of the time spent in a certain sleep-wake stage compared with the placebo group for the compound under investigation is recorded. These new variables are then used in a linear discriminant analysis.

In their paper, Ruigt *et al* (1993) apply this method to a training dataset with 12 antidepressants, 7 antipsychotics, 4 stimulants, 3 anxiolytics, 4 hypnotics and 3 anticonvulsants. The classification results obtained with cross-validation are shown in Table 2.4. In this dataset, stimulants, antipsychotics and anticonvulsants are well differentiated, while the classification of hypnotics and anxiolytics is rather poor. The overall error rate obtained for this dataset is 0.398, pointing already to room for improvement.

In Table 2.5, we apply the method described above to our training dataset with cross validation. The motivating study does not include anticonvulsants and anxiolytics were excluded given the limited amount of information in this class. Ruigt's method performs reasonably well for hypnotics and stimulants, but the performance in antidepressants and antipsychotics is very poor. The resulting overall error rate obtained in this case is even larger (0.438), showing possible deficiencies in the classification methodology proposed.

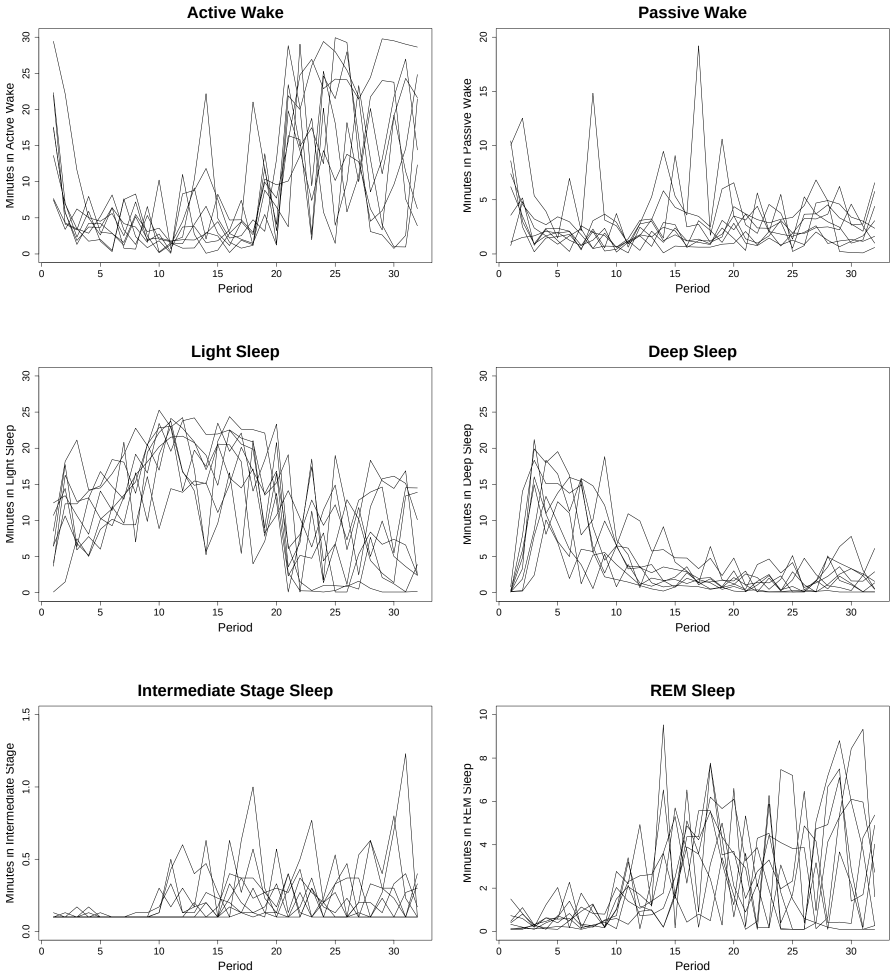
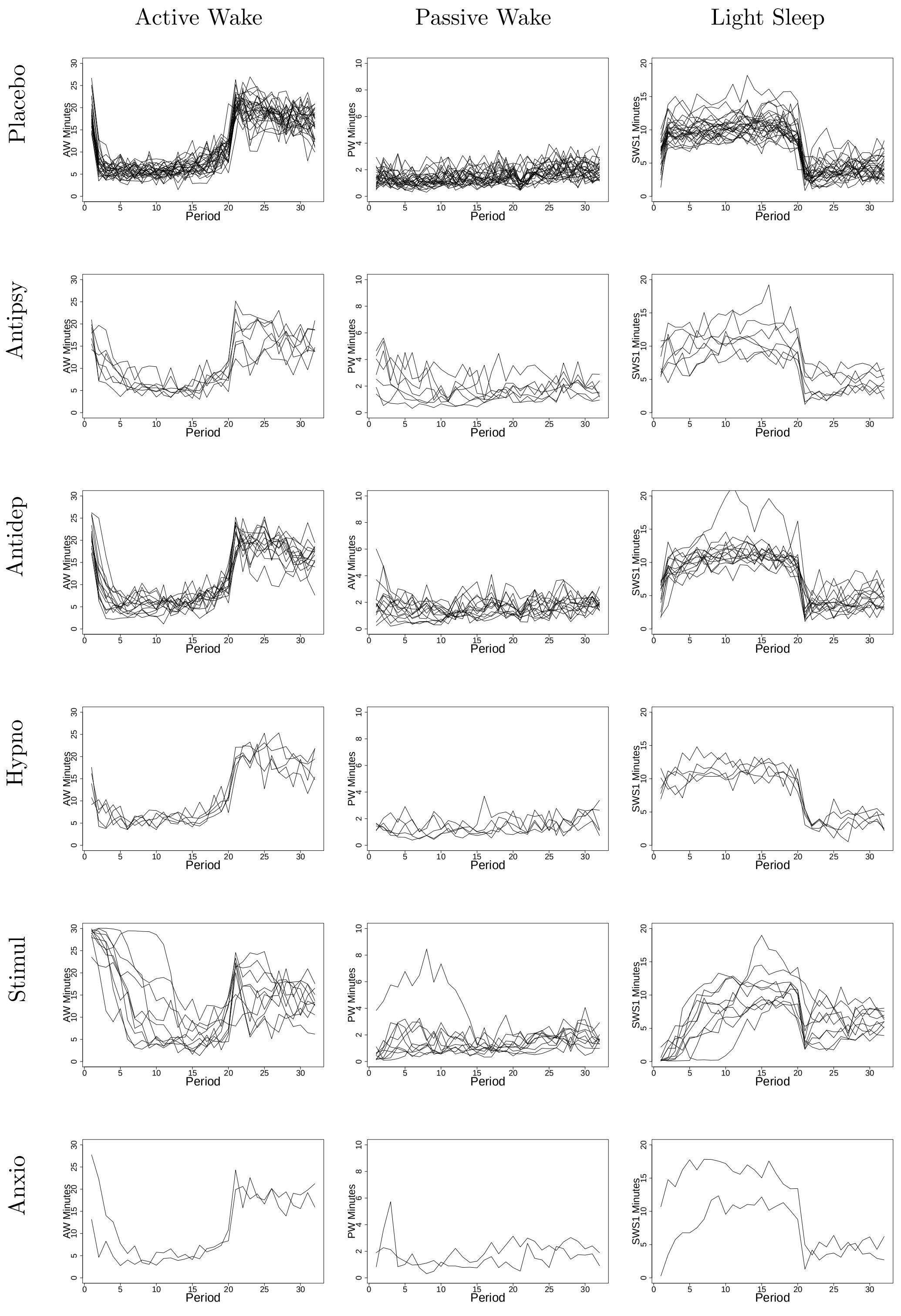


Figure 2.2: *Observed number of minutes spent in each of the six sleep-wake stages for the eight rats receiving Clomipramine (Antidepressant).*


Active Wake
Passive Wake
Light Sleep
Placebo
Antipsy
Antidep
Hypno
Stimul
Anxio
AW Minutes
PW Minutes
SWS1 Minutes
AW Minutes
Period
0
5
10
15
20
25
30
0
2
4
6
8
10
0
5
10
15
20

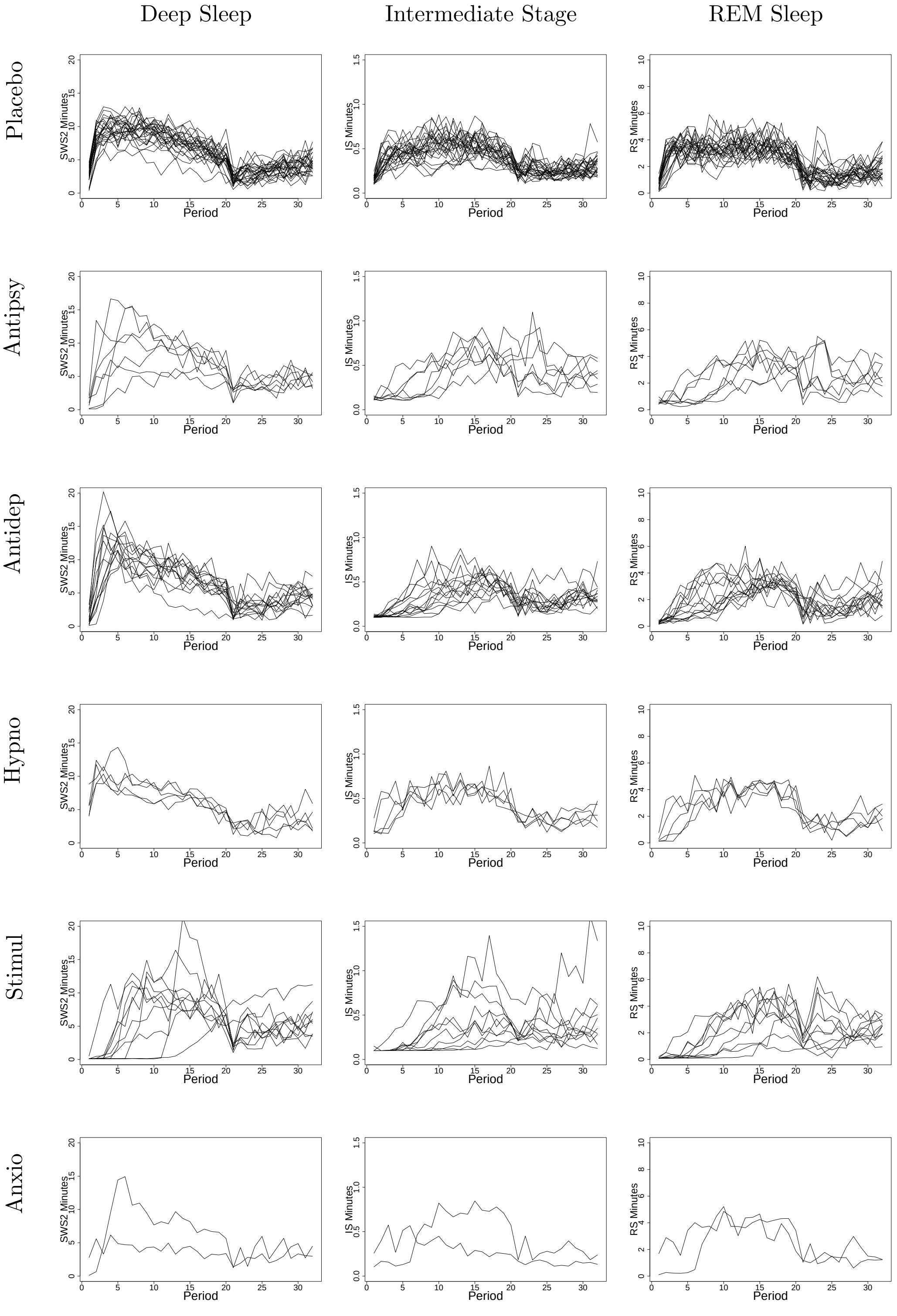


Figure 2.3: *Number of minutes spent in the six sleep-wake stages per class. Each line represents the mean profile over all the rats in one compound-dose combination.*

Table 2.4: *Classification of Ruigt's training dataset with method of Ruigt* et al *(1993) with cross-validation.*

	Predicted Class							
Class	Placebo	Antidep	Antipsy	Hypno	Stimul	Anxio	Anticonv	Total
Antidepressants	0	**7**	0	1	3	1	0	12
Antipsychotics	1	0	**6**	0	0	0	0	7
Hypnotics	3	0	0	**1**	0	0	0	4
Stimulants	0	0	0	0	**4**	0	0	4
Anxiolytics	2	1	0	0	0	**1**	0	4
Anticonvulsants	0	1	0	0	0	0	**2**	3

Table 2.5: *Classification of the EEG training dataset with method of Ruigt* et al *(1993) with cross-validation.*

	Predicted Class				
Class	Antidep	Antipsy	Hypno	Stimul	Total
Antidepressants	**49**	37	19	7	112
Antipsychotics	20	**23**	7	6	56
Hypnotics	4	4	**32**	0	40
Stimulants	5	9	3	**63**	80

3

Visualizing Classes in EEG Data

3.1 Introduction

Ever since the introduction of Shannon's entropy (Shannon, 1948), information theory has been of great theoretical and applied interest. While initially information theory was designed to solve problems in the field of communication theory, it is nowadays used in a much broader range of context. It has found applications in statistical inference (Kullback, 1959), biology (Adami, 2004), quantum information theory (Bennett and Shor, 1998), alongside many other areas.

Informally, mutual information is the information shared between two random variables. It thus measures to what extent knowledge of one of these variables reduces the uncertainty about the other one. In this chapter we apply the concept of mutual information in the field of classification of longitudinal profiles. Our purpose is to propose a simple graphical tool to explore the classes in a multi-class longitudinal classification problem. More precisely, we want to know how much information a new observation has in common with each of the considered classes in the training data set, and use this knowledge to classify the new observation. In this way, the classes can be visualized in a simple plot, showing the densities of a function of the mutual information measure, constraining this measure to lie between 0 and 1, for the class of interest against the remainder of the classes. The measure makes use, in an intuitive way, of the variability between and within classes.

In Section 3.2, the concept of mutual information is briefly reviewed, together with the estimation procedure proposed by Kraskov *et al* (2004) and the use of this to visualize a measure, which can be seen as a distance between and within a particular class and the rest (Wouters *et al*, 2008b). Thereafter, the proposed method is applied to our data, and the results are outlined in Section 3.3 and discussed in Section 3.4.

3.2 Methodology

3.2.1 Mutual Information Concepts

The *entropy* $H(X)$ of a random variable X is the uncertainty of that random variable, defined by

$$H(X) = -\int f(x) \log f(x) dx \tag{3.1}$$

where $f(x)$ is the density function of X. The *mutual information* $I(X, Y)$ of two random variables X and Y, measures the reduction in uncertainty about X due to the knowledge contained in Y.

$$I(X, Y) = H(X) - H(X|Y) \tag{3.2}$$

If X and Y are two continuous random variables, their mutual information can be written as:

$$I(X, Y) = \int\int f(x, y) \log \left(\frac{f(x, y)}{f_x(x) f_y(y)} \right) dxdy \tag{3.3}$$

where $f(x, y)$ is the joint density function of X and Y, and $f_x(x) = \int f(x, y)dy$ and $f_y(y) = \int f(x, y)dx$ are the marginal density functions of X and Y respectively. In other words, mutual information quantifies the distance between the joint distribution of the random variables X and Y on the one hand and what the joint distribution would be if X and Y were independent on the other hand. When X and Y are independent, their mutual information $I(X, Y)$ equals 0. Moreover, mutual information is nonnegative and symmetric, i.e., $I(X, Y) = I(Y, X)$.

To estimate $I(X, Y)$ from a set of data points $\{x_i, y_i\}$ alone, without knowing the densities f, f_x and f_y we will use the method suggested by Kraskov *et al* (2004). They propose two algorithms, both based on entropy estimates from k-nearest neighbour distances. According to Kraskov *et al* (2004), both algorithms perform very similarly with respect to computation time, statistical error and systematic error and score better than conventional estimators in terms of bias. In what follows we will restrict to the algorithm based on rectangular neighbourhoods, but the other method can be used as well.

First note that the mutual information $I(X, Y)$ can be written as

$$I(X, Y) = H(X) + H(Y) - H(X, Y), \tag{3.4}$$

As a consequence estimating $H(X)$, $H(Y)$ and $H(X, Y)$ is enough to get an estimate for $I(X, Y)$. For this purpose we can use the Kozachenko-Leonenko estimate for Shannon entropy (Shannon, 1948):

$$\hat{H}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N} \sum_{i=1}^{N} \log \varepsilon(i), \tag{3.5}$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function, $\varepsilon(i)/2$ is the distance between x_i and its k-th neighbour, d is the dimension of x and c_d is the volume of the d-dimensional unit ball.

As proposed by Kraskov *et al* (2004) we will now denote the edge lengths of the smallest rectangle around point i containing k neighbours by $\varepsilon_x(i)$ and $\varepsilon_y(i)$. The number of points with $||x_i - x_j|| \leq \varepsilon_x(i)/2$ and $||y_i - y_j|| \leq \varepsilon_y(i)/2$ are given by $n_x(i)$ and $n_y(i)$, respectively. The estimate for the mutual information is now

$$\hat{I}(X, Y) = \psi(k) - 1/k - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N) \tag{3.6}$$

where $\langle \ldots \rangle$ denotes averages, both over all $i = 1, \ldots, N$ and over all realizations of the random samples.

The mutual information can be linked to the information-theoretic measure of association defined by

$$R_h^2(X, Y) = 1 - e^{-2I(X,Y)} \tag{3.7}$$

This measure of association ranges from 0 to 1 and equals zero if and only if X and Y are independent.

Intuitively, this measure can be used to have an idea how much information each member of a class, c say, contains about the class itself (for this we will use the mean of all elements in the class c), and how distant they are from the other classes (mean of the compound-dose combinations that do not belong to the class c). This measure can be computed for each of the compound-dose combinations, leading to a measure that can be associated to distance within the class c and distance between the class c and the rest of the classes, respectively (Wouters *et al*, 2008b).

3.2.2 Computational Issues

In our situation, the mutual information is calculated to determine how much information is shared between two classes with respect to one or a group of variables. The six variables we consider are the number of minutes spent in each of the six sleeping stages in 32 subsequent periods of 30 minutes.

For each group of variables, we propose to determine the association between class c and class c', the latter corresponding to all other classes, as

$$R_h^2(c, c') = \frac{1}{2}\left(\frac{1}{n_1}\sum_{s\in c} R_h^2(s, \bar{c}') + \frac{1}{n_2}\sum_{s'\in c'} R_h^2(s', \bar{c})\right), \tag{3.8}$$

where n_1 and n_2 are the number of observations in class c and c' respectively, s are the measurements for subject s that belongs to class c and $\bar{c}$ is the mean of the observations in c, both with respect to the group of variables under consideration.

The association within a certain class c can be determined by (3.8) for each observation within class c, with respect to the mean of the class c, upon replacing c' with c.

Since we have only a small number of observations in each of the classes, we will use bootstrap samples (Efron, 1979) to get a more reliable estimate of the association between and within classes. When calculating the between-class association of c and c', we take 1000 bootstrap samples in c and c' by resampling the six-variate profiles of the rats in class c and c' respectively. For each of these bootstrap samples we calculate the association between all observations in c and the mean of the samples. For the within-class association, we proceed in a similar way. Now, this can be used to estimate the distribution of the within- and between-class distances, which brings forward how much a class is overlapping with the rest of the classes. It also provides information about the variability within a class. For each class, the group of variables that produces the smallest overlap between that class and the rest will be retained.

Eventually, this can also be used to calculate the sensitivity and specificity for each class with respect to the rest of the classes based on this set of variable. To illustrate how this is computed, we focus on the placebo class. The cut-off value v_{pl} for placebo will be defined as the point where the density function of the measure of association $R_h^2(\text{placebo}, \text{placebo})$ crosses the density function of $R_h^2(\text{rest}, \text{placebo})$, and such that the area under each density curve, respectively above and below the cut-off, is maximized. The sensitivity for placebo is now calculated as the area under the placebo density curve above v_{pl}, where the specificity is defined as the area under the density curve for the other classes below v_{pl}. Other cut-off values can be entertained, depending on the objectives of the study.

Table 3.1: *Information-theoretic measure of association $R_h^2(c, \overline{c}')$ (and standard error) for the 5 classes (columns) and the mean of each class (rows) when all six sleeping stages are used.*

Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.79 (0.05)**	0.74 (0.06)	0.75 (0.06)	0.74 (0.06)	0.60 (0.12)
Antipsy	0.74 (0.06)	**0.87 (0.04)**	0.79 (0.05)	0.77 (0.06)	0.73 (0.08)
Antidep	0.75 (0.06)	0.79 (0.05)	**0.82 (0.05)**	0.76 (0.06)	0.69 (0.10)
Hypnot	0.74 (0.06)	0.77 (0.06)	0.76 (0.06)	**0.85 (0.09)**	0.63 (0.13)
Stimul	0.60 (0.12)	0.73 (0.06)	0.69 (0.10)	0.63 (0.13)	**0.77 (0.10)**

3.3 Results

For each of the 5 classes in the EEG dataset, we calculate the mutual information within and between the classes with respect to the 6 sleeping stages, using 1000 bootstrap samples from the training dataset. This means that to calculate the measure of association R_h^2 we are using all 6 sleeping stages from the EEG experiment. In Table 3.1 we see in each row the information-theoretic association between the considered class and the 4 other classes. The within-class association is displayed in boldface. Standard deviations are presented parenthetically.

The difference between within- and between-class association is rather small. This indicates that using all information does not necessarily mean that we will reach good differentiation. It is even possible that using all information is generating too much noise, thereby creating difficulties to separate out the classes.

We know that the sleeping stages influenced by a psychotropic compound can be different for different classes. For example, for stimulants it is expected that the number of minutes spent in Active Wake should be larger than for the other four classes. In a similar fashion, this occurs for other classes, in which other sleeping stages can be influenced by the compound-dose combination. Therefore, instead we perform the same bootstrap exercise, but now on all possible combinations of sleeping stages. For each of these combinations the means and the 90% quantiles of the information-theoretic associations within a class and between this class and the rest is computed. The combination of variables for which the overlap between such two quantiles is the smallest is retained. The sleeping stages used for each class are reported in Table 3.2.

With this combination of sleeping stages, we get the associations as displayed in

Table 3.2: *Sleeping stages used for each class.*

Class	Sleeping Stages
Placebo	Active Wake – Light Sleep
Antipsychotic	Active Wake – Passive Wake – REM Sleep
Antidepressant	Active Wake – Light Sleep
Hypnotic	Active Wake – Light Sleep
Stimulant	Active Wake – Light Sleep – Deep Sleep

Table 3.3: *Information-theoretic measure of association $R_h^2(c_1, \overline{c}_2)$ (and standard error) for the 5 classes (columns) and the mean of each class (rows) using only the selected sleeping stages.*

Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.84 (0.03)**	0.68 (0.09)	0.81 (0.03)	0.70 (0.08)	0.66 (0.08)
Antipsy	0.68 (0.09)	**0.84 (0.05)**	0.74 (0.08)	0.65 (0.11)	0.67 (0.11)
Antidep	0.81 (0.03)	0.74 (0.08)	**0.85 (0.03)**	0.68 (0.12)	0.68 (0.10)
Hypnot	0.70 (0.08)	0.65 (0.11)	0.68 (0.12)	**0.80 (0.09)**	0.48 (0.21)
Stimul	0.66 (0.08)	0.67 (0.11)	0.68 (0.10)	0.48 (0.21)	**0.79 (0.08)**

Table 3.3. In Figure 3.1 the density function for the associations within each class is plotted with a solid line, the density of the association between this class and the rest of the classes is plotted with a dashed line. Below the graphs, the 90% quantiles for within-class (solid) and between-class (dashed) associations are displayed, where the bullet is indicating the median. The wide quantile for the hypnotics reveals that there is a large variability within this group, which is not unexpected given the low number of hypnotic compounds in the training dataset. Also for stimulants, we notice high variability, which can be attributed to the natural variability in this drug class.

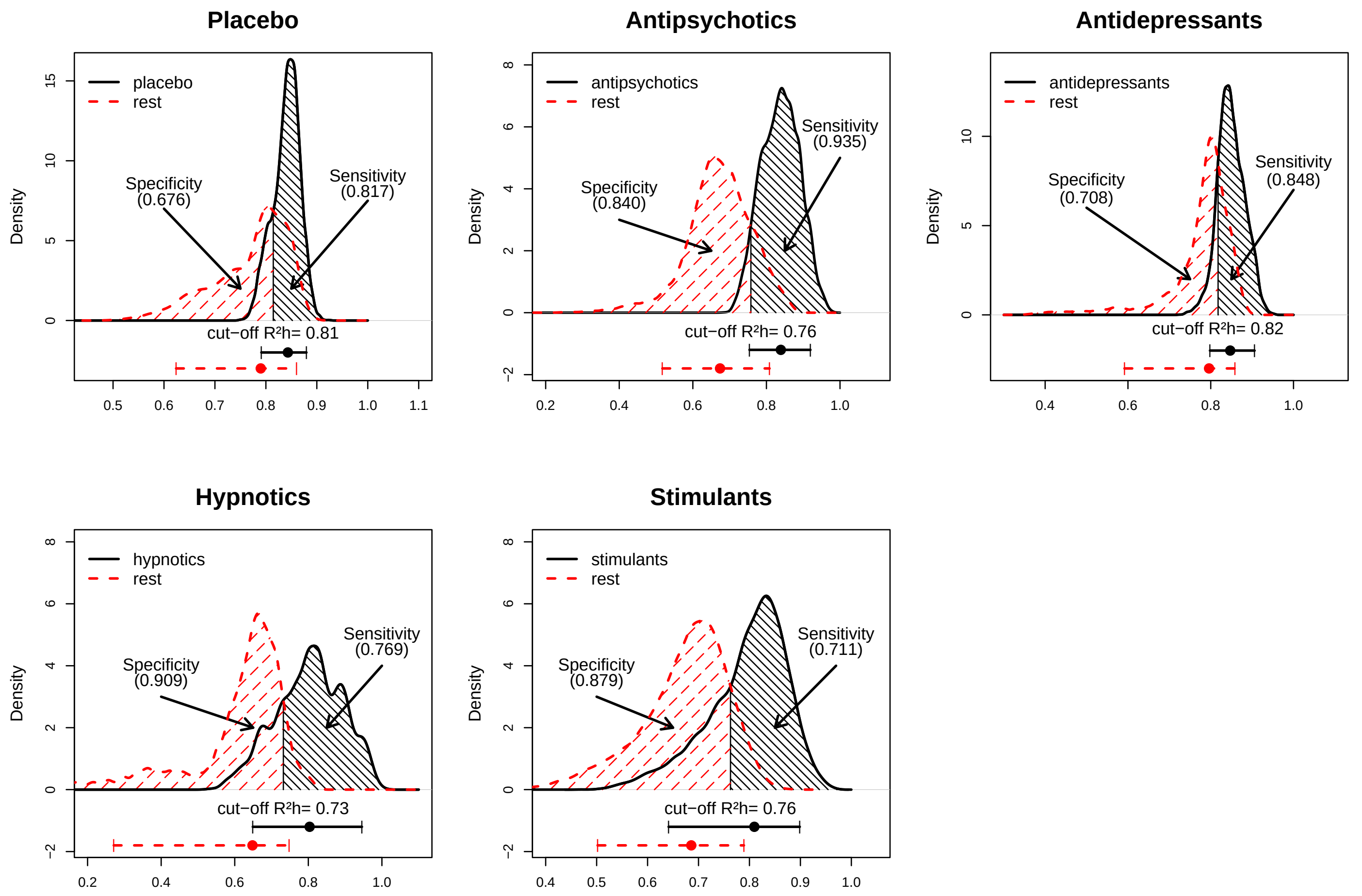


Figure 3.1: *Density plots for the association within (solid) and between (dashed) classes. Below are the 90% quantiles for within- (solid) and between- (dashed) class association.*

Both Table 3.3 and Figure 3.1 reveal that placebo and antidepressant are not clearly separated from the rest. For the other three classes, we get a nice distinction between these classes and the remaining four, although there still is some overlap in the quantiles.

For each class, we can calculate the specificity and sensitivity, as indicated in Figure 3.1. The cut-off value, taken as the value in which both curves intersect, is also marked in the figure. The sensitivity for hypnotics and stimulants are rather low, meaning that we have a high probability of incorrectly classifying an hypnotic or stimulant. This seems to contradict the results we have seen in Table 3.3, but it can be explained by the large variability in these two classes, which is not taken into account in Table 3.3.

On the other hand, for placebo and antidepressants, we get a relatively low specificity, so the probability of classifying a compound wrongly into one of these classes is rather high.

The low specificity for placebo can be explained by the relatively large amount of compounds in this class (40% of the training dataset), leading to a small variability in $R_h^2(\text{placebo}, \text{placebo})$, while the variability in $R_h^2(\text{rest}, \text{placebo})$ is much larger since 4 smaller classes are pooled here. Also antidepressants are well represented in the training dataset (24%), which is again reflected in the small variability in $R_h^2(\text{antidep}, \text{antidep})$, but now the rest of the classes are dominated by placebo which explains the small variability in $R_h^2(\text{rest}, \text{antidep})$. The low specificity therefore reflects mainly the fact that placebo and antidepressants are difficult to separate out.

Let us now turn to the test dataset. For each of the test compounds we calculate the association between this compound and a class mean for all the bootstrap samples in this class. The mean association with that class is indicated in Figure 3.2 with a cross. So for each compound, we get 5 crosses, one for each of the classes. The quantiles as already shown in Figure 3.1 are again plotted here for comparative purposes.

The association between the three placebo test compounds and placebo is indeed high and falls within the interval of the within-class association. But also the association between these compounds and both antidepressant as well as hypnotic is high. Thus, the logical conclusion is that these compounds are not likely to be antipsychotics or stimulants, but no decisive answer can be given on the actual class of the compounds. Both antipsychotic test compounds seem to belong to the hypnotics instead of antipsychotics. The association between test compound 5 and antipsychotic is even very low. For the sixth test compound, the associations with placebo, antidepressant and hypnotic fall nicely within the quantiles for these classes, making it hard to predict the class. A similar problem arises for the antidepressant test compounds 7 and 9, but for these ones the association with antidepressant falls borderline outside the quantile. Test compound 8, which is also an antidepressant, has a low association with all the classes, but can be borderline interpreted as an hypnotic. Based on the quantiles, we would correctly classify test compounds 10 and 11 as hypnotics. Finally, for the stimulant test compounds 12 and 14, we can conclude that they probably belong to stimulants, given the structure brought forward in the graph, while compound 13 would be wrongly classified as an hypnotic.

In the test dataset, the hypnotic and the stimulant compounds have a high association with their own class and a lower association relative to the quantile with the other drug-classes. So, based on the information-theoretic association, we would expect to predict their class correctly. While for the placebo and antidepressant test compounds we see that the association with the actual class is high, the association between the compound and some other classes is also high, indicating potential problems in identifying such compounds. For the antipsychotic test compounds, the association with the antipsychotic training data falls below the lower boundary of the corresponding quantile, while the association with hypnotics is relatively high. This would lead to an incorrect classification.

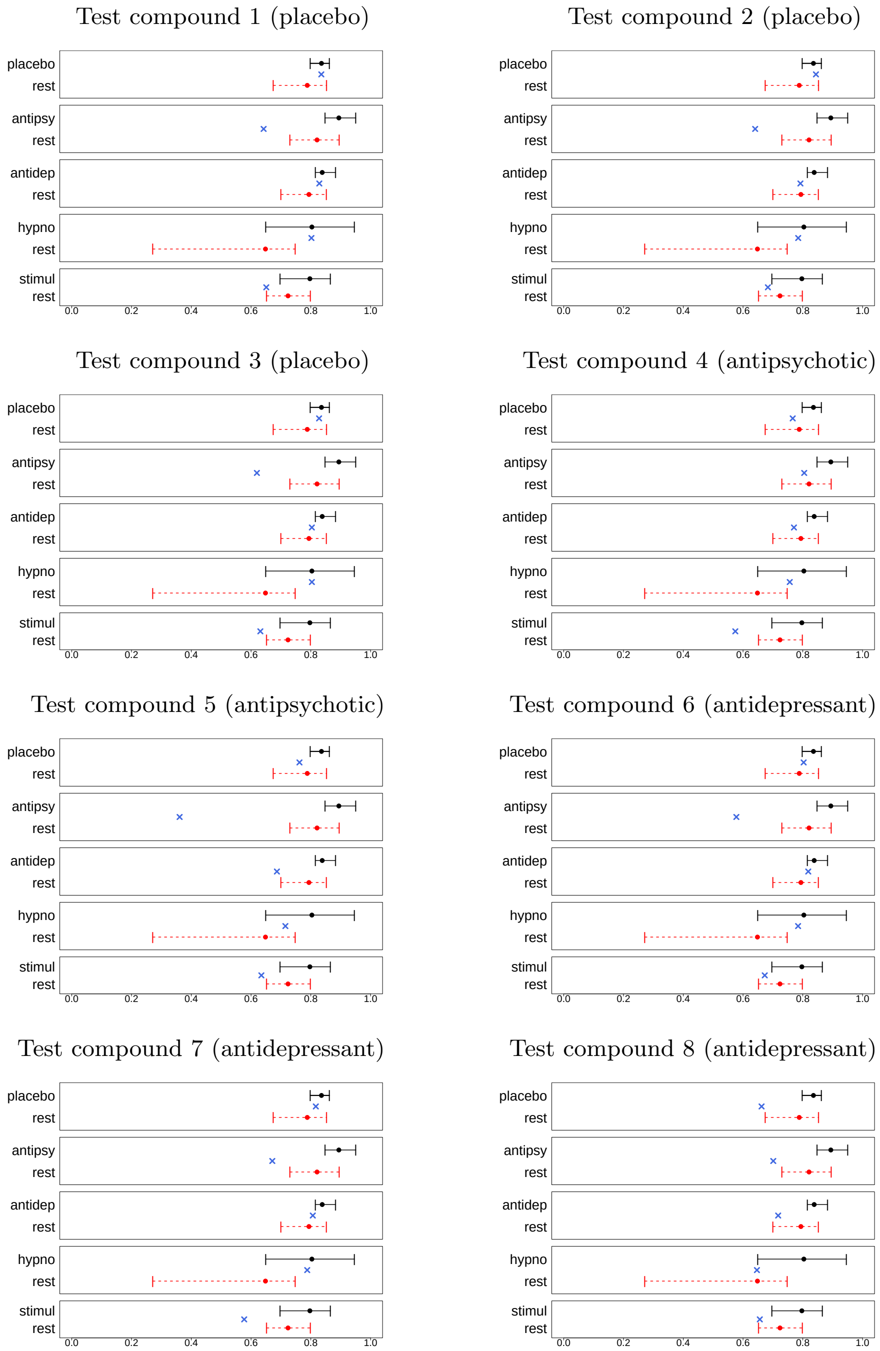
3.4 Discussion

In this chapter we proposed an exploratory tool to visualize classes in EEG data. Rather than the actual prediction of the class of a new compound, we focus on how well separated the classes are in a particular set of data, to be able to explain difficulties in the classification procedure. This method can also be used to detect a-typical samples in both training and test datasets.

The overlapping density plots, together with the sensitivities and specificities for each of the classes already reveal that the classification of EEG data is a difficult task. For hypnotics and stimulants, this can be explained by the high variability in the data. This indicates that elaborate classification techniques, taking into account the longitudinal nature of the data are needed to perform a formal classification.

It is also important to note that considering more information does not automatically lead to improved classification. Here, we have shown that when not all 6 sleeping stages (all information about the psychotropic drug) are used to obtain a measure of the within- and between-distances for each class, a better distinction can be obtained. This also can induce a reduction of the variability associated with each of these measures. For the selection of the sleeping stages to be used in each class, we looked at the overlap in the quantiles. Of course, other selection criteria can be used here. A valuable alternative could be to select those sleeping stages that maximize both specificity and sensitivity for the considered class.

Further potential for the information-theoretic approach is in the screening of new variables. Instead of going through the entire classification process, new variables can be quickly tested for their discriminative potential, by computing the association within and between each of the classes.



Test compound 1 (placebo)

Test compound 2 (placebo)

Test compound 3 (placebo)

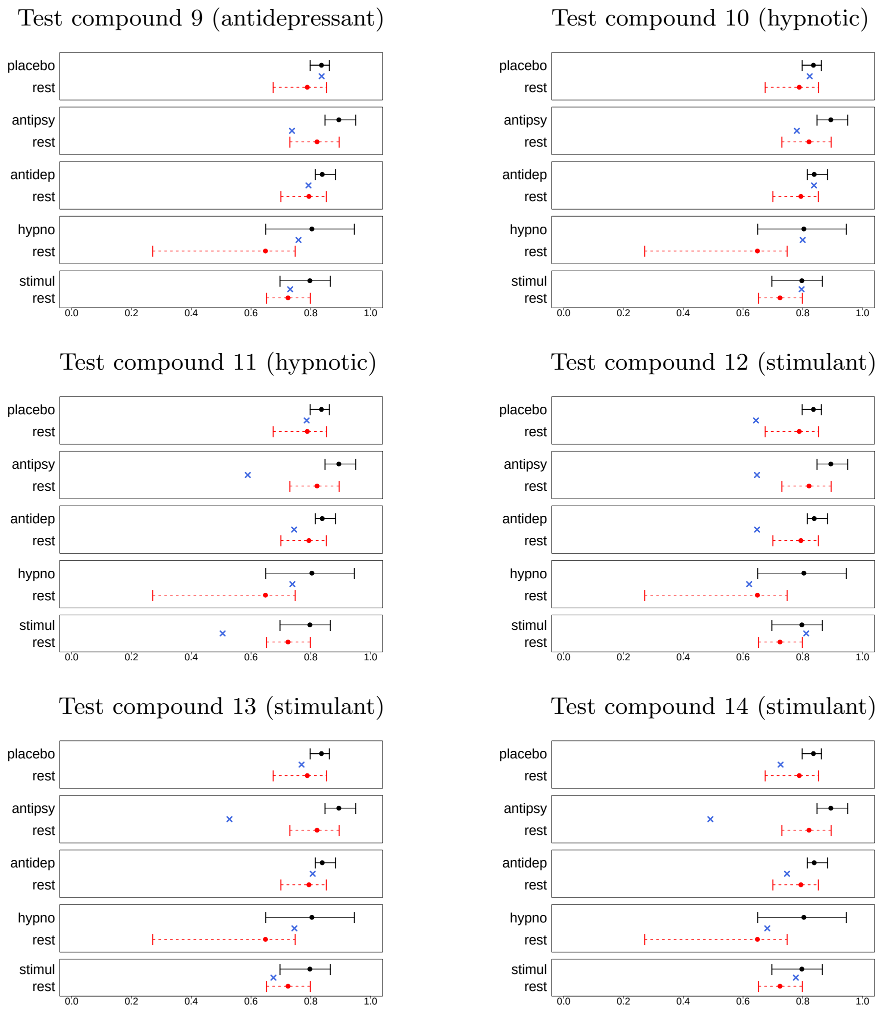
Test compound 4 (antipsychotic)

Test compound 5 (antipsychotic)

Test compound 6 (antidepressant)

Test compound 7 (antidepressant)

Test compound 8 (antidepressant)



Test compound 9 (antidepressant)

Test compound 10 (hypnotic)

Test compound 11 (hypnotic)

Test compound 12 (stimulant)

Test compound 13 (stimulant)

Test compound 14 (stimulant)

Figure 3.2: *Information-theoretic association between the test compounds and the five drug classes (indicated with a cross), together with the quantiles for the association between and within classes.*

4

Modeling EEG Data

This chapter focusses on how to deal with the longitudinal character of the data. For the modeling of this type of data, mixed effects models are a widely used approach (Verbeke and Molenberghs, 2000). To capture the irregular trends in the individual profiles in the EEG dataset, an even more flexible model is needed. Several approaches can be used that allow flexibility in order to cope with the irregularities observed in the mean profiles. In this thesis we will focus on two modeling approaches: (1) a fractional polynomial model (Royston and Altman, 1994) combined with random effects and (2) a mixed model with splines as fixed and random effects (Ruppert *et al*, 2003).

4.1 Methodology

We will briefly introduce the linear mixed model methodology (Verbeke and Molenberghs, 2000) which is the basis of the two modeling approaches used throughout this thesis.

4.1.1 Linear Mixed Model

Let us first introduce some notations. Let $\boldsymbol{Y}_i$ denote the n_i-dimensional vector of measurements available for subject $i = 1, \ldots, N$. A linear mixed effect model or

LMM then assumes that $\boldsymbol{Y}_i$ satisfies

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon_i}, \tag{4.1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of population-average regression coefficients, also called *fixed effects* and $\boldsymbol{b}_i$ is a q-dimensional vector of subject specific regression coefficients, also called *random effects.* The matrices X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates also known as design matrices. The random effects $\boldsymbol{b}_i$ and residual components $\boldsymbol{\varepsilon_i}$ are assumed to be independent with distributions $N(\mathbf{0}, D)$, and $N(\mathbf{0}, \Sigma_i)$, respectively. Note that Σ_i depends on i only through n_i, the number of measurements available for subject i. Thus, in summary,

$$\boldsymbol{Y}_i|\boldsymbol{b}_i \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \Sigma_i). \tag{4.2}$$

Inference is based on the marginal distribution of $\boldsymbol{Y_i}$ which can be expressed as

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i) \tag{4.3}$$

Now that we have introduced the notions of the linear mixed model, we can continue with the description of the fractional polynomial mixed model that we propose to fit to the EEG data.

4.1.2 Fractional Polynomial Mixed Model

Fractional polynomials consider beside the integer powers of a covariate (e.g. time) also fractional powers. In this way a wide range of shapes can be modeled. As soon as non-integer powers are allowed for, the number of potential models is endless and it is wise to consider a priori a sensible model building strategy. This has been provided by Royston and Altman (1994).

A fractional polynomial of degree m (Royston and Altman, 1994) is defined as any function of the form

$$\phi_m(X; \boldsymbol{\beta}, \boldsymbol{p}) = \sum_{k=0}^{m} \beta_k H_k(X), \tag{4.4}$$

where the degree m is a positive integer, $\boldsymbol{p}$ is a real-valued vector of powers with $p_1 \leq \ldots \leq p_m$ and $\beta_0, \beta_1, \ldots, \beta_m$ are real-valued coefficients. We set $p_0 = 0$, $H_0(X) = 1$, and, for $k = 1, \ldots, m$

$$H_k(X) = \begin{cases} \ln X & \text{if } p_k = 0 \text{ and } p_k \neq p_{k-1}, \\ X^{p_k} & \text{if } p_k \neq p_{k-1}, \\ H_{k-1}(X)\ln X & \text{if } p_k = p_{k-1}. \end{cases} \tag{4.5}$$

Royston and Altman (1994) argue that fractional polynomials with degree higher than 2 or 3 are rarely required in practice and that powers can be restricted to the set $\Omega = \{-2,\ -1.5,\ -1,\ -0.5,\ 0,\ 0.5,\ 1,\ 1.5,\ 2\}$. Therefore, second-degree fractional polynomials will be used in combination with the mixed effects model (4.1). For all 36 pairwise combinations of p_1 and p_2 in Ω, a fractional polynomial model is fitted. The Akaike's Information Criteria (AIC), obtained for the 36 models, are sorted and the powers p_1 and p_2 that lead to the smallest AIC value are retained.

In our application, a different fractional polynomial mixed model will be fitted for every compound-dose combination j, to allow for different shapes in different compound-doses. In each of these models, a random effect per subject is included. The model is given by

$$\boldsymbol{Y_{ij}} = (\beta_{0j} + b_{0ij}) + (\beta_{1j} + b_{1ij})H_{j1}(X_{ij}) + (\beta_{2j} + b_{2ij})H_{j2}(X_{ij}) + \boldsymbol{\varepsilon_{ij}}, \tag{4.6}$$

where $\boldsymbol{Y_{ij}}$ is the n-dimensional vector of measurements for subject i in compound-dose combination j. $\boldsymbol{\beta_j}$ is the 3-dimensional vector of fixed effects for compound-dose combination j and $\boldsymbol{b}_{ij}$ is the 3-dimensional vector of subject-specific random effects for compound-dose combination j. The random effects $\boldsymbol{b}_{ij}$ and residual components $\boldsymbol{\varepsilon}_{ij}$ are assumed to be independent with distributions $N(\mathbf{0}, D_j)$, and $N(\mathbf{0}, \Sigma_{ij})$, respectively, where D_j is an unstructured (3×3) matrix and Σ_{ij} is a diagonal $(n \times n)$ matrix.

Since we are using different combinations of powers p_1 and p_2, it is possible to end up with large scale differences between the covariates in the model, especially when the range of values taken by X_{ij} is large, which can then induce computational problems when inverting $X'X$. To avoid this, the covariates $H_{j1}(X_{ij})$ and $H_{j2}(X_{ij})$ are standardized in the following way:

$$X_{ij\ell} = \frac{H_{j\ell}(X_{ij}) - E[H_{j\ell}(X_{ij})]}{\sqrt{\mathrm{Var}[H_{j\ell}(X_{ij})]}}, \qquad \ell = 1, 2. \tag{4.7}$$

which leads to the following model

$$\boldsymbol{Y_{ij}} = (\beta_{0j} + b_{0ij}) + (\beta_{1i} + b_{1ij})X_{ij1} + (\beta_{2j} + b_{2ij})X_{ij2} + \boldsymbol{\varepsilon_{ij}}, \tag{4.8}$$

Now it is clear that for these new standardized variables X_{ijl}, the scale lies between 0 and 1, which solves the computational issues. Centering the predictor variable also reduces the multicollinearity drastically as is stated by Neter *et al* (1996) (P. 296) and tends to avoid computational difficulties.

Fractional Polynomial Mixed Model for EEG Data

Let us now have a look at the parametrization of the fractional polynomial model for the EEG dataset. Since we want to allow for a different shape in each compound-dose combination, we model the number of minutes spent in the 6 sleep-waking stages for every compound-dose combination in the training and the test dataset by a fractional polynomial model with mixed effects. In this way, not only the coefficients, but also the powers p_1 and p_2 can be different for different compound-dose combinations. For example, for the minutes spent in Active Wake in time period k for rat i in compound-dose combination j the model (4.8) becomes

$$\begin{aligned}
&(\text{AW min})_{ijk} = \\
&\left[(\beta_{0j}+b_{0ij})+(\beta_{1j}+b_{1ij})\frac{t_k^{p_{1jl}}-E[\boldsymbol{t}^{p_{1jl}}]}{\sqrt{\text{Var}[\boldsymbol{t}^{p_{1jl}}]}}+(\beta_{2j}+b_{2ij})\frac{t_k^{p_{2jl}}-E[\boldsymbol{t}^{p_{2jl}}]}{\sqrt{\text{Var}[\boldsymbol{t}^{p_{2jl}}]}}\right] I(t_k)+ \\
&\left[(\gamma_{0j}+c_{0ij})+(\gamma_{1j}+c_{1ij})\frac{t_k^{p_{1jd}}-E[\boldsymbol{t}^{p_{1jd}}]}{\sqrt{\text{Var}[\boldsymbol{t}^{p_{1jd}}]}}+(\gamma_{2j}+c_{2ij})\frac{t_k^{p_{2jd}}-E[\boldsymbol{t}^{p_{2jd}}]}{\sqrt{\text{Var}[\boldsymbol{t}^{p_{2jd}}]}}\right](1-I(t_k)) \\
&+\varepsilon_{ijk}, \qquad (4.9)
\end{aligned}$$

where $(\text{AW min})_{ijk}$ is the number of minutes spent in AW for rat i in compound-dose combination j during the k^{th} time period ($i = 1,\dots,8$, $j = 1,\dots,64$ and $k = 1,\dots,32$). The index l refers to the light period, d to the dark period. We standardized the vectors $\boldsymbol{t}^{p_1}$ and $\boldsymbol{t}^{p_2}$, where $\boldsymbol{t}$ is the vector of all time periods, $\boldsymbol{t} = (1,\dots,32)'$. The vectors $\boldsymbol{\beta}_j = (\beta_{0j},\beta_{1j},\beta_{2j})$ and $\boldsymbol{\gamma}_j = (\gamma_{0j},\gamma_{1j},\gamma_{2j})$ are the compound-dose specific regression coefficients for the light and the dark period respectively, while $\boldsymbol{b}_{ij} = (b_{0ij},b_{1ij},b_{2ij})$ and $\boldsymbol{c}_{ij} = (c_{0ij},c_{1ij},c_{2ij})$ are the random effects or rat specific coefficients. The random effects $\boldsymbol{b}_{ij}$ and $\boldsymbol{c}_{ij}$ are assumed to be independent with distributions $N(\mathbf{0},D_j^b)$ and $N(\mathbf{0},D_j^c)$ respectively, where D_j^b and D_j^c are unstructured (3×3) matrices. The residual components ε_{ijk} are also independent with distribution $N(0,\sigma_j^2)$. The function $I(t)$ is an indicator function specified as

$$I(t) = \begin{cases} 1 & \text{if } t \leq 20, \\ 0 & \text{otherwise.} \end{cases}$$

in order to identify the change between light and dark period.

In analogy to equation 4.9 similar models can be defined when one of the powers $p_{1jl}, p_{2jl}, p_{1jd}$ or p_{2jd} equals zero or when $p_{1jl} = p_{2jl}$ or $p_{1jd} = p_{2jd}$.

4.1.3 Splines Model Using Mixed Model Parametrization

Another flexible way to smooth the irregular trends in the data is through cubic splines, which are piecewise third degree polynomials with components smoothly spliced together. The regions defining the pieces are separated by a series of knots $\xi_1, \cdots, \xi_K$, not necessarily equally spaced. Ruppert *et al* (2003) define a cubic spline model with knots $\xi_1, \cdots, \xi_K$ by

$$S(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \beta_{k+3}(x - \xi_k)_+^3. \tag{4.10}$$

where $(x - \xi_k)_+^3$ are the truncated power basis functions, the subscript $+$ is the notation for the positive part of the function. It is clear that $K + 4$ basis functions are needed to describe a cubic spline with K knots, which are the K truncated power basic functions $(x - \xi_k)_+^3$, complemented with the functions 1, x, x^2 and x^3.

A natural cubic spline additionally requires that the function is linear beyond the boundary knots. A natural cubic spline with K knots is represented by K basis functions. One can start from the natural truncated power basis and derive the reduced basis functions $N_i(x)$ by imposing the boundary constraints. In this way, we arrive at

$$N_1(x) = 1, \; N_2(x) = x, \; N_{k+2}(x) = d_k(x) - d_{K-1}(x) \tag{4.11}$$

where $k = 1, \cdots, K - 2$ and

$$d_k(x) = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k} \tag{4.12}$$

While the truncated power basis is simple in concept, it can lead to severe rounding problems when powers of large numbers must be calculated, which makes this basis not too attractive numerically. An alternative choice are B-spline basis functions (Dierckx, 1993), which are defined strictly local. They are non-zero over an interval of at most five knotpoints, meaning that their evaluation rarely gets out of hand.

A simple and straightforward method to fit splines is by considering the coefficient of each knot a fixed effect, usually referred to as *regression spline*. However, this approach tends to overfit the data, leading to a too coarse regression curve. This can be overcome by including splines in the mixed model framework, meaning that each knot point coefficient acts as a random effect (Verbyla *et al* (1999), Ruppert *et al* (2003)). The variance component governing these random effects controls and describes the degree of flexibility and smoothness of the model.

In what follows, we will use the natural cubic splines as fixed and random effects in the mixed effects model 4.1. We consider then the design matrix

$$Z = \begin{bmatrix} N_1(x_1) & \cdots & N_K(x_1) \\ \vdots & \ddots & \vdots \\ N_1(x_n) & \cdots & N_K(x_n) \end{bmatrix} \tag{4.13}$$

The mixed effects model with natural cubic splines as fixed and random effects can be written as

$$\boldsymbol{Y_i} = Z_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon_i}, \tag{4.14}$$

where $\boldsymbol{\beta}$, $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$ are defined as in (4.1).

Splines Model for EEG Data

For each sleeping stage we will now fit a linear mixed model with natural cubic splines as fixed and random effects for each compound-dose combination separately. For the light period we assume five equally spaced internal knots, and four knots in the dark period. For example, the model for the minutes spent in Active Wake for subject i in compound-dose combination j in the light period now becomes

$$(\text{AW min})_{ij} = (\beta_{0j} + b_{0ij} + Z_{\text{light}}\boldsymbol{\beta}_j + Z_{\text{light}}\boldsymbol{b}_{ij} + \boldsymbol{\varepsilon}_{ij}$$

with design matrix Z specified as

$$Z_{\text{light}} = \begin{bmatrix} N_{l1}(t_1) & \cdots & N_{l5}(t_1) \\ \vdots & \ddots & \vdots \\ N_{l1}(t_{20}) & \cdots & N_{l5}(t_{20}) \end{bmatrix}$$

where $N_{l1}, \cdots, N_{l5}$ are the five B-spline basis functions for a natural cubic spline with five knots. Since the time points are equal for all the subjects, we get the same design matrix Z_{light} for all subjects in all compound-dose combinations. Similarly, a model with four knots is specified for the dark period, with basis functions $N_{d1}, \cdots, N_{d4}$. This results in the following model for the minutes spent in Active Wake in time period k for rat i in compound-dose combination j

$$\begin{aligned} (\text{AW min})_{ijk} = {} & [(\beta_{0j} + b_{0ij}) + N_{l1}(\beta_{1j} + b_{1ij}) + N_{l2}(\beta_{2j} + b_{2ij}) + N_{l3}(\beta_{3j} + b_{3ij}) \\ & + N_{l4}(\beta_{4j} + b_{4ij}) + N_{l5}(\beta_{5j} + b_{5ij})]\, I(t_k) \\ & + [(\gamma_{0j} + c_{0ij}) + N_{d1}(\gamma_{1j} + c_{1ij}) + N_{d2}(\gamma_{2j} + c_{2ij}) + N_{d3}(\gamma_{3j} + c_{3ij}) \\ & + N_{d4}(\gamma_{4j} + c_{4ij})]\, (1 - I(t_k)) + \varepsilon_{ijk}, \end{aligned} \tag{4.15}$$

where $I(t)$ is defined as in equation (4.9). The vectors $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \cdots, \beta_{5j})$ and $\boldsymbol{\gamma}_j = (\gamma_{0j}, \gamma_{1j}, \cdots, \gamma_{4j})$ are the compound-dose specific regression coefficients for the light and the dark period respectively, while $\boldsymbol{b}_{ij} = (b_{0ij}, b_{1ij}, \cdots, b_{5ij})$ and $\boldsymbol{c}_{ij} = (c_{0ij}, c_{1ij}, \cdots, c_{4ij})$ are the random effects or rat specific coefficients. The random effects $\boldsymbol{b}_{ij}$ and $\boldsymbol{c}_{ij}$ are assumed to be independent with distributions $N(\mathbf{0}, D_j^b)$ and $N(\mathbf{0}, D_j^c)$ respectively, where D_j^b and D_j^c are unstructured (3×3) matrices. The residual components ε_{ijk} are also independent with distribution $N(0, \sigma_j^2)$.

4.2 Application to the EEG Dataset

We will now apply both modeling approaches to the EEG data, and compare their fitting abilities.

4.2.1 Fractional Polynomial Mixed Model

The model described in Section 4.1.2 was fitted to the data. As an illustration, the fitted models for Active Wake, Light Sleep and Deep Sleep for one compound-dose combination in each of the 5 classes are shown in Figure 4.1. The individual observed profiles for each of the eight rats in the compound-dose combination are represented by grey lines, with the blue dashed line representing the mean of that particular compound-dose combination and the red solid line depicting the fitted fractional polynomial model. For all the classes, the fitted profile nicely follows the mean evolution over time. Since we allowed for different parameters for light and dark period in model 4.9, the jump at time point 20 could be captured. For the other sleeping stages, and the other compound-dose combinations in the dataset, similar model fits were obtained.

4.2.2 Splines Model

The splines model parametrization described in Section 4.1.3 is now used to model the EEG data. Figure 4.2 shows the observed profiles for Active Wake, Light Sleep and Deep Sleep for the eight rats in one compound-dose combination in each class (grey solid lines), together with the mean observed profile (blue dashed line) and the fitted spline model (red solid line). For each of the compound-dose combinations, the fitted profiles follow the mean evolution over time. Also here, the jump at time point 20, is nicely captured, since different model parameters are allowed for the light and the dark period. Similar model fits were obtained for the other sleeping stages and the other compound-dose combinations.

4.3 Discussion

In Figure 4.3, the model fits obtained with the fractional polynomial mixed model and the splines mixed model are compared. For one compound-dose combination in each of the classes, we draw in black the mean observed profile for the number of minutes spent in each of the six sleeping stages. In red is the fitted profile for the fractional polynomial mixed model, while in blue we have plotted the model fit obtained with the splines mixed model.

Both models are following the data very well. Except for REM Sleep in the placebo and antidepressant compound-dose combination, the fits are very similar. In the plots for REM Sleep in placebo and antidepressant, it can be seen that the splines model is trying to capture the irregularities in the profile, while the fractional polynomial mixed model only focusses on the main trend.

By allowing for different parameters for the light and the dark period, we are able to capture the jump at the moment that the light is switched off. However, by doing this we assume indirectly that the sleep wake pattern in the dark period is independent of the pattern in the light period, which might be questionable, but it simplifies the model and the fitting process dramatically, avoiding also computational issues.

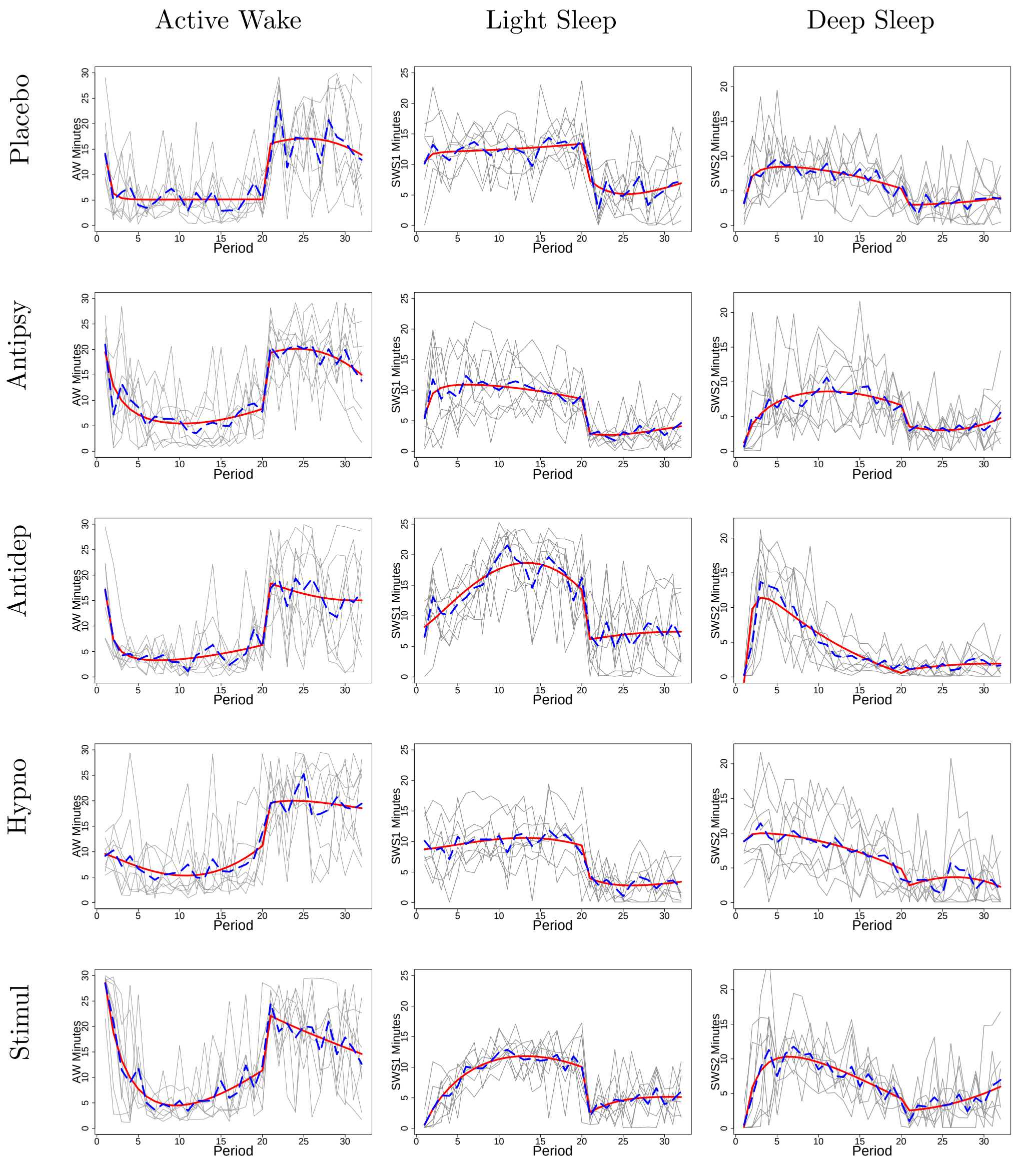


Figure 4.1: *Fractional Polynomial Mixed Model. Observed individual profile for Active Wake, Light and Deep Sleep, for the eight rats in one compound-dose combination in each of the classes in grey solid lines together with the mean profile for that compound-dose combination (blue dashed line) and the fitted fractional polynomial mixed model (red solid line).*

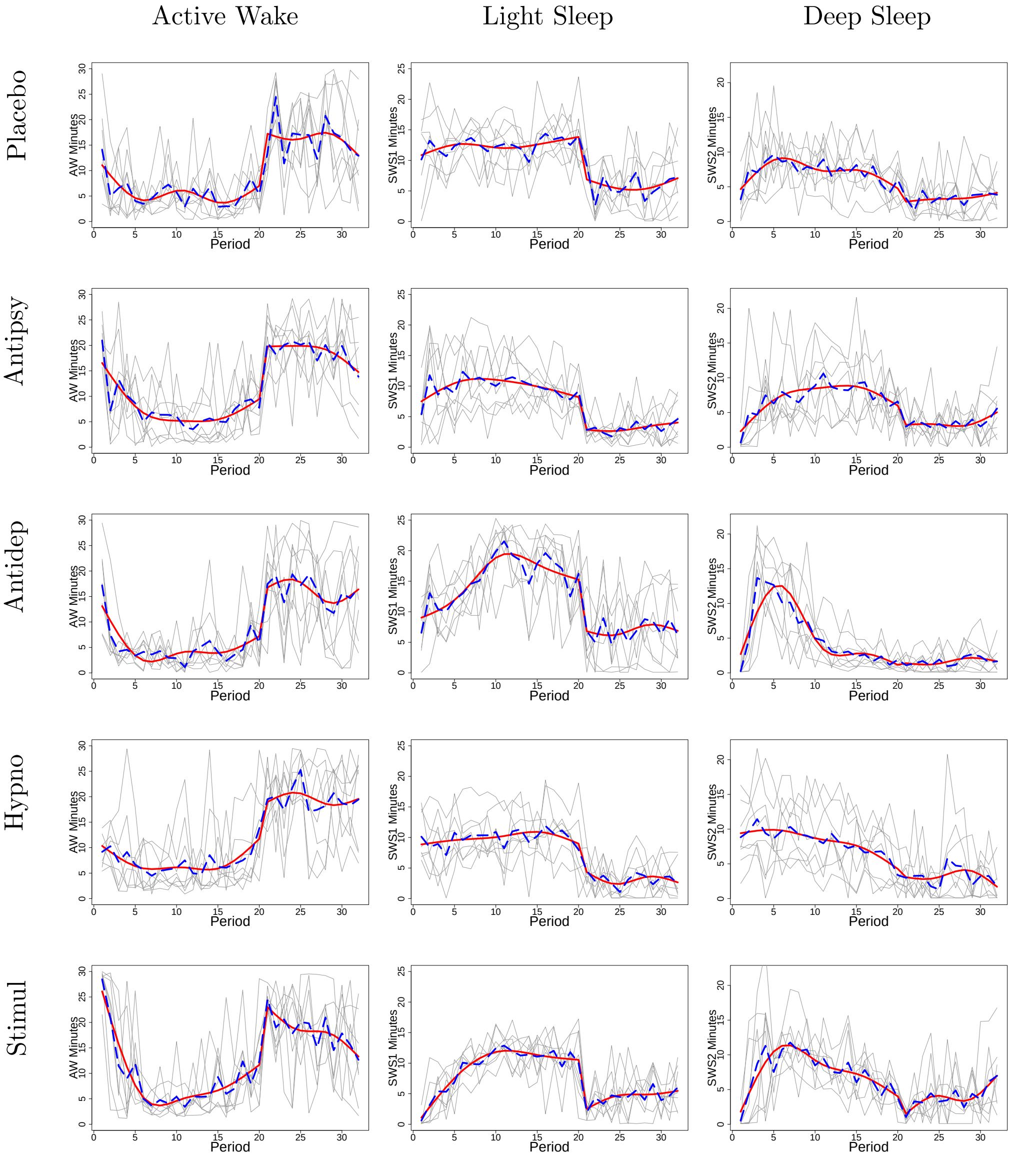


Figure 4.2: *Splines Model. Observed individual profile for Active Wake, Light and Deep Sleep, for the eight rats in one compound-dose combination in each of the classes in grey solid lines together with the mean profile for that compound-dose combination (blue dashed line) and the fitted splines model (red solid line).*

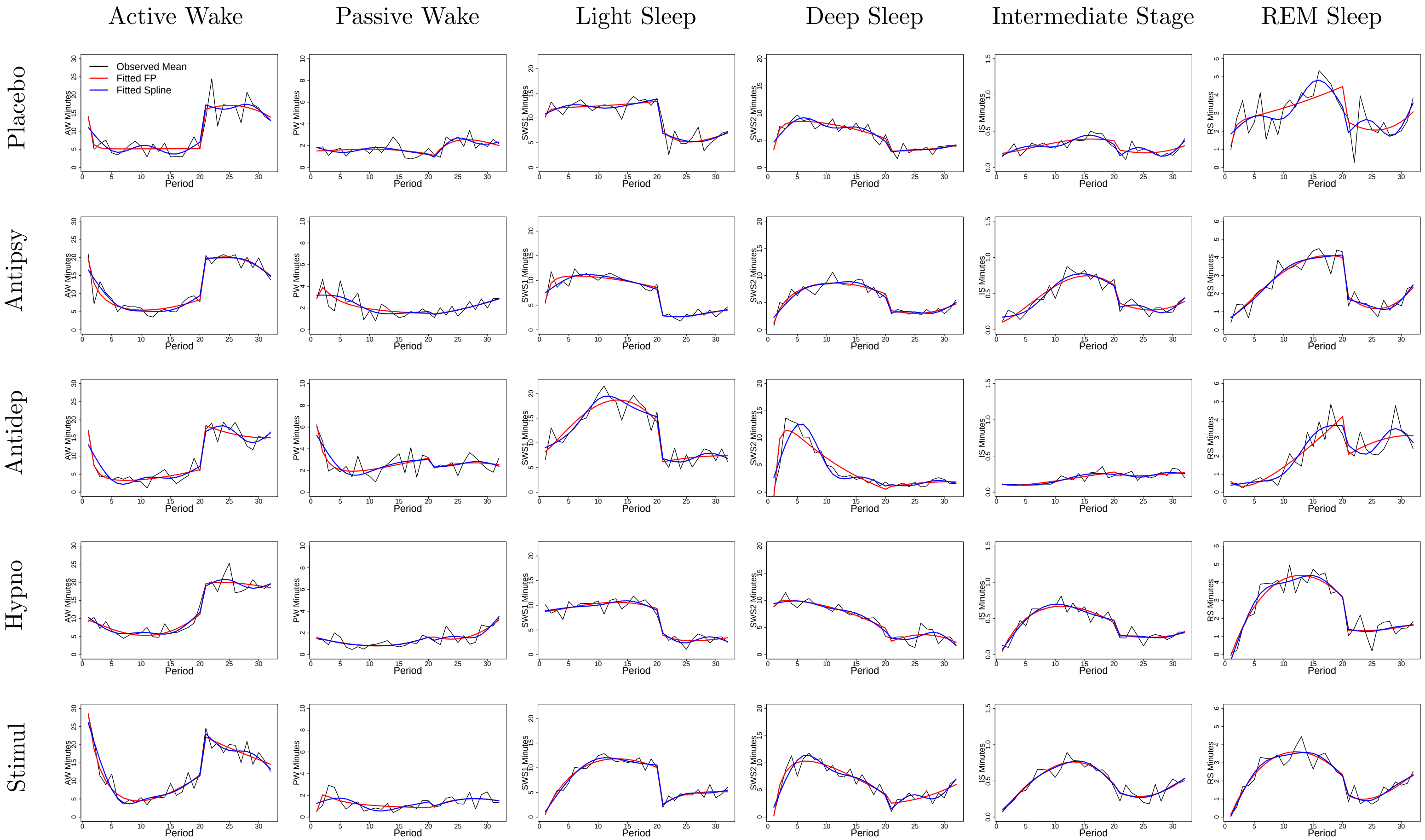


Figure 4.3: *Observed mean profile for one compound-dose combination in each of the classes in black solid lines together with the fitted fractional polynomial mixed model in red and the fitted splines mixed model in blue.*

5

Doubly Hierarchical Supervised Learning Analysis

Developing classification rules for complex data structures, such as multiple-class problems with a longitudinal design, is a non-trivial task and requires appropriately tailored methods. Precisely these features are encountered in EEG experiments, where we have the number of minutes spent in 6 sleeping stages (multivariate), measured at 32 time-intervals (longitudinal) for 5 classes (multiple-class). Classical supervised learning techniques are not suited to handle the combination of a multiple-class problem and a longitudinal design. In order to deal with these features in the construction of a classification rule, we propose a flexible two-step procedure, termed doubly hierarchical supervised learning analysis (DHSLA) which copes with such issues (Wouters *et al*, 2007a).

In the first section of this chapter, the general idea of this two-step procedure is explained. Later, in the next two sections, we elaborate on the first and second step respectively. The final section is devoted to study some computational issues that could arise.

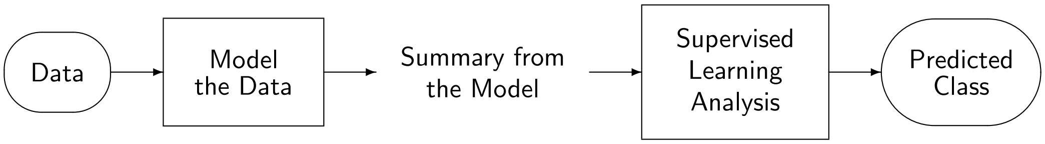


Figure 5.1: *Diagram representing the doubly hierarchical supervised learning analysis.*

5.1 Description of General Two-stage Procedure

To establish classification rules for application with multiple-class longitudinal data we propose a flexible hierarchical supervised learning tool, that allows to take into account the specific nature of the multiple drug classes, as well as the longitudinal aspect of the data. The procedure is schematically represented in Figure 5.1. In the first stage of the DHSLA, the longitudinal profiles are modeled and appropriate summaries are extracted from the model fit. These summary measures are then used, in the second stage, as input for the supervised learning analysis, in view of classifying the data. This second stage proceeds in a hierarchical fashion.

In both the first and the second stage, various techniques can be used. In the two subsequent sections we will elaborate on the two stages in turn and highlight a few implementations of the procedure.

The term doubly hierarchical supervised learning analysis refers to the hierarchical structure present in our data (which will be modelled using hierarchical models) on the one hand, and the hierarchical character of the supervised learning procedure on the other hand.

The evaluation of the DHSLA procedures is done through cross-validation on the two levels in the dataset, i.e. the rat and the compound-dose combination level (Section 5.5). Because of the hierarchical character of the procedures some issues arise regarding the selection of the variables, which is solved using a lack of classification measure, and the calculation of the posterior probabilities. These issues will be dealt with in Section 5.4 – 5.6.

5.2 Phase I: Modeling the Data

In the first phase of the doubly hierarchical supervised learning analysis, the data is modeled using a flexible model. Different modeling techniques can be used here. As we have seen in the previous chapter, both fractional polynomial mixed models and splines mixed models are fitting the data very well. We have chosen to work with the fractional polynomial mixed model, but splines mixed models would have been a sensible choice as well.

5.3 Phase II: Supervised Learning Approach

The continuation of the classification procedure necessitates informative summaries of the highly variable longitudinal profile available for each rat. To this end, the parameters of the models in the first stage, i.e., the collection made up of $\beta_{0j} + b_{0ij}$, $\beta_{1j} + b_{1ij}$, $\beta_{2j} + b_{2ij}$ and the powers corresponding with H_{j1} and H_{j2}, denoted by p_{1j} and p_{2j}, will be used as input in the supervised learning procedure.

To establish and optimize a flexible classification rule, we proceed in a stepwise, hierarchical fashion. In a first step we discriminate, one class from the rest, using the parameters describing the longitudinal profiles. Then, focus shifts to the remaining classes. This process continues until a complete decision tree has been built. The order in which the classes are discriminated is determined based on the performance in the training dataset. Different orders are checked and the one that leads to the best classification results using cross-validation, is retained.

Various supervised learning techniques can be used at this stage, three of which will be considered here: linear (LDA), flexible (FDA), and mixture (MDA) discriminant analysis. We will briefly outline each of the three choices in turn in the next subsections.

We will construct a toy example using a small crossectional part of the EEG dataset, containing 48 rats who received placebo compounds and 80 rats treated with stimulant compounds. The number of minutes spent in Active Wake and Deep Sleep during the first 2 hours are considered as covariates. The three supervised learning approaches will be applied to the toy example to illustrate the discriminant properties of each procedure. The data is shown in Figure 5.2, where rats receiving placebo treatments are represented by red spheres and rats receiving stimulant compounds are displayed as green triangles.

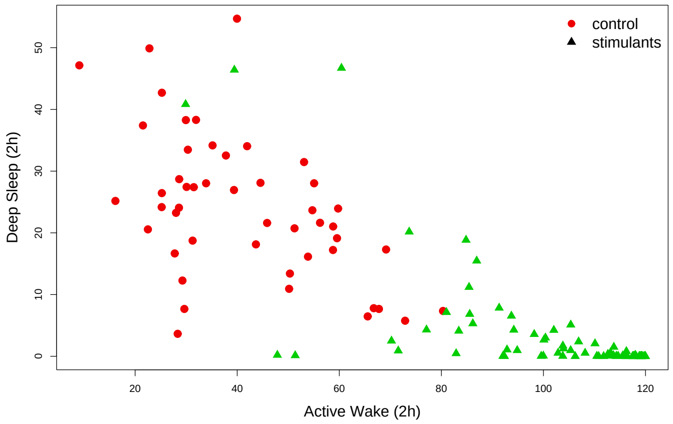


Figure 5.2: *Toy example. Number of minutes spent in Active Wake (X-axis) and Deep Sleep (Y-axis) during the first 2 hours of the observation period, for placebo (spheres) and stimulants (triangles).*

5.3.1 Linear Discriminant Analysis

In linear discriminant analysis (Johnson and Wichern, 1992), each class c is assumed to follow a multivariate normal distribution $N(\boldsymbol{\mu}_c, \Sigma)$ with class-specific mean $\boldsymbol{\mu}_c$ and common variance-covariance matrix Σ, leading to a linear decision rule. In the case of two classes this rule is given by

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln \left[\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right] \tag{5.1}$$

where $c(1|2)$ is the cost of misclassifying a subject from class 2 in class 1 and p_c $(c = 1, 2)$ is the prior probability of belonging to class c. One assigns a new subject with response vector $\boldsymbol{x}$ to class 1 if the inequality is satisfied, and to class 2 otherwise.

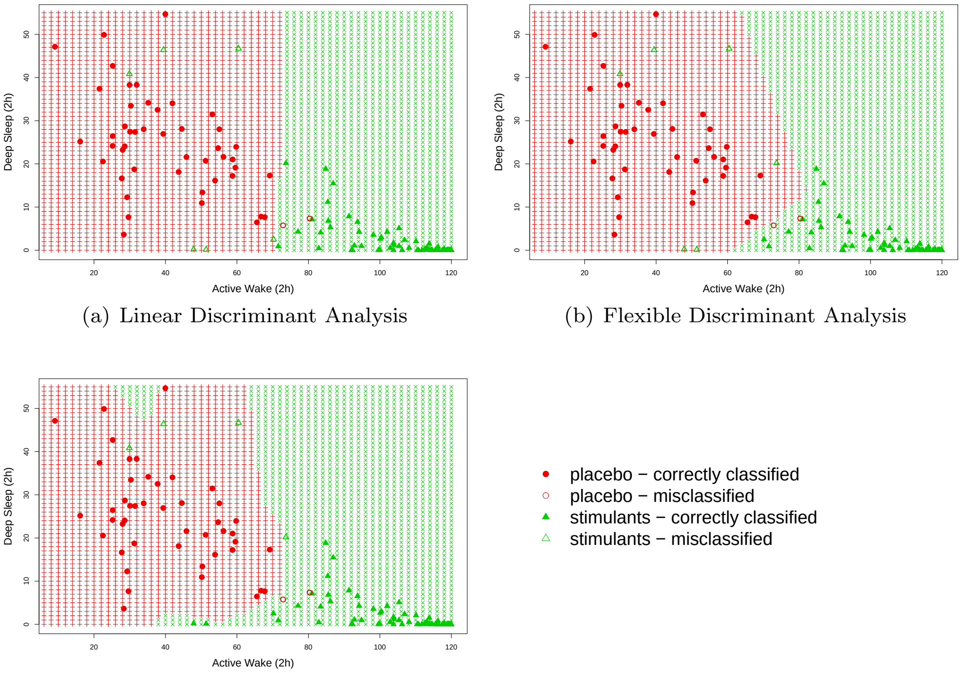
Using Bayes' theorem, the posterior probability of belonging to class k can be calculated as

$$P(c|\boldsymbol{x}) = \frac{p_c f_c(\boldsymbol{x})}{\sum_u p_u f_u(\boldsymbol{x})} \tag{5.2}$$

where $f_c(\boldsymbol{x})$ is the group-specific density estimate at $\boldsymbol{x}$ from class c given by

$$f_c(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\Sigma_c^{-1}| \exp\left(-0.5(\boldsymbol{x} - \mu_c)^T \Sigma_c^{-1} (\boldsymbol{x} - \mu_c)\right)\big) \tag{5.3}$$

In Figure 5.3(a) the classification result obtained with LDA for the toy example is shown. The observations on the left side of the boundary (indicated in red) are



(a) Linear Discriminant Analysis

(b) Flexible Discriminant Analysis

(c) Mixture Discriminant Analysis

Figure 5.3: *Classification result for the toy example obtained with (a) linear discriminant analysis, (b) flexible discriminant analysis, with MARS and (c) mixture discriminant analysis, with 3 prototypes per class. The region classified as placebo is indicated in red, the region classified as stimulants in green.*

classified as placebo, while the ones on the right side are assigned to stimulants. We used cross-validation to classify the subjects in the dataset. In this way, two rats who received a placebo treatment and six rats who received a stimulant were misclassified.

The linearity of rule (5.1) makes it easy to implement and interpret the decision boundaries. Unfortunately, in a number of situations linear decision boundaries are not adequate to separate the classes. To account for this, Hastie, Tibshirani and Friedman (2001) propose generalizations of LDA, such as flexible discriminant analysis (FDA), mixture discriminant analysis (MDA), and penalized discriminant analysis (PDA). In what follows, we will confine attention to FDA, because of its ability to model irregular decision boundaries, and MDA which allows us to use more than one prototype per class.

5.3.2 Flexible Discriminant Analysis

The linear discriminant analysis can be regarded as a sequence of linear regression followed by classification to the closest class centroid in the space of fits. In that case, the linear regression is defined via

$$ASR = \frac{1}{n}\sum_{\ell=1}^{K}\left[\sum_{i=1}^{n}(\theta_\ell(g_i) - x_i^T\beta_\ell)^2\right]. \tag{5.4}$$

where $\theta_1, \cdots, \theta_K$ are independent scorings for the class labels and $\eta_\ell = X^T\beta_\ell$ are the corresponding linear maps. These scores and maps are chosen to minimize the average squared residual (ASR).

The linear regression can now be generalized to a more flexible function (Hastie, Tibshirani and Friedman, 2001). In this more general form, the regression problems are defined via

$$ASR = \frac{1}{n}\sum_{\ell=1}^{L}\left[\sum_{i=1}^{n}(\theta_\ell(g_i) - f(x_i))^2 + \lambda J(f)\right],$$

where J is a regularizing function.

In our particular case, we use Multivariate Adaptive Regression Splines (MARS) models (Friedman, 1991). The input space is partitioned into regions, each with its own linear regression equation. The MARS equation is given by

$$f(x) = \gamma_0 + \sum_{m=1}^{M}\gamma_m h_m(x),$$

where M is the number of non-constant terms in the model and h_m is a basis function in the collection

$$\mathcal{C} = \{(X_j - t)_+, (t - X_j)_+ | t \in \{x_{1j}, x_{2j}, \ldots, x_{nj}\}, j = 1, 2, \ldots, p\},$$

with n the number of observations.

The classification obtained for the toy example with flexible discriminant analysis using MARS is shown in Figure 5.3(b). The resulting boundaries are clearly more flexible than the ones obtained with LDA, but still we were not able to classify correctly two rats from the placebo group and six rats from the stimulant group.

Note that we get indeed more flexible boundaries with FDA, but there is a price to pay. Too flexible models can result in overfitting, and the generalizability of the results may be in doubt. We try to overcome this issue by using cross validation in the selection of the model.

5.3.3 Mixture Discriminant Analysis

Mixture discriminant analysis (Hastie, Tibshirani and Friedman, 2001) is an extension of LDA, to be viewed as a prototype classifier with each class represented by its centroid. We assign an observation to the closest centroid using an appropriate distance measure. In many situations, a single prototype per class is not sufficient, in which case mixture models can be used. Assume classes have several prototypes, thence a Gaussian mixture model for the class c could be considered. The corresponding density is

$$P(X|c) = \sum_{r=1}^{R_c} \pi_{c_r} \phi(X; \mu_{c_r}, \Sigma),$$

where the mixing proportions satisfy $\sum_{r=1}^{R_c} \pi_{c_r} = 1$, R_c is the number of prototypes for class c and Σ the covariance matrix used as a metric throughout. For class c with a priori probabilities Π_c, we estimate the parameters by maximizing the joint log-likelihood:

$$\sum_{c=1}^{K} \sum_{g_i=c} \log \left[\sum_{r=1}^{R_c} \pi_{c_r} \phi(X; \mu_{c_r}, \Sigma) \Pi_c \right].$$

The expectation-maximization (EM) algorithm is a convenient mode to obtain maximum likelihood estimates (Dempster, Laird, and Rubin, 1977). The algorithm consists of iterating between the expectation (E) and maximization (M) steps, until convergence. In our situation, they take the following forms.

E-step: Given the current values for the parameters, compute the weights associated with the subclasses c_r:

$$W(c_r|x_i, g_i) = \frac{\pi_{c_r} \phi(x_i; \mu_{c_r}, \Sigma)}{\sum_{l=1}^{R_c} \pi_{c_l} \phi(x_i; \mu_{c_l}, \Sigma)}. \tag{5.5}$$

M-Step: Compute weighted MLEs for the parameters of each of the component Gaussian densities, within each of the classes, using the weights obtained from (5.5).

In Figure 5.3(c) we see the classification result for the toy example obtained with MDA using a mixture of three normal densities in each class. The result is now two misclassified rats in placebo, and four misclassified rats in stimulants.

As in FDA, also here the concern of overfitting raises, but again we will use cross-validation deal with this issue.

5.4 A Proposal for a Measure of Lack-of-Classification

To determine the goodness of our discriminant analysis, we have to take into account not only the error rate and the posterior probability with respect to the class discriminated in step s, denoted by C_s, but also with respect to the other classes in step s, denoted by C_{-s}. Therefore we calculate Error1, focussing on the false-negative cases, and Error2, which is monitoring the false-positives, as follows:

$$\begin{aligned} \text{Error1}_s &= \text{ERR}_{C_s C_{-s}} + (1 - \text{PP}_{C_s C_s}) \qquad (5.6) \\ \text{Error2}_s &= \text{ERR}_{C_{-s} C_s} + \sum_{k \neq C_s} \text{PP}_{k C_s} \qquad (5.7) \end{aligned}$$

where ERR_{kl} is the misclassification percentage from class k into class l and PP_{kl} is the posterior probability for rats belonging to class k to be classified in class l.

The lack-of-classification measure (LC) in step s is now defined as a weighted sum of Error1 and Error2.

$$\text{LC}_s = w_{s1} \cdot \text{Error1}_s + w_{s2} \cdot \text{Error2}_s. \qquad (5.8)$$

Different weights w_{s1} and w_{s2} can be chosen, depending on the type of application. In our particular case, we chose the weights $w_{s1} = s + 1$ and $w_{s2} = 2 \cdot (g - s)$. Along the process more weight is given to the false-negatives whereas the weight given to the false-positives is decreased. The choice of these weights is based on the fact that the algorithm discriminates in the first steps the classes that are well differentiated from the rest whereas in the final steps the classes are less clearly separated.

The lack-of-classification measure is now standardized and corrected for the number of parameters in the model by multiplying with a decreasing function of the number of sleep-wake stages used, given by $F(\text{ss})$.

$$\text{LC'}_s = 1 - \left(1 - \frac{\text{LC}_s}{2 \cdot w_{s1} + (g - s + 1) \cdot w_{s2}}\right) \cdot F(\text{ss}). \qquad (5.9)$$

Again, different choices can be entertained for $F(ss)$. We choose to proceed with $F(\text{ss}) = 0.999^{\text{ss}}$.

Note that LC'_s is a useful device to ensure that a particular sleeping stage be added, whenever the researcher is quite certain that such a stage would lead to added benefit in terms of classification. Of course, the choice for this particular function is a pragmatic one and, arguably, other functional forms could be entertained as well. The most important thing here is that this choice exhibits good behaviour.

The model leading to the lowest lack-of-classification LC' will be retained.

5.5 Selection Procedure to Retrieve Best Model

We consider two different selection procedures, both based on 10-fold cross-validation, a technique to be described next, inspired by the fact that the dataset can be divided randomly at each of two different hierarchical levels.

In the first approach (Selection Procedure I), we use rats as the unit of analysis. The 472 rats comprising the dataset are then randomly divided into ten groups. For every parameter combination obtained from the fractional polynomial models and for each sleep-wake stage, one of the 10 samples is used as a test dataset, while the remaining 9 samples are assigned the role of training sets. For the test dataset, both the misclassification error and the posterior probabilities are calculated. The combination of sleep-wake stages resulting in the lowest lack of classification measure is retained. This is repeated for every step in the DHSLA.

Selection Procedure II uses 10-fold cross-validation at the compound-dose combination level. We randomly divide the 59 such combinations into ten approximately equal sized groups and then proceed in the same way it was described above.

For each selection procedure, the error count is calculated at both levels, i.e., rat and compound-dose combination. The first is computed as the average of the percentage of misclassified rats in each class ($\text{error}_{\text{rat}}$), while the second uses the percentages of compound-dose combinations that are misclassified in a particular class ($\text{error}_{\text{c-d}}$).

5.6 Adjusted Posterior Probabilities

The posterior probabilities for a hierarchical classification process need to be adjusted. In each step, we have to correct the posterior probabilities of a compound-dose combination for the fact that this combination has not been classified in one of the previous steps.

For the first step, we calculate the posterior probabilities P_1 of belonging to the class that we want to discriminate in step 1 and $Q_1 = 1 - P_1$. Given that s splits have been made, the values of the posterior probabilities at split $s + 1$ are then multiplied with the posterior probabilities of not being classified at the previous steps in the class we were interested to discriminate from the rest. So,

$$P_{s+1}^{\text{adj}} = Q_i^{\text{adj}} P_{s+1} \tag{5.10}$$

$$Q_{s+1}^{\text{adj}} = Q_i^{\text{adj}} Q_{s+1} \tag{5.11}$$

In what follows, all posterior probabilities reported will use the correction outlined here.

6

Comparison of Doubly Hierarchical Supervised Learning Analysis with Different Discriminant Techniques

In this chapter we will apply the doubly hierarchical supervised learning analysis with fractional polynomial mixed models in the first phase and the three different discriminant techniques explained in the previous chapter in the second phase. The procedure is schematically represented in Figure 6.1. The results in this chapter are based on Wouters *et al* (2007b).

6.1 Phase I: Fractional Polynomial Mixed Model

As a prelude to our doubly hierarchical supervised learning analysis, we model, for every compound-dose combination in the training and the test dataset, the number

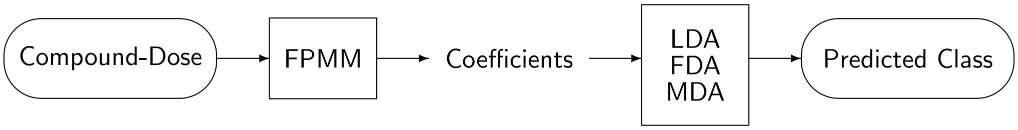


Figure 6.1: *Diagram representing the doubly hierarchical discriminant analysis, when a fractional polynomial mixed model (FPMM) is used in Stage I and linear (LDA), flexible (FDA) or mixture discriminant analysis (MDA) are used in Stage II.*

of minutes spent in the 6 sleep-waking stages by a fractional polynomial model with mixed effects as has been done in Chapter 4. Not only the coefficients, but also the powers p_1 and p_2 can be different for different compound-dose combinations. For the minutes spent in each sleeping stage in time period k for subject i in treatment j we will fit model (4.9)

Given that the drugs are administered at the beginning of the light period and based on experts' belief that the action may be quite different during the initial period, it is sensible to allow for a different, perhaps more pronounced action of the drug during the first three hours after administration. Therefore, we choose to consider a separate model for the first three hours, for example for Active Wake this model becomes

$$(\text{AW min})_{ijk} = (\delta_{0j}+d_{0ij})+(\delta_{1j}+d_{1ij})\frac{t_k^{p_{1jf}} - E[\boldsymbol{t}^{p_{1jf}}]}{\sqrt{\text{Var}[\boldsymbol{t}^{p_{1jf}}]}}+(\delta_{2j}+d_{2ij})\frac{t_k^{p_{2jf}} - E[\boldsymbol{t}^{p_{2jf}}]}{\sqrt{\text{Var}[\boldsymbol{t}^{p_{2jf}}]}}+\epsilon_{ijk}. \tag{6.1}$$

Figure 6.2 shows the fitted models for all the compound-dose combinations in each of the 5 classes. For placebo we see a very consistent course in all the compound-dose combinations. For other classes, especially for the stimulants, there is a wide variability in shapes.

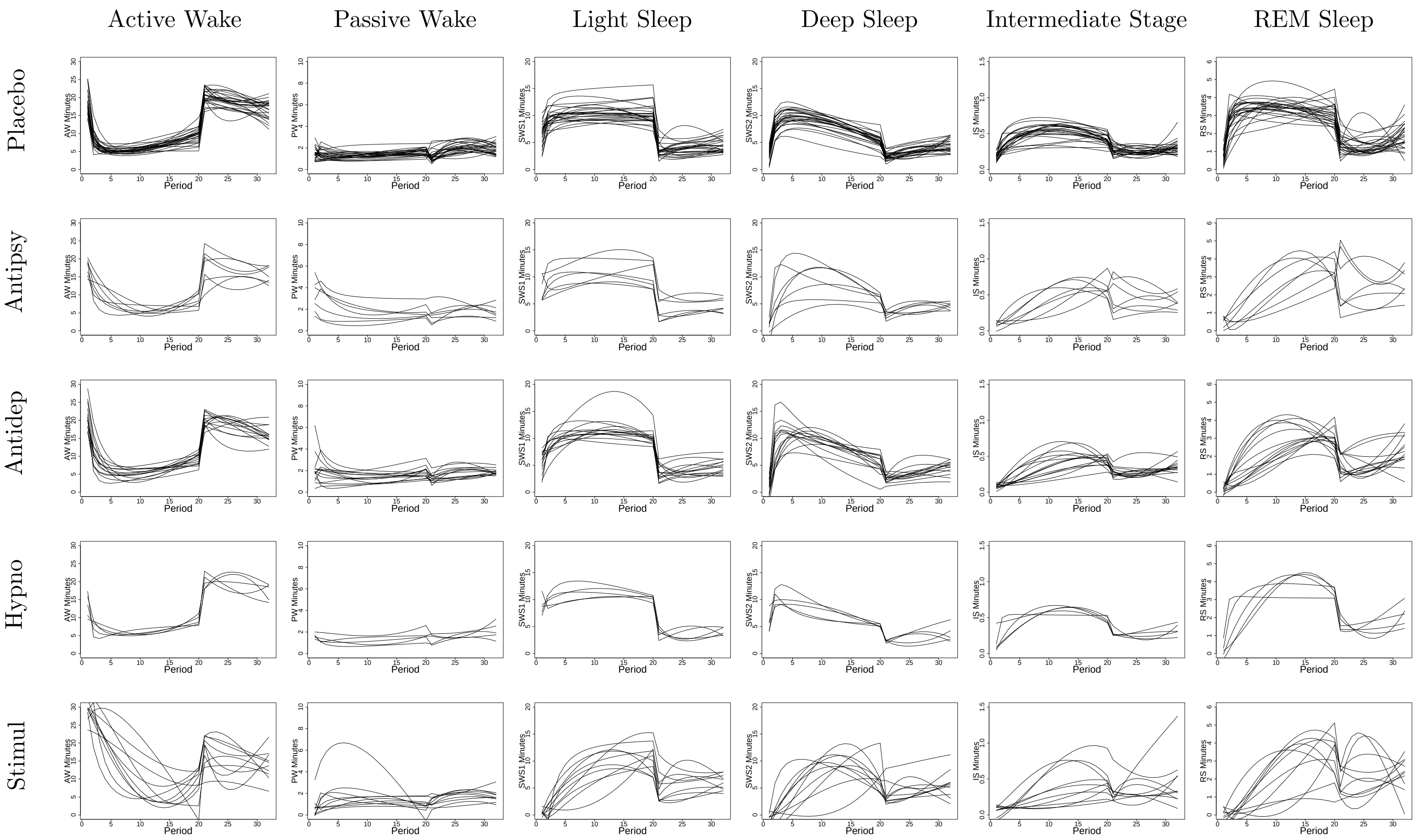


Figure 6.2: *Fitted fractional polynomial mixed model for all the compound-dose combinations in the five drug classes.*

6.2 Phase II: Hierarchical Supervised Learning Analysis

The training dataset is used to build a doubly hierarchical supervised learning rule, following the principle laid out in Chapter 5. For each compound-by-dose combination, we will derive five variables, based on the empirical Bayes estimates, obtained from the fractional polynomial mixed model. We will do so for every response variable in the light period, i.e., $\beta_{0j}+b_{0ij}$, $\beta_{1j}+b_{1ij}$, $\beta_{2j}+b_{2ij}$, p_{1jl} and p_{2jl}, the dark period $\gamma_{0j}+c_{0ij}$, $\gamma_{1j}+c_{1ij}$, $\gamma_{2j}+c_{2ij}$, p_{1jd}, p_{2jd}, and also for the first three hours of the light period $\delta_{0j}+d_{0ij}$, $\delta_{1j}+d_{1ij}$, $\delta_{2j}+d_{2ij}$, p_{1jf} and p_{2jf}, where $j=1,\ldots,59$. For each of these new variables, the number of 'observations' equals the number of rats in the compound-dose combinations.

In all three discriminant procedures, the order in which the classes are separated will be the same, but the sleep-wake stages used in each step are allowed to differ. As schematically presented in Figure 6.3, we sequentially discriminate first stimulants, then antipsychotics, antidepressants and finally hypnotics are separated from placebo. This choice was guided by considering the results from the exploratory phase, supplemented with pharmacological information from the experts. Arguably, in general, such a choice will always have a somewhat subjective component to it and ought to be guided by substantive considerations.

The posterior probabilities will be calculated as indicated in Section 5.6 and schematically presented in Figure 6.3. In this way, we adjust for the fact that we have a hierarchical classification procedure.

Originally, the sleep-waking stages used in each step were determined ad hoc, based on the results of the exploratory analysis and prior experts knowledge. For example, it is known that a stimulant induces Active Wake and reduces Light and Deep Sleep. This information could be used in the first step. Later the selection procedures described in Section 5.5 in combination with the lack of classification measure LC' were used to obtain the sleep-wake stages in each step in a more formal way. Parameters for a certain sleep-waking stage in the light period and the first three hours are never included in the same step, given that they are describing essentially the same period. The fact that a certain sleep-wake stage is retained in a particular step in the DHSLA merely means that the behaviour of the rats for this particular compound-dose combination within this particular class is different from that of rats not belonging to this class, relative to the sleep-wake stage being scrutinized.

In what follows, the results obtained with linear discriminant analysis, flexible

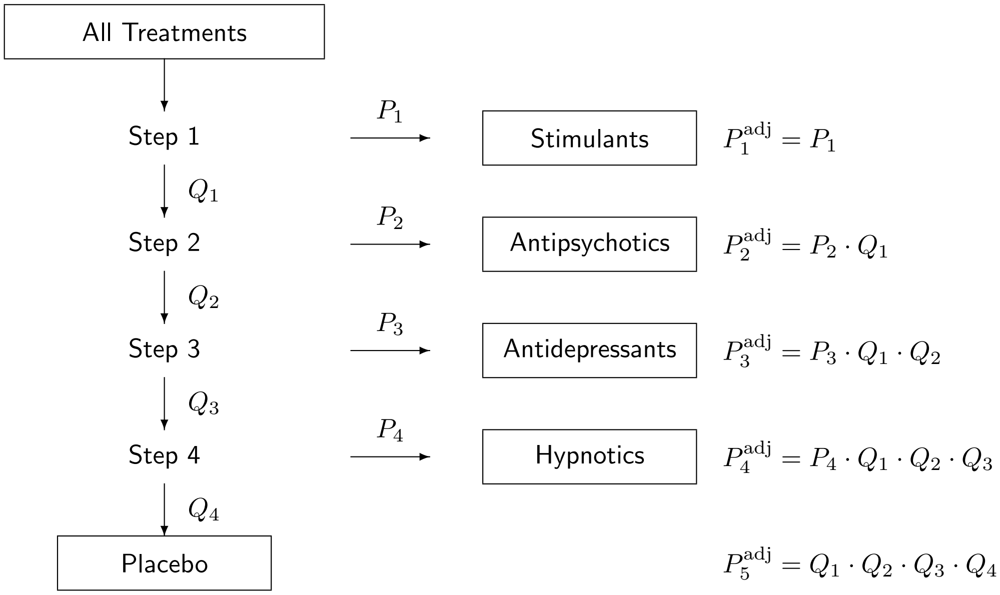

All Treatments
Step 1
P_1
Stimulants
$P_1^{\text{adj}} = P_1$
Q_1
Step 2
P_2
Antipsychotics
$P_2^{\text{adj}} = P_2 \cdot Q_1$
Q_2
Step 3
P_3
Antidepressants
$P_3^{\text{adj}} = P_3 \cdot Q_1 \cdot Q_2$
Q_3
Step 4
P_4
Hypnotics
$P_4^{\text{adj}} = P_4 \cdot Q_1 \cdot Q_2 \cdot Q_3$
Q_4
Placebo
$P_5^{\text{adj}} = Q_1 \cdot Q_2 \cdot Q_3 \cdot Q_4$

Figure 6.3: *Phase II of the Doubly Hierarchical Supervised Learning Analysis. In each step, one class is discriminated from the rest.*

discriminant analysis built on MARS, and mixture discriminant analysis with two subclasses per group are compared with respect to the sleep-wake stages used in each step and the performance with 10-fold cross validation on the rat level (Selection Procedure I) as well as the compound-dose level (Selection Procedure II). Also the classification results in the validation dataset will be compared.

6.2.1 Linear Discriminant Analysis

10 fold cross validation on rat level (Selection Procedure I)

The sleep-wake stages selected in each of the steps of the linear discriminant analysis, by 10-fold cross validation using the rat level are displayed in Table 6.1. As could be expected Active Wake, Light Sleep and Deep Sleep play a role in the discrimination of stimulant compounds. This lines up with expectation because a stimulant generally increases Active Wake and reduces Light and Deep Sleep. When comparing the sleep-wake stages retained for antidepressant and hypnotic with the generally observed

Table 6.1: *Linear Discriminant Analysis. Sleep-waking stages used in each step of the DHSLA with 10-fold cross validation by rat.*

Selection Procedure I

	Stages used in the following periods		
Step	Light Period	Dark Period	First 3 Hours
(1) Stimul	PW SWS2	AW SWS1 RS	AW SWS1
(2) Antipsy	PW SWS2 IS	AW PW SWS2	AW
(3) Antidep	SWS2 IS RS	SWS1 SWS2	AW PW
(4) Hypno	AW SWS2 IS RS		

changes associated to these two classes, see Table 2.3, we see indeed that Passive Wake and REM Sleep for antidepressant, and Deep Sleep for hypnotic were retained. For antipsychotics, we expect to see Light Sleep and Intermediate Stage Sleep, based on Table 2.3, but Light Sleep was not retained in this step.

Table 6.2 shows the number of observations classified in the 5 classes per class, obtained by 10-fold cross validation using the rat level (Selection Procedure I). In the upper panel we see the number of rats classified into each drug class. Except for one antipsychotic rat which is misclassified as antidepressant and two stimulant rats which end up in the placebo and the hypnotic class, all the rats were correctly classified. The error on the rat level, $\text{error}_{\text{rat}}$, is thus 1.6%. In the lower panel of Table 6.2, we classify the compound-dose combination as a whole, by looking at the mean posterior probabilities over the eight rats in that compound-dose. In this case, all the compound-doses were correctly classified, which results in error rate $\text{error}_{\text{c-d}}$ of 0%.

The adjusted posterior probabilities for LDA with cross-validation on the rat level are shown in Table 6.3. The posterior probabilities for correct classification are all above 0.94.

10 fold cross validation on compound-dose level (Selection Procedure II)

The sleep-waking stages selected by 10-fold cross validation on compound-dose level are shown in Table 6.4. Also here we see a lot of similarities with the generally observed changes associated with the concerned classes in Table 2.3. When comparing

Table 6.2: *Linear Discriminant Analysis. Summary of the hierarchical discrimination procedure for the training data set, using 10-fold cross validation on the rat level (number of rats (upper panel) and compound-dose combinations (lower panel) classified in each drug class).*

Selection Procedure I

	Predicted Class					
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul	Total
Placebo	**184**	0	0	0	0	184
Antipsy	0	**55**	1	0	0	56
Antidep	0	0	**112**	0	0	112
Hypnot	0	0	0	**40**	0	40
Stimul	1	0	0	1	**78**	80

	Predicted Class					
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul	Total
Placebo	**23**	0	0	0	0	23
Antipsy	0	**7**	0	0	0	7
Antidep	0	0	**14**	0	0	14
Hypnot	0	0	0	**5**	0	5
Stimul	0	0	0	0	**10**	10

the selected sleep-wake stages with the ones that were retained with 10-fold cross validation on the rat level, we see some resemblances. For example for stimulants, Active Wake during the first three hours, Light Sleep in the dark period and Passive Wake in the light period were retained in both cases. Light Sleep was retained in the first three hours in selection procedure I, while in selection procedure II it was retained in the light period, which is also covering the first three hours. For the other classes we observe some further similarities between the two selection procedures. Passive Wake and Intermediate Stage Sleep in the light period and Active Wake and Deep Sleep in the dark period were selected for the classification of antipsychotics with both procedures. For antidepressants, Intermediate Stage Sleep and REM Sleep in the light period, and Active Wake in either the light or the first three hour period

Table 6.3: *Linear Discriminant Analysis. Adjusted posterior probabilities for the training data set obtained with 10-fold cross-validation on the rat level.*

Selection Procedure I

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.97**	0.02	0.01	0.00	0.00
Antipsy	0.01	**0.94**	0.03	0.00	0.02
Antidep	0.00	0.03	**0.95**	0.01	0.01
Hypnot	0.00	0.00	0.01	**0.99**	0.00
Stimul	0.01	0.01	0.02	0.00	**0.96**

are retained. Finally for hypnotics, Deep Sleep in the light period and Active Wake and Intermediate Stage Sleep in the light period or the first three hours were retained by both selection procedures.

The number of sleep-wake stages retained with both selection procedures is comparable. For step 2 and 3 the same number of sleep-wake stages was selected. For step 1 less sleeping stages were retained with procedure II, while for step 4, much less sleep-wake stages were needed in procedure I.

The results for the 10-fold cross validation using the compound-dose combination level are shown in Tables 6.5 and 6.6.

The number of rats and compound-dose combinations classified in each of the six classes is given for each psychotropic drug class in Table 6.5. The resulting error rate on the rat level is now 9% and the one on the compound-dose level is 6%. Both errors are much higher if we compare with the results obtained with cross validation on the rat level. First, additional randomness is introduced in selection procedure II due to random sampling. Second, the dataset considered contains relatively few compound-by-dose combinations. Third, and very fundamentally, leaving out rats versus leaving out compound-dose combinations does not assess the same aspects. Indeed, by removing rats from a combination, one can still estimate, admittedly somewhat less precise, all parameters associated to such a combination. This is obviously not true when an entire combination is removed. Therefore, both methods focus on different aspects of variability, associated with rat-level and compound-by-dose combination-level replication, respectively. Thus, there is room for both.

Table 6.4: *Linear Discriminant Analysis. Sleep-waking stages used in each step of the DHSLA with 10-fold cross validation by compound-dose combination.*

Selection Procedure II

	Stages used in the following periods		
Step	Light Period	Dark Period	First 3 Hours
(1) Stimul	PW SWS1	SWS1	AW
(2) Antipsy	AW PW SWS1 IS	AW SWS2 IS	
(3) Antidep	AW PW SWS1 IS RS	IS RS	
(4) Hypno	PW SWS2	IS RS	AW SWS1 IS

Table 6.6 presents the adjusted posterior probabilities, showing posterior probabilities for correct classification above 73% for all the classes. For placebo we get even 93%. Although the posterior probabilities and classification percentages are still high, also here we can see that the results obtained by cross validation using the rat-level seem more promising, but might suffer generalizability.

Validition Dataset

Let us now turn to the validation dataset. The posterior probabilities obtained here with the sleep-wake stages selected with 10-fold cross validation on the rat level are presented in the upper panel of Table 6.7, while the lower panel shows the results obtained with the selection procedure on the compound-dose level. Both selection procedures produce poor classification results in the test dataset. For the selection procedure I, placebo and stimulants get still a high posterior probability in the correct class, even 96% for placebo, while the other three classes get a very low posterior probability for the correct class. For selection procedure II, antidepressants and stimulants are doing better than in selection procedure I. But for placebo and hypnotic, the performance is less good.

6.2.2 Flexible Discriminant Analysis

The sleep-wake stages selected per step when flexible discriminant analysis is used in the DHSLA for the two selection procedures described in Section 5.5 are shown in Table 6.8. Similar to the case when the linear discriminant analysis is used in the

Table 6.5: *Linear Discriminant Analysis. Summary of the hierarchical discrimination procedure for the training data set, using cross validation on the compound-dose level (number of rats (upper panel) and compound-dose combinations (lower panel) classified in each drug class).*

Selection Procedure II

	Predicted Class					
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul	Total
Placebo	**176**	0	0	8	0	184
Antipsy	0	**46**	4	6	0	56
Antidep	0	0	**106**	6	0	112
Hypnot	0	0	0	**40**	0	40
Stimul	3	2	9	1	**65**	80

	Predicted Class					
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul	Total
Placebo	**22**	0	0	1	0	23
Antipsy	0	**7**	0	0	0	7
Antidep	0	0	**13**	1	0	14
Hypnot	0	0	0	**5**	0	5
Stimul	0	0	1	0	**9**	10

DHSLA, we see that the number of sleep-wake stages retained does not differ much for both selection procedures. Also, we note that for some classes both selection procedures arrive at selecting the same sleep-wake stages, indicating association between the class and its effect on a particular sleep-wake stage. For example, for stimulants, we retain Light Sleep in the light period or the first three hours, with either selection procedure. For antipsychotics, Passive Wake, Deep Sleep and Intermediate Stage Sleep in the light period and Active Wake, and Intermediate Stage Sleep in the dark period are common to both selection procedures. Intermediate Stage Sleep and REM Sleep in the light period, and Active Wake and Passive wake either in the light period or in the first three hours are retained with both selection procedures for antidepressants. Finally, for hypnotics, Deep Sleep in the light period and Active

Table 6.6: *Linear Discriminant Analysis. Adjusted posterior probabilities for the training data set obtained with 10-fold cross-validation on the compound-dose level.*

Selection Procedure II

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.94**	0.01	0.01	0.04	0.00
Antipsy	0.00	**0.72**	0.10	0.11	0.07
Antidep	0.00	0.18	**0.75**	0.06	0.01
Hypnot	0.00	0.04	0.18	**0.78**	0.00
Stimul	0.04	0.04	0.09	0.04	**0.79**

Wake in the light period or the first three hours is selected with both selection procedures.

Regarding the adjusted posterior probabilities, the upper panel of Table 6.9 displays very high posterior probabilities for the correct classification with flexible discriminant analysis, obtained with Selection Procedure I. For Selection Procedure II, high posterior probabilities, above 70%, for placebo, antipsychotics, antidepressants, and stimulants are obtained, while the adjusted posterior probability obtained for hypnotics is considerably lower.

The adjusted posterior probabilities obtained for the validation dataset using FDA with both selection procedures are shown in Table 6.10. For selection procedure I we get very high posterior probabilities in the correct class for placebo, antipsychotic and stimulants. For antidepressant and hypnotic, the chance of correct classification is rather low. When the classification were selected with cross-validation on the compound-dose level, the posterior probabilities in the correct classes are all above 50% and even around 70% for placebo and hypnotics.

6.2.3 Mixture Discriminant Analysis

The corresponding results for mixture discriminant analysis in the training dataset are presented in Tables 6.11 and 6.12, respectively.

Similar conclusions can be drawn here with respect to the sleep-wake stages retained at each step. The number of sleeping stages retained is similar for both

Table 6.7: *Linear Discriminant Analysis. Adjusted posterior probabilities for the validation data set obtained with 10-fold cross-validation on the rat level (upper panel) and the compound-dose level (lower panel).*

Selection Procedure I

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.96**	0.00	0.04	0.00	0.00
Antipsy	0.38	**0.33**	0.03	0.26	0.00
Antidep	0.11	0.45	**0.16**	0.16	0.12
Hypnot	0.48	0.03	0.31	**0.18**	0.00
Stimul	0.01	0.23	0.07	0.04	**0.65**

Selection Procedure II

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.33**	0.00	0.05	0.62	0.00
Antipsy	0.24	**0.33**	0.00	0.43	0.00
Antidep	0.02	0.38	**0.49**	0.02	0.09
Hypnot	0.93	0.02	0.02	**0.03**	0.00
Stimul	0.00	0.03	0.28	0.00	**0.69**

selection procedures, and some of them are selected by both procedures. For example, for stimulants, Deep Sleep and REM Sleep in the light period, Active and Passive Wake in the dark period and Active Wake and Light Sleep in the first three hours or the light period, are chosen irrespective of the selection procedure used. Also for the other classes in the hierarchical procedure similarities between the two selection procedures are observed.

For the adjusted posterior probabilities, we observe, once more, very promising results for Selection Procedure I. All posterior probabilities for the correct classes are above 98%. For Selection Procedure II, all adjusted posterior probabilities, except for

Table 6.8: *Flexible Discriminant Analysis. Sleep-waking stages used in each step of the DHSLA with 10-fold cross validation by rat (upper panel) and compound-dose combination (lower panel).*

Selection Procedure I

	Stages used in the following periods		
Step	Light Period	Dark Period	First 3 Hours
(1) Stimul	SWS1 SWS2	PW SWS1 SWS2	AW
(2) Antipsy	PW SWS2 IS	AW PW IS	RS
(3) Antidep	SWS2 IS RS	SWS1 SWS2	AW PW
(4) Hypno	AW QW SWS2 RS		

Selection Procedure II

	Stages used in the following periods		
Step	Light Period	Dark Period	First 3 Hours
(1) Stimul		AW RS	AW PW SWS1
(2) Antipsy	PW SWS2 IS RS	AW SWS2 IS	
(3) Antidep	AW PW SWS1 IS RS	IS RS	
(4) Hypno	SWS1 SWS2	SWS2 IS RS	AW IS

antipsychotics, are above 83%. For antipsychotics, we obtain a posterior probability for the correct class of 61%.

The adjusted posterior probabilities for the validation dataset, obtained with mixture discriminant analysis with selection procedure I and II are presented in Table 6.13. For the selection based on 10-fold cross validation on the rat level, a high posterior probability for placebo in the placebo class was achieved, while for antipsychotics, antidepressants and stimulants, the posterior probability for the correct class is around 45% and for hypnotics only 26%. For selection procedure II, a very high posterior probability was obtained in the antipsychotic and stimulant class. Placebo, antidepressants and hypnotics present much smaller posterior probabilities.

6.3 Discussion

In this chapter we applied our doubly hierarchical supervised learning analysis to classify compounds with psychotropic potential into standard classes as they have been defined by Deniker (1982) by using the longitudinal sleep-wake pattern collected on rats. The doubly hierarchical supervised learning analysis was used with a fractional polynomial mixed model in the first stage of the procedure and a stepwise linear, flexible or mixture discriminant analysis in the second stage. Selection of the variables used in each step of the hierarchical discriminant procedure was done using either the individual or the compound-by-dose combination as unit of analysis.

The number of sleep-wake stages used by the procedure at each step is very stable across the three discriminant techniques for both selection procedures. Some sleep-wake stages are retained in all analyses irrespective of the discriminant analysis or the selection procedure. The sleep-wake stages that are generally observed in practice to be influenced by a certain psychotropic drug class, as was seen in Table 2.3, were indeed retained for the corresponding step in most of the procedures.

It appears that the level on which the cross-validation is performed plays an important role in the selection of the sleep-wake stages. For antipsychotics (step two), Passive Wake is selected in the dark period for all three discriminant techniques for Selection Procedure I, but does not show up in any of the analyses when Selection Procedure II is applied. The same is observed for Deep Sleep in the light period and Light and Deep Sleep in the dark period when antidepressants are discriminated, or for REM Sleep in the light period for the classification of hypnotics in the last step. On the other hand, we have some sleep wake stages that are needed in all three analyses when using Selection Procedure II, but not at all when Selection Procedure I is used, such as Intermediate Stage Sleep in the dark period for the classification of antidepressants and REM Sleep in the dark period for the classification of hypnotics.

Variables that appear in a certain step for all three discriminant analyses, regardless of the selection procedure, can be seen as important variables for the discrimination of that particular class from the rest. The first three hours are part of the light period; therefore we will consider a variable as common when it is used either in the light period or in the first three hours. In general, Active Wake in either the light or the first three hour period is showing up in the first step, designed to discriminate stimulants from the rest. This agrees with expectation, because stimulants generally increase the wakefulness. Passive Wake in the light period appears in all six analyses in the second step to classify antipsychotics. Active Wake, Intermediate Stage Sleep and REM Sleep in the light or the first three hour period seem to be crucial in the classification of an antidepressant. Finally for hypnotics, we see that Active Wake, Deep Sleep and REM Sleep in either the light period or the first three hours are retained in all the analyses.

An overview of the error rates in the training dataset on rat and compound-dose level, obtained in the six analyses are summarized in the left panel of Table 6.14. The error rates obtained with cross validation on the rat level are lower than those obtained with cross validation on the compound-dose level. This is not surprising because in the first approach all compound-dose combinations are still represented in the training dataset. On the other hand, when leaving out a whole combination, we get problems for classes where only a few combinations were available (e.g. Hypnotics and Antipsychotic). The two approaches can not be compared, and should be seen next to each other as two different views on the same problem.

In the training dataset, the three discriminant techniques produce comparable results in terms of adjusted posterior probabilities and error rates. Therefore we are inclined to recommend the use of linear discriminant analysis in similar settings also in view of its simplicity. However when looking at the performance in the test dataset, we see that flexible and mixture discriminant analysis are performing better in terms of both posterior probabilities and error rates.

In general almost all methods have some difficulty to discriminate between placebo and antidepressant components and, to a lesser extent, between hypnotic and antidepressant drugs. We can find some evidence for that when comparing the profiles in the test dataset with the profiles of the reference compounds in the training dataset. In Figures A.1 – A.5 in Appendix A, the test compounds in each class are plotted versus the compounds in the five classes in the training dataset. Ideally the test compounds of a certain class should nicely fit within the range and shapes of the respective training compounds. This is certainly the case for the stimulant and placebo test compounds (Figure A.1 and A.5, which is reflected in their good classification results with almost all the methods. When looking at Figure A.2, we would expect difficulties to classify the antipsychotic test compound, since they show a different behaviour for Active Wake and Light Sleep compared to the training compounds in the antipsychotic class. This was indeed the case for the DHSLA with linear discriminant analysis, but in flexible and mixture discriminant analysis, these two sleeping stages were not used in the discrimination of antipsychotic (step 2), and therefore the compounds could be classified well. For the antidepressant and hypnotic test compounds we see in Figure A.3 and A.4 that they fit indeed within the range and shape of the antidepressant compounds in the training dataset, but they also show similarities with the other four classes, which makes it of course difficult to classify them correctly.

One possible drawback of this method is the selection of the level to be used in the cross validation. Another could be the fact that several thousand models can be used to discriminate a particular class from the rest and we are just selecting the model which perform best in terms of misclassification error and posterior probability. Introducing then model selection uncertainty.

As a final remark, the methods developed here are tightly linked to the motivating problem, coming from the wish to classify potentially active psychotropic compounds or, rather, compound-by-dose combinations. It is evident that the methodology can be used in a variety of similar preclinical and clinical settings, across the widest range of therapeutic areas. The method has been tuned for this particular dataset, but the general idea as presented in Figure 5.1 can be applied to any other dataset. The techniques used in both the first and second phase of the procedure can be tuned to the situation at hand.

Table 6.9: *Flexible Discriminant Analysis. Adjusted posterior probabilities for the training data set obtained with 10-fold cross-validation on the rat level (upper panel) and the compound-dose level (lower panel).*

Selection Procedure I

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.99**	0.00	0.01	0.00	0.00
Antipsy	0.01	**0.98**	0.01	0.00	0.00
Antidep	0.02	0.02	**0.95**	0.01	0.00
Hypnot	0.00	0.01	0.00	**0.90**	0.00
Stimul	0.00	0.01	0.00	0.00	**0.99**

Selection Procedure II

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.74**	0.07	0.09	0.10	0.00
Antipsy	0.12	**0.72**	0.13	0.01	0.03
Antidep	0.00	0.10	**0.79**	0.11	0.01
Hypnot	0.13	0.03	0.29	**0.57**	0.00
Stimul	0.04	0.06	0.03	0.08	**0.79**

Table 6.10: *Flexible Discriminant Analysis. Adjusted posterior probabilities for the validation data set obtained with 10-fold cross-validation on the rat level (upper panel) and the compound-dose level (lower panel).*

Selection Procedure I

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.99**	0.01	0.00	0.00	0.00
Antipsy	0.12	**0.82**	0.01	0.05	0.01
Antidep	0.11	0.27	**0.24**	0.22	0.16
Hypnot	0.56	0.17	0.23	**0.04**	0.00
Stimul	0.00	0.14	0.01	0.00	**0.84**

Selection Procedure II

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.67**	0.00	0.01	0.32	0.00
Antipsy	0.49	**0.51**	0.00	0.00	0.00
Antidep	0.02	0.26	**0.61**	0.04	0.07
Hypnot	0.00	0.00	0.27	**0.73**	0.00
Stimul	0.01	0.19	0.17	0.00	**0.63**

Table 6.11: *Mixture Discriminant Analysis. Sleep-waking stages used in each step of the DHSLA with 10-fold cross validation by rat (upper panel) and compound-dose combination (lower panel).*

Selection Procedure I

	Stages used in the following periods		
Step	Light Period	Dark Period	First 3 Hours
(1) Stimul	PW SWS2 RS	AW PW	AW SWS1
(2) Antipsy	PW SWS2	PW SWS1 IS	AW RS
(3) Antidep	AW SWS1 SWS2 IS RS	SWS1 SWS2	
(4) Hypno	AW SWS1 SWS2 IS RS	IS	

Selection Procedure II

	Stages used in the following periods		
Step	Light Period	Dark Period	First 3 Hours
(1) Stimul	AW SWS1 SWS2 RS	AW PW SWS1	
(2) Antipsy	PW SWS2 IS	IS	
(3) Antidep	IS RS	PW IS	AW PW SWS1
(4) Hypno	PW SWS2	IS RS	AW SWS1 IS

Table 6.12: *Mixture Discriminant Analysis. Adjusted posterior probabilities for the training data set obtained with 10-fold cross-validation on the rat level (upper panel) and the compound-dose level (lower panel).*

Selection Procedure I

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.99**	0.01	0.00	0.00	0.00
Antipsy	0.00	**0.98**	0.00	0.02	0.00
Antidep	0.00	0.00	**0.99**	0.01	0.00
Hypnot	0.00	0.00	0.01	**0.99**	0.00
Stimul	0.01	0.00	0.00	0.00	**0.99**

Selection Procedure II

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.85**	0.02	0.09	0.04	0.00
Antipsy	0.12	**0.61**	0.09	0.12	0.06
Antidep	0.03	0.05	**0.85**	0.04	0.03
Hypnot	0.00	0.01	0.08	**0.91**	0.00
Stimul	0.00	0.11	0.06	0.00	**0.83**

Table 6.13: *Mixture Discriminant Analysis. Adjusted posterior probabilities for the validation data set obtained with 10-fold cross-validation on the rat level (upper panel) and the compound-dose level (lower panel).*

Selection Procedure I

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.76**	0.20	0.04	0.00	0.00
Antipsy	0.02	**0.49**	0.04	0.45	0.00
Antidep	0.06	0.20	**0.43**	0.08	0.23
Hypnot	0.35	0.31	0.08	**0.26**	0.00
Stimul	0.00	0.25	0.00	0.00	**0.42**

Selection Procedure II

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.32**	0.04	0.22	0.42	0.00
Antipsy	0.10	**0.83**	0.08	0.00	0.00
Antidep	0.01	0.24	**0.25**	0.28	0.22
Hypnot	0.52	0.00	0.40	**0.08**	0.00
Stimul	0.00	0.00	0.00	0.00	**1.00**

Table 6.14: *Linear Discriminant Analysis. Summary table for the error rates obtained in train and validation dataset, with both selection procedures.*

			Train		Test	
Technique	Selection Procedure		$\mathrm{error}_{\mathrm{rat}}$	$\mathrm{error}_{\mathrm{cv}}$	$\mathrm{error}_{\mathrm{rat}}$	$\mathrm{error}_{\mathrm{cv}}$
LDA	I	Rat level	0.009	0.000	0.583	0.600
	II	Compound-dose level	0.093	0.043	0.650	0.600
FDA	I	Rat level	0.005	0.000	0.385	0.350
	II	Compound-dose level	0.161	0.129	0.404	0.383
MDA	I	Rat level	0.000	0.000	0.483	0.500
	II	Compound-dose level	0.123	0.094	0.475	0.483

7

A Model Averaging Approach for Doubly Hierarchical Supervised Learning Analysis

As we have seen in the previous chapters, the classification problem under consideration poses several challenges. First, one has to address how to use all the features in the data at hand to establish a classification rule. Second, we have to select the sleep-wake stage and the period (light, dark or first 3 hours) to be used to establish such discrimination rule, given the fact that maybe not all are needed. Third, and closely linked to the previous issue, given that an exhaustive search needs to be carried out, a selection bias may also be introduced and could play an important role on the performance of the discriminant procedure used. While the first two challenges are already dealt with in chapter 5 and 6, the third one is still an open problem. This chapter is devoted to study this third issue in more detail and proposes an approach based on the model averaging ideas used on regression models (Burnham and Anderson, 2002). The lack of fit measure defined in chapter 5, is used to calculate the weights in the model average.

In Section 7.1, the methodology is explained, starting from the general form of the doubly hierarchical supervised learning analysis as explained in chapter 5. Thereafter the model-averaging principle is modified to be used with linear discriminant analysis. Finally, the results obtained with model averaging are shown in Section 7.2 and compared to the initial classification results.

7.1 Methodology to Study Model Selection Bias in the Context of Doubly Hierarchical Supervised Learning Analysis

We start again from the doubly hierarchical supervised learning analysis. In the first phase we will use a fractional polynomial mixed model. For the second phase, we saw in Chapter 6 that LDA, FDA and MDA produce comparable results in the training dataset with respect to posterior probabilities and error counts. Therefore, the linear discriminant analysis will be preferred in view of its simplicity.

However, the doubly hierarchical supervised learning analysis might suffer from model selection bias. To avoid this we can base the classification in stage II on more than one model. Barnard (1963) provided the first mention of model combination in the statistical literature in a paper studying airline passenger data. Bates and Granger (1969) stimulated the contribution of articles in the economics literature about the combination of predictions from different forecasting models. Later several articles appear and in the late 90s, George (1998) reviews bayesian model selection and discusses bayesian model averaging (BMA) in the context of decision theory. Draper (1995), Chatfield (1995), and Kass and Raftery (1995) all review BMA and the costs of ignoring model uncertainty. Many model averaging approaches have been proposed in the literature, Hoeting *et al.* (1999) wrote a tutorial pointing out the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are, proposing a bayesian model averaging which provides a coherent mechanism for accounting for this model uncertainty. Also several frequentist approaches for model averaging have been presented in the literature, Hjort and Claeskens (2003) build a general large-sample likelihood apparatus in which limiting distributions and risk properties of estimators-post-selection as well as of model average estimators are precisely described, also explicitly taking modelling bias into account. Williams and Christian (2006) introduce frequentist model-averaged estimators for univariate twin data analysis that use information-theoretic criteria to assign model weights. Burnham and Anderson (2002) also proposed model averaging to deal with model selection bias in the case of regression models. We will use this

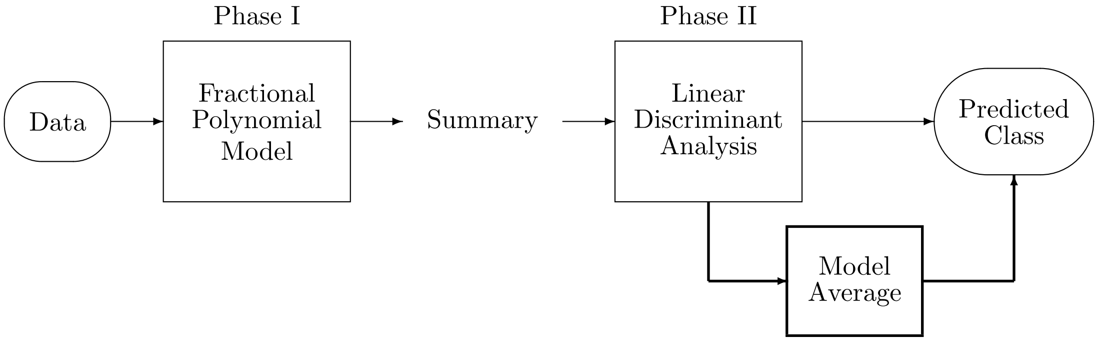


Figure 7.1: *Diagram representing doubly hierarchical supervised learning analysis with model averaging, when a fractional polynomial mixed model (FPMM) is used in Stage I and linear (LDA) discriminant analysis is used in Stage II.*

last approach and adapt it to fit in the discriminant analysis framework.

The current procedure, consisting of the doubly hierarchical supervised learning analysis, extended with model averaging, is schematically presented in Figure 7.1.

Before we turn to the model averaging in discriminant analysis in Section 7.1.2, we briefly review the model averaging approach as proposed by Burnham and Anderson (2002) in Section 7.1.1.

7.1.1 Model Averaging in the Context of Regression Models

Let us first have a look at the model averaging approach proposed by Burnham and Anderson (2002). We illustrate this approach in the case of a linear regression problem. In many cases, one has a large number of closely related models. Defining a best model is often not satisfactory since this choice can vary from dataset to dataset, collected under the same underlying process. In order to get a more stabilized inference, Burnham and Anderson (2002) suggest to use model averaging.

Assume we have a linear regression model m given by

$$Y_i = \beta_0^{(m)} + \sum_{j=1}^{n^{(m)}} \beta_j^{(m)} x_{ij} + \epsilon_i^{(m)} \tag{7.1}$$

For each model m, the AIC (Akaike, 1973) is calculated, and the difference with the minimum AIC over all possible models is computed

$$\Delta_m = \mathrm{AIC}_m - \mathrm{AIC}_{min}. \tag{7.2}$$

To calculate the new coefficients $\hat{\hat{\beta}}_j$ for the model average over R models, β_j is averaged over all the models in which x_j appears.

$$\hat{\hat{\beta}}_j = \frac{\sum_{m=1}^{R} w_m I_j(m) \hat{\beta}_j^{(m)}}{w_+(j)} \tag{7.3}$$

where

$$\begin{aligned} w_m &= \frac{\exp(-\Delta_m/2)}{\sum_{r=1}^{R} \exp(-\Delta_r/2)} & (7.4) \\ w_+ &= \sum_{m=1}^{R} w_m I_j(m) & (7.5) \end{aligned}$$

and

$$I_j(m) = \begin{cases} 1 & \text{if predictor } x_j \text{ is in model } m, \\ 0 & \text{otherwise.} \end{cases}$$

Inferences will now be made based on model

$$Y_i = \hat{\hat{\beta}}_0 + \sum_{j=1}^{n} \hat{\hat{\beta}}_j x_{ij} + \epsilon_{ij} \tag{7.6}$$

This approach has both practical and philosophical advantages. Burnham and Anderson (2002) argue that where a model averaged estimator can be used it often has reduced bias and better precision compared to $\hat{\beta}$ from the selected best model. Model averaging has been used in the context of regression in several applications (e.g. Faes *et al.* (2007), Hansen (2007)).

7.1.2 A Novel Proposal of Model Averaging for Linear Discriminant Analysis

We can now extend the model averaging of Burnham and Anderson to the case of linear discriminant analysis (Wouters *et al*, 2008a). Let us focus on step s, for each subject i we use a set of p measures $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$. We assume now that each class c has an underlying multivariate normal distribution with mean μ_c and common variance-covariance matrix Σ.

$$\text{Class } c \sim f_c(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c)] \tag{7.7}$$

Since the first part of equation 7.7 is independent of the class and since we assume equal variance covariance matrix, this density can be seen as a linear function of $\boldsymbol{x}$ with coefficients $\alpha_j, j = 1, \ldots, p$.

$$f_c(\boldsymbol{x}) \sim \exp[\sum_{j=1}^{p} \alpha_{jc} x_j] \tag{7.8}$$

The posterior probability of belonging to class c when $\boldsymbol{x}$ was observed is given by:

$$P(c|\boldsymbol{x}) = \frac{p_c f_c(\boldsymbol{x})}{\sum_{l=1}^{g_s} p_l f_l(\boldsymbol{x})} \tag{7.9}$$

where p_c is the prior probability for class c and g_s is the total number of classes in step s. In our situation, we can assume that all classes are equally likely to occur, which is translated in equal prior probabilities $p_c = 1/g_s$. Together with equation 7.8, this reduces the posterior probabilities to

$$P(\pi_c|\boldsymbol{x}) = \frac{1}{g_s} \frac{\exp[\sum_{j=1}^{p} \alpha_{jc} x_j]}{\sum_{l=1}^{g_s} \exp[\sum_{j=1}^{p} \alpha_{jl} x_j]} \tag{7.10}$$

We use the lack of classification measure LC' defined in chapter 5, equations (5.6) – (5.9), to determine the goodness of a model.

$$\text{LC'}_s = 1 - \left(1 - \frac{\text{LC}_s}{2 \cdot w_{s1} + (g - s + 1) \cdot w_{s2}}\right) \cdot F(\text{ss}). \tag{7.11}$$

While before, the model with the lowest lack of classification was retained, we focus now on the R models with the lowest LC'. For these R models we calculate weights $w^{(m)}$ in analogy to the weights defined in equation (7.4), where the bracketed upper index is referring to the model under consideration. The AIC is replaced by the lack of classification measure LC' which gives us

$$w_s^{(m)} = \frac{\exp(-\Delta_s^{(m)}/2))}{\sum_{r=1}^{R} \exp(-\Delta_s^{(r)}/2)}, \tag{7.12}$$

where

$$\Delta_s^{(m)} = \text{LC'}_s^{(m)} - \min_r(\text{LC'}_s^{(r)}).$$

The coefficients α_{jc} in the discriminant analysis equation (7.10) are now averaged over the R best models as follows

$$\hat{\bar{\alpha}}_{jc} = \frac{\sum_{m=1}^{R} w^{(m)} I_j(m) \hat{\alpha}_{jc}^{(m)}}{w_+(j)}, \tag{7.13}$$

where

$$w_+(j) = \sum_{m=1}^{R} w^{(m)} I_j(m),$$

and

$$I_j(m) = \begin{cases} 1 & \text{if predictor } x_j \text{ is in model } m, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\hat{\alpha}_{jc}^{(m)}$ denotes the estimator of α_{jc} based on model m. The notation $w_+(j)$ is the sum of the weights over all models in the set where predictor variable j is explicitly in the model.

7.2 Results - EEG Data

In chapter 6, a fractional polynomial model was built for each compound-dose combination and each sleep-wake stage for the light and the dark period separately as well as for the first three hour period. The parameters of these 18 models were used in the second step in a stepwise discriminant analysis. For all possible combinations of these 18 groups of parameters, a discriminant analysis with 10-fold cross-validation on rat level and compound-dose level (Selection Procedures I and II) was performed in each step. The model with the lowest lack of classification measure LC' was retained. When parameters for a certain sleep-wake stage in the light period are used in a model, then the model does not contain the parameters for that sleep-wake stage during the first three hours and vice versa, because they are both partly describing the same time period.

The results obtained with linear discriminant analysis in Stage II of the DHSLA were summarized in Tables 6.1– 6.7. For both selection procedures I and II, we found that the adjusted posterior probabilities for the correct classes are very high in the training dataset. In the test dataset we obtained a high posterior probability for placebos and for stimulants with selection procedure I, while the probabilities for the other three classes are much lower. For selection procedure II, only stimulants can be classified well. The error rate for both selection procedures was about 60 percent. Although the procedure is doing very well in the training dataset, with 10 fold cross validation, we get surprisingly bad results in the test dataset. One possible reason for this is the model selection bias. To solve this we use the modified model averaging approach as described in section 7.1.

The model averaging approach will be evaluated using the training and the test dataset. Cross-validation, at rat- and compound-dose level, will only be used to

Table 7.1: *Linear Discriminant Analysis. Summary table for the error rates obtained in train and validation dataset, with both selection procedures.*

		Train		Test	
Selection Procedure		$\text{error}_{\text{rat}}$	error_{cv}	$\text{error}_{\text{rat}}$	error_{cv}
I	Rat level	0.001	0.000	0.631	0.600
II	Compound-dose level	0.015	0.000	0.650	0.600

calculate the lack-of-classification measures and to select the best model. Once the order of the models is obtained, the train-test setting will be used to evaluate the performance of the model averaging approach. Using cross-validation at this stage will sharply increase the computation time and is therefore not desirable. In what follows, we will still distinguish between selection procedure I and II, but this is only referring to the selection, and the order, of the models and not to the classification results themselves.

7.2.1 Selection Procedure I: Based on Rat-Level

To be able to compare the classification results, the error rates for the training and test data set and the posterior probabilities obtained in the test dataset, are recapitulated in Table 7.1 and 7.2 respectively. Since the error rate in the training dataset is already very low, there is not much room for improvement here. Therefore, we will focus in this chapter on the classification of the test dataset. The results obtained in the training dataset can be found in Appendix B.

In Table 7.3, the adjusted posterior probabilities obtained by averaging over the 25 best models are presented. We see here that for antidepressant and hypnotics, the classification has not improved, but for antipsychotics and stimulants we get a much better performance. Placebo was already well classified with only one model, and is still getting a high posterior probability for the correct class when 25 models are combined. The error count on the compound-dose level is here 0.400.

In the upper left panel of Figure 7.2 the error rates, on the compound-dose level, for the test dataset, obtained with model averaging over the 1, 10, 25, 50, 100, and 200 best models for selection procedure I are graphically displayed. The error rate can be reduced to 40 percent when 10 or more models are combined. Using 200 models seems to introduce too much noise, leading to a slightly higher error rate. The error

Table 7.2: *Linear Discriminant Analysis. Adjusted posterior probabilities for the test data set obtained without model averaging when selection procedure I is used.*

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.96**	0.00	0.04	0.00	0.00
Antipsy	0.38	**0.33**	0.03	0.26	0.00
Antidep	0.11	0.45	**0.16**	0.16	0.12
Hypnot	0.33	0.02	0.42	**0.23**	0.00
Stimul	0.01	0.23	0.07	0.04	**0.65**

Table 7.3: *Model Average. Adjusted posterior probabilities for the validation data set obtained with model averaging with the 25 best models when selection procedure I is used.*

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.96**	0.00	0.04	0.00	0.00
Antipsy	0.16	**0.80**	0.00	0.04	0.00
Antidep	0.04	0.46	**0.07**	0.19	0.24
Hypnot	0.33	0.01	0.37	**0.29**	0.00
Stimul	0.00	0.19	0.02	0.00	**0.79**

rate could be reduced even further by restricting to the models with only 4 or only 5 sleep-wake stages as can be seen in the other panels of Figure 7.2.

When restricting to the models with 7 sleep-wake stages, we see that the error rate stabilizes after 100 models. For the model averaging restricted to 4, 5, or 6 sleep-wake stages we still have a decreasing trend when going from 100 to 200 models, but adding more models did not lead to a further decrease (results not shown).

In Figure 7.3, the adjusted posterior probabilities for the correct classification in the five classes are plotted for model averaging on 1, 10, 25, 50, 100, and 200 models. As reference, a horizontal line is drawn at the initial value, obtained with only the best

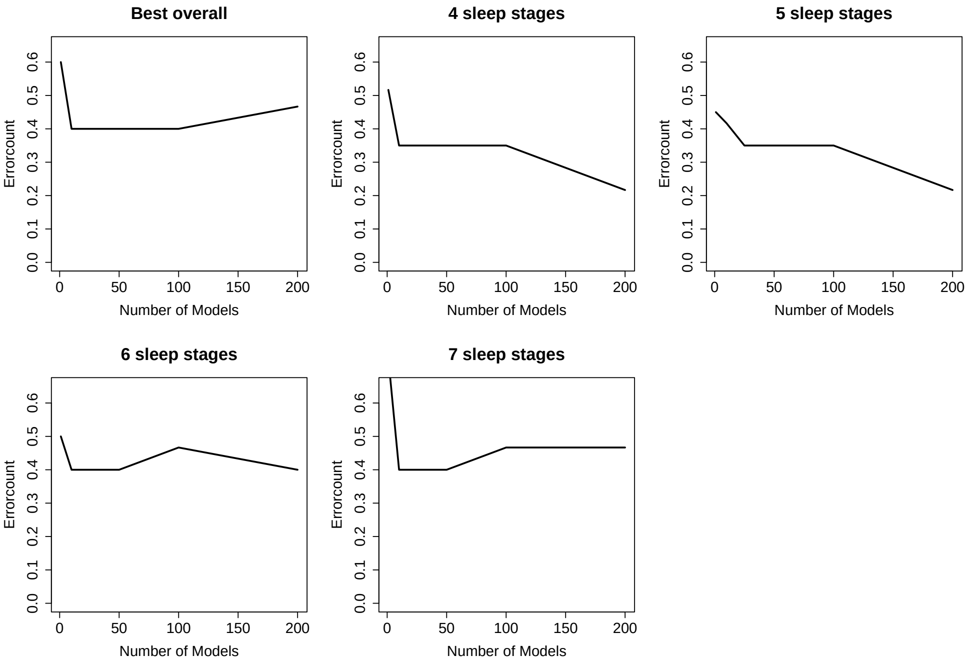


Figure 7.2: *Model Averaging. Error rates in the test dataset obtained with modelaveraging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure I.*

model. For placebos, antipsychotics, hypnotics and stimulants, an improvement is obtained by combining 10 models or more. Including more than 100 models does not improve the posterior probabilities anymore. For antidepressants, model averaging does not lead to higher posterior probabilities for correct classification.

7.2.2 Selection Procedure II: Based on Compound-Dose Level

Also for selection procedure II, i.e. cross-validation on the level of the compound-dose combination, we recapitulate the results obtained with the best combination of sleep-wake stages in the light, dark and first period in Tables 7.1 and 7.4.

In Table 7.5 we see the adjusted posterior probabilities obtained with model averaging over the 25 best models when using selection procedure II. Also here we see a major increase in posterior probability for the correct class in placebo, antipsychotic and stimulants. For antidepressant and hypnotic the posterior probabilities could not be improved with 25 models. The error rate on the compound-dose level is 0.417. The

Table 7.4: *Linear Discriminant Analysis. Adjusted posterior probabilities for the test data set obtained without model averaging when selection procedure II is used.*

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.33**	0.01	0.05	0.61	0.00
Antipsy	0.24	**0.34**	0.00	0.42	0.00
Antidep	0.02	0.38	**0.49**	0.02	0.09
Hypnot	0.93	0.02	0.02	**0.03**	0.00
Stimul	0.01	0.02	0.28	0.00	**0.69**

Table 7.5: *Model Average. Adjusted posterior probabilities for the validation data set obtained with model averaging with the 25 best models when selection procedure II is used.*

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.89**	0.00	0.11	0.00	0.00
Antipsy	0.32	**0.62**	0.02	0.04	0.00
Antidep	0.05	0.51	**0.21**	0.12	0.11
Hypnot	0.70	0.03	0.20	**0.07**	0.00
Stimul	0.01	0.08	0.20	0.00	**0.71**

error rates on the compound dose level for the test dataset are shown in Figure 7.4. With model averaging over all the models, irrespectively of the number of sleep-wake stages used, the error rate is even reduced to 26% (upper left panel). When restricting to the models with only 4, 5, 6, or 7 sleep-wake stages, the error rate converges to the same value of 0.26. In all of these cases, more than 100 models was not needed.

In Figure 7.5, the adjusted posterior probabilities in the five drug classes obtained with model averaging are presented. Again a reference line is drawn at the value of the error rate obtained with only one model. For the adjusted posterior probabilities for placebos, antipsychotics, hypnotics and stimulants combining 10 models or more results in a large improvement in posterior probabilities for correct classification. Including more than 100 models does not improve the posterior probabilities anymore. For antidepressants the posterior probabilities obtained with model averaging are even lower than the initial ones.

7.3 Concluding Remarks

When applying the doubly hierarchical supervised learning analysis to the EEG data, we could see that the misclassification error was very low in the training dataset, for both selection procedures I and II. However, for the test dataset, the error rate turned out to be around 0.60 in both cases. One of the reasons for this can be the model selection bias. Burnham and Anderson (2002) proposed a solution by model averaging for this in the case of regression problems. In this chapter, we have modified this model averaging approach to fit in our DHSLA procedure.

In general, model averaging improved the classification results. This can be seen in the decreased error rates and in the posterior probabilities for correct classification for almost all classes. If there is room for improvement, model averaging will enhance the classification. But we can not expect miracles: if classes are poorly separated, the model averaging will hardly improve the results. This has been observed for antidepressants and to a lesser extent for hypnotics, which has already been pointed out in the previous chapters as well.

For Selection Procedure I, restricting to the models with only four or only five sleep-wake stages leads to the best classification results. More than 200 models were not needed. For Selection Procedure II, it does not matter whether or not one restricts to the models with only a fixed number of sleep-wake stages. In all situations, the error rate converges to a value around 0.26 for 100 models or more.

When comparing the best results obtained for Selection Procedures I and II, we can see that they perform similarly in terms of adjusted posterior probabilities and error rates.

Adding more sleep-wake stages does not necessarily lead to better classification results, this has also been discussed already in previous chapters.

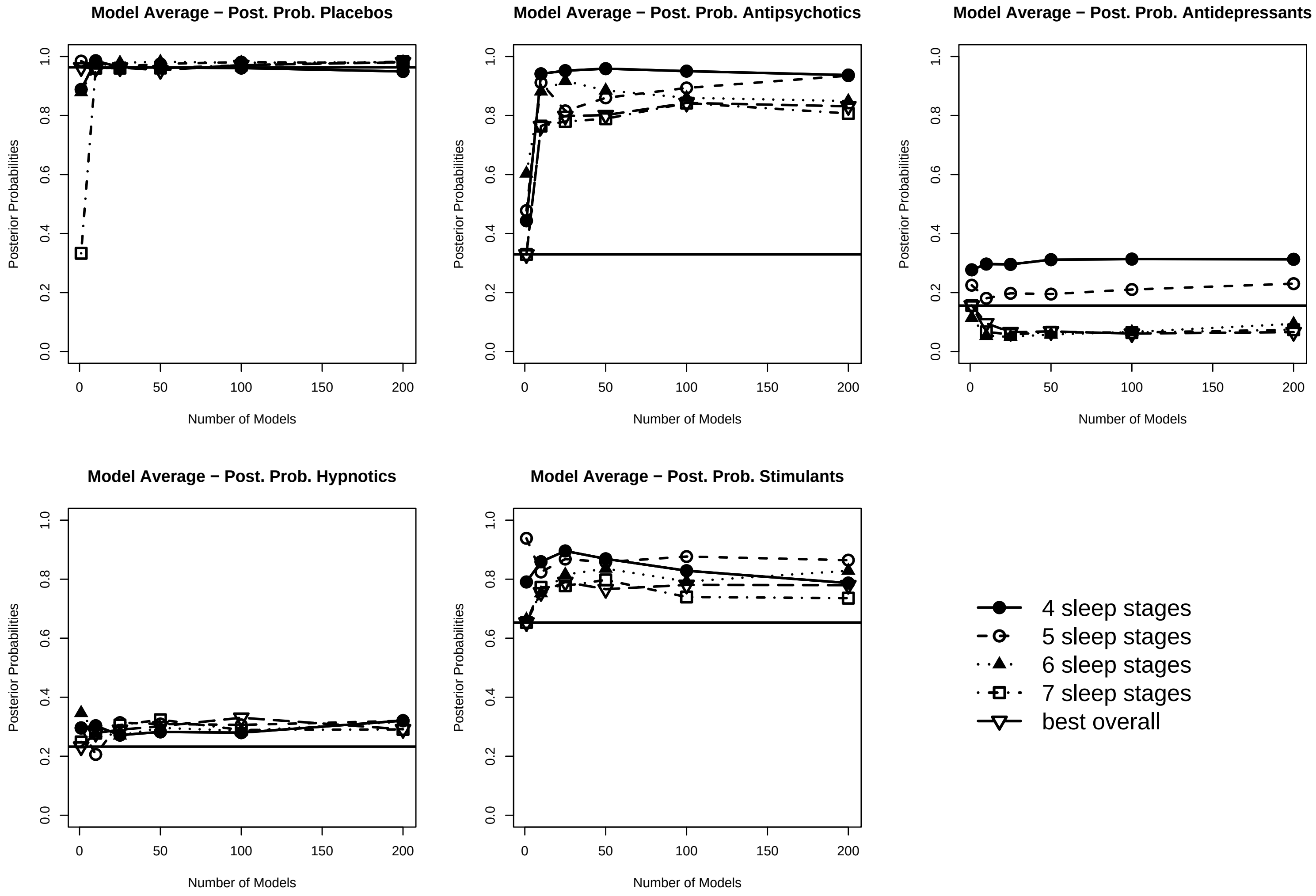


Figure 7.3: *Model Averaging. Adjusted posterior probabilities in the test dataset obtained with modelaveraging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure I.*

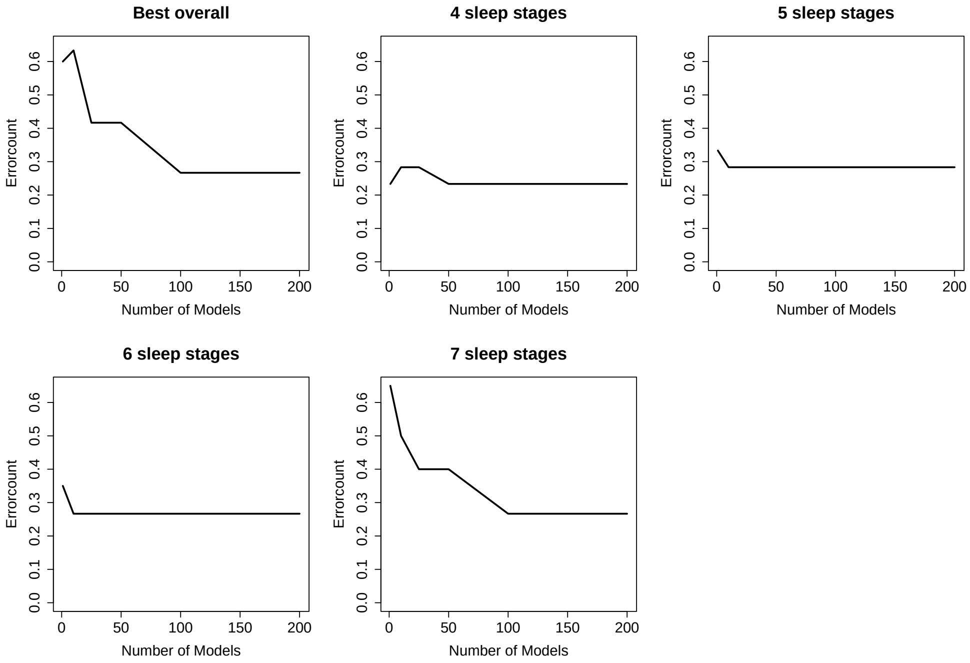


Figure 7.4: *Model Averaging. Error rates in the test dataset obtained with modelaveraging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure II.*

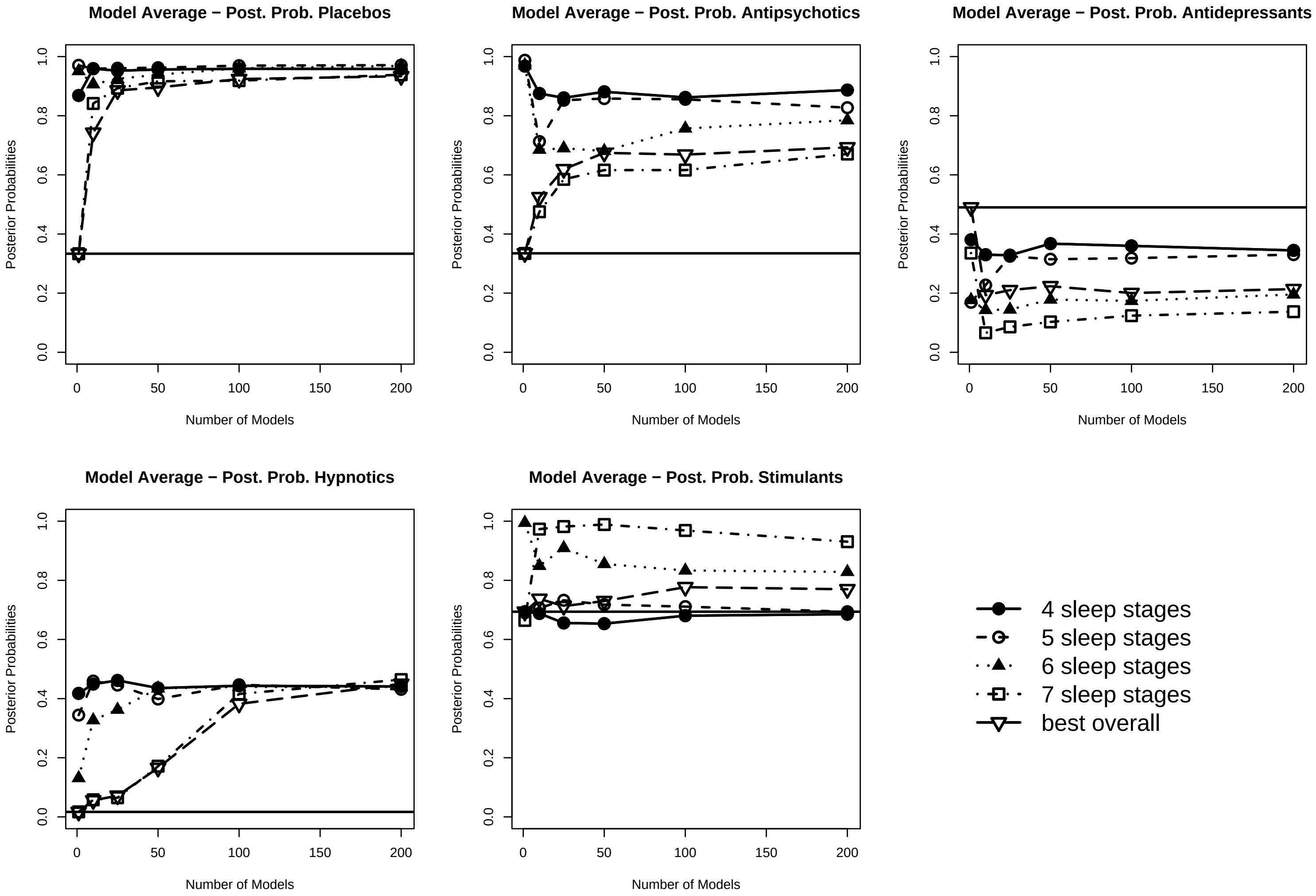


Figure 7.5: *Model Averaging. Adjusted posterior probabilities in the test dataset obtained with modelaveraging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure II.*

8

Multivariate Functional Linear Discriminant Analysis in Combination with Pseudo-likelihood Techniques

In the previous chapters we have presented a two-step strategy to deal with classification problems where the predictor variables are longitudinal. This procedure first models the data and uses the model parameters in a discriminant analysis, but the correlation between the predictor variables is ignored. In this chapter, we propose a discriminant procedure that is taking into account both the longitudinal and the multivariate aspect of the data in a single stage.

James and Hastie (2001) propose a functional linear discriminant analysis (FLDA) to deal with classification of univariate longitudinal predictor variables. This method generalizes linear discriminant analysis to functional data and possesses all the usual LDA tools, including a low-dimensional graphical summary of the data, and classification of new curves. This approach will be described in Section 8.1.

In Section 8.2, we will propose an extension of this method for the case where several longitudinal profiles are recorded for the same individual by combining the functional linear discriminant analysis with a pseudo-likelihood modeling approach (Fieuws and Verbeke, 2006).

The performance of this procedure is established through simulation studies, using different number of classes and observations on the one hand, and through application to the EEG dataset on the other hand. The results obtained here are presented in Section 8.3 and 8.4.

8.1 Univariate Functional Linear Discriminant Analysis

James and Hastie (2001) introduce a generalization of linear discriminant analysis to the case of longitudinal data by taking into account the covariance structure in the data when calculating the distance between observations. In order to be able to handle irregularly sampled curves, they propose to use a linear mixed model with splines as fixed and random effects (see Section 4.1.3) to estimate the covariance structure. Later, this covariance matrix is used to classify the observations based on the mahalanobis distance. Assume we have a n_{ic}-dimensional vector of observations $\boldsymbol{Y}_{ic}$ for subject i in class c and let $\boldsymbol{s}(t)$ denote a natural cubic spline basis with dimension q (see Section 4.1.3). A linear mixed model with the basis functions $\boldsymbol{s}(t)$ as fixed and random effects is now fitted to the data

$$\boldsymbol{Y}_{ic} = S_{ic}\boldsymbol{\beta}_c + S_{ic}\boldsymbol{b}_{ic} + \boldsymbol{\varepsilon}_{ic}$$

where S_{ic} is a $q \times n_{ic}$-dimensional matrix of basis functions

$$S_{ic} = (\boldsymbol{s}(t_{ic1}), \cdots, \boldsymbol{s}(t_{icn_{ic}}))^T,$$

$\boldsymbol{\beta}_c$ and $\boldsymbol{b}_{ic}$ are q-dimensional vectors of fixed and random effects respectively. The random effects $\boldsymbol{b}_{ic}$ and the residual components $\boldsymbol{\varepsilon}_{ic}$ are assumed to be independent with distributions $N(\boldsymbol{0}, D_c)$ and $N(\boldsymbol{0}, \Sigma_{ic})$ respectively, which leads us to the following model

$$\boldsymbol{Y}_{ic} \sim N(S_{ic}\boldsymbol{\beta}_c, \Gamma_{ic})$$

where

$$\Gamma_{ic} = \Sigma_{ic} + S_{\text{grid}} D_c S_{\text{grid}}^T$$

where S_{grid} is the natural cubic spline basis matrix evaluated in a fine lattice of points, which is in practice the set of all time points present in the data. Once the parameters in the model have been estimated, we can use the estimated covariance matrix to classify a new curve.

For a new curve $\boldsymbol{y}$ we can now calculate the distance to class c as follows

$$d(\boldsymbol{y}, c) = (\boldsymbol{y} - \bar{\boldsymbol{Y}}_c)\Gamma_c(\boldsymbol{y} - \bar{\boldsymbol{Y}}_c)^T \tag{8.1}$$

where $\bar{\boldsymbol{Y}}_c$ is the mean in class c. The curve $\boldsymbol{y}$ is now classified to the class c for which the distance $d(\boldsymbol{y}, c)$ is minimal.

In the particular case where all the subjects are measured at the same timepoints, the basis matrix S_{ic} is the same for all the subjects and can be denoted by S. The distance between $\boldsymbol{y}$ and class c now simplifies to

$$d(\boldsymbol{y}, c) = (\boldsymbol{y} - \bar{\boldsymbol{Y}}_c)(\Sigma_c + SD_cS^T)(\boldsymbol{y} - \bar{\boldsymbol{Y}}_c)^T \tag{8.2}$$

8.2 Multivariate Extension of Functional Linear Discriminant Analysis

Now that we have briefly described the FLDA methodology, we will present in this section an extension of the functional linear discriminant analysis of James and Hastie (2001) to the case of multivariate longitudinal data. When dealing with more than one longitudinal variable, there are two sources of associations we need to account for. The first one is the correlation between the timepoints, which FLDA nicely deals with. The second one is the correlation between the longitudinal variables. A fully multivariate model would be the natural choice, but given the complexity of the data, computational issues are commonly present during such modeling exercise. Therefore, we are proposing to use a pseudo-likelihood modeling approach (Fieuws and Verbeke, 2006) combined with smoothing techniques such as splines, in order to allow for mean shape flexibility to model the multivariate longitudinal characteristics.

8.2.1 Pseudo-Likelihood Approach to Model Multivariate Longitudinal Profiles

Let us first introduce the pseudo-likelihood modeling approach proposed by Fieuws and Verbeke (2006). Suppose we want to model m different outcomes jointly. The linear mixed model for one single n_i-dimensional outcome $\boldsymbol{Y}_i$ for subject i is given by

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{8.3}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects and $\boldsymbol{b}_i$ the vector of random effects. The matrices X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates. The random effects $\boldsymbol{b}_i$ and the residual components $\boldsymbol{\varepsilon}_i$ are assumed to be independent with distributions $N(\mathbf{0}, D)$ and $N(\mathbf{0}, \Sigma_i)$ respectively. Σ_i only depends on i through its dimension n_i, meaning that the parameters in Σ_i are common to all subjects. Thus, in summary,

$$\boldsymbol{Y}_i|\boldsymbol{b}_i \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \Sigma_i) \tag{8.4}$$

or marginally

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i) \tag{8.5}$$

Inference is based on maximizing the marginal log-likelihood function $l(\boldsymbol{Y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector containing all parameters (fixed effects and covariance parameters).

The random effects model (8.3) can be easily extended to jointly model m outcomes $\boldsymbol{Y}_{1i}, \cdots, \boldsymbol{Y}_{mi}$ assuming a mixed model for each outcome, and combining these univariate models through the specification of a joint multivariate distribution for all random effects. However, as the number of outcomes and/or the number of random effects per outcome increases, the dimension of the joint covariance matrix of the random effects $\boldsymbol{b}_i$ grows, leading to computational problems.

To solve the computational issues, Fieuws and Verbeke (2006) propose to reduce the dimensionality by fitting all pairwise models separately. In this way, the number of parameters to be estimated is decreased and thus computational burden is avoided.

For a pair of outcomes r, s we fit the following model

$$\begin{pmatrix} Y_{ri} \\ Y_{si} \end{pmatrix} = \begin{pmatrix} X_{ri} & 0 \\ 0 & X_{si} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_r \\ \boldsymbol{\beta}_s \end{pmatrix} + \begin{pmatrix} Z_{ri} & 0 \\ 0 & Z_{si} \end{pmatrix} \begin{pmatrix} \boldsymbol{b}_{ri} \\ \boldsymbol{b}_{si} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{ri} \\ \boldsymbol{\varepsilon}_{si} \end{pmatrix}$$

$$\text{with } \begin{pmatrix} \boldsymbol{b}_{ri} \\ \boldsymbol{b}_{si} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} D_r & D_{rs} \\ D_{rs} & D_s \end{pmatrix} \right)$$

$$\text{and } \begin{pmatrix} \boldsymbol{\varepsilon}_{ri} \\ \boldsymbol{\varepsilon}_{si} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_r & \Sigma_{rs} \\ \Sigma_{rs} & \Sigma_s \end{pmatrix} \right) \tag{8.6}$$

for $r = 1, \cdots, m-1,\ s = r+1, \cdots, m$. The matrices X_{ri}, X_{si}, Z_{ri} and Z_{si} are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates. The variance-covariance matrices of the random effects D_r, D_{rs} and D_s are unstructured $(q \times q)$ matrices and the variance matrices of the residuals Σ_r, Σ_{rs} and Σ_s are usually assumed to be diagonal $(n_i \times n_i)$ matrices with equal variances on the diagonal, i.e. $\sigma_r I_{n_i}$, $\sigma_{rs} I_{n_i}$ and $\sigma_s I_{n_i}$. For each pair of outcomes, the log likelihood will be maximized.

$$\sum_{i=1}^{N} l_{rsi}(\boldsymbol{Y}_{ri}, \boldsymbol{Y}_{si} | \boldsymbol{\Theta}_{r,s}) \tag{8.7}$$

where N denotes the total number of subjects and $\boldsymbol{\Theta}_{r,s}$ is the vector of all parameters in the bivariate model for pair (r,s).

Once all pairwise models are fitted, for some parameters a unique estimate is obtained, e.g. for the pairwise covariance matrices of the random effects D_{rs}, while other parameters are estimated multiple times, for example the covariance matrix of the random effects from the same outcome D_r. Fieuws and Verbeke propose to estimate the overall covariance matrix D^* of the random effects as a block matrix, with $D_{r,s}$ in the off-diagonal block (r,s), and the mean of the matrices D_r, coming from all the pairs consisting of Y_r in the r-th diagonal block, for $r = 1, \cdots, m-1$ and $s = r+1, \cdots, m$. In a similar way an estimate for the overall residual covariance matrix Σ^* is obtained by averaging Σ_r over all pairs containing Y_r and putting Σ_{rs} on the off-diagonal positions.

Standard errors can also be obtained using results from pseudo-likelihood theory, ample details can be found in Fieuws and Verbeke (2006).

8.2.2 Multivariate Functional Linear Discriminant Analysis

Using the pseudo-likelihood methodology described in the previous section, we can now extend the functional linear discriminant analysis to the case of multivariate longitudinal data (Wouters *et al*, 2008c). For notational simplicity we will assume that all subjects are measured at the same timepoints, but the classification rule can easily be extended to the case of different timepoints per subject.

Fitting the model

First a joint pseudo-likelihood model, as in equations (8.6) and (8.7), will be fitted for each class. In analogy to James and Hastie (2001) we choose to use a natural q-dimensional spline basis for the fixed as well as the random effects. Let $\boldsymbol{Y}_{ric}$ and $\boldsymbol{Y}_{sic}$ denote the n-dimensional vector of measurements for response variable r and s

respectively for subject i in class c.

$$\begin{pmatrix} \boldsymbol{Y}_{ric} \\ \boldsymbol{Y}_{sic} \end{pmatrix} = \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{rc} \\ \boldsymbol{\beta}_{sc} \end{pmatrix} + \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} \boldsymbol{b}_{ric} \\ \boldsymbol{b}_{sic} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{ric} \\ \boldsymbol{\varepsilon}_{sic} \end{pmatrix}$$

$$\text{with } \begin{pmatrix} \boldsymbol{b}_{ric} \\ \boldsymbol{b}_{sic} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} D_{rc} & D_{rsc} \\ D_{rsc} & D_{sc} \end{pmatrix} \right) \tag{8.8}$$

$$\text{and } \begin{pmatrix} \boldsymbol{\varepsilon}_{ric} \\ \boldsymbol{\varepsilon}_{sic} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{rc} & \Sigma_{rsc} \\ \Sigma_{rsc} & \Sigma_{sc} \end{pmatrix} \right)$$

The fixed effects β_c in class c will be estimated by β_c^* which is an $(n \cdot q)$ vector consisting of the averages over all the pairs. The covariance matrices D_c^* and Σ_c^* for the random effects and the residual components in class c, can be estimated as described in section 8.2.1, leading us to the overall covariance matrix Γ_c^* in class c

$$\Gamma_c^* = \Sigma_c^* + S_{\text{full}} D_c^* S_{\text{full}}^T$$

where

$$S_{\text{full}} = \begin{pmatrix} S & 0 & \cdots & 0 \\ 0 & S & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & S \end{pmatrix} \tag{8.9}$$

where S is a q-dimensional natural spline basis matrix.

Model Illustration

As an illustration we write down the model for the particular case of three n-dimensional responses V_1, V_2 and V_3. Fitting a pseudo-likelihood model with q knots for class c comes down to fitting the following three pairwise models

$$\begin{cases} \begin{pmatrix} \boldsymbol{V}_{1ic} \\ \boldsymbol{V}_{2ic} \end{pmatrix} = \begin{pmatrix} S_q & 0 \\ 0 & S_q \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(1)}_{1c} \\ \boldsymbol{\beta}^{(1)}_{2c} \end{pmatrix} + \begin{pmatrix} S_q & 0 \\ 0 & S_q \end{pmatrix} \begin{pmatrix} \boldsymbol{b}^{(1)}_{1ic} \\ \boldsymbol{b}^{(1)}_{2ic} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}^{(1)}_{1ic} \\ \boldsymbol{\varepsilon}^{(1)}_{2ic} \end{pmatrix} \\ \\ \begin{pmatrix} \boldsymbol{V}_{1ic} \\ \boldsymbol{V}_{3ic} \end{pmatrix} = \begin{pmatrix} S_q & 0 \\ 0 & S_q \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(2)}_{1c} \\ \boldsymbol{\beta}^{(2)}_{3c} \end{pmatrix} + \begin{pmatrix} S_q & 0 \\ 0 & S_q \end{pmatrix} \begin{pmatrix} \boldsymbol{b}^{(2)}_{1ic} \\ \boldsymbol{b}^{(2)}_{3ic} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}^{(2)}_{1ic} \\ \boldsymbol{\varepsilon}^{(2)}_{3ic} \end{pmatrix} \\ \\ \begin{pmatrix} \boldsymbol{V}_{2ic} \\ \boldsymbol{V}_{3ic} \end{pmatrix} = \begin{pmatrix} S_q & 0 \\ 0 & S_q \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(3)}_{2c} \\ \boldsymbol{\beta}^{(3)}_{3c} \end{pmatrix} + \begin{pmatrix} S_q & 0 \\ 0 & S_q \end{pmatrix} \begin{pmatrix} \boldsymbol{b}^{(3)}_{2ic} \\ \boldsymbol{b}^{(3)}_{3ic} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}^{(3)}_{2ic} \\ \boldsymbol{\varepsilon}^{(3)}_{3ic} \end{pmatrix} \end{cases}$$

$$\text{with} \quad \begin{pmatrix} \boldsymbol{b}^{(1)}_{1ic} \\ \boldsymbol{b}^{(1)}_{2ic} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} D^{(1)}_{1c} & D^{(1)}_{12c} \\ D^{(1)}_{12c} & D^{(1)}_{2c} \end{pmatrix} \right)$$

$$\begin{pmatrix} \boldsymbol{b}^{(2)}_{1ic} \\ \boldsymbol{b}^{(2)}_{3ic} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} D^{(2)}_{1c} & D^{(2)}_{13c} \\ D^{(2)}_{13c} & D^{(2)}_{3c} \end{pmatrix} \right)$$

$$\begin{pmatrix} \boldsymbol{b}^{(3)}_{2ic} \\ \boldsymbol{b}^{(3)}_{3ic} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} D^{(3)}_{2c} & D^{(3)}_{23c} \\ D^{(3)}_{23c} & D^{(3)}_{3c} \end{pmatrix} \right)$$

$$\text{and} \quad \begin{pmatrix} \boldsymbol{\varepsilon}^{(1)}_{1ic} \\ \boldsymbol{\varepsilon}^{(1)}_{2ic} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^{(1)}_{1c} I_n & \sigma^{(1)}_{12c} I_n \\ \sigma^{(1)}_{12c} I_n & \sigma^{(1)}_{2c} I_n \end{pmatrix} \right)$$

$$\begin{pmatrix} \boldsymbol{\varepsilon}^{(2)}_{1ic} \\ \boldsymbol{\varepsilon}^{(2)}_{3ic} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^{(2)}_{1c} I_n & \sigma^{(2)}_{13c} I_n \\ \sigma^{(2)}_{13c} I_n & \sigma^{(2)}_{3c} I_n \end{pmatrix} \right)$$

$$\begin{pmatrix} \boldsymbol{\varepsilon}^{(3)}_{2ic} \\ \boldsymbol{\varepsilon}^{(3)}_{3ic} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^{(3)}_{2c} I_n & \sigma^{(3)}_{23c} I_n \\ \sigma^{(3)}_{23c} I_n & \sigma^{(3)}_{3c} I_n \end{pmatrix} \right)$$

where S_q is the $(n \times q)$ natural cubic splines basis matrix with q knots, and all the D-matrices are unstructured $(q \times q)$ matrices.

The final model then becomes

$$\begin{pmatrix}\boldsymbol{V}_{1ic}\\ \boldsymbol{V}_{2ic}\\ \boldsymbol{V}_{3ic}\end{pmatrix} = \begin{pmatrix}S_q & 0 & 0\\ 0 & S_q & 0\\ 0 & 0 & S_q\end{pmatrix}\begin{pmatrix}\beta^*_{1c}\\ \beta^*_{2c}\\ \beta^*_{3c}\end{pmatrix} + \begin{pmatrix}S & 0 & 0\\ 0 & S & 0\\ 0 & 0 & S\end{pmatrix}\begin{pmatrix}\boldsymbol{b}_{1ic}\\ \boldsymbol{b}_{2ic}\\ \boldsymbol{b}_{3ic}\end{pmatrix} + \begin{pmatrix}\boldsymbol{\varepsilon}_{1ic}\\ \boldsymbol{\varepsilon}_{2ic}\\ \boldsymbol{\varepsilon}_{3ic}\end{pmatrix} \tag{8.10}$$

with

$$\begin{cases}\beta^*_{1c} & = & \frac{\boldsymbol{\beta}^{(1)}_{1c}+\boldsymbol{\beta}^{(2)}_{1c}}{2}\\ \beta^*_{2c} & = & \frac{\boldsymbol{\beta}^{(1)}_{2c}+\boldsymbol{\beta}^{(3)}_{2c}}{2}\\ \beta^*_{3c} & = & \frac{\boldsymbol{\beta}^{(2)}_{3c}+\boldsymbol{\beta}^{(3)}_{3c}}{2}\end{cases}$$

$$\begin{pmatrix}\boldsymbol{b}_{1ic}\\ \boldsymbol{b}_{2ic}\\ \boldsymbol{b}_{3ic}\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}D^*_{1c} = \frac{D^{(1)}_{1c}+D^{(2)}_{1c}}{2} & D^*_{12c} = D^{(1)}_{12c} & D^*_{13c} = D^{(2)}_{13c}\\ D^*_{12c} = D^{(1)}_{12c} & D^*_{2c} = \frac{D^{(1)}_{2c}+D^{(3)}_{2c}}{2} & D^*_{23c} = D^{(3)}_{23c}\\ D^*_{13c} = D^{(1)}_{13c} & D^*_{23c} = D^{(3)}_{23c} & D^*_{3c} = \frac{D^{(2)}_{3c}+D^{(3)}_{3c}}{2}\end{pmatrix}\right)$$

$$\begin{pmatrix}\boldsymbol{\varepsilon}_{1ic}\\ \boldsymbol{\varepsilon}_{2ic}\\ \boldsymbol{\varepsilon}_{3ic}\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}(\sigma^*_{1c} = \frac{\sigma^{(1)}_{1c}+\sigma^{(2)}_{1c}}{2})I_n & (\sigma^*_{12c} = \sigma^{(1)}_{12c})I_n & (\sigma^*_{13c} = \sigma^{(2)}_{13c})I_n\\ (\sigma^*_{12c} = \sigma^{(1)}_{12c})I_n & (\sigma^*_{2c} = \frac{\sigma^{(1)}_{2c}+\sigma^{(3)}_{2c}}{2})I_n & (\sigma^*_{23c} = \sigma^{(3)}_{23c})I_n\\ (\sigma^*_{13c} = \sigma^{(2)}_{13c})I_n & (\sigma^*_{23c} = \sigma^{(3)}_{23c})I_n & (\sigma^*_{3c} = \frac{\sigma^{(3)}_{3c}+\sigma^{(3)}_{3c}}{2})I_n\end{pmatrix}\right)$$

Classification

We can now basically use the classification rule proposed by James and Hastie, now incorporating the variance-covariance matrix obtained from the pseudo-likelihood model. For a new observation $\boldsymbol{Y}_{\text{new}}$, which is a $m{\cdot}n$ vector consisting of m longitudinal curves measured at n timepoints, the functional distance $d(\boldsymbol{Y}_{\text{new}}, c)$ to class c is calculated using the mean and covariance structure estimated in the pseudo-likelihood models as follows

$$d(\boldsymbol{Y}_{\text{new}}, c) = (\boldsymbol{Y}_{\text{new}} - S_{\text{full}}\beta^*_c)\,\Gamma^*_c\,(\boldsymbol{Y}_{\text{new}} - S_{\text{full}}\beta^*_c)^T\,. \tag{8.11}$$

The new observation $\boldsymbol{Y}_{\text{new}}$ will now be classified into the class for which the functional distance is minimal.

8.2.3 Computational Issues

Some computational issues arise when modeling multivariate outcomes using pseudo-likelihood methodology. Since the covariance matrix Γ^*_c is not guaranteed to be positive definite, problems can occur in the calculation of the functional distance. To solve this issue, we followed two strategies.

Rousseeuw and Molenberghs (1993) Strategy

A first option is to correct the covariance matrix to be positive definite as proposed by Rousseeuw and Molenberghs (1993). The covariance matrix Γ_c^* can be written as

$$\Gamma_c^* = P_c \Lambda_c P_c^T$$

where Λ_c is a diagonal matrix, containing the eigenvalues of Γ_c and P is an orthogonal matrix of the corresponding eigenvectors. In a positive definite matrix all the eigenvalues are positive. To transform a non-positive definite matrix to a positive definite one, Rousseeuw and Molenberghs (1993) therefore proposed to replace the negative values in Λ_c by a small positive value, which gives us a diagonal matrix $\Lambda_{c,\text{modif}}$. The modified covariance matrix $\Gamma^*_{c,\text{modif}}$ can now be calculated by

$$\Gamma^*_{c,\text{modif}} = P_c \Lambda_{c,\text{modif}} P_c^T$$

This modified covariance matrix is now positive definite and it can be used instead of the original covariance matrix in equation (8.11), which becomes now

$$d_{\text{modif}}(\boldsymbol{Y}_{\text{new}}, c) = (\boldsymbol{Y}_{\text{new}} - S_{\text{full}}\beta_c^*)\,\Gamma^*_{c,\text{modif}}\,(\boldsymbol{Y}_{\text{new}} - S_{\text{full}}\beta_c^*)^T\,. \tag{8.12}$$

Pairwise Strategy

For each of the pairwise models, the estimated variance covariance matrices are known to be positive definite. Therefore we can calculate the distance between a new observation and a particular class for each pair of covariates. For example for covariates r and s, we compute the distance to class c as

$$d_{rs}(\boldsymbol{Y}_{\text{new}}, c) =$$

$$\left(\begin{pmatrix} Y_{\text{new},r} \\ Y_{\text{new},s} \end{pmatrix} - \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta}_{rc} \\ \boldsymbol{\beta}_{sc} \end{pmatrix}\right)\Gamma_{rs,c}\left(\begin{pmatrix} Y_{\text{new},r} \\ Y_{\text{new},s} \end{pmatrix} - \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta}_{rc} \\ \boldsymbol{\beta}_{sc} \end{pmatrix}\right)^T$$

with

$$\Gamma_{rs,c} = \begin{pmatrix} \Sigma_{rc} & \Sigma_{rsc} \\ \Sigma_{rsc} & \Sigma_{sc} \end{pmatrix} + \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}\begin{pmatrix} D_{rc} & D_{rsc} \\ D_{rsc} & D_{sc} \end{pmatrix}\begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix}^T$$

where $\boldsymbol{\beta}_{rc}, \boldsymbol{\beta}_{sc}, \Sigma_{rc}, \Sigma_{sc}, \Sigma_{rsc}, D_{rc}, D_{sc}, D_{rsc}$ and S are defined as in equation 8.8. The distance between the new observation and class c is calculated as the average of all the pairwise distances.

$$d_{\text{pair}}(\boldsymbol{Y}_{\text{new}}, c) = \frac{1}{m(m-1)/2}\sum_{r=1}^{m-1}\sum_{s=r+1}^{m} d_{rs}(\boldsymbol{Y}_{\text{new}}, c) \tag{8.13}$$

8.3 Simulation Study

Before we turn to the classification of the EEG dataset, the performance of the method will be evaluated trough a simulation study. Since we want to mimic a real-life application, we will use the EEG dataset as a basis for the generation of the data. Several settings, regarding the number of classes and the number of subjects in each class will be considered.

We will describe the three different settings used in the simulation study in Section 8.3.1. The results are reported and discussed in Section 8.3.2.

8.3.1 Simulation Setting

The simulation setting is based on the EEG dataset and the estimated parameters obtained for this data. A three-variate longitudinal profile is generated using the parameters in the pseudo-likelihood model with splines for the covariates Active Wake, Light Sleep and Deep Sleep of the EEG-dataset.

Let us start with the first setting. In this setting, we restrict to only two classes, based on parameters obtained for antipsychotics and stimulants in the EEG dataset. For both classes, a pseudo-likelihood model using splines with five knots as random and fixed effects, is fitted to the three covariates in the light period. For antipsychotics, the model is written as follows

$$\begin{pmatrix} \text{AW min}_{i\text{ap}} \\ \text{SWS1 min}_{i\text{ap}} \\ \text{SWS2 min}_{i\text{ap}} \end{pmatrix} = \begin{pmatrix} S_5 & 0 & 0 \\ 0 & S_5 & 0 \\ 0 & 0 & S_5 \end{pmatrix} \begin{pmatrix} \beta^*_{1\text{ap}} \\ \beta^*_{2\text{ap}} \\ \beta^*_{3\text{ap}} \end{pmatrix} + \begin{pmatrix} S_5 & 0 & 0 \\ 0 & S_5 & 0 \\ 0 & 0 & S_5 \end{pmatrix} \begin{pmatrix} \boldsymbol{b}^*_{1i\text{ap}} \\ \boldsymbol{b}^*_{2i\text{ap}} \\ \boldsymbol{b}^*_{3i\text{ap}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}^*_{1i\text{ap}} \\ \boldsymbol{\varepsilon}^*_{2i\text{ap}} \\ \boldsymbol{\varepsilon}^*_{3i\text{ap}} \end{pmatrix}$$

with

$$\begin{pmatrix} \boldsymbol{b}^*_{1i\text{ap}} \\ \boldsymbol{b}^*_{2i\text{ap}} \\ \boldsymbol{b}^*_{3i\text{ap}} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} D^*_{1\text{ap}} & D^*_{12\text{ap}} & D^*_{13\text{ap}} \\ D^*_{12\text{ap}} & D^*_{2\text{ap}} & D^*_{23\text{ap}} \\ D^*_{13\text{ap}} & D^*_{23\text{ap}} & D^*_{3\text{ap}} \end{pmatrix} \right)$$

$$\begin{pmatrix} \boldsymbol{\varepsilon}^*_{1i\text{ap}} \\ \boldsymbol{\varepsilon}^*_{2i\text{ap}} \\ \boldsymbol{\varepsilon}^*_{3i\text{ap}} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^*_{1\text{ap}} I_n & \sigma^*_{12\text{ap}} I_n & \sigma^*_{13\text{ap}} I_n \\ \sigma^*_{12\text{ap}} I_n & \sigma^*_{2\text{ap}} I_n & \sigma^*_{23\text{ap}} I_n \\ \sigma^*_{13\text{ap}} I_n & \sigma^*_{23\text{ap}} I_n & \sigma^*_{3\text{ap}} I_n \end{pmatrix} \right)$$

where all the parameters are defined as in model (8.10). For class C_1 we now generate a new dataset with M subjects by sampling from the normal distribution $N(\mu^*_{\text{ap}}, \Gamma^*_{\text{ap}})$, where μ^*_{ap} and Γ^*_{ap} are defined as follows

$$\mu^*_{\text{ap}} = \begin{pmatrix} S_5 & 0 & 0 \\ 0 & S_5 & 0 \\ 0 & 0 & S_5 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^*_{1\text{ap}} \\ \boldsymbol{\beta}^*_{2\text{ap}} \\ \boldsymbol{\beta}^*_{3\text{ap}} \end{pmatrix}$$

$$\Gamma^*_{\text{ap}} = \begin{pmatrix} \sigma^*_{1\text{ap}} I_n & \sigma^*_{12\text{ap}} I_n & \sigma^*_{13\text{ap}} I_n \\ \sigma^*_{12\text{ap}} I_n & \sigma^*_{2\text{ap}} I_n & \sigma^*_{23\text{ap}} I_n \\ \sigma^*_{13\text{ap}} I_n & \sigma^*_{23\text{ap}} I_n & \sigma^*_{3\text{ap}} I_n \end{pmatrix} + \begin{pmatrix} S_5 & 0 & 0 \\ 0 & S_5 & 0 \\ 0 & 0 & S_5 \end{pmatrix} \begin{pmatrix} D^*_{1\text{ap}} & D^*_{12\text{ap}} & D^*_{13\text{ap}} \\ D^*_{12\text{ap}} & D^*_{2\text{ap}} & D^*_{23\text{ap}} \\ D^*_{13\text{ap}} & D^*_{23\text{ap}} & D^*_{3\text{ap}} \end{pmatrix} \begin{pmatrix} S_5 & 0 & 0 \\ 0 & S_5 & 0 \\ 0 & 0 & S_5 \end{pmatrix}$$

Basically, this means that in each simulated dataset, we have M three-variate longitudinal profiles, at 20 timepoints. Similarly, we generate datasets with M profiles in class C_2 by sampling from the normal distribution $N(\mu^*_{\text{st}}, \Gamma^*_{\text{st}})$. The number of samples M in each generated dataset will take values 20, 40, 60 80 and 100. The simulated training dataset now consists of the data generated in both classes. An independent test dataset with 10 profiles in each class is generated from the same distribution. This test dataset will be classified using MFLDA with splines with two or three knots based on the training dataset.

Setting two is essentially the same as setting one, but now one extra class C_3, based on antidepressants in the EEG dataset, is added. In setting three, we start from the three classes of setting two and add another class C_4, based on hypnotics.

Given the complexity and the computational time needed for such exercise, we will run 100 simulations for each setting and for each number of observations.

Let us now turn to the classification results obtained for each of the three settings.

8.3.2 Simulation Results

In order to illustrate the three simulation settings, one dataset was selected randomly from the simulated data in each setting. For each of these datasets, 10 profiles were randomly picked and displayed in Figures 8.1 − 8.3. As in the EEG dataset, the profiles are highly irregular and the classes are difficult to differentiate at sight.

In Table 8.1 the overall error rates in setting 1, 2 and 3 are displayed. Between parentheses are the corresponding standard deviations. For each setting, four different analyses were performed, MFLDA with two or three knots using the modified distance measure to correct for positive definiteness or using the pairwise strategy.

It can be seen that the error rates increase when introducing more classes, while they decrease when more subjects are added to the training datasets. The MFLDA using three knots performs better in terms of misclassification rate than the MFLDA

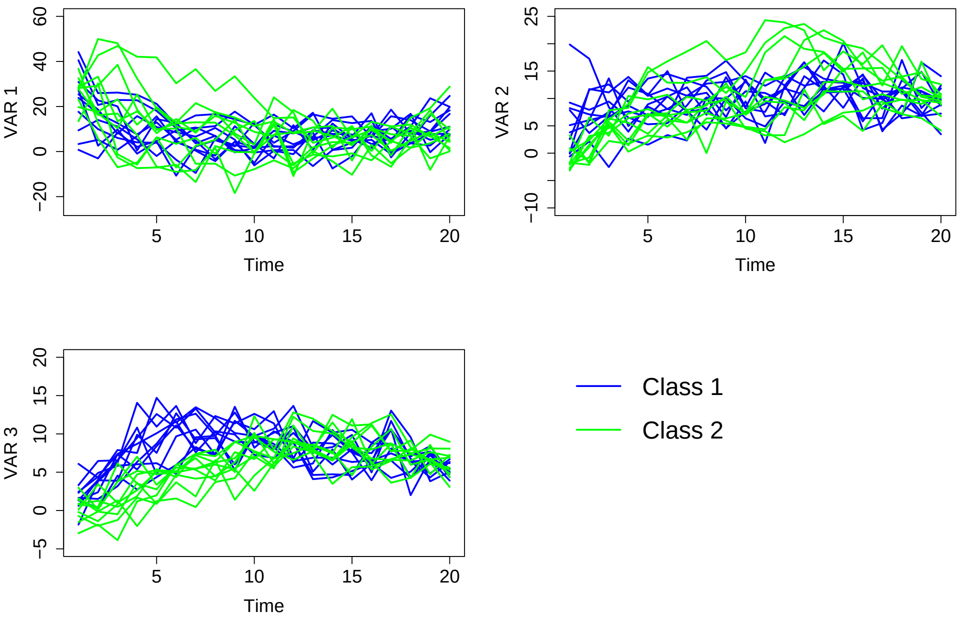


Figure 8.1: *Generated data in simulation setting 1 for the three variables based on the estimated parameters from the EEG dataset for Active Wake, Light Sleep and Deep Sleep respectively, in the classes antipsychotics (class C_1) and stimulants (class C_2).*

with only two knots. This is expected since the model with two knots smooths out more drastically the trend in the datasets compared to the model with three knots, and thus is less efficient to estimate the variance-covariance matrix, which is a key feature in the classification procedure. Nevertheless, it is important to note that the MFLDA based on only two knots is still performing well in setting 1 and 2 and should not be set aside since it is computationally less demanding than MFLDA with three knots and the performance of both methods is comparable.

Comparing the results obtained with the correction for positive definiteness and with the pairwise strategy shows us that in general the error rates obtained with the latter method is giving a lower overall error rate than strategy one. Only when four classes are considered (setting 3), MFLDA with the pairwise strategy encounters more problems to discriminate between the classes. The standard deviations for the error rates obtained with the pairwise strategy are in general smaller than the ones obtained with the correction for positive definiteness, producing then narrower

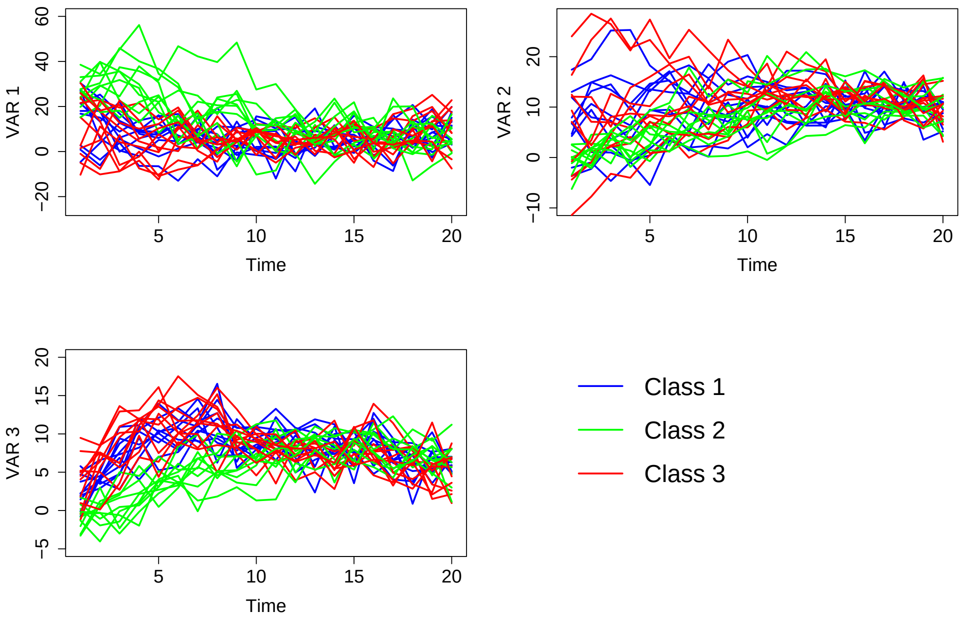


Figure 8.2: *Generated data in simulation setting 2 for the three variables based on the estimated parameters from the EEG dataset for Active Wake, Light Sleep and Deep Sleep respectively, in the classes antipsychotics (class C_1), stimulants (class C_2) and antidepressants (class C_3).*

confidence intervals, and thus more informative results.

In Tables 8.2 – 8.4 the error rates (and corresponding standard deviations) for each of the classes in the three settings are reported. For setting 1 the error rates per class and the corresponding standard deviations decrease with increasing sample size.

For setting 2, we notice that in all four classification methods, there is a problem to differentiate class 3 from the rest. This problem is more pronounced with two knots compared to three knots. With low sample size (less than 60 subjects in the training dataset), also class 1 is poorly discriminated. Note also that when only 2 knots are used, the classification of class 3 does not improve with increasing sample size, while class 1 does improve. When 3 knots are used, both for class 1 and class 3 the error rate reduces with increasing sample size, but the pairwise strategy performs better than the correction for positive definiteness strategy, even when the sample size is small.

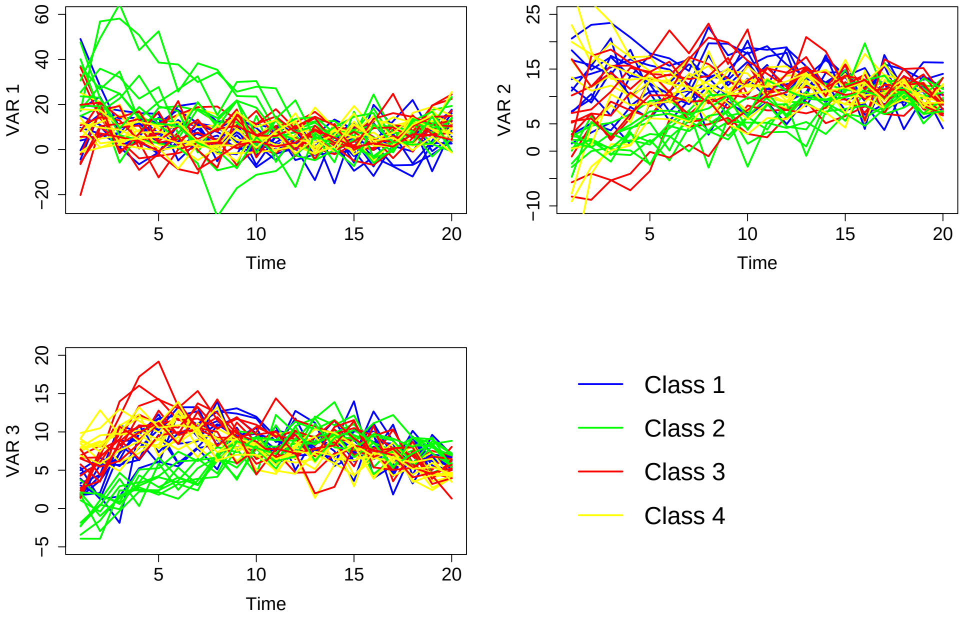


Figure 8.3: *Generated data in simulation setting 3 for the three variables based on the estimated parameters from the EEG dataset for Active Wake, Light Sleep and Deep Sleep respectively, in the classes antipsychotics (class C_1), stimulants (class C_2), antidepressants (class C_3) and hypnotics (class C_4).*

In setting 3, the error rates in class 1, 2, and 3 are similar as in setting 2, but the error rates obtained in class 4 are dramatically high, especially for the analysis with 2 knots using pairwise strategy. Since the profiles for class 3 and 4 were generated based on antidepressants and hypnotics in the EEG dataset respectively, this problem is not unexpected and it confirms our belief that hypnotics and antidepressants are hardly separable. Even though the pairwise strategy produces higher error rates for class 4, both strategies reduce the error rate with about 30% when comparing a model using 3 knots and 2 knots.

To have a further idea on the separability of the classes in the EEG dataset, we repeated the simulation exercise with setting 1 for all the pairs of classes. The results of these analyses are shown in Appendix C and can be summarized as follows. Antipsychotics and stimulants were differentiated well in all settings. Placebos are hardly differentiable from antipsychotics and antidepressants. Also between hypnotics and antidepressants the classification is poor. This last one explains the classification results in setting 3 of the simulation study. If we can not distinguish between antidepressants and hypnotics when there are no other compounds in the dataset, we cannot expect to get a good classification result with two additional classes.

Table 8.1: *Overall misclassification error (and empirical standard deviations) obtained in the simulation studies for the three settings (2, 3 and 4 classes) with MFLDA with two (left panel) or three knots (right panel).*

		2 knots		3 knots	
	Number of Subjects	Correction Positive Definite	Pairwise Classification	Correction Positive Definite	Pairwise Classification
Setting 1	20	0.065 (0.082)	0.013 (0.026)	0.035 (0.061)	0.005 (0.019)
	40	0.060 (0.079)	0.012 (0.026)	0.008 (0.037)	0.002 (0.009)
	60	0.021 (0.037)	0.011 (0.023)	0.001 (0.007)	0.002 (0.009)
	80	0.029 (0.057)	0.008 (0.020)	0.001 (0.007)	0.001 (0.005)
	100	0.016 (0.032)	0.007 (0.022)	0.001 (0.008)	0.002 (0.010)
Setting 2	20	0.201 (0.079)	0.169 (0.086)	0.213 (0.074)	0.128 (0.064)
	40	0.175 (0.087)	0.144 (0.074)	0.151 (0.079)	0.094 (0.058)
	60	0.149 (0.066)	0.141 (0.073)	0.124 (0.073)	0.076 (0.052)
	80	0.140 (0.059)	0.132 (0.063)	0.119 (0.069)	0.087 (0.051)
	100	0.141 (0.072)	0.131 (0.070)	0.092 (0.059)	0.069 (0.048)
Setting 3	20	0.295 (0.086)	0.346 (0.073)	0.290 (0.083)	0.251 (0.077)
	40	0.282 (0.072)	0.329 (0.055)	0.221 (0.084)	0.229 (0.064)
	60	0.271 (0.068)	0.333 (0.056)	0.194 (0.074)	0.222 (0.060)
	80	0.269 (0.071)	0.326 (0.049)	0.201 (0.066)	0.215 (0.063)
	100	0.271 (0.072)	0.328 (0.056)	0.188 (0.069)	0.219 (0.057)

Table 8.2: *Error rates for the simulation study with 2 classes and corresponding empirical standard deviations (between parentheses) obtained with MFLDA with two knots (upper panel) and three knots (lower panel) using the correction for positive definiteness or the pairwise distances.*

Setting 1 − 2 knots

	Correction PD		Pairwise Strategy	
M	Class C_1	Class C_2	Class C_1	Class C_2
20	0.112 (0.163)	0.018 (0.066)	0.018 (0.041)	0.008 (0.037)
40	0.120 (0.158)	0.001 (0.010)	0.020 (0.049)	0.004 (0.019)
60	0.041 (0.074)	0.001 (0.010)	0.020 (0.045)	0.002 (0.014)
80	0.057 (0.114)	0.002 (0.014)	0.014 (0.037)	0.002 (0.014)
100	0.032 (0.063)	0.000 (0.000)	0.014 (0.045)	0.000 (0.000)

Setting 1 − 3 knots

	Correction PD		Pairwise Strategy	
M	Class C_1	Class C_2	Class C_1	Class C_2
20	0.062 (0.120)	0.008 (0.034)	0.004 (0.024)	0.005 (0.026)
40	0.014 (0.074)	0.002 (0.014)	0.000 (0.000)	0.003 (0.017)
60	0.000 (0.000)	0.002 (0.014)	0.000 (0.000)	0.004 (0.019)
80	0.001 (0.010)	0.001 (0.010)	0.000 (0.000)	0.001 (0.010)
100	0.000 (0.000)	0.003 (0.017)	0.000 (0.000)	0.005 (0.022)

Table 8.3: *Error rates for the simulation study with 3 classes and corresponding empirical standard deviations (between parentheses) obtained with MFLDA with two knots (upper panel) and three knots (lower panel) using the correction for positive definiteness or the pairwise distances.*

Setting 2 – 2 knots

	Correction PD			Pairwise Strategy		
M	Class C_1	Class C_2	Class C_3	Class C_1	Class C_2	Class C_3
20	0.321 (0.255)	0.003 (0.017)	0.278 (0.249)	0.169 (0.211)	0.006 (0.042)	0.333 (0.282)
40	0.267 (0.268)	0.002 (0.014)	0.256 (0.255)	0.107 (0.141)	0.001 (0.010)	0.325 (0.240)
60	0.172 (0.202)	0.002 (0.014)	0.272 (0.217)	0.093 (0.137)	0.002 (0.014)	0.327 (0.237)
80	0.148 (0.177)	0.000 (0.000)	0.272 (0.201)	0.043 (0.073)	0.001 (0.010)	0.352 (0.202)
100	0.138 (0.187)	0.003 (0.017)	0.282 (0.217)	0.064 (0.113)	0.003 (0.017)	0.325 (0.227)

Setting 2 – 3 knots

	Correction PD			Pairwise Strategy		
M	Class C_1	Class C_2	Class C_3	Class C_1	Class C_2	Class C_3
20	0.256 (0.241)	0.013 (0.056)	0.370 (0.270)	0.168 (0.169)	0.013 (0.034)	0.202 (0.169)
40	0.123 (0.180)	0.003 (0.017)	0.326 (0.249)	0.125 (0.155)	0.004 (0.020)	0.152 (0.155)
60	0.071 (0.131)	0.003 (0.017)	0.297 (0.227)	0.109 (0.133)	0.006 (0.024)	0.114 (0.132)
80	0.065 (0.102)	0.002 (0.014)	0.290 (0.220)	0.103 (0.129)	0.005 (0.022)	0.152 (0.152)
100	0.045 (0.078)	0.004 (0.019)	0.228 (0.187)	0.074 (0.097)	0.006 (0.024)	0.126 (0.129)

Table 8.4: *Error rates for the simulation study with 4 classes and corresponding empirical standard deviations (between parentheses) obtained with MFLDA with two knots (upper panel) and three knots (lower panel) using the correction for positive definiteness or the pairwise distances.*

Setting 3 – 2 knots

	Correction PD				Pairwise Strategy			
M	Class C_1	Class C_2	Class C_3	Class C_4	Class C_1	Class C_2	Class C_3	Class C_4
20	0.340 (0.276)	0.006 (0.024)	0.278 (0.253)	0.557 (0.273)	0.156 (0.213)	0.007 (0.026)	0.355 (0.302)	0.865 (0.156)
40	0.227 (0.242)	0.000 (0.000)	0.276 (0.234)	0.623 (0.228)	0.106 (0.141)	0.003 (0.017)	0.278 (0.228)	0.929 (0.107)
60	0.190 (0.221)	0.000 (0.000)	0.265 (0.217)	0.630 (0.195)	0.088 (0.154)	0.002 (0.014)	0.310 (0.211)	0.931 (0.103)
80	0.156 (0.193)	0.001 (0.010)	0.261 (0.207)	0.660 (0.213)	0.079 (0.114)	0.004 (0.019)	0.266 (0.190)	0.954 (0.066)
100	0.114 (0.180)	0.001 (0.010)	0.299 (0.194)	0.669 (0.205)	0.044 (0.078)	0.002 (0.014)	0.344 (0.078)	0.922 (0.112)

Setting 3 – 3 knots

	Correction PD				Pairwise Strategy			
M	Class C_1	Class C_2	Class C_3	Class C_4	Class C_1	Class C_2	Class C_3	Class C_4
20	0.281 (0.255)	0.008 (0.031)	0.339 (0.250)	0.531 (0.299)	0.223 (0.208)	0.008 (0.027)	0.182 (0.179)	0.590 (0.234)
40	0.104 (0.169)	0.005 (0.022)	0.337 (0.244)	0.440 (0.279)	0.126 (0.139)	0.007 (0.026)	0.178 (0.142)	0.607 (0.216)
60	0.050 (0.096)	0.003 (0.017)	0.298 (0.219)	0.427 (0.255)	0.106 (0.113)	0.002 (0.014)	0.132 (0.129)	0.649 (0.201)
80	0.078 (0.131)	0.003 (0.017)	0.255 (0.240)	0.469 (0.289)	0.102 (0.115)	0.002 (0.014)	0.142 (0.138)	0.615 (0.205)
100	0.051 (0.073)	0.002 (0.014)	0.230 (0.191)	0.470 (0.254)	0.077 (0.098)	0.003 (0.017)	0.136 (0.125)	0.659 (0.179)

8.4 Application: The EEG Dataset

8.4.1 Pseudo-Likelihood Model

Before we turn to the classification results obtained with the multivariate functional discriminant analysis, let us first zoom in on the pairwise modeling results.

When dealing with all data, obtained in the light and the dark period, we will need to fit all pairwise combinations of 6 responses in the light period and 6 in the dark period, leading to 66 pairwise models. This is computationally very demanding and will therefore not be presented here. Instead we will restrict to the light period, since we know that most of the action of the drugs will be located there.

For subject i in class c the model is specified as

$$\left\{\begin{array}{rcl} \text{AW min}_{ic}(t) & = & S_{\text{eeg}}\boldsymbol{\beta}_{\text{AW},c} + S_{\text{eeg}}\boldsymbol{b}_{\text{AW},ic} + \varepsilon_{\text{AW},ic}(t) \\ \text{PW min}_{ic}(t) & = & S_{\text{eeg}}\boldsymbol{\beta}_{\text{PW},c} + S_{\text{eeg}}\boldsymbol{b}_{\text{PW},ic} + \varepsilon_{\text{PW},ic}(t) \\ \text{SWS1 min}_{ic}(t) & = & S_{\text{eeg}}\boldsymbol{\beta}_{\text{SWS1},c} + S_{\text{eeg}}\boldsymbol{b}_{\text{SWS1},ic} + \varepsilon_{\text{SWS1},ic}(t) \\ \text{SWS2 min}_{ic}(t) & = & S_{\text{eeg}}\boldsymbol{\beta}_{\text{SWS2},c} + S_{\text{eeg}}\boldsymbol{b}_{\text{SWS2},ic} + \varepsilon_{\text{SWS2},ic}(t) \\ \text{IS min}_{ic}(t) & = & S_{\text{eeg}}\boldsymbol{\beta}_{\text{IS},c} + S_{\text{eeg}}\boldsymbol{b}_{\text{IS},ic} + \varepsilon_{\text{IS},ic}(t) \\ \text{RS min}_{ic}(t) & = & S_{\text{eeg}}\boldsymbol{\beta}_{\text{RS},c} + S_{\text{eeg}}\boldsymbol{b}_{\text{RS},ic} + \varepsilon_{\text{RS},ic}(t) \end{array}\right. \tag{8.14}$$

where $t = 1, \cdots, 20$ is the vector of timepoints in the light period and S_{eeg} is a natural cubic spline basis matrix for t with 5 knotpoints, $\boldsymbol{\beta}_{\text{AW},c}, \cdots, \boldsymbol{\beta}_{\text{RS},c}$, $\boldsymbol{b}_{\text{AW},ic}, \cdots, \boldsymbol{b}_{\text{RS},ic}$ are 5-dimensional vectors of coefficients. The fixed effects vector $\boldsymbol{\beta}_c = \boldsymbol{\beta}_{AW,c}, \cdots, \boldsymbol{\beta}_{RS,c}$ describes the average evolution. The random effects $\boldsymbol{b}_{AW,ic}, \cdots, \boldsymbol{b}_{RS,ic}$ follow a 30-dimensional joint normal distribution with mean 0 and covariance matrix D_c. The error components $\varepsilon_{AW,ic}, \cdots, \varepsilon_{RS,ic}$ follow a 6-dimensional normal distribution with mean 0 and covariance matrix Σ_c.

The pairwise fitting described in Section 8.2.1 results in 5 estimates for each fixed effect, each variance of the random effects and each variance of the error components. The mean of these estimates is computed to get an estimate of the parameters. For the covariances between each two random effects and between each two error components, one single estimate is obtained and thus there is no need for further adjustment.

In Figure 8.4 the model obtained with the pseudo-likelihood pairwise modeling approach in the light period is shown together with the mean profiles for the treatments in the five drug classes. All graphs show that the fitted line follows the trends present in the data.

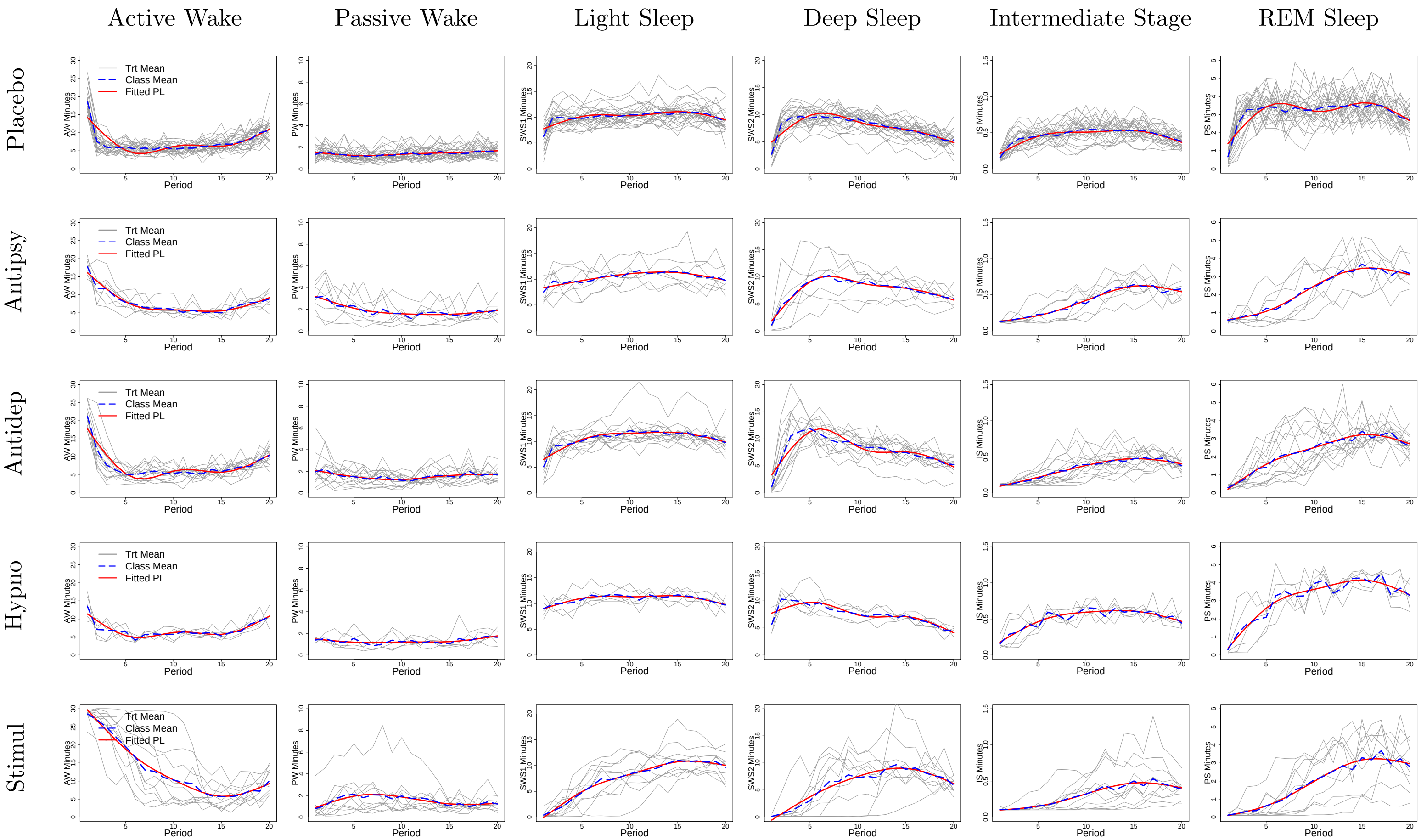


Figure 8.4: *Observed mean profile per treatment (grey) with observed mean profile per class (blue dashed line) and fitted pseudo-likelihood model with spline basis for each sleeping stage and each class (red solid line).*

8.4.2 Multivariate Functional Linear Discriminant Analysis

Stepwise MFLDA

Given the results of the simulation study that 2 classes were easier to classify, we will develop a stepwise classification rule in analogy to the doubly hierarchical supervised learning analysis. In this way, the variability in the data is reduced along the process, making it easier to classify the classes in the last steps of the procedure. Also, in each step, different sleeping stages can be used in the classification rule, making it possible to fine-tune the procedure to the considered classes.

To select the sleeping stages to be used in step s, the classification performance in the training dataset is evaluated by means of an error measure taking into account the misclassifications in the class to be discriminated in step s as well as the misclassified observations in the other classes. The error measure, in analogy with the measurement proposed in Chapter 5, is defined as

$$\text{Error}_s = w_{s1}\text{ERR}_{C_sC_{-s}} + w_{s2}\text{ERR}_{C_{-s}C_s} \tag{8.15}$$

where $\text{ERR}_{C_sC_{-s}}$ is the percentage of misclassified observations in the class discriminated in step s, while $\text{ERR}_{C_{-s}C_s}$ is the percentage of observations in the other classes that are misclassified in the class discriminated in step s. The weights w_{s1} and w_{s2} can be chosen according to the situation. In analogy to the DHSLA we choose here $w_{s1} = s + 1$ and $w_{s2} = 2 \cdot (5 - s)$. The combination of sleeping stages resulting in the smallest value for Error_s is retained in step s.

For that combination of sleeping stages, the error rate is calculated on the rat level ($\text{error}_{\text{rat}}$) as well as on the compound-dose level (error_{cd}). The error rate on the rat level is calculated with the percentage of rats misclassified in each class. For the error rate on the compound-dose level, we focus on the classification of a compound-dose, which is defined as the class for which the average distance over all the rats in the concerned compound-dose is minimal.

For both levels, rat and compound-dose combination, the order in which the drug classes need to be discriminated is determined as the one for which the final error ($\text{error}_{\text{rat}}$ or error_{cd}) is minimal.

Classification Results

Now we can turn to the actual classification of the EEG dataset with multivariate functional linear discriminant analysis. We start with the fitting process for each class and calculate the distance between an observation in the training or test dataset and each of the classes. The observation will be classified to the class for which this distance is minimal.

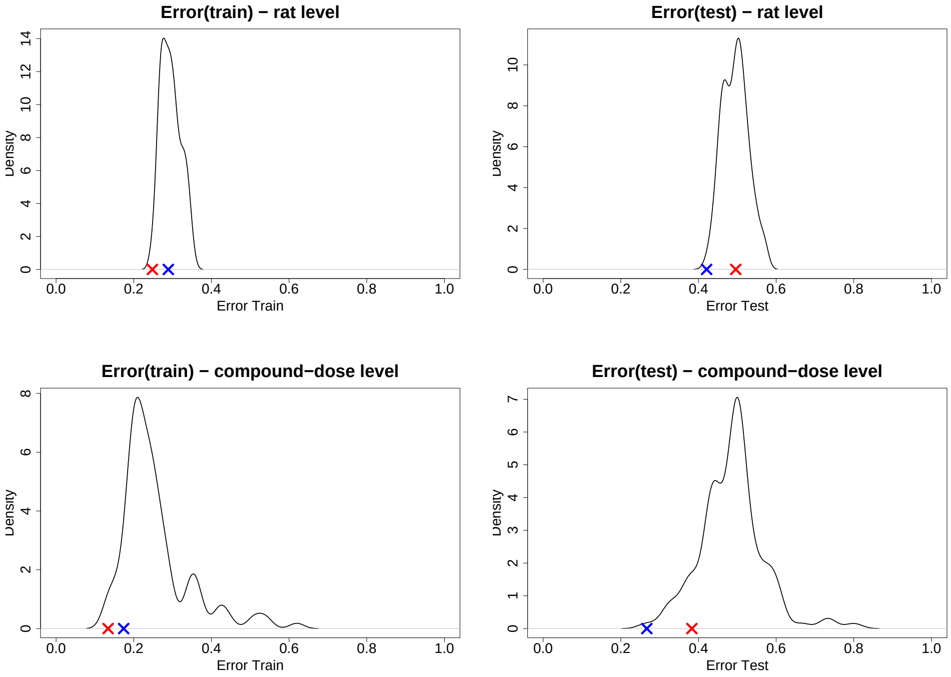
We will start with the results obtained with the modified distance measure d_{modif}. Afterwards the results with the pairwise distance measure d_{pair} are discussed. For both measures, we will first focus on the results obtained when the selection procedure of the sleeping stages is based on the rat level, followed by the results on the compound-dose level.

In the stepwise procedure, there are 120 possible sequences in which the classes can be differentiated. To determine which order is best, we calculate the error in the training dataset on the rat-level and the compound-dose level for each of the 120 orders. The order that leads to the smallest error will be retained. This is done for the classification with both distance measures. In the upper panels of Figure 8.5, the density function of the error rates in the training and test dataset on the rat level with the modified distance measure is plotted. The red cross marks the minimal error rate in the training dataset, and the corresponding error rate in the test dataset, while the blue cross indicates the minimum error rate in the test dataset and the corresponding error rate in the training dataset. The spread in the error rates is rather small, meaning that the order in which the classes are separated has not a big impact on the final classification result. When choosing the order that leads to the smallest error rate in the training dataset, the error in the test dataset is also relatively small.

The order chosen for the classification with the modified distance measure on the rat level is shown in Table 8.5. The sleeping stages selected in each of the steps are also presented in the table. Although the order we selected now is different from the order used in the DHSLA, we can still compare the sleeping stages we retain for a certain class with the ones that were retained for the DHSLA. Notice hereby that most sleeping stages that were selected in all the DHSLA analyses for a certain class, are also retained now. For example, for antidepressants we selected Intermediate Stage Sleep and REM Sleep, for hypnotics Deep Sleep and REM Sleep and for stimulants we selected Active Wake.

In Table 8.6 the classification result obtained in the training (upper panel) and test dataset (lower panel) with the correction for positive definiteness strategy are shown. The percentages in this table are the proportions of rats classified in each class. For the training dataset we get a high percentage of correctly classified rats in placebo, antipsychotics and stimulants and to less extent also for hypnotics. For antidepressants, the error rate is somewhat larger. The overall error rate on the rat level in the training dataset is 0.248. In the test dataset, only placebos could be

Figure 8.5: *Density of the error rates in training and test dataset obtained with MFLDA using the modified distance measure.*



classified well. For the other four classes the classification percentage in the correct class was below, or equal to, 50%. The overall error rate here is 0.496.

If instead we focus on the level of compound-by-dose combination, we note somewhat different results. The lower panel of Figure 8.5 shows the densities of the error rates on the compound-dose level in training and test dataset. Again the red and blue cross correspond to the order with the minimal error rate in training and test dataset respectively. We see immediately that the spread in the error rates is much larger compared to the rat level. The order corresponding to the red cross is presented in Table 8.7 together with the sleeping stages selected in each of the steps. Also here, we see similarities with the sleeping stages retained in the DHSLA analyses, e.g. Active Wake for stimulants, Passive Wake for antipsychotics and REM Sleep for hypnotics.

Table 8.8 shows the classification result on the compound-dose level obtained with the modified distance measure. In the training dataset, we get again high classification

Table 8.5: *Sleeping Stages used in each step of the MFLDA with modified distance measure, on rat level.*

Step	Sleeping Stages
Step 1: Placebo	AW - PW - SWS2 - RS
Step 2: Antidepressants	SWS2 - IS - RS
Step 3: Hypnotics	SWS1 - SWS2 - IS - RS
Step 4: Stimulants	AW - PW - SWS2 - RS

percentages in the correct classes for all the drug classes except antidepressants. Most antidepressants were classified as antipsychotic. The overall error rate on the compound-dose level in the training dataset is 0.134. In the test dataset, both placebo and antipsychotics could be well classified now, but there are still problems to classify antidepressants, hypnotics and stimulants correctly. The overall error rate on the compound-dose level is here 0.383.

The same analyses can be done with the pairwise strategy. In Figure 8.6, the density function for the error rates in training and test dataset on the rat level (upper panels) and compound-by-dose combination level(lower panels) are presented. Also here we observe that the variability in error rate is much larger for the compound-dose level, compared to the rat level. But in both scenarios, we can see that when choosing the order with the lowest error rate in the training dataset, the error rate in the test dataset is relatively small as well.

Let us first focus on the rat level. The selected order and the corresponding sleeping stages for each of the steps are shown in Table 8.9. As for the DHSLA, Active Wake is retained to classify stimulants, and Active Wake, Intermediate Stage Sleep and REM Sleep are retained for the classification of antidepressants. Table 8.10 describes the results obtained on the rat level when MFLDA is used with the pairwise distance calculation. In the training dataset we get now high classification percentages for the correct class in all five classes, while in the test dataset, antidepressants, hypnotics and stimulants are still poorly classified. The overall error rate on the rat level is here 0.199 in the training dataset and 0.442 in the test dataset.

On the level of the compound-by-dose combination, the order of the steps and the sleeping stages to be used in each of the steps are displayed in Table 8.11. Also here the sleeping stages selected in all the DHSLA analyses for the discrimination of a

Table 8.6: *Classification percentages for the EEG training dataset and test dataset, with modified distance measure, on rat level ($error_{rat}(train) = 0.248$, $error_{cd}(train) = 0.174$, $error_{rat}(test) = 0.496$, $error_{cd}(test) = 0.267$).*

Training Dataset

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypno	Stimul
Placebo	83.15%	5.43%	6.52%	4.89%	0.00%
Antipsy	10.71%	80.36%	5.36%	1.78%	1.78%
Antidep	9.82%	21.43%	59.82%	8.93%	0.00%
Hypno	25.00%	2.50%	2.50%	70.00%	0.00%
Stimul	3.75%	7.50%	6.25%	0.00%	82.50%

Test Dataset

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypno	Stimul
Placebo	83.33%	0.00%	4.17%	12.50%	0.00%
Antipsy	0.00%	37.50%	50.00%	0.00%	12.50%
Antidep	25.00%	25.00%	43.75%	6.25%	0.00%
Hypno	50.00%	8.33%	0.00%	41.67%	0.00%
Stimul	0.00%	20.83%	33.33%	0.00%	45.83%

certain class are retained here, such as Active Wake and REM Sleep for hypnotics and antidepressants, Active Wake for stimulants and Passive Wake for antipsychotics. The classification results for the training and the test dataset can be found in Table 8.12. Also here we get a high correct classification percentage for all five classes in the training dataset. The overall error rate on the compound-dose level in the training dataset is even 0.034. For the test dataset, the percentage of correctly classified rats in the placebo group is somewhat smaller than in the previous analyses, but this is amply compensated in the other classes, where we get now classification percentages in the correct class above 55% for antipsychotics, antidepressants and stimulants. The

Table 8.7: *Sleeping Stages used in each step of the MFLDA with modified distance measure, on the compound-by-dose combination level.*

Step	Sleeping Stages
Step 1: Stimulants	AW -PW - RS
Step 2: Antipsychotics	PW - SWS1 - SWS2 - IS - RS
Step 3: Placebo	PW - SWS1 - SWS2 - IS - RS
Step 4: Hypnotics	SWS2 - IS - PS

overall error rate on the compound-dose level in the test dataset is now 0.452.

To be able to compare the results in the four different analyses (with the modified and pairwise strategy on both rat and compound-dose level), we summarize the error rates obtained on the rat and the compound-dose level for the four procedures in Table 8.13. We can see here that for both levels of analysis, the procedure based on the pairwise strategy in general produces smaller error rates.

8.5 Concluding Remarks

In this chapter a novel extension of the functional linear discriminant analysis, called MFLDA, is introduced. Since a pseudo-likelihood model is used to model the multivariate longitudinal data, the estimated variance-covariance matrix is not guaranteed to be positive definite. To overcome this problem we followed two different strategies. On the one hand a correction, proposed by Rousseeuw and Molenberghs (1993), was applied to the variance-covariance matrix to ensure the positive definiteness of this matrix. On the other hand, the distance between two subjects was calculated for each pair of covariates in the dataset and the final distance between the subjects was defined as the mean of all pairwise distances. The performance of MFLDA with both approaches was evaluated on the EEG dataset and through simulations.

For the EEG dataset, a stepwise classification procedure was performed, where in each step one class is separated from the remaining classes using MFLDA. The selection of the class to be discriminated in each step and the selection of the sleeping stages to be used in each of the steps is done on the rat level and on the compound-by-dose combination level, leading to four different classifications. For all four analyses,

Table 8.8: *Classification percentages for the EEG training dataset and test dataset, with the modified distance measure, on the compound-dose level ($error_{rat}(train) = 0.306$, $error_{cd}(train) = 0.134$, $error_{rat}(test) = 0.471$, $error_{cd}(test) = 0.383$).*

Training Dataset

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypno	Stimul
Placebo	80.98%	13.59%	2.17%	2.72%	0.54%
Antipsy	5.36%	91.07%	0.00%	0.00%	3.57%
Antidep	10.71%	52.68%	31.25%	4.46%	0.89%
Hypno	22.50%	17.50%	0.00%	60.00%	0.00%
Stimul	5.00%	10.00%	1.25%	0.00%	83.75%

Test Dataset

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypno	Stimul
Placebo	87.50%	8.33%	0.00%	4.17%	0.00%
Antipsy	0.00%	87.50%	6.25%	0.00%	6.25%
Antidep	18.75%	56.25%	18.75%	3.12%	3.12%
Hypno	66.67%	8.33%	0.00%	25.00%	0.00%
Stimul	4.17%	33.33%	16.67%	0.00%	45.83%

the training dataset could be classified well, being slightly worse than with the DHSLA. The classification of the test dataset was even better than the one obtained with DHSLA (without model averaging). The MFLDA with the pairwise strategy leaded in general to a smaller error rate in training and test dataset compared to the MFLDA with the correction for positive definiteness.

The same was seen in the simulation studies with two or three classes. In the setting with four classes, MFLDA with the correction for positive definiteness produces a smaller error rate than MFLDA with the pairwise strategy. In this setting, the fourth class, based on hypnotics, and to lesser extent the third class, based on

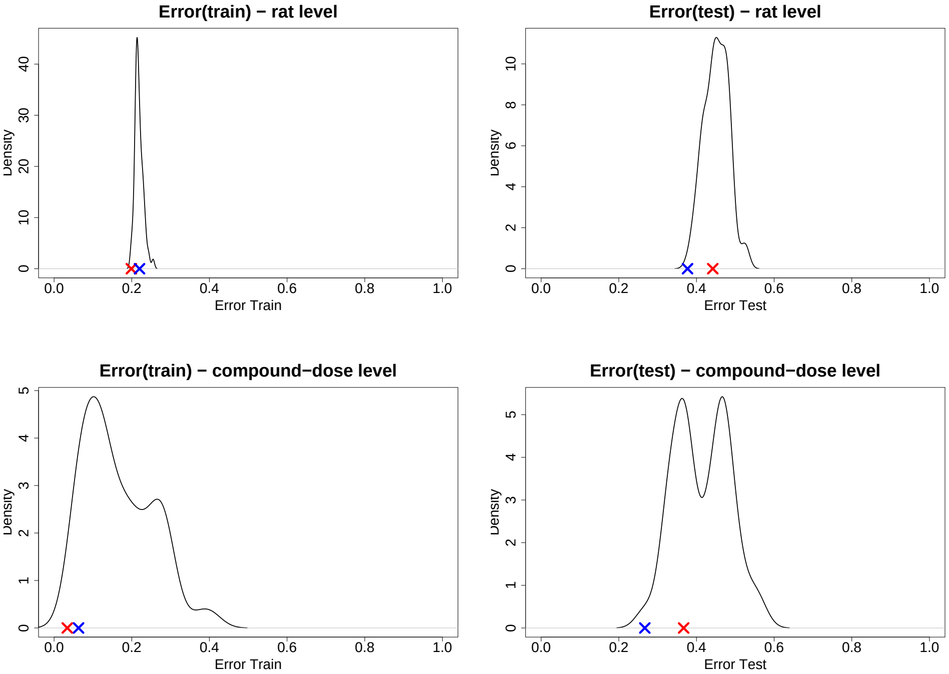


Figure 8.6: *Density of the error rates in training and test dataset obtained with MFLDA using pairwise distances.*

antidepressants, were hard to separate out. This lines up with expectation, since also in the original EEG dataset, these two classes are difficult to discriminate.

Increasing the size of the training dataset resulted in a decreasing error rate in the simulation studies with 2 or 3 classes. In the study with four classes, a higher sample size could not improve the classification results much, which enhances the impression that hypnotics and antidepressants are problematic to separate out.

The number of knots in the splines models had an influence on the classification results. For all three settings, MFLDA with three knots gave a better classification than the one with only two knots. This could be expected since the variance-covariance structure is modeled better when three knots are used. A higher number of knots or a different pseudo-likelihood model, e.g. based on fractional polynomials, was not incorporated in the simulation study, but can be considered as well. The number of knots, and in general the pseudo-likelihood model to be used in the MFLDA, should be carefully determined depending on the application at hand.

Table 8.9: *Sleeping Stages used in each step of the MFLDA with pairwise distance calculation, on the rat level.*

Step	Sleeping Stages
Step 1: Placebo	AW - SWS1 - SWS2 - RS
Step 2: Stimulants	AW - PW - SWS1 - RS
Step 3: Antidepressants	AW - SWS2 - IS - RS
Step 4: Antipsychotics	SWS1 - SWS2 - IS - RS

Table 8.10: *Classification percentages for the EEG training dataset and test dataset, with pairwise distance calculation, on the rat level (*$error_{rat}(train) = 0.199$*,* $error_{cd}(train) = 0.274$*,* $error_{rat}(test) = 0.442$*,* $error_{cd}(test) = 0.383$*).*

Training Dataset

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypno	Stimul
Placebo	85.87%	6.52%	7.07%	0.54%	0.00%
Antipsy	5.36%	89.29%	1.78%	1.78%	1.78%
Antidep	4.46%	23.21%	65.18%	6.25%	0.89%
Hypno	17.50%	2.50%	5.00%	75.00%	0.00%
Stimul	3.75%	5.00%	6.25%	0.00%	85.00%

Test Dataset

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypno	Stimul
Placebo	87.50%	4.17%	4.17%	4.17%	0.00%
Antipsy	0.00%	56.25%	43.75%	0.00%	6.25%
Antidep	25.00%	25.00%	43.75%	3.12%	3.12%
Hypno	41.67%	8.33%	8.33%	41.67%	0.00%
Stimul	4.17%	25.00%	20.83%	0.00%	50.00%

Table 8.11: *Sleeping Stages used in each step of the MFLDA with pairwise distance calculation, on the compound-dose level.*

Step	Sleeping Stages
Step 1: Hypnotics	AW - SWS1 - IS - RS
Step 2: Antidepressants	AW - SWS2 - RS
Step 3: Stimulants	AW - SWS2
Step 4: Antipsychotics	AW -PW - SWS1 - SWS2 - IS - PS

Table 8.12: *Classification percentages for the EEG training dataset and test dataset, with pairwise distance calculation, on the compound-dose level ($error_{rat}(train) = 0.236$, $error_{cd}(train) = 0.034$, $error_{rat}(test) = 0.452$, $error_{cd}(test) = 0.367$).*

Training Dataset

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypno	Stimul
Placebo	78.26%	11.41%	5.43%	3.26%	1.63%
Antipsy	3.57%	67.86%	19.64%	7.14%	1.79%
Antidep	7.14%	13.39%	66.96%	6.25%	6.25%
Hypno	12.50%	2.50%	2.50%	82.50%	0.00%
Stimul	2.50%	3.75%	7.50%	0.00%	86.25%

Test Dataset

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypno	Stimul
Placebo	66.67%	8.33%	4.17%	20.83%	0.00%
Antipsy	0.00%	56.25%	31.25%	0.00%	12.50%
Antidep	21.87%	15.63%	59.38%	3.12%	0.00%
Hypno	41.67%	0.00%	25.00%	33.33%	0.00%
Stimul	0.00%	12.50%	29.17%	0.00%	58.33%

Table 8.13: *Multivariate Functional Linear Discriminant Analysis. Summary table for the error rates obtained in train and validation dataset.*

			Train		Test	
Classification	Selection Procedure		$\text{error}_{\text{rat}}$	error_{cv}	$\text{error}_{\text{rat}}$	error_{cv}
PD Correction	I	Rat level	0.248	0.174	0.496	0.267
	II	Compound-dose level	0.306	0.134	0.471	0.383
Pairwise Distances	I	Rat level	0.199	0.274	0.442	0.383
	II	Compound-dose level	0.236	0.034	0.452	0.367

9

Concluding Remarks and Further Research

9.1 Concluding Remarks

In this thesis, we have focussed on the classification of multiple class, multivariate longitudinal data. This research was driven by a study conducted to classify psychotropic drugs based on electro-encephalogram or EEG data. For each of the compound-by-dose combinations in the five psychotropic drug classes, data on the sleep-wake behaviour of rats were collected during a 16 hours period. The sleep-wake behaviour was summarized into six standard sleep-wake stages, resulting in a six-variate longitudinal profile per rat.

From a statistical point of view, analyzing EEG data poses important challenge, because of the high-dimensionality and the longitudinal character of the data. The longitudinal profiles are usually highly irregular and the variability between and within subjects are relatively high.

9.1.1 Exploratory Tools

For the visualization of multiple class multivariate longitudinal data, we proposed a graphical tool to explore characteristics of classes in the data at hand, using the so-called mutual information measure. Rather than the actual prediction of the class of a new compound, we focus on how well separated the classes are in a particular set of data, to be able to explain difficulties in the classification procedure.

The mutual information measure quantifies the amount of information a new observation has in common with each of the classes in the dataset. In this way, the classes can be visualized in a simple plot, showing the densities of the mutual information measures, for the class of interest against the remainder of the classes. With these density plots, the level of overlap between one class and all other classes can be measured.

The overlapping quantiles, together with the specifities and sensitivities revealed that classifying EEG data is a difficult task, indicating the need for elaborate classification techniques, which take into account the longitudinal nature of the data and also possible association between the six-variate longitudinal variables.

9.1.2 Doubly Hierarchical Supervised Learning Analysis

In this thesis, we proposed a general and simple procedure that can be applied to establish classification rules for application with multiple class longitudinal data, called doubly hierarchical supervised learning analysis (DHSLA). This flexible procedure takes into account the specific nature of the multiple drug classes, as well as the longitudinal character of the data. The method consists of two stages, in stage one the longitudinal profiles are modelled using a flexible modeling technique to account for the irregularities in the profiles, then a summary extracted from this model is used in stage two in a stepwise classification process.

Several variations to this procedure were applied to the EEG data. In stage one, a fractional polynomial mixed model was always used, while in stage two, linear, flexible and mixture discriminant analysis were incorporated. The three different techniques were found to produce comparable results in the training dataset with respect to adjusted posterior probabilities and error rates. In the test dataset, DHSLA with flexible and mixture discriminant analysis were performing slightly better than DHSLA with linear discriminant analysis. Especially placebo, antidepressants and hypnotic were poorly discriminated.

Concerns regarding model selection bias arose when applying the DHSLA procedure. Therefore an extension of model averaging to the case of linear discriminant analysis was developed. This novel approach was then integrated in the second stage of the DHSLA in order to get more robust classification results. This leaded to a considerable improvement of the classification in the train as well as the test dataset, but still antidepressant and hypnotic were difficult to separate.

For the selection of the variables to be used in each step of the stepwise classification procedure, we concentrated on two levels, the rat and the compound-dose level. Both approaches must be seen next to each other, since they both shed a different light on the classification problem at hand. When the interest of the researcher is in the classification of rats rather than treatments, the rat level should be the level to focus on and vice versa.

9.1.3 Multivariate Functional Linear Discriminant Analysis

Although in the DHSLA the longitudinal character of the data is taken into account, the multivariate aspect is still ignored, thus we are implicitly assuming that each longitudinal variable is independent of all other ones. In order to deal with both aspects of the data we extended the functional linear discriminant analysis of James and Hastie (2001) for the case where several longitudinal profiles are recorded for the same individual. When dealing with multivariate longitudinal data, the correlation between the variables must be taken into account in our classification. A fully multivariate model would have been the natural choice, but given the complexity of the data, computational issues are commonly present during such modeling exercise. Therefore, we proposed to use a pseudo-likelihood modeling approach (Fieuws and Verbeke, 2006) combined with smoothing techniques such as splines, to model the multivariate longitudinal characteristics.

While the computational issues are overcome by this modeling approach, others arise, since the fitted covariance matrix obtained from the modeling approach is not always positive definite. To solve this issue, we proposed to use a modification of the variance covariance matrix which is positive definite or to calculate the average of the distance between a new observation and the classes in the training dataset for each pair of variables. Both strategies were followed and compared.

The performance of the MFLDA is established through application to the EEG dataset as well as through simulation studies, using different number of classes and observations. The classification is evaluated based on the error rate using a train-test setting.

It is also important to highlight that when we compare DHSLA, without considering model averaging, and MFLDA, the latter performs slightly worse in the training dataset, but better in the test dataset. The MFLDA with pairwise strategy was in general performing better than MFLDA with the correction for positive definiteness counterpart. Also here, hypnotics and antidepressants poses difficulties to differentiate observations from both classes.

The simulation studies reveal that the number of subjects in the training dataset and the model used to fit the longitudinal profiles are crucial. When more subjects are included in the training dataset, and when the variance-covariance structure in the training dataset is properly estimated, better classification results are be obtained.

9.2 Further Research

The research presented in this thesis allows to indicate several interesting topics for further investigation in the classification of multiple-class multivariate longitudinal data.

A first issue raised by clinical experts, is the need for a quick and simple method for the screening of potentially new variables that may improve the classification. A possible solution for this could be the mutual information measure, which can be used to determine the discriminative property for each new variable.

A second important issue is the methods to be used in each stage of the DHSLA. In this thesis, we restricted to a fractional polynomial mixed model in stage 1 and a linear, flexible and mixture discriminant analysis in stage 2. In both stages other techniques can be considered as well. For instance a splines mixed model or other flexible modeling technique can be used in stage 1 and different supervised learning methods can be applied in stage 2, such as non-parametric discriminant analysis, support vector machines, neural networks, random forests, ... The performance of these methods can be further evaluated through simulation studies.

Also in MFLDA, the use of different discriminant techniques instead of linear discriminant analysis can be investigated. A few potential choices are flexible, mixture and penalized discriminant analysis, but also other techniques can be investigated as well.

A final issue is the discrimination of hypnotics and antidepressants in the EEG dataset. In general all methods have some difficulties to classify these two classes. While at first sight this is a drawback, we should approach such a conclusion with due caution. First, it is conceivable that a given component at a certain dose has more than one modalities of activity. Second, the very classification into psychotropic classes, while generally used, remains arbitrary and should perhaps be called into question. At least, a revision might be in place.

References

Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews*, **1**, 3–22.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd International Symposium on Information Theory*, 267–281, Budapest.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders*, 4th edition. Washington, DC: American Psychiatric Association.

Barnard, G. A. (1963). New methods of quality control. *Journal of the Royal Statistical Society, Series A*, **126**, 255–258.

Bennett, C.H., and Shor, P.W. (1998). Quantum information theory. *IEEE Transactions on Information Theory*, **44**, 2724–2742.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, **20**, 451-468.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1), 5 – 32.

Breslow, N.E., Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association,**88**, 9–25.

Burnham, K.P. and Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*, 2nd edition. New York: Springer.

Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series A*, **158**, 419-466.

Cohen, D. and Cailloux-Cohen, S. (1995). *Guide critique des médicaments de l'âme.* Québec: Les Editions de l'Homme.

Dempster, A.P., Laird, N.M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Deniker, P. (1982). Vers une classification automatique des psychotropes à travers un fichier informatisé de leurs propriétés. *Annales Médico-psychologiques*, **1**, 25–27.

Depoortere, H., Francon, D., Van Luijtelaar, E.L.J.M., Drinkenburg, W.H.I.M. and Coenen, A.M.L. (1995). Differential effects of midazolam and zolpidem on sleep-wake states and epileptic activity in WAG/Rij rats. *Pharamacology, Biochemistry and Behavior*, **51**, 571–576.

Dierckx, P. (1993) *Curve and surface fitting with splines.* Oxford: Clarendon Press.

Diggle, P.J., Heagerty, P., Liang, K-Y., and Zeger, S.L. (2002). *Analysis of longitudinal data.* New York: Oxford University Press.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, **57**, 75–97.

Drinkenburg, W.H.I.M. and Ahnaou, A. (2004). The use of pEEG in preclinical models in drug discovery. *Essentials and Applications of EEG Research in Preclinical and Clinical Pharmacology.*, Berlin, International Pharmaco-EEG Group, 131–148.

Edgar, D.M. (2002). Signature profiles in sleep-wake drug discovery. *Methods and Findings in Experimental and Clinical Pharmacology*, **24** (Suppl. D), 71–72.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **9,** 1218-1228.

Faes, C., Aerts, M., Geys, H., Molenberghs, G. (2007). Model averaging using fractional polynomials to estimate a safe level of exposure. *Risk Analysis*, **27**(1), 111-123.

Fieuws, S., Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, **62**, 424–431.

Fink, M. (1959). EEG and behavioural effects of psychopharmacology agents. *Neuropsychopharmacology*, **1**, 441–446.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenic (London)*, **7**, 179–188.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.

George, E.I. (1998). Bayesian model selection. *Encyclopedia of Statistical Sciences, Update Volume 3, (eds. S. Kotz, C. Read and D. Banks)*, New York: Wiley, 39–46.

Hansen, B.E. (2007). Least squares model averaging. *Econometrica*, **75**(4), 1175–1189.

Hastie, T. J., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, **89**, 1255–1270.

Hastie, T.J., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics*, **23**, 73–102.

Hastie, T. J. and Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society, Series B*, **58**, 158–176.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction.* New York: Springer.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*, 2nd edition. Upper Saddle River, N.J.: Prentice Hall.

Hjort N.L., Claeskens G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, **98**, 879-899.

Hoeting, J. A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**, 382–417.

James, G.M., Hastie, T.J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society. Series B*, **63**(3), 533–550.

Johnson, R.A. and Wichern, D.W. (1992). *Applied multivariate statistical analysis*, 3rd edition. Englewood Cliffs: Prentice-Hall.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Kraskov, A., Stögbauer, H. and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, **69**, 66–138.

Krijzer, F., Koopman, P. and Olivier, B. (1993). Classification of psychotropic drugs based on pharmaco-electrocorticographic studies in vigilance-controlled rats. *Neuropsychobiology*, **28**(3), 122–137.

Kullback, S. (1959). *Information Theory and Statistics.* New York: Wiley.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Mucci, A., Volpe, U., Merlotti, E., Bucci, P., Galderise, S. (2006). Pharmaco-EEG in psychiatry. *Clinical EEG and Neuroscience*, **37**, 81–98.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data.* New York: Springer.

Murck, H., Nickel, T., Künzel, H., Antonijevic, I.A., Schill, J., Zobel, A., Steiger, A., Sonntag, A. and Holsboer, F. (2003). State markers of depression in sleep EEG: Dependency on drug and gender in patients treated with tianeptine or paroxetine. *Neuropsychopharmacology*, **28**, 348–358.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. (1996). *Applied linear statistical models*, 4th edition. Chicago: Irwin.

Oughourlian, J. M. (1984). *La personne du toxicomane. Psychosociologie des toxicomanies actuelles dans la jeunesse occidentale.* Toulouse: Privat.

Rousseeuw, P.J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics: Theory and Methods*, **22**(4), 965–984.

Royston, P, Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with Discussion). *Applied Statistics*, **43**, 429–467.

Ruigt, G.S., Engelen, S., Gerrits, A., and Verbon, F. (1993). Computer-based prediction of psychotropic drug classes based on a discriminant analysis of drug effects on rat sleep. *Neuropsychobiology*, **28**, 138–153.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression.* New York: Cambridge University Press.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423 & 623-656.

Staner, L.P.J., Luthringer, R., and Macher, J-P. (2004). Antidepressant induced alteration of sleep EEG as a surrogate marker of drug activity: A window to the neurobiology of depression. *The World Journal of Biological Psychiatry*, **5**, Supp. 1, 8034.

Uchida, M., Suzuki, M., Shimizu, K. (2007). Effects of urocortin, corticotropin-releasing factor (CRF) receptor agonist, and astressin, CRF receptor antagonist, on the sleep-wake pattern: Analysis by radiotelemetry in conscious rats. *Biological and Pharmaceutical Bulletin*, **30**(10), 1895–1897.

Vapnik, V. (1998). *Statistical learning theory.* New York: Wiley.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

Verbyla, A.P., Cullis, B.R., Kenward, M.G., and Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics*, **48**, 269–311.

Williams,C.J. and Christian, J.C. (2006). Frequentist model-averaged estimators and tests for univariate twin models. *Behavior Genetics*, **37**, 687–696.

Wouters, K., Ahnaou, A., Cortiñas, J., Molenberghs, G., Geys, H., Bijnens, L., and Drinkenbrug, W.H.I.M. (2007a). Psychotropic drug classification based on sleep-wake behaviour of rats. *Journal of the Royal Statistical Society, Series C*, **56**(2), 223–234.

Wouters, K., Cortinas, J., Molenberghs, G., Ahnaou, A., Drinkenburg, W.H.I.M and Bijnens, L (2007b). A comparison of doubly hierarchical discriminant analyses for multiple class longitudinal data from EEG experiments. *Journal of Biopharmaceutical Statistics*, **18**, 000–000.

Wouters, K., Cortinas, J., Molenberghs, G., Geys, H., Ahnaou, A., Drinkenburg, W.H.I.M. and Bijnens, L. (2008a). Correction for model selection bias using a modified model averaging approach for supervised learning methods applied to EEG experiments. *Manuscript submitted for publication.*

Wouters, K., Cortinas, J., Molenberghs, G., Geys, H., Ahnaou, A., Drinkenburg, W.H.I.M. and Bijnens, L. (2008b). Mutual information as a tool to visualize classes in EEG data. *Manuscript submitted for publication.*

Wouters, K., Cortinas, J., Molenberghs, G., Geys, H., Ahnaou, A., Drinkenburg, W.H.I.M. and Bijnens, L. (2008c). Multivariate Functional Linear Discriminant Analysis Based on Pairwise Pseudo-Likelihood Modeling Combined with Splines. *Manuscript submitted for publication.*

Zarifian, E. (1988). *Les jardiniers de la folie.* Paris: Odile Jacob.

Zarifian, E. (1996). *Le prix du bien-être. Psychotropes et société.* Paris: Odile Jacob.

A

Validation Data Set – Fractional Polynomial Mixed Model

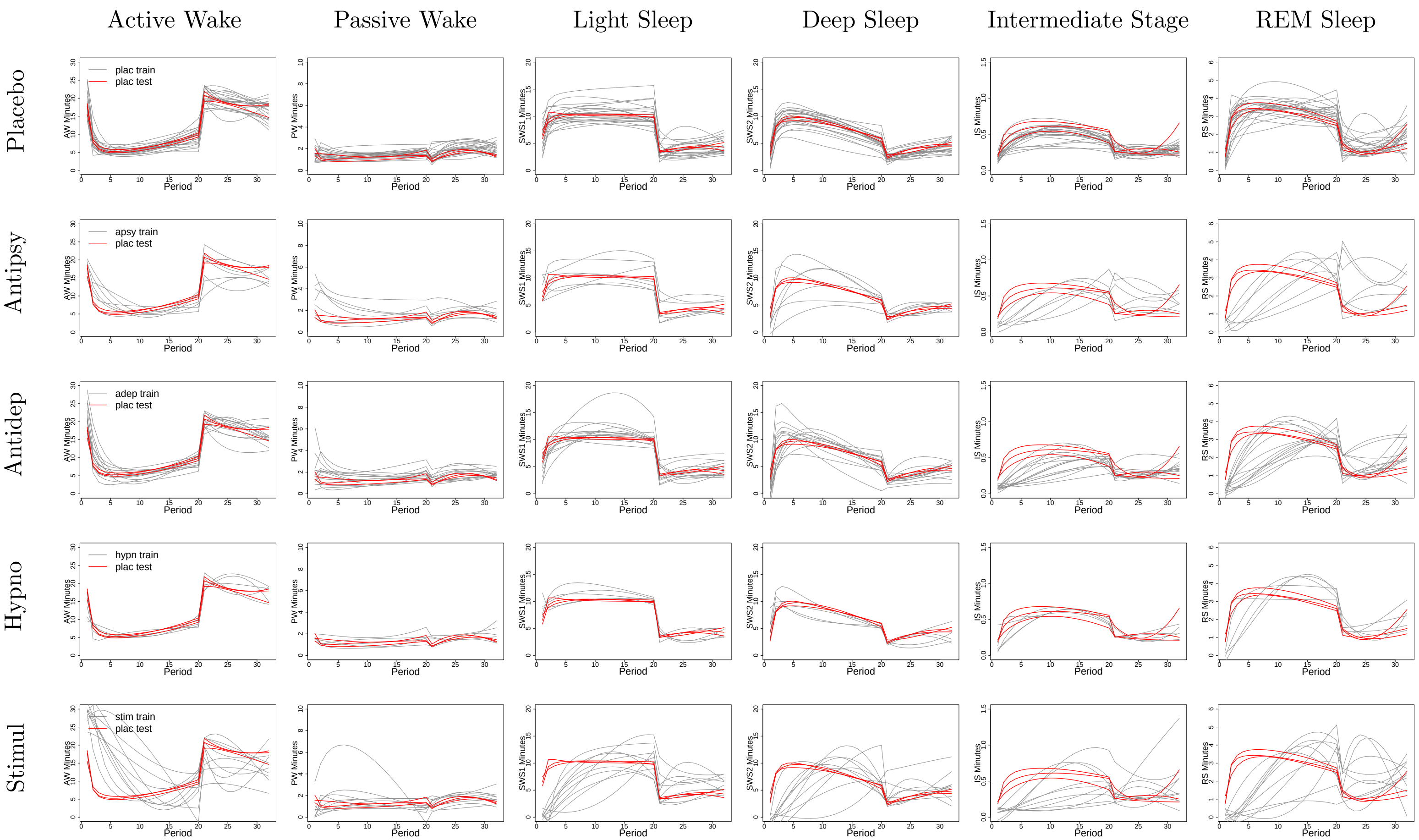


Figure A.1: *Fitted fractional polynomial mixed model for all placebo the compound-dose combinations in the validation dataset, together with the fitted fractional polynomial mixed models in the five drug classes in the training dataset.*

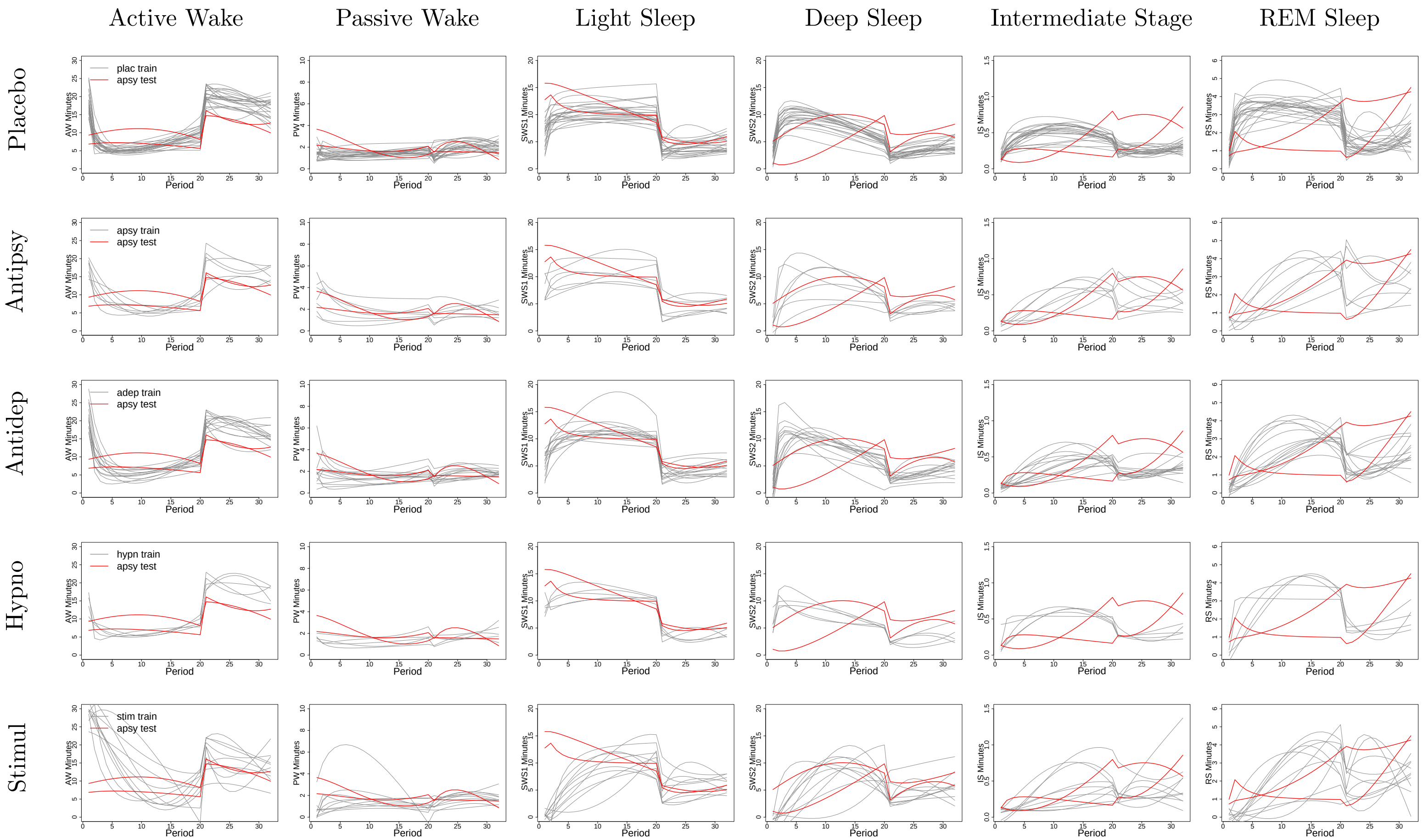


Figure A.2: *Fitted fractional polynomial mixed model for all placebo the compound-dose combinations in the validation dataset, together with the fitted fractional polynomial mixed models in the five drug classes in the training dataset.*

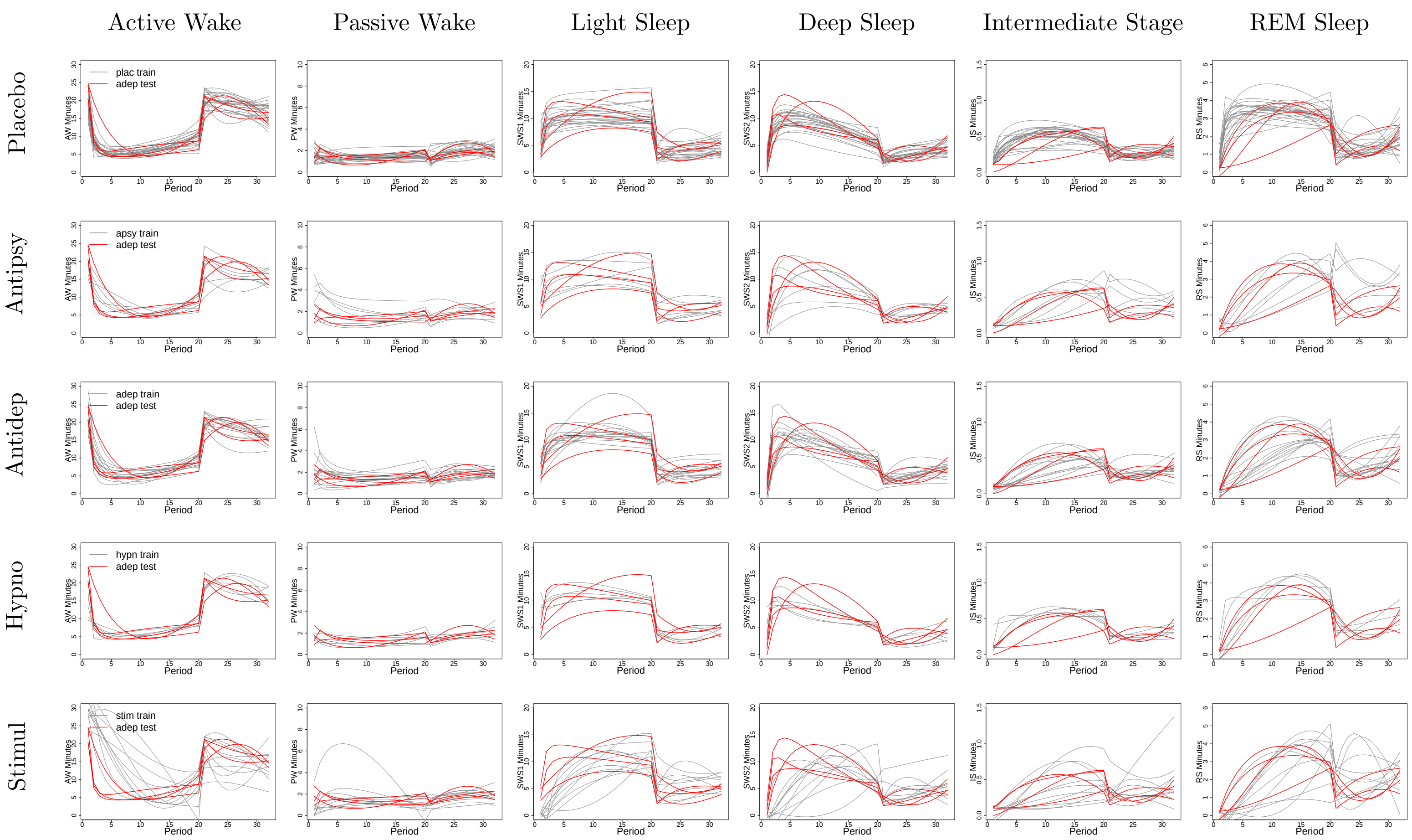


Figure A.3: *Fitted fractional polynomial mixed model for all placebo the compound-dose combinations in the validation dataset, together with the fitted fractional polynomial mixed models in the five drug classes in the training dataset.*

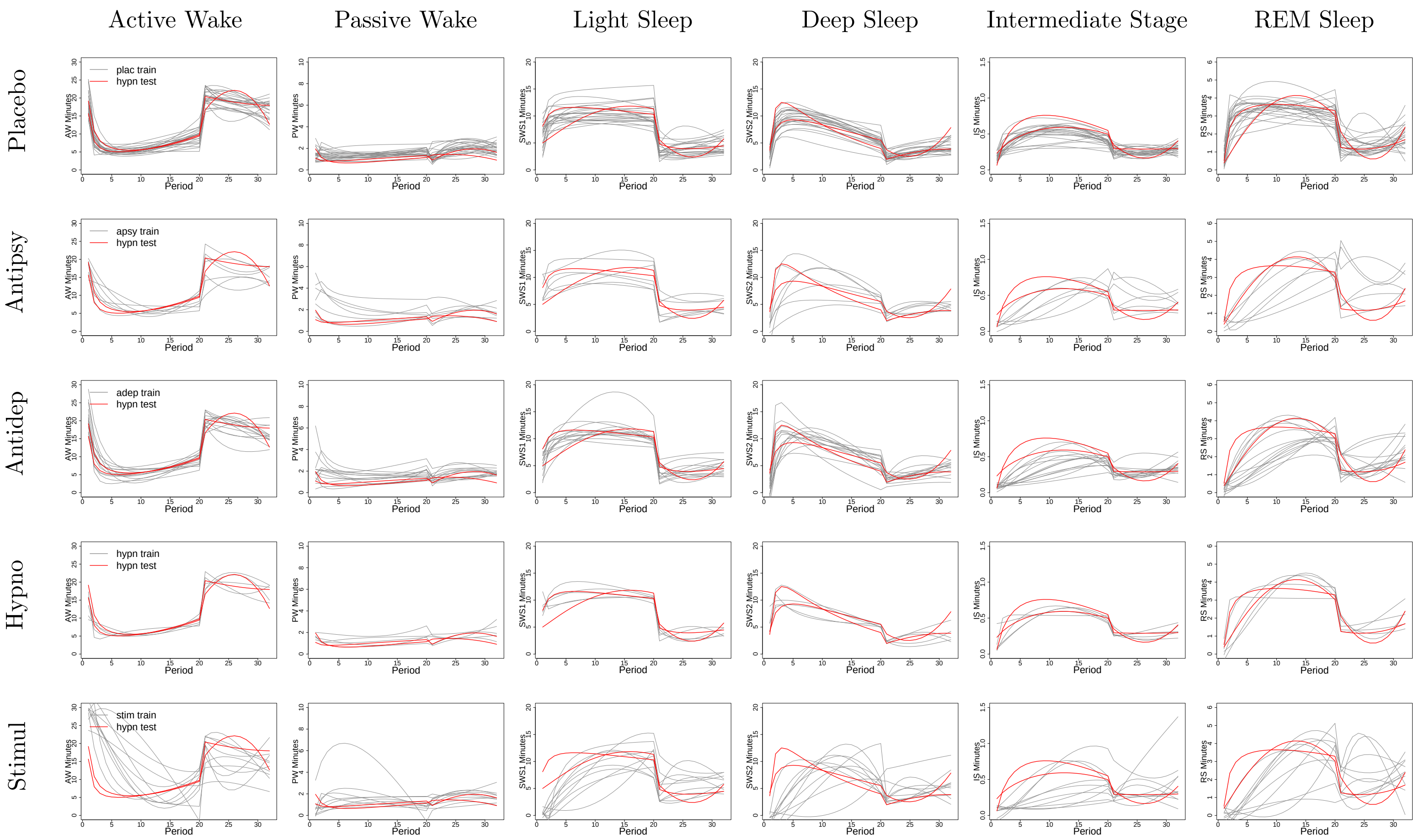


Figure A.4: *Fitted fractional polynomial mixed model for all placebo the compound-dose combinations in the validation dataset, together with the fitted fractional polynomial mixed models in the five drug classes in the training dataset.*

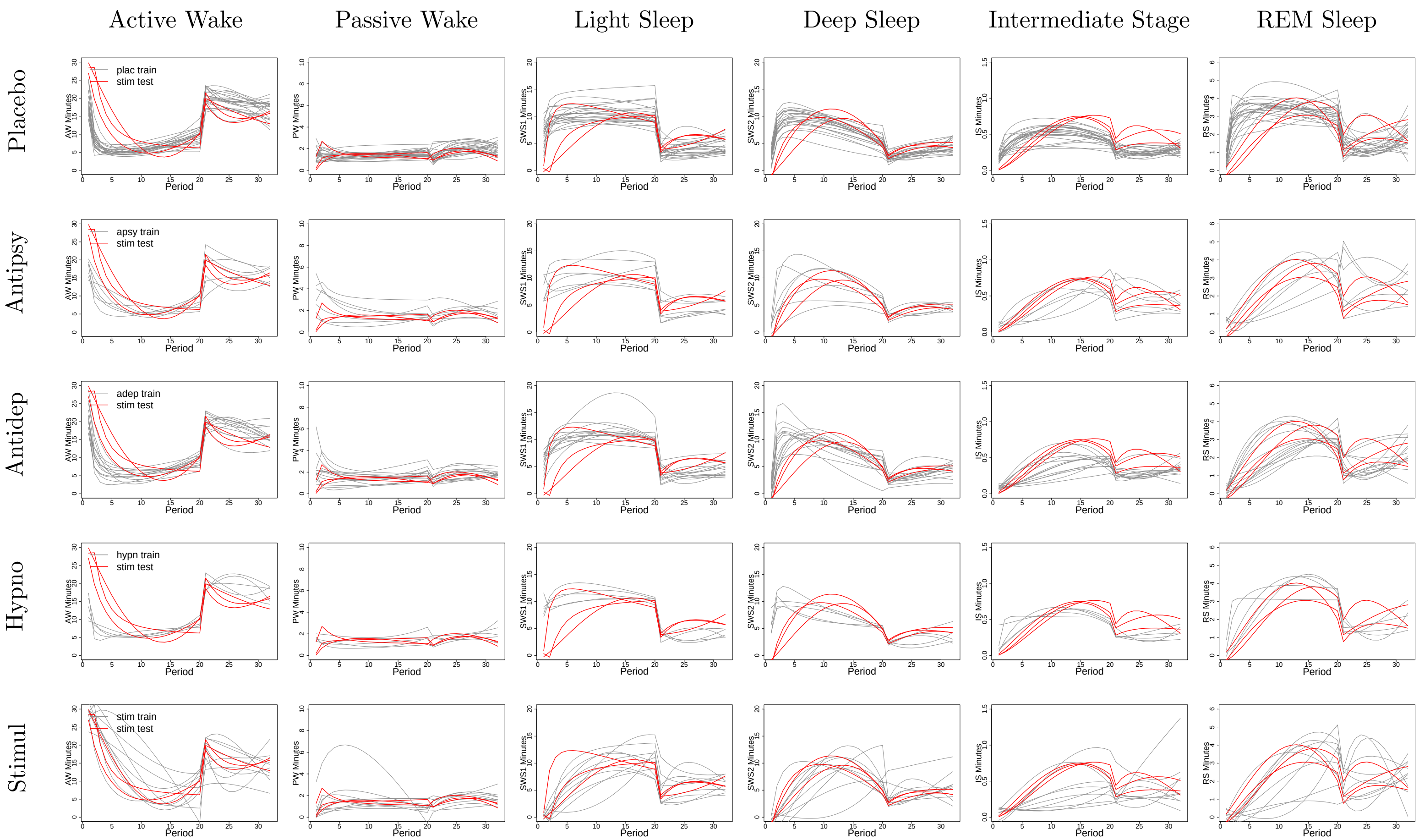


Figure A.5: *Fitted fractional polynomial mixed model for all placebo the compound-dose combinations in the validation dataset, together with the fitted fractional polynomial mixed models in the five drug classes in the training dataset.*

B

Model Average: Results for the Training Dataset

B.1 Selection Procedure I

Table B.1: *Model Average. Adjusted posterior probabilities for the training data set obtained with model averaging with the 25 best models when selection procedure I is used.*

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.99**	0.01	0.00	0.00	0.00
Antipsy	0.00	**0.97**	0.01	0.00	0.02
Antidep	0.00	0.02	**0.98**	0.00	0.00
Hypnot	0.00	0.00	0.00	**1.00**	0.00
Stimul	0.01	0.01	0.00	0.00	**0.99**

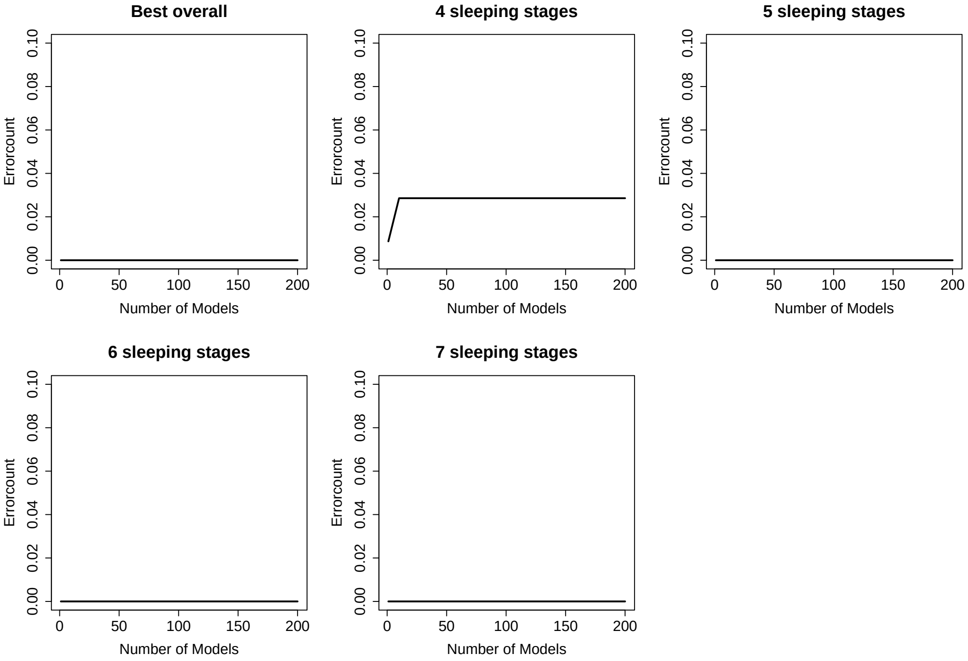


Figure B.1: *Model Averaging. Error rates in the test dataset obtained with modelaveraging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure I.*

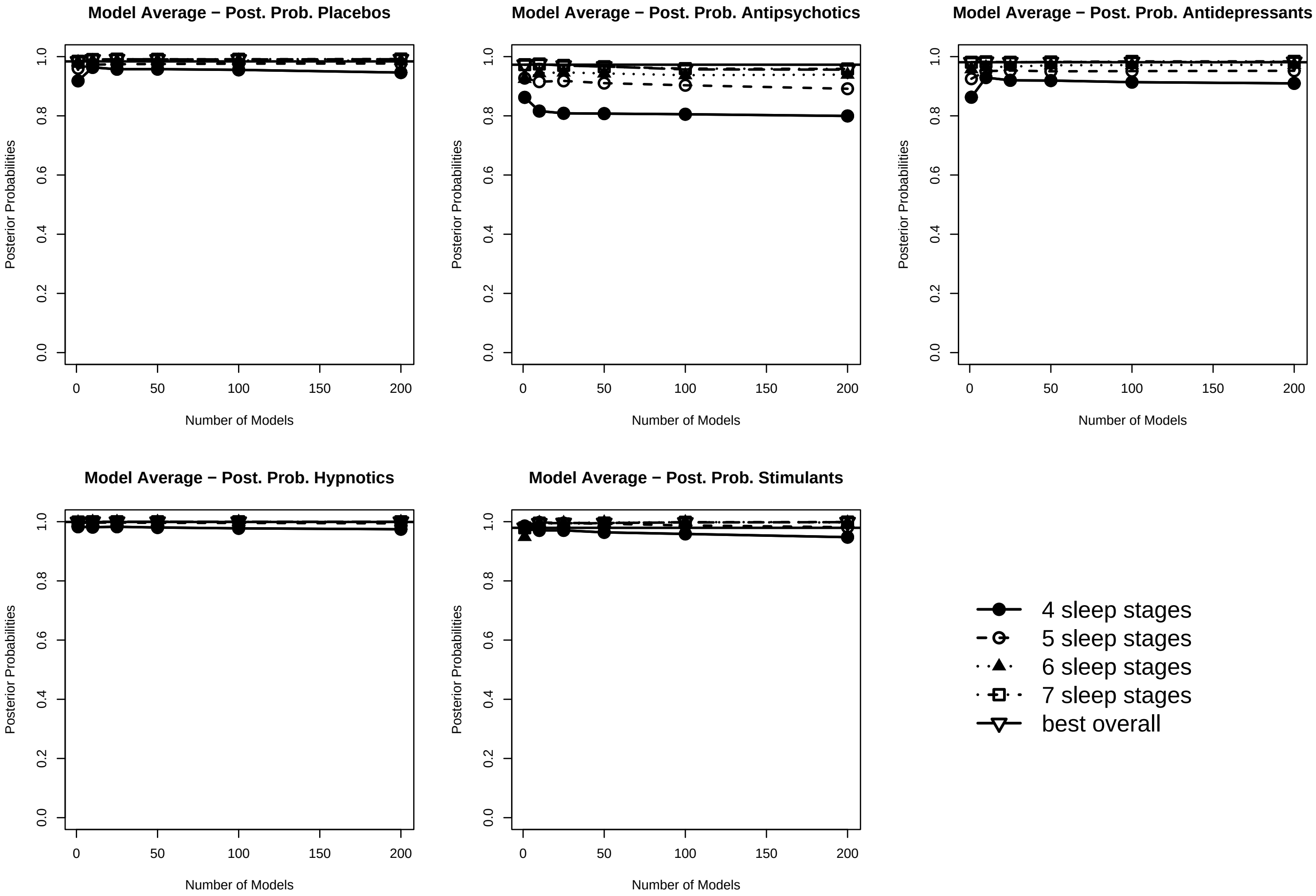


Figure B.2: *Model Averaging. Adjusted posterior probabilities in the test dataset obtained with modelaveraging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure I.*

B.2 Selection Procedure II

Table B.2: *Model Average. Adjusted posterior probabilities for the training data set obtained with model averaging with the 25 best models when selection procedure II is used.*

	Predicted Class				
Class	Placebo	Antipsy	Antidep	Hypnot	Stimul
Placebo	**0.99**	0.01	0.00	0.00	0.00
Antipsy	0.00	**0.94**	0.01	0.05	0.00
Antidep	0.00	0.02	**0.97**	0.01	0.00
Hypnot	0.00	0.00	0.00	**1.00**	0.00
Stimul	0.01	0.01	0.04	0.01	**0.93**

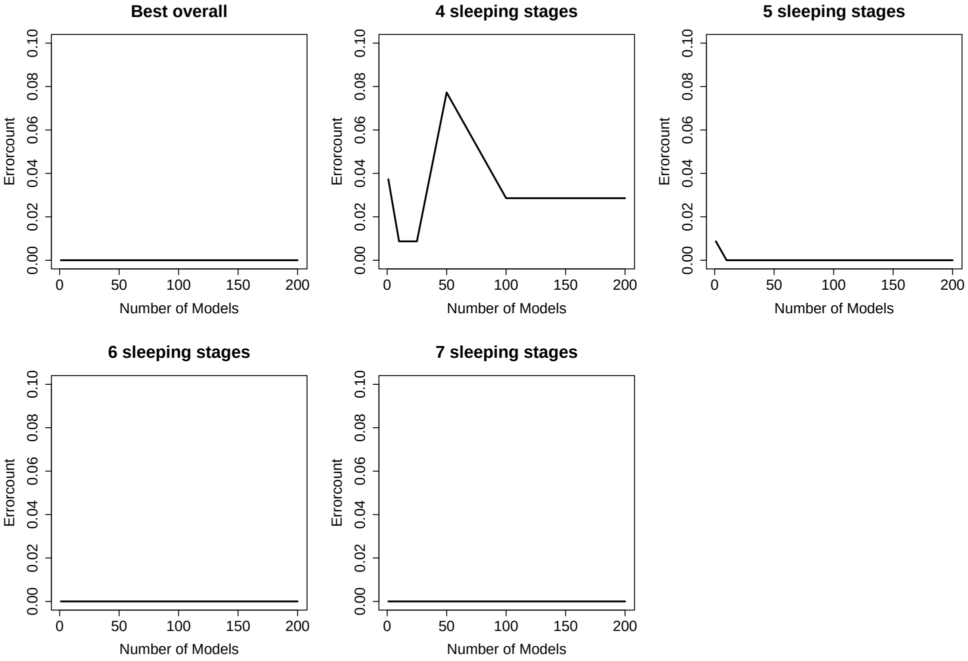


Figure B.3: *Model Averaging. Error rates in the training dataset obtained with modelaveraging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure II.*

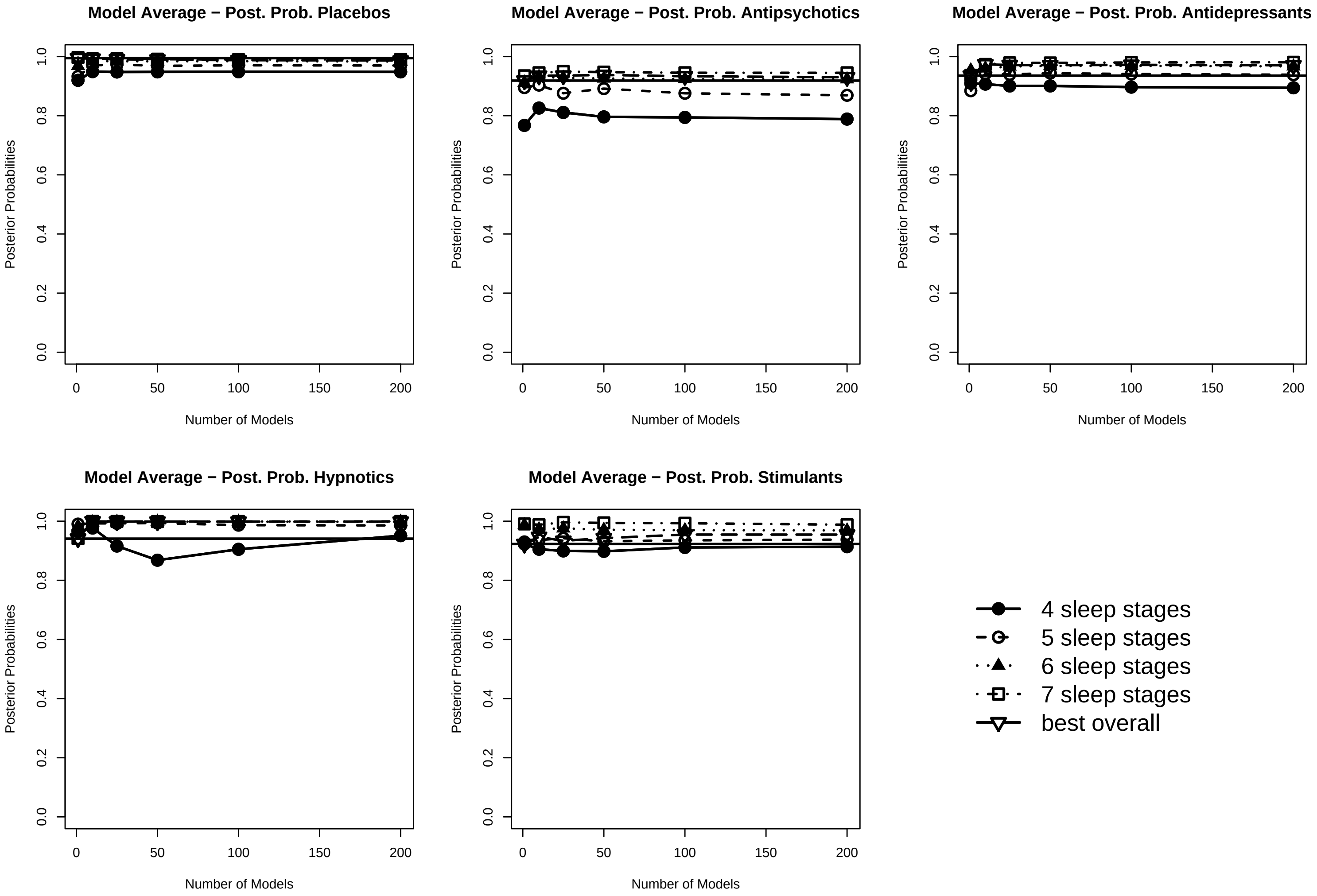


Figure B.4: *Model Averaging. Adjusted posterior probabilities in the training dataset obtained with modelaveraging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure II.*

C

Multivariate Functional Linear Discriminant Analysis – Simulation Study 2 Classes

Table C.1: *Simulation study based on two classes: Error rates and corresponding standard deviations (between parentheses) obtained with MFLDA with two knots using the correction for positive definiteness or the pairwise distances.*

Placebo vs Antipsychotic

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.434 (0.293)	0.173 (0.189)	0.834 (0.176)	0.000 (0.000)
40	0.560 (0.304)	0.123 (0.196)	0.931 (0.112)	0.000 (0.000)
60	0.710 (0.263)	0.038 (0.105)	0.951 (0.088)	0.000 (0.000)
80	0.631 (0.296)	0.046 (0.134)	0.951 (0.076)	0.000 (0.000)
100	0.679 (0.284)	0.045 (0.131)	0.959 (0.085)	0.000 (0.000)

Table C.2: *Simulation study based on two classes (continuation): Error rates and corresponding standard deviations (between parentheses) obtained with MFLDA with two knots using the correction for positive definiteness or the pairwise distances.*

Placebo vs Antidepressant

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.800 (0.201)	0.110 (0.168)	0.966 (0.076)	0.011 (0.040)
40	0.874 (0.183)	0.051 (0.153)	0.995 (0.022)	0.000 (0.000)
60	0.938 (0.097)	0.033 (0.074)	0.998 (0.014)	0.000 (0.000)
80	0.919 (0.154)	0.043 (0.142)	0.998 (0.014)	0.000 (0.000)
100	0.973 (0.060)	0.003 (0.017)	0.999 (0.010)	0.000 (0.000)

Placebo vs Hypnotic

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.150 (0.198)	0.201 (0.208)	0.106 (0.154)	0.295 (0.249)
40	0.095 (0.116)	0.173 (0.153)	0.081 (0.112)	0.253 (0.195)
60	0.059 (0.091)	0.160 (0.141)	0.049 (0.076)	0.232 (0.173)
80	0.072 (0.095)	0.167 (0.136)	0.062 (0.087)	0.252 (0.177)
100	0.047 (0.072)	0.147 (0.134)	0.036 (0.063)	0.235 (0.073)

Placebo vs Stimulant

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.005 (0.026)	0.004 (0.040)	0.013 (0.037)	0.000 (0.000)
40	0.006 (0.024)	0.000 (0.000)	0.013 (0.037)	0.000 (0.000)
60	0.008 (0.027)	0.000 (0.000)	0.014 (0.035)	0.000 (0.000)
80	0.005 (0.022)	0.000 (0.000)	0.009 (0.029)	0.000 (0.000)
100	0.009 (0.029)	0.000 (0.000)	0.019 (0.048)	0.000 (0.000)

Table C.3: *Simulation study based on two classes (continuation): Error rates and corresponding standard deviations (between parentheses) obtained with MFLDA with two knots using the correction for positive definiteness or the pairwise distances.*

Antipsychotic vs Antidepressant

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.300 (0.262)	0.267 (0.245)	0.145 (0.209)	0.348 (0.276)
40	0.235 (0.252)	0.281 (0.238)	0.108 (0.158)	0.345 (0.251)
60	0.142 (0.184)	0.275 (0.237)	0.070 (0.098)	0.335 (0.230)
80	0.170 (0.227)	0.250 (0.207)	0.054 (0.099)	0.314 (0.185)
100	0.096 (0.145)	0.289 (0.214)	0.056 (0.099)	0.306 (0.187)

Antipsychotic vs Hypnotic

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.153 (0.206)	0.136 (0.192)	0.000 (0.000)	0.438 (0.266)
40	0.055 (0.120)	0.170 (0.165)	0.000 (0.000)	0.484 (0.256)
60	0.035 (0.098)	0.208 (0.193)	0.000 (0.000)	0.529 (0.227)
80	0.046 (0.119)	0.223 (0.201)	0.000 (0.000)	0.547 (0.216)
100	0.015 (0.063)	0.229 (0.161)	0.000 (0.000)	0.512 (0.195)

Antipsychotic vs Stimulant

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.112 (0.163)	0.018 (0.066)	0.018 (0.041)	0.008 (0.037)
40	0.120 (0.157)	0.001 (0.010)	0.020 (0.049)	0.004 (0.020)
60	0.041 (0.074)	0.001 (0.010)	0.020 (0.045)	0.002 (0.014)
80	0.057 (0.114)	0.002 (0.014)	0.014 (0.038)	0.002 (0.014)
100	0.032 (0.063)	0.000 (0.000)	0.014 (0.045)	0.000 (0.000)

Table C.4: *Simulation study based on two classes (continuation): Error rates and corresponding standard deviations (between parentheses) obtained with MFLDA with two knots using the correction for positive definiteness or the pairwise distances.*

Antidepressant vs Hypnotic

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.049 (0.107)	0.527 (0.243)	0.003 (0.022)	0.861 (0.174)
40	0.031 (0.080)	0.580 (0.255)	0.001 (0.010)	0.901 (0.131)
60	0.026 (0.081)	0.638 (0.239)	0.000 (0.000)	0.928 (0.103)
80	0.011 (0.037)	0.631 (0.237)	0.000 (0.000)	0.932 (0.103)
100	0.009 (0.029)	0.632 (0.211)	0.000 (0.000)	0.929 (0.107)

Antidepressant vs Stimulant

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.020 (0.071)	0.004 (0.024)	0.001 (0.010)	0.001 (0.010)
40	0.028 (0.113)	0.000 (0.000)	0.001 (0.010)	0.000 (0.000)
60	0.004 (0.024)	0.000 (0.000)	0.005 (0.022)	0.000 (0.000)
80	0.006 (0.031)	0.000 (0.000)	0.001 (0.010)	0.000 (0.000)
100	0.008 (0.052)	0.000 (0.000)	0.001 (0.010)	0.000 (0.000)

Antipsychotic vs Stimulant

	Correction PD		Pairwise Distance	
N	Class 1	Class 2	Class 1	Class 2
20	0.005 (0.022)	0.001 (0.010)	0.011 (0.034)	0.000 (0.000)
40	0.006 (0.028)	0.000 (0.000)	0.002 (0.014)	0.000 (0.000)
60	0.001 (0.010)	0.000 (0.000)	0.002 (0.014)	0.000 (0.000)
80	0.001 (0.010)	0.000 (0.000)	0.002 (0.014)	0.000 (0.000)
100	0.001 (0.010)	0.000 (0.000)	0.002 (0.014)	0.000 (0.000)