Correction for Model Selection Bias Using a Modified Model Averaging Approach for Supervised Learning Methods Applied to EEG Experiments

Peer-reviewed author version

WOUTERS, Kristien; CORTINAS ABRAHANTES, Jose; MOLENBERGHS, Geert; GEYS, Helena; BIJNENS, Luc; Ahnaou, Abdellah & Drinkenburg, W.H.I.M. (2010)
Correction for Model Selection Bias Using a Modified Model Averaging Approach for Supervised Learning Methods Applied to EEG Experiments. In: JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 20 (4). p. 768-786.

DOI: 10.1080/10543401003618744
Handle: http://hdl.handle.net/1942/11001

Correction for Model Selection Bias Using a Modified Model Averaging Approach for Supervised Learning Methods Applied to EEG Experiments

Kristien Wouters[1], José Cortiñas Abrahantes[1], Geert Molenberghs[1],
Helena Geys[1,2], Abdellah Ahnaou[2], Wilhelmus H.I.M. Drinkenburg[2], and Luc Bijnens[2]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Diepenbeek, Belgium

[2] Johnson & Johnson Pharmaceutical Research and Development, Beerse, Belgium

Abstract

This paper proposes a modified model averaging approach for linear discriminant analysis. This approach is used in combination with a doubly hierarchical supervised learning analysis and applied to preclinical pharmaco-electroencephalographical data for classification of psychotropic drugs. Classification of a test dataset was highly improved with this method.

Keywords

EEG, fractional polynomials, linear discriminant analysis, linear mixed model, model average, supervised learning.

1 Introduction

Discriminant analysis is a well-known procedure which dates back to the first half of the last century (Fisher, 1936). Since then, several procedures have been proposed and 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 just a few examples to mention. Nowadays, data-mining procedures such as random forests (Breiman, 2001), neural networks (Haykin, 1999) and support vector machines (Vapnik, 1998) are gaining popularity in the supervised learning field and their good performances have been shown in several applications.

In cases with complex data structures, such as multiple-class problems within a multivariate longitudinal design (in the sense of several longitudinal profiles recorded for the same individual), developing clas-

sification rules is not a trivial task and tailored methods are required to cope with these requirements. The motivation for our research is found in pharmaco-electroencephalographic (pEEG) experiments, conducted to establish classification rules for psychotropic drug classes. For this purpose, six sleep-wake stages are monitored in rats during a total period of 16 hours, for each rat, which will be used to establish classification rules in a multiple class problem. As rats are nocturnal rodents it should be noticed that the recording period included 10 hours under lights-on conditions and 6 hours under lights-off conditions. Classical supervised learning analysis is not suited to handle the combination of these features. A flexible two-step procedure called doubly hierarchical discriminant analysis (Wouters *et al.*, 2007) has been proposed to deal with such problems.

The 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 or dark) to be used to establish such discrimination rule, given the fact that maybe not all are needed. Thirdly, 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 (Wouters *et al.*, 2007), the third one is still an open problem. This paper is devoted to study this third issue in more detail and proposed an approach based on the model averaging ideas used on regression models (Burnham and Anderson, 2002).

In this paper, we propose a modification of the model averaging used in regression problems to the particular case of classification problems. A lack-of-classification measure will be defined, which is afterwards used to calculate the weights in the model average.

In the next section, the data are described and some background on the experiments is provided. In Section 3, the methodology is explained, starting from the general form of the doubly hierarchical supervised learning analysis. 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 4 and compared to the initial classification results.

2 Data Description

Many different recording technologies exist today for measuring brain activity. A "graphical" record of electrical activity of the brain with a high temporal resolution, a so-called electro-encephalogram (EEG), is one of them. EEG experiments have been used for many pre-clinical and clinical (research) purposes. We are interested in particular in EEG studies aiming at characterizing psychotropic drug effects on the basis of spectral EEG analysis. Classifying drugs solely based on chemical structure would create numerous categories, which would not necessarily be indicative of their therapeutic use. New chemical entities are classified according to their potential therapeutic activity as early as possible in the drug discovery process. The pharmaceutical industry currently applies the categorization proposed by Deniker (1982), Oughourlian (1984) and Cohen and Cailloux-Cohen (1995), where psychotropic compounds are divided into 5 major classes, according to their main indication in psychiatry: antidepressants, antipsychotics, anxiolytics, hypnotics, and stimulants. Availability of an advanced classification model or tool that uses a standardized physiological read-out, such as the EEG, would greatly aid efficient determination of psychoactive properties of newly synthesized chemicals.

Pharmaco-electroencephalographical studies aim to characterize psychotropic drug effects, usually on the basis of spectral EEGs, which reflect cortical brain activity. Frequency measurements, in Hertz, range from below 3.5 Hz per second (so-called delta activity), over 4–7.5 Hz/s (theta activity), 8–12 Hz/s (alpha activity), and finally above 13 Hz/s (beta and gamma activity). EEG registrations are reliably carried out in humans and laboratory animals 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), which records muscle activity. It clearly defines states of vigilance that can be separated out and used to classify psychotropic agents. Typically, six sleep-wake stages are distinguished, irrespective of the treatment received: (1) *active wake (AW)*, characterized by movement, theta activity, and high EMG; (2) *passive wake (PW)*, without movement, more variable, low amplitude EEG; (3) *light sleep (SWS1)*, characterized by interspersed delta activity with sometimes EEG spindles (short lasting burst of phasic brain activity, indicative of transitions in neuronal synchronization); (4) *deep sleep (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; (6) *Rapid Eye Movement or REM Sleep (REM)*, with theta activity and very low EMG.

The study considered here includes 26 psychoactive agents at 4 different doses, including a zero dose. For each of these compounds 8 rats were assigned to each of the 4 doses. The brain signals of the rats are recorded for 16 hours, divided into a light period of 10 hours and a period of darkness of 6 hours. The treatment is administered at the beginning of the light period and after each experiment 3 weeks of washout period are considered before using the same rat in another experiment. Several hypnogram parameters are used to assess the effects of the compounds on sleep-waking behavior. For every interval of 30 minutes the time spent in each of the six sleep-wake stages is measured (in minutes).

From these data, a training and test dataset are constructed. For all compound-dose combinations in both datasets we know exactly to which class they belong. The compound-dose combinations in the training dataset are extensively used in clinical practice. The training dataset consists of 59 compound-dose combinations: 23 placebos, 14 antidepressants, 7 antipsychotics, 5 hypnotics, and 10 stimulants. The test dataset, consists of 3 placebos, 4 antidepressants, 2 antipsychotics, 2 hypnotics and 3 stimulants.

Figure 1, About Here.

Table 1, About Here.

By way of illustration, the number of minutes spent in the six sleeping stages for 8 rats who got Clomipramine, which belongs to the antidepressant class, are plotted in Figure 1. As we can see, the profiles are very irregular, exhibiting high variability within and between rats. Table 1 contains the mean number of minutes spent in each of the six sleeping stages for every drug class. In brackets are the standard deviations. Again, a high variability (or large standard deviation) is seen for each sleeping stage and each drug class. The number of minutes spent in the sleeping stages is very similar across the drug classes. Only for stimulants do we see an increase in Active Wake and a decrease in Light Sleep. It is obvious that we need subtle techniques, taking into account the evolution over time, to set up a classification rule.

3 Methodology

Figure 2, About Here.

The doubly hierarchical supervised learning analysis (DHSLA), as has been proposed by Wouters *et al.*

(2007), is schematically represented in Figure 2. In the first stage, the longitudinal profiles are modeled and appropriate summaries extracted from the model fit (Section 3.1.1). In the second stage, these summary measures are used as input for the supervised learning analysis, in view of classifying the data. This second stage proceeds in a hierarchical fashion (Section 3.1.2).

Various flexible modeling techniques can be considered in the first stage of the procedure. A random-splines approach (Verbyla *et al.* (1999), Ruppert, Wand and Carroll (2003)) and a fractional polynomial mixed model (Royston and Altman, 1994) were used (Wouters *et al.*, 2007). While they perform similar in terms of model fit, the fractional polynomial mixed model is preferred because it is computationally less demanding and requires fewer parameters.

For the second stage, several supervised learning techniques can be used. Wouters *et al.* (2008) compared three of them: Linear (LDA), flexible (FDA), and mixture (MDA) discriminant analysis. As all three discriminant procedures produce comparable results with respect to posterior probabilities and error counts, the linear discriminant analysis is recommended, 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) write 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 modeling 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 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 3.

Before we turn to the model averaging in discriminant analysis in Section 3.2.2, we briefly review the doubly hierarchical supervised learning analysis in Section 3.1 and the model averaging approach in the context of regression as proposed by Burnham and Anderson (2002) in Section 3.2.1.

Figure 3, About Here.

3.1 Doubly Hierarchical Supervised Learning Analysis

3.1.1 Stage I: Modeling the Longitudinal Data

In the first stage, we model the longitudinal data so as to obtain relevant summaries from the profiles. Given the characteristic of the data that the outcomes are constrained to the period 0–30 minutes, imposing these on the model, or switching to a multinomial model is an obvious choice. The complexity and the hierarchical structure combined, however, does pose insurmountable convergence problems when such models are used. Thus, a modeling approach that allows for capturing complexities and intricacies in the data, while lending itself easily to the obtention of simple summaries is to be preferred. While several approaches are possible, we will use so-called fractional polynomial mixed models (FPMM). Linear mixed effects models (LMM) are a widely used tool for modeling longitudinal data (Verbeke and Molenberghs, 2000). To capture the irregular trends in our profiles, we combine the LMM with the use of fractional polynomial functions (Royston and Altman, 1994). The details of this approach can be found in Wouters *et al.* (2007). In our case, for each compound-dose combination and each sleep-wake stage, separate second-degree fractional polynomial mixed models are fitted to the light and dark periods. Not only the coefficients, but also the fractional powers, denoted by subscripted p's, are allowed to differ across compound-dose combinations. For example, for the minutes spent in Active Wake in time period k for subject j in compound-dose combination i the fractional polynomial mixed model, leading to the largest likelihood as proposed by Royston and Altman (1994), laid out in Verbeke and Molenberghs (2000), and

applied by Wouters *et al.* (2007), becomes

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

where $(\text{AW min})_{ijk}$ is the number of minutes spent in Active Wake for rat j in compound-dose combination i during the k^{th} time period ($i = 1, \ldots, 59$; $j = 1, \ldots, 8$; $k = 1, \ldots, 32$). 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, $\boldsymbol{t} = (1, \ldots, 32)'$ and p_1 and p_2 are the fractional powers. The vectors $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})$ and $\boldsymbol{\gamma}_i = (\gamma_{0i}, \gamma_{1i}, \gamma_{2i})$ are the compound-dose specific regression coefficients for the light and the dark periods, 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 $N(\boldsymbol{0}, \boldsymbol{D}_i^b)$ and $N(\boldsymbol{0}, \boldsymbol{D}_i^c)$, respectively, where $\boldsymbol{D}_i^b$ and $\boldsymbol{D}_i^c$ are unstructured 3×3 matrices. The residual components ε_i are also independent with distribution $N(0, \sigma_i^2)$. The function $I(t)$ is an indicator function specified as $I(t) = 1$ if $t \leq 20$ and 0 otherwise. This modeling approach allows for a jump between the light and dark periods, as well as for a difference in model shape, in agreement with the biology of the experiment.

Given that the drugs are administered at the beginning of the light period and based on the available expertise on drug pharmaco-kinetics and -dynamics, the action may be quite different during the initial period. Therefore, it is sensible to allow for a different, perhaps more pronounced action of the drug during the first three hours after administration. The three-hour threshold is based on expert opinion. In general, the drug action is most pronounced during the first three hours. Of course, this can be slightly different for different drugs. A smooth decay of the action may be more appropriate, but this will enhance the complexity of the models. This smooth function will lead to a different threshold for each compound-dose combination, which makes it more difficult to compare the different compound-dose combinations. Consequently, we allow for a separate model for the first three hours:

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

In this way, a flexible model combined with random effects for each effect in the model (i.e., the intercept

and both variables associated with time), in order to account for the association between time points, is the preferred choice.

3.1.2 Stage II: Hierarchical Supervised Learning Analysis

The continuation of the classification procedure necessitates informative summaries of the highly variable longitudinal profiles available for each rat. To this end, the parameters of the models in the first stage, i.e., the collection made up of $\beta_{0i} + b_{0ij}$, $\beta_{1i} + b_{1ij}$, $\beta_{2i} + b_{2ij}$, p_{1il}, p_{2il}, $\gamma_{0i} + c_{0ij}$, $\gamma_{1i} + c_{1ij}$, $\gamma_{2i} + c_{2ij}$, p_{1id}, p_{2id}, $\delta_{0i} + d_{0ij}$, $\delta_{1i} + d_{1ij}$, $\delta_{2i} + d_{2ij}$, p_{1if}, and p_{2if}, will be used as input in the supervised learning procedure. A small extract from the dataset containing the parameter estimates is shown in Table 2.

Table 2, About Here.

To establish and optimize a flexible classification rule, we proceed in a stepwise, hierarchical way. In the first step we discriminate, for example, stimulants from the other psychotropic classes, using the parameters describing the longitudinal profile pertaining to some of the sleep-wake stages for the three different periods considered (first 3 hours, light period, and dark period). Then, focus shifts to the remaining four classes. This process continues until a complete decision tree, or classification tree, has been built. The selection procedure of the sleep-wake stages in each step will be explained in detail in Section 3.1.4

3.1.3 Lack-of-Classification Measure

To determine the performance of our classification rule, 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:

$$\begin{aligned} \texttt{Error1}_s &= \mathrm{ERR}_{C_s C_{-s}} + (1 - \mathrm{PP}_{C_s C_s}), \\ \texttt{Error2}_s &= \mathrm{ERR}_{C_{-s} C_s} + \textstyle\sum_{k \neq C_s} \mathrm{PP}_{k C_s}. \end{aligned}$$

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`:

$$LC_s = w_{s1} \cdot \texttt{Error1}_s + w_{s2} \cdot \texttt{Error2}_s.$$

Different weights w_{s1} and w_{s2} can be chosen, depending on the type of application. In our particular case, we choose the weights $w_{s1} = s + 1$ and $w_{s2} = 2 \cdot (g - s)$, where g is the total number of classes in the training dataset. Along the process more weight is given to 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 between 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 such that it takes values between 0 and 1. In addition, it is 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 $F(ss)$:

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

Again, different choices can be made for $F(ss)$. We choose to proceed with $F(\text{ss}) = 0.999^{(\text{ss})}$. With this choice, an extra sleeping stage is added to the model when the lack of classification is decreased with 0.1% (this corresponds to approximately an increase in posterior probability of 0.05). As such, LC' 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 model leading to the lowest lack-of-classification LC' will be retained.

3.1.4 Selection Procedure

To arrive at an adequate estimate of the error rate, cross validation can be used. Since we have a hierarchy in our data, the cross validation can be applied at two different levels, the level of the rat and the level of the compound-dose. Model selection will therefore be conducted at both levels of cross validation, as described in the next paragraphs, and the results obtained under both scenarios will be compared.

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 (8 groups of 47 rats and 2 groups of 48 rats). 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 compound-dose combinations into ten approximately equal sized groups and then proceed in the same way as described above.

The posterior probabilities for belonging to each of the five drug classes must be adjusted for the fact that we are using a hierarchical procedure. The adjusted posterior probabilities are therefore determined in an iterative way. At the first split of the agents into two subclasses, posterior probabilities are calculated for each of them. Generally, given that k splits have been made, the values of the posterior probabilities at split $k+1$ are multiplied with the posterior probabilities of not being classified at the previous steps in the class we aimed to discriminate from the rest. More detail can be found in Wouters *et al.* (2007).

For each selection procedure, the error rate is calculated as the average of the percentages of compound-dose combinations that are misclassified in a particular class.

3.2 Model Averaging

3.2.1 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

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

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

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

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

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

where

$$w_m = \frac{\exp(-\Delta_m/2)}{\sum_{r=1}^{R}\exp(-\Delta_r/2)} \quad (2)$$

$$w_+ = \sum_{m=1}^{R} w_m I_j(m) \quad (3)$$

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

$$Y_i = \hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x_{ij} + \epsilon_{ij} \quad (4)$$

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)).

3.2.2 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. 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.

$$\text{Class } c \sim f_c(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_c)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c)] \quad (5)$$

Since the first part of equation (5) 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 $\alpha_j,\ j = 1, \ldots, p$.

$$f_c(\boldsymbol{x}) \sim \exp[\sum_{j=1}^{p} \alpha_{jc} x_j] \quad (6)$$

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

$$P(c|\boldsymbol{x}) = \frac{p_c f_c(\boldsymbol{x})}{\sum_{l=1}^{g_s} p_l f_l(\boldsymbol{x})} \quad (7)$$

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 (6), this reduces the posterior probabilities to

$$P(c|\boldsymbol{x}) = \frac{\exp[\sum_{j=1}^{p} \alpha_{jc} x_j]}{\sum_{l=1}^{g_s} \exp[\sum_{j=1}^{p} \alpha_{jl} x_j]} \tag{8}$$

We will use the lack-of-classification measure LC' defined in Section 3.1.3, equation (1), to determine the classification performance of a model.

$$\text{LC'}_s = 1 - \left(1 - \frac{\text{LC}_s}{2 \cdot w_{s1} + (g - s + 1) \cdot w_{s2}}\right) \cdot F(\text{ss}). \tag{9}$$

The lack-of-classification measure can be seen as a function that contains information coming from the likelihood function through the posterior probabilities in Error_1 and Error_2, which simply are functions of the likelihood and the prior probabilities. In both Error_1 and Error_2, the misclassification rate is taken into account to penalize those rules with high posterior probabilities but which also present larger classification errors. In addition, we control the complexity of the model using $F(ss)$. In some sense, the lack-of classification measure bears similarities with the AIC, which controls the complexity of the model through a penalization of the likelihood function by the number of parameters used in the regression model. As such the AIC tries to find a trade off between the likelihood (data) and the complexity of the model. With the lack-of-classification measure, we not only control the model complexity but also the classification performance, the main interest of our measure. Thus, we can now follow the same strategy as described in Section 3.2.1.

While before, the one model with the lowest lack-of-classification was retained, we focus now on the R models with the lowest lack-of-classification measure. For these R models we calculate weights $w^{(m)}$ in analogy to the weights defined in equation (2), 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

$$w_s^{(m)} = \frac{\exp(-\Delta_s^{(m)}/2)}{\sum_{r=1}^{R} \exp(-\Delta_s^{(r)}/2)}, \tag{10}$$

where

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

The coefficients α_{jc} in the discriminant analysis equation (8) are now averaged over the R best models as follows

$$\hat{\bar{\alpha}}_{jc} = \frac{\sum_{m=1}^{R} w^{(m)} I_j(m) \hat{\alpha}_{jc}^{(m)}}{w_+(j)}, \tag{11}$$

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.

4 Results

A fractional polynomial model is 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 are 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) is performed in each step. When parameters for a certain sleep-wake stage in the light period are used in a model, the same 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.

Initially, the model with the lowest lack-of-classification measure LC' is retained. Table 3 shows the sleep-wake stages used in each step of the discriminant analysis with both selection procedures. The adjusted posterior probabilities for the training and test datasets obtained with selection procedure I are displayed in Table 4, the ones obtained with selection procedure II can be found in Table 5.

For both selection procedures, we see that the adjusted posterior probabilities for the correct classes are very high in the training dataset. The probabilities are higher for selection procedure I than for selection procedure II. This can be explained as follows: when leaving out 10 percent of the compound-dose combinations, we obtain a substantial decrease of information in the training dataset, while leaving out 10 percent of the rats still leaves information on all compound-dose combinations in the training dataset, in turn leading to a better classification of the training dataset. Intuitively, we would expect the same to be true in the test dataset. However, there is an extra complication here. When new compound-dose combinations are to be classified, it is unlikely that the same compound-dose combinations are already

in the training dataset. Therefore, selection procedure II is closer to reality than selection procedure I. This is why both selection methods will be evaluated next to each other.

In the test dataset, we obtain 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 is 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 3.2.2.

Table 3, About Here.

Table 4, About Here.

Table 5, About Here.

The model averaging will be evaluated using the training and the test dataset. The adjusted posterior probabilities and error rate for the training dataset were improved by combining several models (classification results not shown), but much more interesting are the results obtained for the test dataset. Only the results for the test dataset are displayed here.

In the upper left panel of Figure 4 the error rates 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 rate could be reduced even further. For example, one could consider all possible model combination with only 4 sleeping stages. In this way, the error rate is reduced to 25 percent, when 200 models are used, as can be seen in the second panel of Figure 4. The same is true for a model average using only the combinations with 5 sleeping stages. 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.

In Figure 5, 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 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.

Figure 4, About Here.

Figure 5, About Here.

Similar graphs are obtained for selection procedure II as shown in Figures 6 and 7. Here, the error rates are even reduced to 26%. When reducing 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.

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.

Figure 6, About Here.

Figure 7, About Here.

5 Discussion

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 paper, we 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. Of course, the classification cannot

outperform the data: when classes are poorly separated (e.g., antidepressants), model averaging will hardly improve the results.

For Selection Procedure I, restriction 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 0.25 for 100 models or more. In general we suggest to use model averaging with 100 models in order to get nice classification results.

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 extra sleep-wake stages does not necessarily lead to better classification results.

An important issue in the proposed classification procedure is the normality assumption behind the linear discriminant analysis. It is important to highlight that first and foremost we could rely on asymptotic normality because we are using maximum likelihood estimates in the second stage, and they are theorically normally distributed. Of course, for the case of the powers, we use a grid search procedure for their estimation, but they could be seen as well as a form of discretization of what could be a maximum likelihood estimate. Another point is that the estimates are coming from the same drug class, so they are expected to have similar behavior and therefore the range of estimates will be limited. This could be seen as stacking vectors of normally distributed variables, having similar mean and variance-covariance matrix, which in principle should not destroy the normality property. Furthermore, it has been shown before that linear discriminant analysis is quite robust against violations of the assumption of normality (Lachenbruch, Sneeringer and Revo, 1973; Krzanowski, 1977; Pohar, Blas and Turk, 2004). In any case. other more flexible discriminant techniques such as mixture discriminant analysis and non-parametric discriminant analysis have been investigated as classification tool in the second phase of the doubly hierarchical supervised learning analysis (see also previously published work in Wouters *et al.* (2008)), but applying model averaging in the context of mixture or non-parametric discriminant analysis is still a topic for further research.

References

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

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

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.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A practical Information-Theoretic Approach (2nd Edition).* Springer-Verlag, New-York.

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.

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

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

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.

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

George, E.I. (1998). Bayesian model selection. *Encyclopedia of Statistical Sciences, Update Volume 3, (eds. S. Kotz, C. Read and D. Banks)*, Wiley, N.Y., 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.

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

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

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

James, G. and Hastie, T. (2001). Functional Linear Discriminant Analysis for Irregularly Sampled Curves. *Journal of the Royal Statistical Society Series B*, **63**: 533–550.

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

Krzanowski, W.J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics*,**19**, 191–200.

Lachenbruch, P.A., Sneeringer, C., and Revo, L.T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics*,**1**, 39–56.

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

Pohar, M., Blas, M., and Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodoloki zvezki*,**1**, 143–161.

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

Ruppert, D, Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

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

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* Springer Series in Statistics, Springer-Verlag, New-York.

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. (2007). Psychotropic drug classification based on sleep-wake behaviour of rats. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **56 (2)**: 223–234.

Wouters, K., Cortiñas, J., Molenberghs, G., Ahnaou, A., Bijnens, L., and Drinkenbrug, W.H.I.M. (2008). A Comparison of Doubly Hierarchical Supervised Learning Procedures for Multiple Class Longitudinal Data from EEG Experiments. *Journal of Biopharmaceutical Statistics* **18 (6)**: 1120–1135.

Table 1: *Mean number of minutes spent in each of the sleeping stages, per drug class.*

Drugclass	Active Wake	Passive Wake	Light Sleep	Deep Sleep	Intermediate Stage	REM Sleep
Placebo	11.50 (8.35)	1.57 (1.72)	7.83 (4.86)	6.17 (4.44)	0.39 (0.29)	2.51 (2.03)
Antipsychotics	11.00 (7.87)	1.87 (2.12)	8.15 (5.43)	6.28 (4.99)	0.42 (0.35)	2.37 (2.13)
Antidepressants	11.47 (8.51)	1.63 (1.79)	8.18 (5.36)	6.36 (5.05)	0.33 (0.28)	2.07 (1.93)
Hypnotics	11.57 (8.62)	1.46 (1.64)	8.16 (5.36)	5.87 (4.52)	0.42 (0.32)	2.57 (2.14)
Stimulants	13.71 (9.81)	1.59 (1.99)	6.90 (5.20)	5.63 (5.34)	0.34 (0.37)	2.01 (2.19)

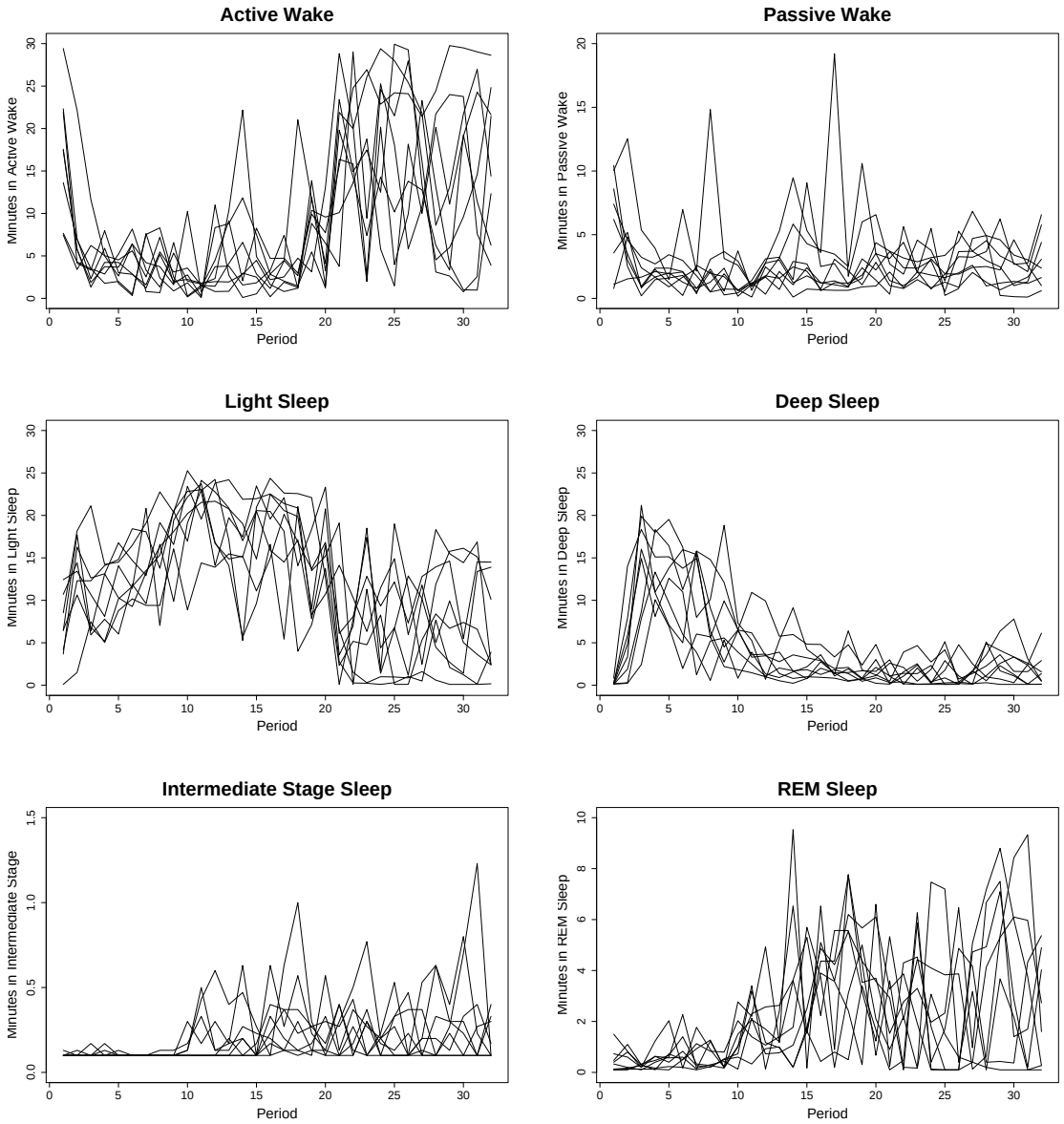

Figure 1: *Observed number of minutes spent in each of the six sleep-wake stages for the eight rats receiving Clomipramine (Antidepressant).*

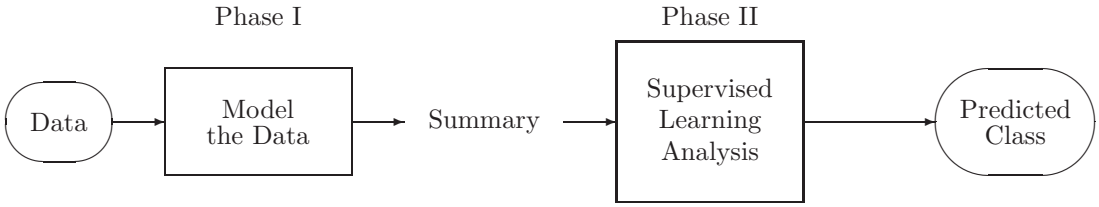

Figure 2: *Diagram representing Doubly Hierarchical Supervised Learning Analysis (DHSLA).*

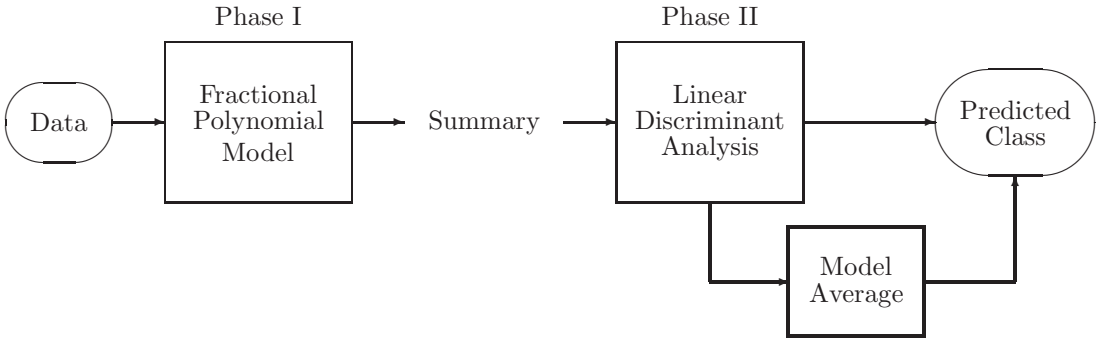

Figure 3: *Diagram representing doubly hierarchical supervised learning analysis with model averaging, when a fractional polynomial mixed model (FPMM) is used in Stage I and linear (LDA) discriminant analysis is used in Stage II.*

Table 2: Parameter estimates for the fractional polynomial mixed model for Active Wake in the light period for a compound-dose combination, belonging to the antidepressant class.

Subject	Treatment	Intercept $\beta_{0i} + b_{0ij}$	Coefficient 1 $\beta_{1i} + b_{1ij}$	Coefficient 2 $\beta_{2i} + b_{2ij}$	p_1	p_2
1	161	5.306	2.801	1.770	-1.5	2
2	161	3.556	1.804	1.034	-1.5	2
3	161	4.519	3.198	0.815	-1.5	2
4	161	4.892	1.783	1.882	-1.5	2
5	161	5.913	3.650	1.221	-1.5	2
6	161	6.029	6.538	1.617	-1.5	2
7	161	5.648	3.990	1.305	-1.5	2
8	161	4.682	3.183	0.898	-1.5	2

Table 3: Linear Discriminant Analysis. Sleep-wake stages used in each step of the doubly hierarchical discriminant analysis with linear discriminant analysis for both selection procedures.

LDA - Selection Procedure I			
Step	Light period	Dark period	First 3 hours
(1) Stimul	PW SWS2	AW SWS1 REM	AW SWS1
(2) Antipsy	PW SWS2 IS	AW PW SWS2	AW
(3) Antidep	SWS2 IS REM	SWS1 SWS2	AW PW
(4) Hypno	AW SWS2 IS REM		

LDA - Selection Procedure II			
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 REM	IS REM	
(4) Hypno	PW SWS2	IS REM	AW SWS1 IS

Table 4: Adjusted posterior probabilities obtained when FPMM and linear discriminant analysis with Selection Procedure I is applied to the training dataset (upper panel) and the test dataset (lower panel).

Selection Procedure I - Training dataset (error = 0.009)					
Drug class	Placebo	Antidep	Antipsy	Hypnotic	Stimulant
Placebo	**0.97**	0.01	0.02	0.00	0.00
Antidepressant	0.00	**0.95**	0.03	0.01	0.01
Antipsychotic	0.01	0.03	**0.94**	0.00	0.02
Hypnotic	0.00	0.01	0.00	**0.99**	0.00
Stimulant	0.01	0.02	0.01	0.00	**0.96**

Selection Procedure I - Test dataset (error = 0.583)					
Drug class	Placebo	Antidep	Antipsy	Hypnotic	Stimulant
Placebo	**0.96**	0.04	0.00	0.00	0.00
Antidepressant	0.11	**0.16**	0.45	0.16	0.12
Antipsychotic	0.38	0.03	**0.33**	0.26	0.00
Hypnotic	0.48	0.31	0.03	**0.18**	0.00
Stimulant	0.01	0.07	0.23	0.04	**0.65**

Table 5: Adjusted posterior probabilities obtained when FPMM and linear discriminant analysis with Selection Procedure II is applied to the training dataset (upper panel) and the test dataset (lower panel).

Selection Procedure II - Training dataset (error = 0.043)					
Drug class	Placebo	Antidep	Antipsy	Hypnotic	Stimulant
Placebo	**0.94**	0.01	0.01	0.04	0.00
Antidepressant	0.00	**0.75**	0.18	0.06	0.01
Antipsychotic	0.00	0.10	**0.72**	0.11	0.07
Hypnotic	0.00	0.04	0.18	**0.78**	0.000
Stimulant	0.04	0.09	0.04	0.04	**0.79**
Selection Procedure II - Test dataset (error = 0.600)					
Drug class	Placebo	Antidep	Antipsy	Hypnotic	Stimulant
Placebo	**0.33**	0.05	0.00	0.62	0.00
Antidepressant	0.02	**0.49**	0.38	0.02	0.09
Antipsychotic	0.24	0.00	**0.33**	0.43	0.00
Hypnotic	0.93	0.02	0.02	**0.03**	0.000
Stimulant	0.00	0.03	0.28	0.00	**0.69**

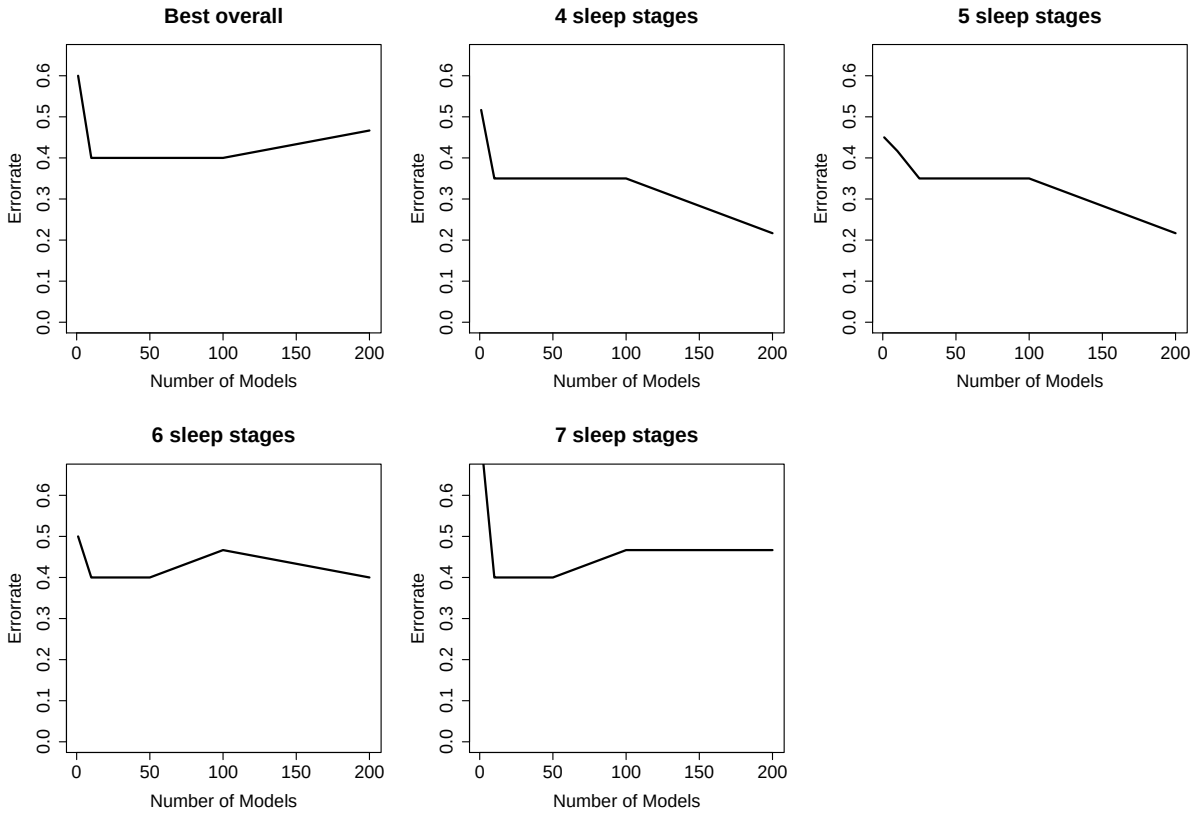

Figure 4: Model Averaging. Error rates in the test dataset obtained with model averaging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure I.

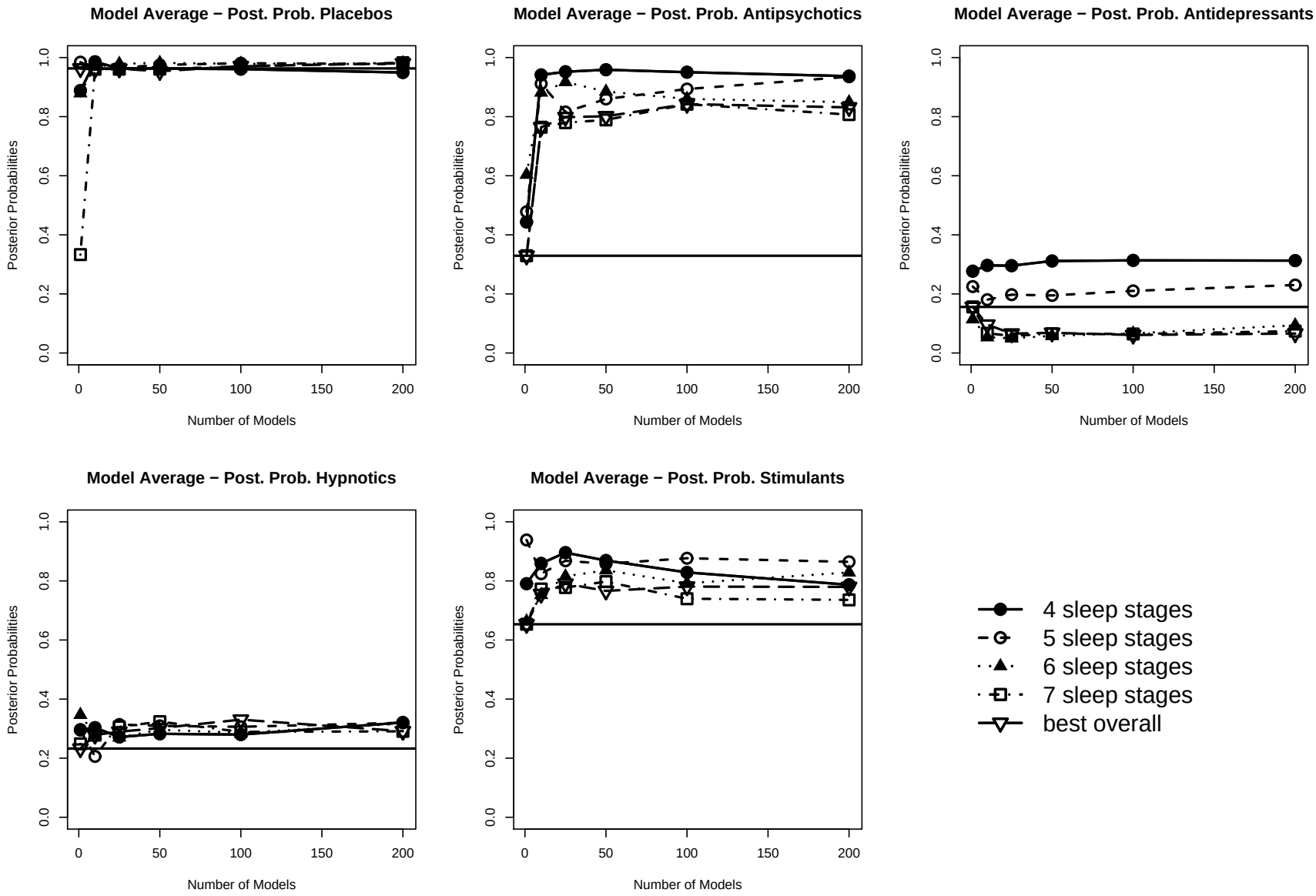

Figure 5: Model Averaging. Adjusted posterior probabilities in the test dataset obtained with model averaging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure I.

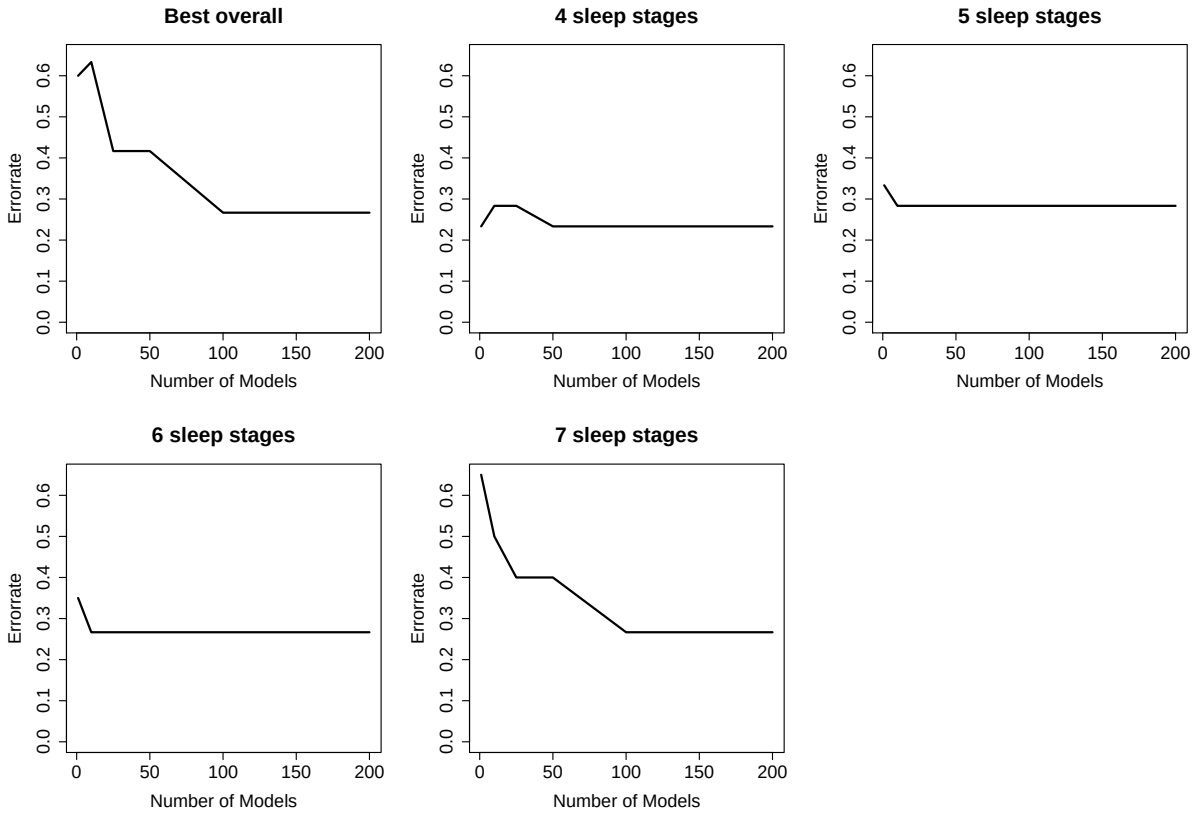

Figure 6: Model Averaging. Error rates in the test dataset obtained with model averaging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure II.

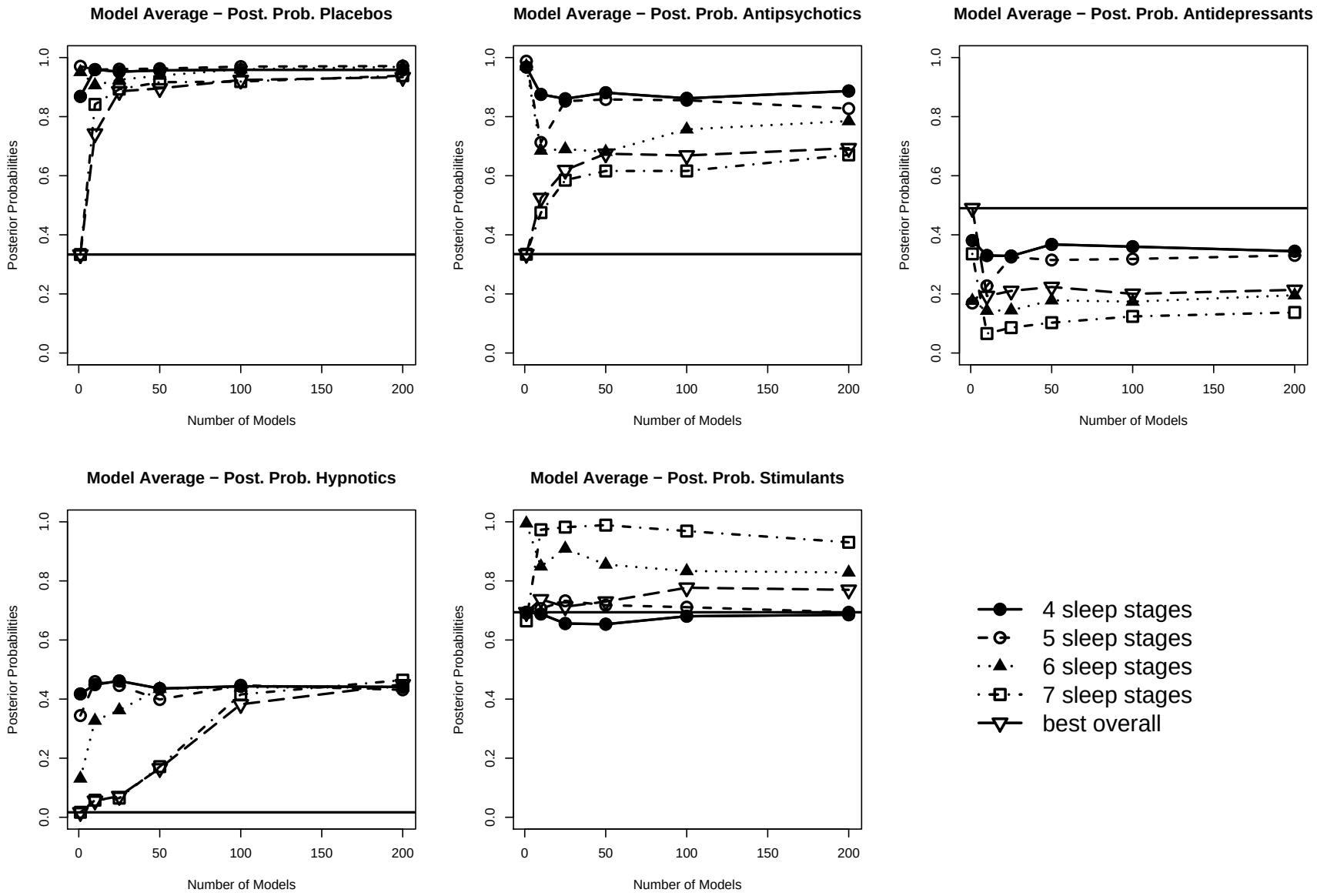

Figure 7: Model Averaging. Adjusted posterior probabilities in the test dataset obtained with model averaging for 1, 10, 25, 50, 100 and 200 models, applied to DHSLA with Selection Procedure II.